PAY-FOR-PERFORMANCE FOR HEALTH SERVICE PROVIDERS

Effectiveness, Design, Context, and Implementation

YEWANDE K. OGUNDEJI

DOCTOR OF PHILOSOPHY

UNIVERSITY OF YORK

HEALTH SCIENCES

MARCH 2015

Abstract

Countries are increasingly implementing pay for performance (P4P) as a way to improve health services. The evidence base is conflicting and difficult to interpret. It is necessary to more systematically explore evaluations of P4P schemes in order to synthesize more useful evidence to inform the use of P4P schemes in health care.

This thesis starts with a literature review, which shows that the results of evaluations of P4P schemes are heterogeneous, which may possibly be explained by differences in programme design, context, implementation, and evaluation study design. I sought to find ways to better analyse and make sense of these evaluations using two approaches. A quantitative approach was used to systematically explore the heterogeneity. I developed and tested a theoretical typology to categorise P4P schemes by their design features. This typology considers who receives the incentive, type of incentive, size of incentive, and perceived risk of not earning the incentive. I then used the typology to quantitatively explore the influence of P4P design features and evaluation designs on it effectiveness using meta-regression and multilevel logistic regression analyses. I also undertook a formative evaluation of a pilot P4P scheme in Nigeria (a case study). This used semi-structured in-depth interviews with 36 purposively sampled health workers to explore how contextual and implementation factors (e.g. delay in incentive payment) influenced the impact of the scheme.

This research presents three notable and novel contributions to knowledge about P4P in healthcare. First a useful typology was developed, which can be used to help categorize, think about, structure and report P4P schemes in a standardized and theoretically informed way. Second, I show that P4P schemes with design features such as payment to groups, large incentive size (>5% of salary or usual budget), and low perceived risk of not earning the incentive are more likely to be effective compared to schemes characterized by payment to individuals, small incentives, and high perceived risk of not earning the incentive. In addition, I demonstrate that P4P evaluations without adequate controls over-estimate the effectiveness of P4P. Third, I show that contextual factors such as incentive payment delays, poor health worker understanding of the P4P scheme, and poor infrastructure affect the effectiveness of the Nigerian P4P scheme and need to be addressed in its future development.

Contents

Abstract **3**

List of Tables **9**

List of Figures **11**

List of Appendices **12**

Acknowledgements **14**

Author's declaration **15**

Chapter 1 Introduction **16**

1.1. Background **16**

1.2. How P4P works **17**

1.2.1. Underpinning theory (agency theory) 17

1.2.2. P4P as a strategy to improve motivation of clinicians 19

1.2.3. Practicalities of pay for performance scheme 20

1.3. Thesis context **21**

1.3.1. P4P in Nigeria 22

1.4. Relevance and contribution of thesis **23**

1.4.1. Aim 24

1.4.2. Objectives 24

1.5. Thesis outline **24**

Chapter 2 Pay for performance (a review of reviews) **28**

2.0. Introduction **28**

2.1. Information sources **28**

2.1.1. Search strategy (databases and search terms to identify reviews) 29

2.2. Identification of primary studies (from the reviews and other sources) and data extraction **30**

2.3. Synthesis of results (narrative review and meta-analysis) **32**

2.4. Results **33**

2.4.1. Description of studies (reviews and primary studies) 33

2.4.2. Overview of evidence (narrative review) 37

2.5. Meta-analyses results **44**

2.5.1. Meta-analysis results for all 36 evaluation studies 45

2.5.2. RCT studies (meta-analysis results) 46

2.5.3. P4P Evaluation studies with no control groups (meta-analysis results) 48

2.5.4. Quasi-experimental studies (meta-analysis results) 48

2.5.5. Subgroup analyses for domain of performance (process vs. outcomes) 49

2.6. Discussion **51**

2.6.1. Rigour of evaluation 51

2.6.2. Heterogeneity 52

2.6.3. Exploring heterogeneity 53

2.7. What this chapter adds **55**

Chapter 3 Developing a framework to categorise P4P schemes (A P4P typology) **56**

3.0. Introduction **56**
3.1. Aims and Objectives **57**
3.2. Methods **58**
3.2.1. Identification of potential design features 59
3.2.2. Identification and exploration of relevant theories to P4P design features 60
3.2.3. Combining the features in a multidimensional space 72
3.2.4. Reducing the Typology 73
3.2.5. Piloting the typology 77
3.3. Results (refining, retesting, and labelling the typology) **79**
3.3.1. Refining the typology 79
3.3.2. Retesting the typology 82
3.3.3. Labeling the types on the P4P typology (preparing the typology for use in exploring heterogeneity) 85
3.4. Discussion **87**
3.5. What this chapter adds **91**

Chapter 4 Assessing the reliability of the Typology as a tool to categorise P4P schemes **92**
4.1. Introduction **92**
4.2. Methods **95**
4.2.1. Statistical test for estimating inter rater reliability (kappa) 95
4.2.2. Number of reports of P4P schemes 99
4.2.3. Number of raters 101
4.2.4. Recruiting the raters (and ethics approval) 102
4.2.5. Rater training 103
4.3. Results **109**
4.3.1. Rater characteristics 109
4.3.2. Ease of use of the P4P typology 110
4.3.3. Inter-rater reliability (kappa) of each item on the P4P typology 111
4.3.4. Sources of disagreement 111
4.4. Discussion **113**
Limitations 115
4.5. What this chapter adds **115**

Chapter 5 Exploring the heterogeneity of the results of evaluations of P4P in health care **116**
5.0. Introduction **116**
5.1. Aims and objectives **117**
5.2. Methods **117**
5.2.1. Identification of studies and data extraction 117
5.2.2. Regression models 119
5.2.3. Model specification 126
5.2.4. Statistical analyses 127
5.3. Results **129**
5.3.1. Descriptive statistics 129
5.3.2. Meta-regression 132

5.3.3. Multilevel logistic regression results 133
5.3.4. Sensitivity analyses 135
5.4. Discussion 138
5.5. What this chapter adds 146

Chapter 6 The Nigerian Health System 148
6.0. Introduction 148
6.1. Country overview and organization of the Nigerian healthcare system 150
6.2. Health system challenges 152
6.2.1. Health status overview (child and maternal health) 152
6.2.2. Distribution of utilisation of maternal and child health services in Nigeria 154
6.2.3. Health inequalities 156
6.2.4. Cost of health care 156
6.2.5. Current state of maternal and child health services (quality of healthcare) 157
6.3 Past Health Reforms 158
6.3.1. National Health Insurance Scheme (NHIS) 158
6.3.2. National Immunization Coverage Scheme (NICS) 159
6.3.3. Midwives Service Scheme (MSS) 159
6.4. Evidence of the impact of past health reforms on maternal and child health outcomes 162
6.5. Failure of past health reforms to meet MDG health targets in Nigeria 163
6.5.1. Misappropriation of funds (lack of transparency) in the Nigerian health system 163
6.5.2. Poor governance and Lack of accountability 164
6.6. P4P as a strategy to improve the Nigerian healthcare system 165
6.7. Summary 167

Chapter 7 Overview of the Nigerian P4P Scheme 168
7.1 Aim of the Nigerian P4P scheme 168
7.2. Phases of the Nigerian P4P Scheme 168
7.2.1. Pre-Implementation phase 169
7.2.2. Pre-pilot phase 170
7.2.3. Pilot phase of the Nigerian P4P 171
7.3. Approval to conduct research on the Nigerian P4P pre-pilot 172
7.4. Design Features of the Nigerian P4P scheme 173
7.4.1. Who receives the incentives? (And timing of payment) 173
7.4.2. Type of incentives: Fines or Bonuses 173
7.4.3. Performance measure and domain of performance 174
7.4.4. Size of incentives and payment mechanism 179
7.4.5. Review of design features of the Nigerian P4P scheme 179
7.5. Early findings of the Nigerian P4P scheme 181
7.5.1. Method of estimating change in utilisation of health services 181
7.5.2. Results 182
7.6. Discussion of the early results of the Nigerian P4P scheme 186
7.6.1. Risk (uncertainty of earning the incentive) 187
7.6.2. The role of health facility managers 190
7.6.3. Health worker understanding the P4P scheme 190

7.6.4. Infrastructure (readiness to implement the P4P programme) ... 191
7.7. What this chapter adds (rationale for exploration of contextual and implementation factors in the Nigerian P4P scheme) ... 191

Chapter 8 Methods of the formative evaluation of the Nigerian P4P scheme ... 193
8.1. Aims and objectives ... 194
8.2. Study design: a qualitative approach ... 194
8.3. Method of data collection (semi-structured face to face interviews) ... 195
8.4. Ethics approval ... 196
8.5. Data collection ... 197
8.5.1. Developing the interview questions ... 197
8.5.2. Piloting the interview questions ... 198
8.6. Setting ... 200
8.6.1. Sample size ... 201
8.6.2. Sampling strategy ... 202
8.6.3. Selecting the health facilities ... 203
8.6.4. Identification and approaching potential participants ... 204
8.6.5. Interview sessions ... 205
8.7. Data Analysis ... 206
8.7.1. Stage 1: Familiarization with the data ... 207
8.7.2. Stage 2: Identification of the thematic framework ... 208
8.7.3. Stage 3: Indexing (coding) ... 210
8.7.4. Stage 4: Charting ... 213
8.7.5. Stage 5: Mapping and interpretation ... 213
8.8. Trustworthiness of the research ... 216
8.8.1. Credibility ... 217
8.8.2. Confirmability ... 219
8.8.3. Dependability ... 220
8.8.4. Transferability ... 221
8.9. Summary ... 221

Chapter 9 Views and experiences of health workers in the Nigerian P4P scheme 222
9.1. Overview of Participants ... 222
9.2. Findings ... 223
9.2.1. Theme 1: Uncertainty of earning the incentive ... 224
9.2.2. Theme 2: Health worker understanding of the P4P scheme ... 230
9.2.3. Theme 3: Management and administration of the P4P scheme (role of the health facility manager) ... 233
9.2.4. Theme 4: Motivation ... 237
9.3. Discussion ... 242
9.3.1. Key findings ... 242
9.3.3. Strengths and limitations ... 248
9.4. Recommendations for the implementation of the Nigerian P4P pilot ... 250
9.5. Recommendations for research in P4P in Nigeria ... 253
9.6. What this chapter adds ... 253

Chapter 10 Discussion and Conclusions ... 255

10.1. Background and significance of the thesis **255**
10.2. Summary of research and findings **256**
10.3. Strengths and limitations **264**
10.4. Implications for policy and practice **267**
10.5. Recommendations for future research **269**
10.6. Conclusion **271**

Appendices **273**

List of Keywords **458**

List of Abbreviations **459**

References **460**

List of Tables

Table 1.1 Chapters addressing thesis objectives 27
Table 2.1 Search strategy output for CRD database 30
Table 2.2 Search output for the updated review by Van Herck et al. (2010) 31
Table 2.3 Quality of identified reviews using the AMSTAR checklist 34
Table 3.1 Collapsed design features to form a conceptual variable 'Risk' 75
Table 3.2 P4P Typology 77
Table 3.3 Criteria for categorisation of design variables in the P4P typology 80
Table 3.4 Results of applying the typology to P4P schemes identified from the review by Eijkenaar et al. (2012) 83
Table 3.5 Summary of design features presented in the P4P evaluation study by Kirschner et al. (2013) 85
Table 3.6 Labelling the types in the P4P typology 87
Table 4.1 Guidelines for interpreting kappa 97
Table 4.2 Number of P4P reports needed to estimate Cohen's kappa 100
Table 4.3 Kappa values for each item on the typology (pilot test) 104
Table 4.4 Ratings for individual studies by two raters 105
Table 4.5 Guidelines for use of the P4P typology 106
Table 4.6 An example of disagreement between raters 110
Table 4.7 Kappa results for each item on the P4P typology 111
Table 4.8 An example of source of disagreement between raters 112
Table 4.9 Sources of disagreement on judging item 4 ('risk') (Werner et al. 2011) 113
Table 5.1 Distribution of multiple evaluations OF P4P schemes 119
Table 5.2 Formulae used in converting effect estimates to standardised mean difference 121
Table 5.3 Formulae for computing summary effect of multiple outcomes within scheme 123
Table 5.4 Correlation (r) values used in the estimation of a summary measure of effect for multiple outcomes within P4P schemes 124
Table 5.5 Outcome variable specification for multilevel logistic regression model 126
Table 5.6 Summary of statistical models 128
Table 5.7 Characteristics of included studies 131
Table 5.8 Meta-regression coefficients for Model A (Outcome variable: P4P effect estimate) 132
Table 5.9 Regression coefficients for multilevel logistic regression (Model B) 134
Table 5.10 Random effects parameters of the multilevel logistic regression model 135
Table 5.11 Results for change in correlation values to account for multiple outcomes within schemes in the meta-regression model 136
Table 5.12 Results for change in categorisation of Binary outcomes in the multilevel logistic regression model 137
Table 7.1 Incentivised health services 174
Table 7.2 Incentivised quality indicators 175
Table 7.3 Individual evaluation tool for health workers in the Nigerian P4P scheme 177
Table 7.4 Key design features in the Nigerian P4P scheme 180
Table 8.1 Preliminary interview questions for health workers in the Nigerian P4P pre-pilot .. 199

Table 8.2 Ranks of the health facilities in each State 203
Table 8.3 An example of development of categories for under Theme: motivation 211
Table 8.4 Final coding index 212
Table 8.5 A sample of the framework matrix 214
Table 9.1 Overview of participants 222
Table 9.2 Health workers' views and experiences with uncertainty of earning the incentive 225
Table 9.3 Health workers' views and experiences regarding the individual assessment tool 227
Table 9.4 Comparison between participant clusters (Theme: Uncertainty of earning the incentive) 229
Table 9.5 View and experiences regarding health worker understanding of the P4P scheme 230
Table 9.6 Comparison between participant clusters (Theme: Health worker understanding of the P4P scheme) 232
Table 9.7 Health workers' views and experiences regarding Management and administration of the P4P scheme (role of the health facility manager) 234
Table 9.8. Comparison between participant clusters (Theme: Management and administration of the P4P scheme/role of the health facility manager) 236
Table 9.9 Participant quotes on motivating factors improving performance 238
Table 9.10 Participants' views and experiences with demotivating factors decreasing performance 239
Table 9.11 Comparison between participant clusters (Theme: Motivation) 241

List of Figures

Figure 1.1 Illustration of the Agency Theory 18
Figure 1.2 P4P States in Nigeria (Abuja is the Federal Capital Territory) 23
Figure 2.1 Flow chart of identification of included studies 36
Figure 2.2 Forest plot with pooled estimate of all 36 studies (and subgroup analyses by evaluation design) 47
Figure 2.3 Funnel plot of all 36 pooled P4P studies 48
Figure 2.4 Forest plot showing subgroup analyses by quasi-experimental evaluation design.... 49
Figure 2.5 Forest plot showing subgroup analyses by domain of performance 50
Figure 3.1 Flow chart of methods used to develop the P4P typology 59
Figure 3.2 Illustration of physician target income relative to performance 65
Figure 4.1 Predicted kappa for two categories, 'yes' and 'no', by probability of a 'yes' and probability observer will be correct (Source: Bland, 2008) 98
Figure 5.1 Illustration of the multilevel structure of the data 118
Figure 5.2 Distribution of number of outcomes per P4P evaluations 119
Figure 6.1 Structure and function(s) of the Nigerian healthcare system 151
Figure 6.2 Sources of maternal and child deaths in Nigeria (Hogan et al., 2014). 153
Figure 6.3 Utilisation of maternal and child health services in Nigeria (Source: Nigeria countdown to 2015, 2012) 155
Figure 6.4 Inequalities in health outcomes and health service utilisation (National Bureau of Statistics, 2011) 156
Figure 6.5 Change in utilisation of maternal health services (Abimbola et al., 2012) 160
Figure 6.6 Change in Maternal mortality rates (MMR) and Neonatal mortality rates (NMR) (Abimbola et al., 2012) 161
Figure 6.7 Reduction of Child and Maternal mortality rates in Nigeria (Nigeria count down to 2015, 2012) 162
Figure 6.8 Reduction in child and maternal mortality rates in Rwanda 163
Figure 7.1 Geographical Map of Nigeria indicating the three P4P and control States 169
Figure 7.2 Nigerian P4P pre-pilot implementation design 170
Figure 7.3 Timeline of the Nigerian P4P scheme 171
Figure 7.4 Impact evaluation design of the Nigerian P4P pilot (Source: NPHCDA, 2012) 172
Figure 7.5 Change in utilisation of incentivised maternal and child health services in Adamawa State (Fufore LGA) from December 2011-December 2012 183
Figure 7.6 Change in utilisation of incentivised maternal and child health services in Nassarawa State (Wamba LGA) from December 2011-December 2012 184
Figure 7.7 Change in utilisation of incentivised maternal and child health services in Ondo State (Ondo east LGA) from December 2011-December 2012 185
Figure 8.1 Overview of the Nigerian P4P pre-pilot 201
Figure 8.2 Sample of transcribed interview with initial impressions of categories 209
Figure 8.3 Sample chart of the relationship between general performance and contextual and implementation 216
Figure 8.4 Sample chart illustrating the contextual and implementation factors linked to top performing health facilities 216

List of Appendices

A1. Search strategy output for Cochrane database 273
A2. Search strategy output for PubMed database 273
A3. Summary of identified reviews 274
A4. List of identified primary studies 288
A5. List of excluded studies.....292
A6. Abstract of systematic review of economic evaluations of P4P 303
A7. Funnel plot for RCT evaluations of P4P 303
A8. Funnel plot for quasi-experimental evaluations of P4P 304
A9. Funnel plot for evaluations of P4P with no control group 304
B1. Search strategy output for economic theories to inform the P4P typology.....305
B2. Preliminary criteria for identified variable to be potentially included in the typology...306
B3. Constructing P4P typology.....307
B4. Application of the typology on identified P4P scheme 308
C1. Summary of existing methods (from studies consulted) for assessing inter-rater reliability for categorization tools in health care 318
C2. Ethics approval for study of the inter-rater reliability of the P4P typology 320
C3. Letter to potential raters participating in the inter-rater reliability study 322
C4. Participant Information Sheet for the P4P typology inter-rater reliability study 323
C5. Participant consent form for the P4P typology inter-rater reliability study 325
C6. Initial guideline for use of the P4P typology.....326
C7. P4P studies used in testing the inter-rater reliability of the P4P typology 328
C8. A sample rating template by volunteer users of the P4P typology 329
C9. Questionnaire for information about participants of the P4P typology inter-rater reliability study 336
C10. Raters report of time and ease of use of the typology.....336
D1. Extraction of information from evaluations of P4P schemes.....337
D2. Formulas and calculations used to convert effect estimates of P4P to standardized mean difference.....397
D3. List of included studies in the meta-regression analyses 399
D4 Extraction of raw numbers used in the meta-analyses and meta-regression.....401
D5. Estimates of effect sizes, standard errors, and study characteristics used in Meta-regression analyses.....413

D6. Statistical output for the multifactorial multilevel logistic regression analysis (Model B variant 2) 415

EI. Detailed description of incentivized health services in the Nigerian P4P scheme 416

E2. Quality checklist of the Nigerian P4P scheme explained 419

E3. A sample performance verification and tariff for each incentivized health service in the Nigerian P4P scheme 442

E4. Baseline trends for incentivized maternal and child health services in health facilities in Adamawa, Nassarawa, and Ondo States 443

FI. Consideration of other methods of data collection in the qualitative study of the formative evaluation of the Nigerian P4P scheme 444

F2. Evidence of Ethics Approval for the qualitative study of the formative evaluation of the Nigerian P4P scheme 445

F3. Original Information sheet for potential participants in the qualitative study of the formative evaluation of the Nigerian P4P scheme 446

F4. Participant Consent form for participants in the qualitative study of the formative evaluation of the Nigerian P4P scheme 448

F5. Amended information sheet for participants qualitative study of the formative evaluation of the Nigerian P4P scheme 449

F6. Final/refined Interview Questions for the qualitative study of the formative evaluation of the Nigerian P4P scheme 450

F7. A list of the health facilities in each State (and their characteristics) positioned according to performance 452

F8. Appointment forms for potential participants in the qualitative study of the formative evaluation of the Nigerian P4P scheme 454

F9. Factors linked with worst performing facilities (sample chart) 454

F10. Relationship between themes, categories and concepts (sample chart) 445

G1. Executive summary of report submitted to the NPHCDA on the formative evaluation of the Nigerian P4P scheme 456

Acknowledgements

Foremost, I would like to express my immense gratitude to God who has sustained me wholly throughout the period of this degree.

I would like to express my deepest gratitude and appreciation to my supervisor Professor Trevor Sheldon for his continuous support throughout my research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I would also like to thank members of my thesis advisory panel: Dr Cath Jackson for proofreading my thesis several times and for her useful and encouraging comments on the qualitative part of the thesis; Professor Martin Bland for countless Friday afternoons spent teaching me advanced statistical techniques; and Professor Alan Maynard for his encouragement, insight, support, and brilliant suggestions.

This thesis was completed with the support of a studentship by the Department of Health Science –University of York, for which I am truly grateful.

My sincere thanks go to Dr Lekan Olubajo and the National Primary Health Care Development Agency (NPHCDA) for offering me the opportunity to work on the Nigerian PBF project. In addition, I would like to thank the health workers that participated in data collection for my research on the Nigerian PBF project.

I thank my all friends (Tari, Wonu, Karisha, and Matt) for their kind words, encouragement, and support towards my goal. I thank Dammy and Dave for proofreading my thesis and listening to my ideas.

Last but not least, I would like to express my immense gratitude to my family for being my pillar of support and safety through this PhD journey. Words cannot express how grateful I am for all of the sacrifices that you've made on my behalf. I thank my parents Mr and Mrs Ogundeji, for supporting me financially, spiritually, and emotionally. I thank my brothers Kayode and Adesola for being my safety net, and for their words of encouragement.

Author's declaration

I hereby declare that I am the sole author if this thesis and that this thesis is original and has not been previously published or submitted for a degree in this or any other institution. To the best of my knowledge this thesis does not contain material previously published or written by another person except where due reference or acknowledgment is made in the text.

I declare that the author solely undertook all the data collection and analyses. I further confirm that data collection that involved health workers and young health service researchers has been conducted with the ethical approval of all relevant bodies and such approvals are acknowledged throughout this thesis.

I confirm that there are no known conflicts of interest associated with this thesis and there has been no financial support for this work that could have influenced its outcomes.

Chapter 1 Introduction

This chapter provides background and context for the thesis. This includes definitions, theoretical framework, and how financial incentives (pay for performance) work in healthcare. It also describes the significance of the research (how it can improve understanding), the aim and focus of the thesis, and the thesis outline.

1.1. Background

The use of financial incentives targeted at health service providers has increasingly been adopted internationally to improve health services across different contexts and different clinical areas (Eldridge and Palmer, 2009). Many low and middle-income countries (LMICs) are experimenting with these sorts of incentive schemes, even though evidence of effectiveness appear mixed (Eichler, 2006, Oxman and Fretheim, 2009b).

These incentive-based interventions take varied but broadly similar forms, which are characterized by different terminologies. The commonly used terminologies are Performance based financing (PBF), Results based financing (RBF), and Pay for Performance (P4P), which are often used synonymously. Despite the different labels, these incentive-based interventions all have the same systematic structure in which payment of incentives is made, conditional on meeting pre-set targets (Hahn, 2006).

This thesis shall adopt the term P4P to refer to such incentive-based interventions.

Eichler (2006, p.5), gives a typical working definition of P4P as "*a system of health financing that employs the transfer of money or/and material goods conditional on taking a measurable action or achieving a predetermined goal".*

A more detailed definition is that of McNamara (2006, p.5*S*), who defined P4P as:

"*A strategy to improve health care delivery that relies on the use of market or purchaser power…depending on the context, refers to financial incentives that reward providers for the achievement of a range of payer objectives, including delivery efficiencies, submission of data and measures to payer, and improved quality and patient safety*"

1.2. How P4P works

It has been argued that the usual method of paying health service providers to deliver health services with a focus on inputs does not deliver the necessary outcomes and that the health care providers may not be sufficiently motivated to perform better or improve quality of care, because money flows are not linked to results (Hecht et al., 2004, Eldridge and Palmer, 2009, Appleby et al., 2012, Rusa and Fritsche, 2010). This has led to the heightened interest in P4P where payment is linked to performance. In this section I describe how P4P might work, using the agency theory and motivation. I also outline the practicalities of P4P, before going on in the final sections of this chapter to describe the relevance and contribution of this thesis.

1.2.1. Underpinning theory (agency theory)

Agency theory is concerned with resolving conflict of interest between principals and agents through the agency contract (Eldridge and Palmer, 2009). In healthcare, the typical principal(s) are patients, the government, health insurance companies, health maintenance organizations (HMOs), and development partners or international donors. The agents are the providers of health services, which includes local health administrations, health facilities, private sector contractors, non-profit organisations, and individual health workers (Mehrotra et al., 2010).

This principal-agent relationship entails a principal contracting an agent to execute a task on his behalf in exchange for a reward which generates utility (Eldridge and Palmer, 2009). In this type of agreement, both the principal and the agent gain utility when the principal is not able or willing to perform the task himself and the agent possesses excellent knowledge and technical capacity in that specific field to carry out the task (see Figure 1.1).

The principal-agent relationship can work well if the principal can judge fully how well the agent is acting on their behalf (e.g. delivery of high quality and appropriate health services, which benefit the principals or those on whose behalf they are acting) (Folland et al., 1993). However, in a typical principal-agent relationship there is asymmetry of information, which is when the providers of services (agents) have more information than the consumers or the commissioners of services (principals) about the market transaction (Kinoti, 2011, Hahn, 2006). Health service providers generally have superior knowledge than the patients, on diagnosis and treatment of the patients'

condition or disease. The supplier in this case knows the problem and the solution better than the consumer (or purchasers) hence the need for the principal-agent relationship. This causes some problems, which include:

1. Moral hazard: this is when the agent takes advantage of the fact that he has more information than the principal (asymmetry of information) and makes decisions not aligned with the interest of the principal often at the expense of the principal (Nilakant and Rao, 1994).
2. Difficulty in verifying the performance of the agent (Eisenhardt, 1989).

This could potentially lead to asymmetry of power in health care provision in which it appears that the health services providers have most of the power. This makes the purchasers and funders of healthcare weak, causing major policy problems in health care.

1.2.2.1. The agency contract (pay for performance)

According to the agency theory, the two problems associated with the principal-agent relationship can be resolved by the agency contract. This agency contract is a type of pay for performance plan that helps reconcile the asymmetry of information between the principal and the agent by aligning both parties' interests (see Figure 1.1). In the agency contract, the agent's pay is linked to performance, which is assessed by measuring the things that reflect the principal's goals. Paying for performance allows for provision of objective measurements, which allows the principals (health service purchasers) to know if the agents (health service providers) have provided the quantity and quality of care required, and to adjust the payment accordingly.

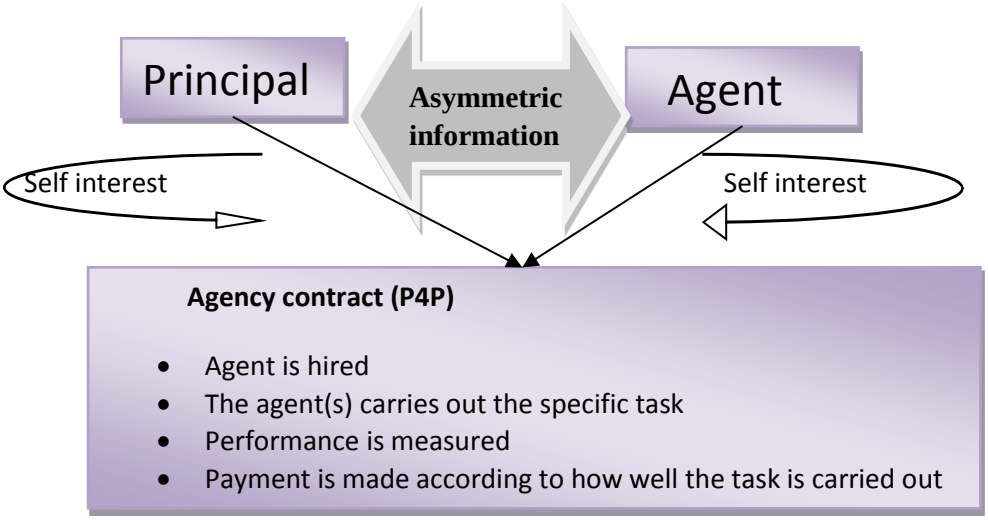


Figure 1.1 Illustration of the Agency Theory

According to Nilakant and Rao (1994), the potential of success of P4P schemes is related to the degree to which the payment plan can align the goals of the principal and the agent and change the behaviour of the agent, meaning the design of the performance-based plan is likely to be important to the effectiveness of the P4P scheme. The P4P system might not be effective if, for instance, the measures of performance are inaccurate, the system of monitoring is flawed, or if the incentives are not sufficient to motivate the agent to work in alignment with the principals' goals (Eisenhardt, 1989, Shetterly, 2000).

1.2.2. P4P as a strategy to improve motivation of clinicians

Another perspective on how P4P might work relates to the effects of financial incentive on motivation (Kohn, 1987). Motivation is defined as the tendency to initiate and sustain effort towards a goal, which can perhaps be measured by how hard one works (Clark and Estes, 2002). Motivation is broadly classified into two types: intrinsic (an inherent satisfaction that stems from the job done or task carried out e.g. altruism, compassion) and extrinsic (where individuals act as a result of external factors e.g. money, recognition) (Ryan and Deci, 2000).

The general idea connecting P4P and motivation is that financial incentives are able to appeal to extrinsic motivation of individuals to work harder or longer (Kohn, 1987). This assumption seems logical for individuals who want to increase their income and are willing to do more work for it. This assumption might be more applicable in contexts where clinicians/health workers are poorly paid and often have to work under poor working conditions. In such cases, financial incentives from P4P could indeed go a long way in increasing extrinsic motivation through bonuses (to supplement salaries) and improved infrastructure, which might have otherwise hindered delivery of quality health services (Willis-Shattuck et al., 2008, Luoma, 2005).

Some researchers however, have argued the use of financial incentives could crowd out intrinsic motivation in clinicians (Deci et al., 1999, Cameron et al., 2001). It is argued that a continual use of incentives may violate clinicians' sense of professionalism and altruism (Wynia, 2009). This may result in the clinicians being solely reliant on incentives to do anything, which could in turn be very costly and inefficient (Gneezy and Rustichini, 2000). In addition, Kohn (1987) argued that the continued use of financial incentive as a means on motivation could backfire because if the incentive

comes to be seen as the main reason one is engaging in an activity, the activity can be viewed as less enjoyable in its own right thereby crowding out intrinsic motivation. This means that the perceived increase in quality could recede after the removal of incentives.

Others argue that intrinsic and extrinsic motivation are intertwined and cannot easily be separated, and that extrinsic motivation can help sustain or support intrinsic motivation especially in LMICs (Willis-Shattuck et al., 2008, Luoma, 2005, Covington and Müeller, 2001). For example, in the case of clinicians who may be motivated by the pride they get in doing their job (or job satisfaction) but who have no basic infrastructure like laboratory or diagnostic tools in their health centres might be forced to treat patients based on trial and error which could lead to poor health outcomes and in turn dampen the intrinsic motivation of the clinician. However, if financial incentives are used (a form of extrinsic motivation), some of the incentives earned might go into purchase of equipment needed thereby sustaining/supporting the intrinsic motivation.

1.2.3. Practicalities of pay for performance scheme

Agency theory described in the previous section highlights the relationship between incentive and performance or behaviour. However, P4P schemes in healthcare implemented on a large scale are often complex undertakings, and their potential to improve performance in health services might depend on other important factors that are able to affect motivation such as job satisfaction or adequate/safe working environment (contextual factors) (Henderson and Tulloch, 2008, Willis-Shattuck et al., 2008, Robyn et al., 2014).

Furthermore, the P4P approach is based on strict rules of accountability (setting targets, and measuring and verifying duties), which often replaces a system based on trust whereby agents are paid a salary and assumed/assured to act on behalf of the principal. Thus, making P4P a more costly approach compared to a system based on trust. Also, P4P introduces the possibility of creating perverse incentives where agents/health service providers are not willing to do anything unless they are paid (O'Neill, 2004). Other negative or unintended consequences of P4P in health care include ‘cherry picking’ of healthier patients, neglect of incentivised activities, and the potential of patients undermining the trust of health professionals as a results of the perception of perverse incentives (a case where patients think health professionals are delivering

certain health services because of the incentive and not because it is needed) (Powell et al., 2012, Chen et al., 2011, Doran et al., 2011).

The success of P4P schemes also depend on implementation factors such as adequate regulatory authority for monitoring, evaluation, and appropriateness of quality indicators (Glasziou et al., 2012, Campbell et al., 2011). Furthermore, literature suggests that design features and contexts are likely to influence the impact of the P4P schemes (Epstein, 2012, Eijkenaar, 2013, Van Herck et al., 2010, Mehrotra et al., 2010, McDonald and Roland, 2009). Despite this, little attention has been given to exploring and understanding of the differences within the schemes and how this can inform development of more effective and cost effective P4P programmes.

1.3. Thesis context

Even though P4P has been gaining global popularity (Epstein, 2012), it is not a new strategy in healthcare and it has been experimented with as far back as the early 1970's when the Korean government offered financial incentives to community distributors to recruit more users of family planning schemes (Cometto, 2008).

P4P schemes could be implemented either on their own or as part of a wider quality improvement strategy in health care. Where quality health care is defined by six characteristics: safe, effective, person-centred, timely, efficient (avoiding waste), and equitable (Institute of Medicine, 2001).

Other quality improvement strategies include non-financial incentives (public/peer reporting), performance profiling, technical assistance, tiered networks, and integrated care management (Roland and Campbell, 2014, Boaden et al., 2008, Grol, 2013).

Some have argued that a combination of P4P and other strategies (e.g. case management, audit, and feedback) might be more effective in improving quality of care (Mauger et al., 2014, Tricco et al., 2014). Whilst critics of P4P have argued that less expensive, strategies for quality improvement such as peer or public reporting might be enough to improve quality of care, with similar outcomes to P4P (Llanos and Rothstein, 2007, Jha et al., 2012). Despite the debate surrounding P4P, it is increasingly being adopted in many countries.

So far, P4P schemes have been implemented to address a number of dimensions of quality in health care. This includes safety, reduction in medical error, reducing variation in clinical practice, cost containment, efficiency, utilization of health services, and delivery of services in a timely manner (Chaix-Couturier et al., 2000). Examples of such P4P schemes include Premier hospital alliance (USA), Advancing Quality (AQ) incentive programme, Quality and Outcomes Framework (UK), P4P in Rwanda, Cordaid P4P (Tanzania), and Iranian P4P (Eijkenaar, 2012, Canavan et al., 2008, Aryankhesal et al., 2013).

1.3.1. P4P in Nigeria

There has been a surge of implementation of P4P in LMICs, primarily funded by international donor organisations such as the World Bank and Save The Children Organisation. Nigeria is one of the countries where P4P has recently been introduced (NPHCDA, 2012).

Over several decades, Nigeria has suffered from a weak healthcare system, characterised by lack of coordination, lack of transparency, poor infrastructure, and high health worker absenteeism (Abdulraheem et al., 2012, Akinwale, 2010, Asuzu, 2005). This, along with other factors, is reflected in poor health outcomes such as a maternal mortality rate of 840 per 100, 000 live births (ninth highest globally) and a child mortality rate of 138 per 1000 live births (WHO, 2012).

P4P in Nigeria was implemented in 2011 as a strategy to improve maternal and child health outcomes. This P4P scheme incentivises health service providers to improve quality and increase utilisation of basic maternal and child health service in primary health care facilities in seven out of 36 States: Zamfara, Kaduna, Katsina, Kano, Ondo, Nassarawa, and Adamawa (Figure 1.2).

The rationale for introducing P4P to the Nigerian health system was that it had potential to directly address some of the weakness in the Nigerian health system by payments aligned to performance, potentially increasing health worker motivation. Also, through the monitoring and evaluation required by these schemes, it would have the potential to improve transparency and accountability (Nair, 2012, Nair, 2011).

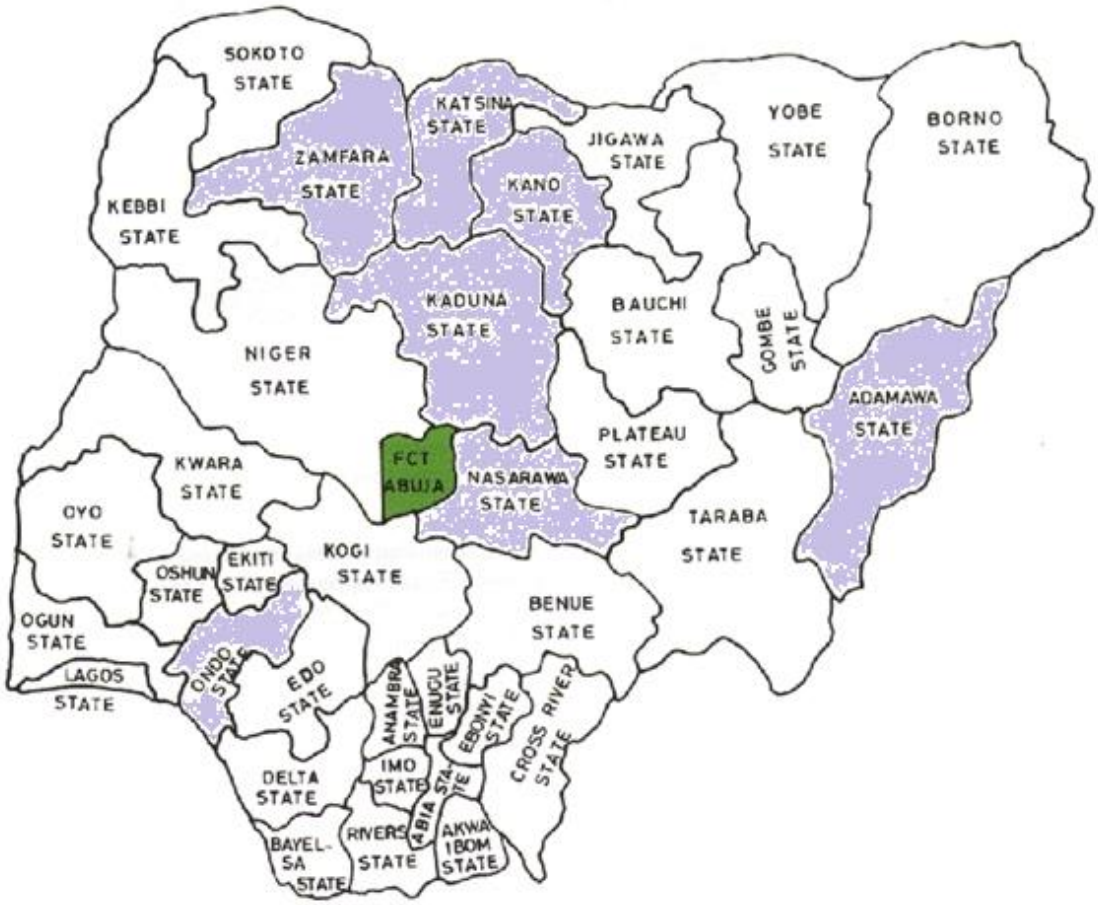


Figure 1.2 P4P States in Nigeria (Abuja is the Federal Capital Territory)

1.4. Relevance and contribution of thesis

In contrast to what the increasing popularity of P4P might suggest, its effectiveness and cost-effectiveness has not been convincingly demonstrated, despite over a decade of experimentation.

The evidence base on what works, what does not and why is mixed and fragmented (Epstein, 2012, Van Herck et al., 2010). Furthermore, theoretical evidence on how individuals respond to incentives is not sufficiently used to inform these schemes, and literature on the best ways to maximize the effectiveness of P4P schemes through the exploration of design features, contexts, and implementation factors) is sparse (Mehrotra et al., 2010, Eijkenaar, 2013). P4P in Nigeria is a relatively new cost intensive project, which is lacking strong evidence to inform its implementation. Therefore, to avoid wastage of scarce resources, there is the need for careful consideration of literature, theory, and empirical evidence to make sense of the available evidence and contribute to understanding of the influence of design on P4P. Exploratory

research of P4P in Nigeria is also required to provide a better understanding of the influence of design features, contexts, and implementation on P4P schemes in Nigeria.

In this thesis I reflect critically on the design, context, and implementation of P4P in healthcare. Specifically, this thesis contributes to (1) ways of categorising P4P scheme based on a conceptual understanding of factors that might influence their effectiveness, (2) improving the evidence base by empirically testing what factors influence the effectiveness of P4P schemes, (3) the understanding of the influence of design features, contextual factors, and implementation factors on the effectiveness of the Nigerian P4P scheme, to inform, improve, and strengthen the design and implementation of the Nigerian P4P scheme and provide some guidance for the scaling up of the scheme over the next five years.

1.4.1. Aim

The aim of this thesis was to better understand important aspects of design, context, and implementation, and to consider their implications on the effectiveness P4P in health care.

1.4.2. Objectives

In order to achieve the aim, I set the following relevant objectives:

1. To review the available evidence on P4P and identify the key findings and gaps in the evidence
2. To develop a reliable framework to categorize P4P schemes in a systematic way to aid evidence synthesis
3. To systematically and quantitatively explore which design features are critical to the effectiveness of P4P
4. To explore the impact of design features, contexts, and implementation factors on the effectiveness of the Nigerian P4P scheme (a formative evaluation).

1.5. Thesis outline

This thesis is divided into two main parts. The first part (chapters two to five) focuses on addressing the first three objectives and the second part (chapter six to nine) focuses on addressing objective four (see Table 1.1).

In chapter two, I conduct a review of literature exploring the effects of P4P in healthcare in different countries in order to summarise the key findings and the limitations of the available evidence. I summarise empirical evidence in distinct relevant sections as identified from literature. This includes effectiveness, country specific evidence, cost effectiveness, and unintended consequences. I further explore and assess the available evidence by conducting a meta-analysis to estimate the pooled effectiveness of P4P. I also use subgroup analyses to explore the impact of different evaluation designs (e.g. RCTs and quasi-experimental studies) and the domain of performance used to measure effect (e.g. process or outcomes). I then discuss shortcomings of the available evidence such as poor evaluation design (lack of adequate control group) and heterogeneity, which makes evidence synthesis difficult. Finally, I identify the need to develop a way to investigate this heterogeneity to help make better sense of the evidence.

Chapter three describes the development of a theoretically informed framework (a P4P typology) to help categorize the P4P schemes based on design features, in order to explore heterogeneity to aid in evidence synthesis. The typology was constructed using theory from behavioural science and economics and empirical evidence relevant to understanding how people respond to incentives. The developed typology consists of four relevant design features of P4P: who receives the incentive, type of incentive, size of incentive, and uncertainty of earning the incentive.

In chapter four, I test the inter-rater reliability of the P4P typology using multiple raters. I also assess the ease of use of the P4P typology as a tool to categorise descriptions of P4P schemes in literature.

In chapter five, I use the developed P4P typology to systematically categorize P4P schemes reported in the review of evaluations. I then carry out quantitative analyses to systematically explore the influence of design features on effectiveness of P4P. I apply several regression models to investigate the relationship between the estimates of effect sizes of P4P and the design features (using the typology), while adjusting for the degree to which evaluation studies had adequate controls.

In Chapter six, I introduce the Nigerian health care system to provide a context for the formative evaluation of the P4P scheme in Nigeria. I describe the extent of the issues in

the health system, such as poor maternal and child health indicators and low quality of healthcare. I also discuss previously implemented health reforms to tackle the issues, drawing out core underlying challenges of the health system.

Chapter seven describes and reviews the design features of the P4P scheme in Nigeria. I also present, review, and discuss the early findings of the schemes; highlighting potentially relevant contextual and implementation factors influencing the results of the scheme. This includes: delay in payment, communication, managerial competence, health worker understanding of the scheme, and infrastructure.

In chapter eight, I report the aims and methods of an interview study I conducted in P4P sites in Nigeria to explore the influence of contextual and implementation factors on the effectiveness of the scheme. I also justify the methods of data collection and analysis used. This involves the use of semi-structured interviews (an in-depth qualitative approach) to explore health workers' views and experiences on potentially relevant contextual and implementation factors in the Nigerian P4P scheme.

Chapter nine presents the findings of my formative evaluation of the Nigerian P4P scheme. I provide extensive participants' quotes from multiple perspectives, discuss the findings, strengths and weaknesses of the study, and provide policy and research recommendations to inform and improve the Nigerian P4P scheme.

In chapter ten, I summarise the findings of this thesis, identify contribution to knowledge, reflect on the strengths and limitations, and discuss the implications of the findings for policy, practice and research with respect to developing, designing, and implementing effective P4P in healthcare.

Table 1.1 illustrates the objectives of the thesis and what chapters address them.

Table 1.1 Chapters addressing thesis objectives

Objectives	**Addressed in chapter**
To review the available evidence on P4P and identify the shortcomings of the evidence	2
To develop a reliable framework to categorize P4P schemes in a systematic way to aid evidence synthesis	3 and 4
To systematically explore what design features are critical to the effectiveness of P4P	5
To explore the impact of design features, contexts, and implementation factors on effectiveness of the Nigerian P4P scheme	6-9

Chapter 2 Pay for performance (a review of reviews)

2.0. Introduction

There are several systematic reviews assessing the impact and effectiveness of P4P in health care e.g. (Witter et al., 2012, De Bruin et al., 2011, Reda et al., 2009, Hamilton et al., 2013). Despite this, evidence regarding the use of P4P in health care is inconclusive, thereby making it of limited use in informing policy and practice (Van Herck et al., 2010, Eijkenaar et al., 2013). It is therefore, important to identify and understand the available literature in order to facilitate and improve the synthesis of evidence, which could be more useful in informing policy and practice of the use of P4P in health care.

The aim of this chapter is to summarise the available evidence and identify features of relevant literature on the effects of P4P for health service providers in healthcare by conducting a review of reviews and a meta-analysis (using primary studies identified from the reviews).

In the following sections, I describe my information sources (including search terms, search strategies, and databases), eligibility criteria, identification and extraction of the relevant studies, synthesis of results, and the findings of the review and meta-analysis.

2.1. Information sources

First, to conduct the review of reviews, I systematically searched and identified published reviews of P4P in literature. Second, to conduct the meta-analysis, primary studies were identified from the reviews identified. Additional primary studies were identified from other sources, such as the bibliography of the studies found and databases of organisations experienced with the use of P4P in healthcare e.g. Cordaid, Global Alliance for Vaccines and Immunisations (GAVI), and The World Bank. In the following sections, I first describe the search strategy for identifying the reviews, before going on to describe how the primary studies were identified from the reviews and additional sources.

2.1.1. Search strategy (databases and search terms to identify reviews)

I searched five electronic databases: Centre for Reviews and Disseminations (CRD), [Database of Abstracts and Reviews of Effect (DARE), National Health Service Economic Evaluation Database (NHS EED), Health Technology Assessment (HTA)], Cochrane reviews and PubMed. DARE, NHSEED, HTA, (Centre for Reviews and Dissemination), and Cochrane reviews were selected for the search because of the quality of the reviews produced by these databases. PubMed database was searched to ensure relevant reviews were not excluded.

I searched using keywords (commonly used in P4P literature), such as financial incentives, performance based financing, and pay for performance. There were no language and publication date restrictions. I conducted the through a period of two years and four months (January 2012-June 2014). See Table 2.1 for the search strategy output for the CRD database (outputs from other databases are shown in Appendix A1 and A2).

I included only reviews that assessed the impact or effects of P4P on health provider's quality of health care (see appendix A3 for a summary of included reviews). Other than that, there were no strict inclusion and exclusion criteria, as I was interested in summarising all relevant evidence on the impact or effects of P4P in health care.

Finally, I assessed the quality of the identified reviews using AMSTAR (a measurement tool to assess the methodological quality of systematic reviews).

Table 2.1 Search strategy output for CRD database

Database	**Centre for Reviews and Dissemination (CRD)**
Host	http://www.york.ac.uk/inst/crd/
Date of search	January 2012-June 2014 last search date: 26/6/14
Years covered	1990-June 2014 (no date restrictions)
Search Strategy	Key word search: Financial incentives, Pay for performance, Performance based financing (Pay for performance) OR (financial incentives) OR (performance based financing) IN DARE, NHSEED, HTA
Language restrictions	None
Number of citations	70
Number of relevant reviews	8: Huang et al., 2013, Reda et al., 2012, Chaix-couturier et al., 2012, Hamilton et al., 2013, Witter et al., 2012, Scott et al., 2011, Petersen et al., 2006, Houle et al., 2012

2.2. Identification of primary studies (from the reviews and other sources) and data extraction

To identify primary studies from the reviews, there were no strict exclusion criteria or date restrictions. The only inclusion criterion for primary studies was that it should be an evaluation study of P4P. Studies describing P4P schemes and other incentive schemes not targeted at health service providers were excluded (see Appendix A4 for included and A5 for excluded studies).

In addition, I updated one of the systematic reviews identified (identified in the previous search) to identify current or new primary studies conducted after publication of these reviews. I selected the review by Van Herck and colleagues (2010), which scored 11/11 on the AMSTAR checklist indicating a well conducted review (see Appendix A3 for grades of other identified reviews) (Shea et al., 2007). In addition, the review was comprehensive, in that it was not specific to one area of health service delivery, but included P4P studies across different health care interventions including, smoking cessation, disease management, cancer screening, prescription behaviour, and cost containment. I updated this review by extending their search terms from the date of their last search (2009) to 2014 (see Table 2.2).

Table 2.2 Search output for the updated review by Van Herck et al. (2010)

Database	**Medline**
Host	http://www.ncbi.nlm.nih.gov/sites/entrez (Pubmed)
Date of search	26/6/2014
Years covered	01/07/2009 to 28/07/2014
Search Strategy	("Salaries and Fringe Benefits"[Majr] OR "Reimbursement, Incentive"[Majr] OR "Fees and Charges"[Majr] OR p4q OR p4p OR pay* OR incentive* OR bonus*) AND ("Treatment Outcome"[Majr] OR "Medical Errors"[Majr] OR "Quality Control"[Majr] OR "Cost-Benefit Analysis"[Majr] OR "Safety"[Majr] OR "Health Services Accessibility"[Majr] OR quality OR outcome* OR performance OR error* OR safety* OR access* OR equity OR effectiveness) AND ("Hospitals"[Majr] OR "Physicians"[Majr] OR hospital* OR physician* OR practitioner*) AND (hasabstract[text] AND ("2009/07/01"[EDat]:"2014/07/28"[EDat]) AND (Humans[Mesh]) AND (Clinical Trial[ptyp] OR Randomised Controlled Trial[ptyp] OR Case Reports[ptyp] OR Clinical Trial, Phase I[ptyp] OR Clinical Trial, Phase II[ptyp] OR Clinical Trial, Phase III[ptyp] OR Clinical Trial, Phase IV[ptyp] OR Comparative Study[ptyp] OR Controlled Clinical Trial[ptyp] OR Evaluation Studies[ptyp] OR Technical Report[ptyp] OR Validation Studies[ptyp]))
Language restrictions	None
Number of citations	1356

There was no detailed assessment for risk of bias for the primary studies identified (to be included in the meta-analysis) due to the poor reporting quality of the primary studies. This would have hindered objective judgement of the risk of bias in most studies. What was clear, however, from the reviews was that a majority of the P4P evaluations were studies with poor methodological quality, and that the P4P literature was populated with evaluation studies without adequate control groups (see Appendix A3 for details), which was taken into consideration in the synthesis of results of this review (see section 2.3).

Data from primary studies was extracted in a standardized way. A uniform template was used for all studies to extract information on intervention area/summary, outcomes, and effect estimates of individual studies for all outcomes considered: whether P4P had a positive effect or not, evaluation design (whether there was an adequate control group or not). I also extracted information on sample size and raw numbers of events, from

studies that reported them (see Appendix D1 and D4[1] for characteristics of included studies).

2.3. Synthesis of results (narrative review and meta-analysis)

To summarise the key findings of the evidence and to identify its shortcomings, I conducted both a narrative review and meta-analyses (quantitative pooling of effect). The narrative review examined the identified reviews and the additional primary studies. I illustrated the results using summaries of the findings of the reviews and some primary studies that provided a background or examples that elucidated on the nature of the evidence. I summarised evidence on clinical effectiveness of P4P, country specific evidence, sustainability of effect of P4P, cost effectiveness, and unintended consequences, while drawing out the issues with the evidence.

The meta-analysis included primary studies identified from the reviews and additional primary studies identified from other sources. I first pooled together all relevant included studies. I then stratified the analyses by evaluation design (presenting separate pooled estimates of RCTs, quasi-experimental studies, and studies with no adequate control group). This was because there is a considerable literature that suggests results or effects of intervention from studies with poor evaluation designs (lacking convincing control groups) are more susceptible to bias, and likely to over-estimate the effect of interventions (Shadish et al., 2002, Tilling et al., 2005, Eccles et al., 2000, Ferriter and Huband, 2005). Thus stratification of the meta-analyses by rigour of evaluation could help shed some light in the context of the evaluation of P4P in healthcare.

For the meta-analyses, a random effects model was used because of the heterogeneity between the primary studies (differences in objectives, designs, countries, contexts, and size). A random, effects model was used as opposed to a fixed effect model because a fixed effects model assumes that the observed differences among study results are due solely to the play of chance, i.e. that there is no heterogeneity. The random effects model on the other hand assumes that the effects being estimated in the different studies

[1] The data extracted here are used substantially in other chapters, where they have other elements added unto them. Therefore in order to avoid repetition of large sized tables, the appendices referred to here (and in some other parts of this chapter) are not in chronological order but in the order of where they are seen or used subsequently in the thesis.

are not identical, but considers the differences as if they were random (i.e. it assumes that there is heterogeneity, which is considered in the model) (Engels et al., 2000, Borenstein et al., 2009).
In the meta-analysis, heterogeneity was also measured and quantified using the I^2 test statistic, which describes the percentage of the variability in effect estimates that is due to heterogeneity rather than chance (Higgins and Green, 2011). Finally, publication bias was assessed by a visual inspection of a funnel plot (Sterne et al., 2011).

STATA statistical package (version 12) was used to perform the random effects meta-analysis using the 'metan' command. I took two things into consideration for the analysis. First, the outcome measures reported in the primary studies were in different forms, which included: odds ratio, percentage point differences, means, and mean differences. It was not possible to combine all these measures of effect in one meta-analysis. In addition, pooling estimates for the different measures would have been difficult to interpret. Therefore, I converted the outcome measures into a common standardised measure (standardised mean difference and associated standard errors), which required extraction of additional data such as absolute differences (percentages or numbers), sample size, standard deviations or standard errors or variance (Bland, 2000, Borenstein et al., 2009). Second, some of the primary studies reported multiple principal outcome measures. If these were all included in the analyses, it would overestimate the amount of independent information included, which could produce possibly biased estimates (Moerbeek, 2004, Snijders and Bosker, 2012). I therefore computed a summary effect and its associated standard error for studies that reported multiple outcomes using the formulas suggested by Borenstein et.al. (2009) (details of these calculations are described in chapter 5).

2.4. Results

The first two sections below provide a description of reviews identifies and an overview of their effect findings in the form of a narrative review. Following that, the findings from the meta-analyses are presented.

2.4.1. Description of studies (reviews and primary studies)

I identified 15 relevant reviews in total (see Table 2.3). The identified reviews varied in

methodological quality (see Appendix A3 for summary of the reviews). The publication dates of the reviews ranged from 2000 to 2013. Using the AMSTAR criteria, all the reviews apart from one were moderate to high quality (see Table 2.3). Two reviews scored four and below: low quality, five reviews scored between five and seven: moderate quality, and eight reviews scored between eight and twelve: high quality (see Appendix A3).

Table 2.3 Quality of identified reviews using the AMSTAR checklist

Reviews	AMSTAR score	Quality
Van Herck et al., 2010	11	High
Witter et al., 2012	11	
Reda et al., 2009	10	
Houle et al., 2012	10	
Hamilton et al., 2013	9	
Gillam et al., 2012	9	
Scott et al., 2011	9	
Huang et al., 2013	8	
Petersen et al., 2006	7	Moderate
De Bruin et al., 2011	6	
Eijkenaar, 2012	6	
Chaix-Couturier et al., 2000	6	
Christianson et al., 2008	5	
Oxman and Fretheim, 2009	4	Low
Canavan et al., 2008	3	

I identified 326 primary studies (excluding duplicates) spanning from 1990-2013 from 15 reviews, from other sources (bibliography etc.), and from updating the review conducted by Van Herck and colleagues (2010) (see Figure 2.1). I screened out P4P studies not targeted at health service providers (12) and studies that I could not find the full text articles (13). I assessed 301 studies for eligibility: I excluded non-evaluation P4P studies (146)[2], descriptions of P4P schemes (36), and studies with unclear and/or poorly reported outcomes (12). In total, I identified 102 primary studies (including nine

[2] Studies that did not specifically evaluate the effects of P4P on healthcare quality/cost/performance/outcomes e.g. implementation studies or studies exploring the take up of p4p

qualitative studies) (see Figure 2.1). Out of the primary studies identified, only nine studies were randomised controlled trials (RCTs), 63 other studies had an adequate control group (quasi-experimental), and 30 studies had no control group.

For the meta-analyses, only 36 studies were included (6 RCTs, 20 quasi-experimental studies, and 10 pre-post studies with no control group) (see Figure 2.1). The reduced number of studies included in the meta-analyses was due to poor reporting of important data to aid conversion of the different reported outcome measures into a standard measure to enable the inclusion of as many as possible studies in the meta-analysis (see section 2.3.). The authors of the other studies were not contacted for more detailed information due to time constraints. Nevertheless, it was considered that the included studies provided some indication of the nature of the available evidence.

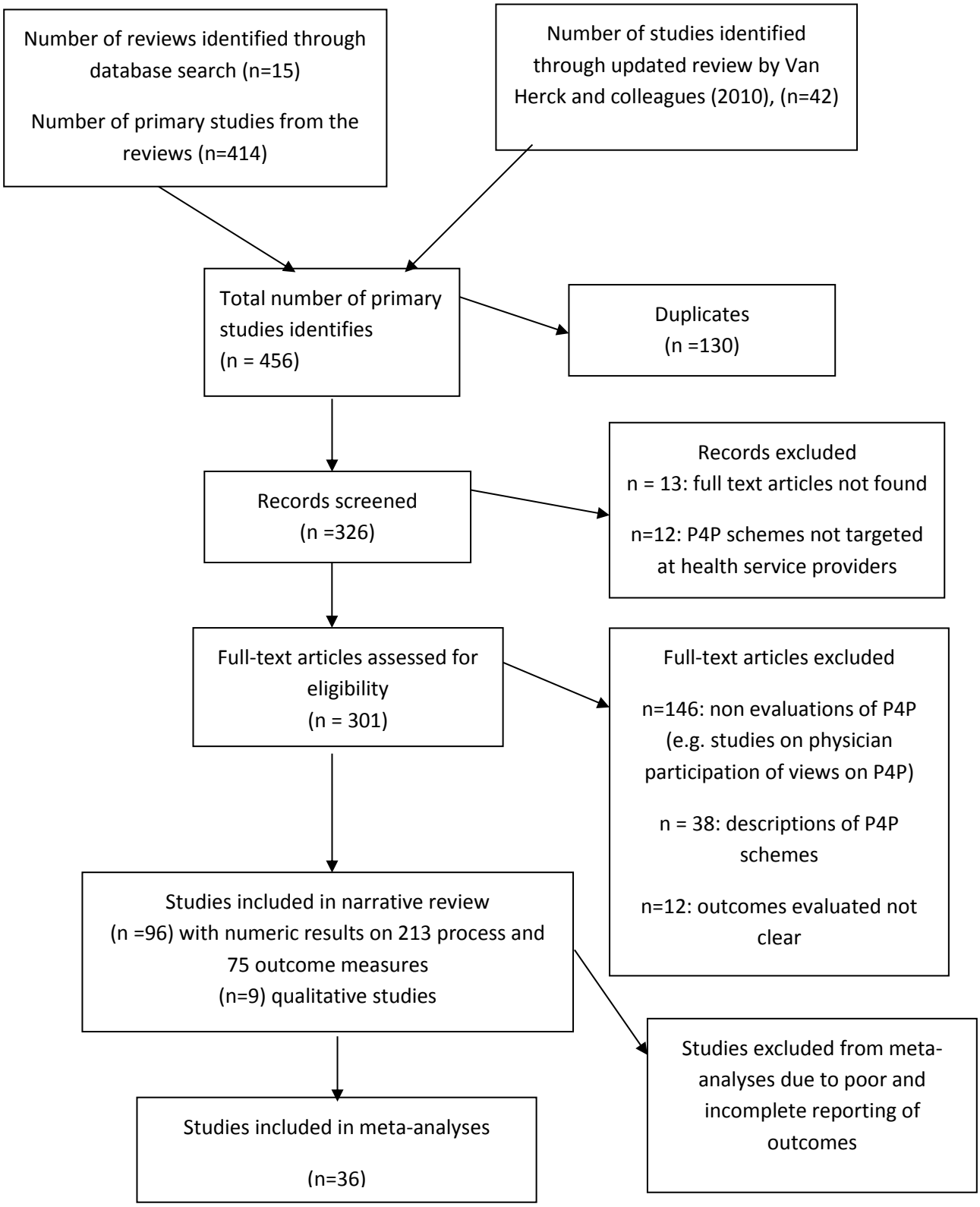


Figure 2.1 Flow chart of identification of included studies

2.4.2. Overview of evidence (narrative review)

In this section, drawing from the identified reviews and primary studies, I summarise the available evidence and highlight the problems therein. First, I summarise the evidence on the general effectiveness of P4P. Then, I summarise the effectiveness of P4P on the most common areas in which it is used health care: smoking cessation and chronic disease management, pointing out variation in effects of P4P on process and outcome measures. Afterwards, I summarise country specific evidence, including the quality and outcomes framework (QOF) in the UK and some evidence in low and middle-income countries (LMIC), highlighting differences in results according to the rigour of evaluation. Finally, I summarise the available evidence on cost-effectiveness, sustainability of the effect of P4P, and the unintended consequences of P4P, highlighting the limited evidence base.

2.4.2.1. Evidence on general effectiveness

An early review by Chaix-Couturier and colleagues in 2002, found that financial incentives could be used to reduce the use of health care resources by increasing compliance with practice guidelines, but that it is more effective to use combinations of financial and non-financial incentives. Other systematic reviews that assessed the impact of financial incentive on health service providers for health quality measures also found mixed results. For example, Petersen et al., (2006), found that about half of the studies included in their review reported mixed results on the impact of financial incentives on quality measures. About 20% of their included studies showed no statistically significant results (for indicators assessed) while the other 20% showed positive impact, with one study even reporting negative effects of P4P on quality measures. Similarly, the review by Christianson et al., (2008) found improvements in some of the quality measures assessed, but the degree of contribution of P4P was not clear because the financial incentives were typically implemented in conjunction with other quality improvement efforts and there were no convincing comparison groups.

More recent reviews show similar mixed results. The review by Scott et al., (2011) found positive but modest effects in a few measures of quality of care, provided by primary care physicians. Houle et al., (2012) also found that financial incentives modestly improved preventive activities, such as immunization rates, but there was little evidence that it was effective for other activities such as mammography referrals and

cancer screening. Further showing the variations in the effect of P4P schemes was a very comprehensive systematic review by Van Herck and colleagues (2010), which assessed impact evaluations of P4P schemes, as well as evidence on the impact of design choices and contextual mediators on the effectiveness of P4P. They found that financial incentives result in the full spectrum of possible effects for specific targets, from absent or negligible to strongly beneficial and that the effects findings of P4P are likely to relate to context.

Finally, the study by Basinga et al. (2011), which assessed the effects of P4P on maternal and child health services in Rwanda (using an RCT design) also found that P4P resulted in significant improvements in institutional deliveries and preventive (child) visits, but had no effect on prenatal visits (pregnant women) and childhood immunizations.

It is noteworthy that despite the overwhelming evidence of varied/mixed effects of P4P, none of these reviews explored this variation using statistical methods.

2.4.2.2. Evidence on management of chronic diseases and smoking cessation (process vs. outcomes)

Two of the systematic reviews assessed the effect of financial incentives on chronic care. De Bruin et al., (2011) assessed the effectiveness of P4P schemes used to stimulate delivery of chronic care through disease management (by health service providers) with regards to quality and costs. They found that most studies showed positive effects of P4P on healthcare quality. However, five out of the eight P4P schemes were part of a larger scheme of interventions to improve quality of care and it was not clear how much of the improvement observed is attributed to P4P. The review by Huang et al., (2013), further showed the inconsistency of effects of P4P on diabetes treatment and management, (e.g., patients with records of total cholesterol or blood pressure). The authors also found that process indicators such as recording of blood pressure and cholesterol levels had higher rates of improvement than outcome indicators (intermediate) such as cholesterol and blood pressure reduction.

In the same way, evidence on smoking cessation interventions was mixed and suggests that P4P might be more effective for process measures compared to outcomes. A review by Hamilton et al. (2013) assessing the impact of financial incentives to health service providers on smoking cessation found that P4P improved some process indicators such

as recording smoking status, advice and referrals but not for outcome measures such as smoking quit rates. Reda and colleagues (2009), also found no evidence of effectiveness of P4P on smoking cessation interventions (both processes and outcomes).

Following this, a descriptive analysis of data extracted from primary studies showed that P4P had a positive effect on 148 out of the 213 reported process measures (70%), as opposed to the positive effect on 41 out of 75 reported outcome measures (55%) (see Appendix D1). The findings demonstrate that evidence of whether pay for performance will lead to better patient outcomes is unclear (Huang et al., 2013, Hamilton et al., 2013).

Understandably, some outcome measures would be dependent on patient behaviour as well as the quality of health care. Therefore, it might be more difficult for financial incentives to health service providers to have a positive impact on those measures. In addition, some incentivised processes might not necessarily impact directly or at all on patient outcomes. It is important that process of care measures used in P4P schemes should be chosen based on good and robust evidence that improving these processes leads to improved health outcomes (Oxman and Fretheim, 2009b, Oxman and Fretheim, 2009a).

2.4.2.3. Rigour of evaluation

All the reviews identified reported that there were large numbers of primary studies with poor evaluation quality (lacking adequate controls) (see Appendix A3). A number of reviews concluded that the evidence base was too weak (due to poor evaluation quality) to draw reliable and valid conclusions or that the validity of the effect of financial incentives on healthcare is limited (Chaix-Couturier et al., 2000, Witter et al., 2012, Gillam et al., 2012, Canavan et al., 2008). In the following paragraphs, I describe evidence on rigour of evaluations using country specific examples of P4P such as the Quality and Outcomes Framework (QOF) in the UK, and developing countries such as Rwanda.

The QOF is one of the largest and most evaluated financial incentive programmes. The systematic review by Gillam et al., (2012) assessed the impact of the QOF on quality measures. They found that the QOF programme improved the incentivised activities in the first year of the programme at a faster rate than the pre-intervention trend. They also

found negative effects such as worsening quality measures for non-incentivised conditions, decline in person-centeredness of consultations, and decline in patients' satisfaction with continuity of care. The conclusion of this review was limited because of the lack of adequate control groups in the evaluations of the QOF programme. A few researchers have evaluated the QOF scheme using a convincing control group. An example is the study by Serumaga et al. (2010) assessing the impact of the QOF on management and outcomes of hypertension using an interrupted time series design (a quasi-experimental design). The study found that improvements in hypertension management and outcomes were as a result of gradual improvements before the introduction of P4P and that P4P had no effect on hypertension management and outcomes. On the other hand, retrospective and cross sectional studies assessing the impact of the QOF on hypertension management and outcomes concluded that the introduction of P4P improved treatment and management (Ryan and Doran, 2012, Simpson et al., 2011).

In the same way, evaluations suggest that P4P might be effective for some quality measures in LMICs, especially in Rwanda, Haiti, and Burundi (Oxman and Fretheim, 2009a, Witter et al., 2012, Canavan et al., 2008). However, the review by Oxman and Fretheim (2009a) demonstrated that it was difficult to disentangle the effects of financial incentives from other quality improvement measures. Similarly, Canavan et al., (2008) found that that P4P evaluations showed remarkable improvements in health indicators (utilization, coverage and emergency referral) in Afghanistan, Democratic Republic of Congo, Rwanda, and Haiti but that it was not clear the extent of attribution of improvements to financial incentives because of the presence of confounding contextual factors e.g. differences in infrastructure. Likewise, the review by Witter et al., (2012) found mixed results from their review of the effectiveness of P4P in LMIC. They found that P4P was effective for some quality measures but not others. The high and moderate quality studies included in this review reported that some quality indicators improved while there was no improvement in others. Two of the studies showed significant improvement for the intervention group, while two showed no significant difference.

2.4.2.4 Evidence on Cost Effectiveness

The cost effectiveness of these schemes is very important and should be central to the debate together with effectiveness because implementing a P4P programme can be quite

costly (the costs to be considered includes the incentive cost, administrative costs, monitoring and evaluation costs).

There were two published systematic reviews with an explicit focus on economic evaluations of P4P schemes. One of the reviews (Emmert et al., 2012), considered costs and consequences of the P4P intervention, and included nine studies. Out of these nine studies included, only three were considered to be full economic evaluations with good methodological quality, and these reported that P4P was not cost effective. For example, the study by Nahra et al. (2006), assessed the cost effectiveness of a P4P programme focusing on improving the quality of heart care (process measures) in the hospital setting over a period of four years. It found a cost per quality adjusted life year (QALY) of £30,081, which was above the cost-effectiveness threshold of around £25,000 as suggested by The National Institute for Health and Care Excellence (NICE) (McCabe et al., 2008).
The other six studies that were considered as partial economic evaluations demonstrated mixed results of cost effectiveness, with poor methodological quality to draw valid or strong conclusions.

A more recently published review by Meacock et al., (2013) focused on assessing the cost effectiveness of the Advancing Quality (AQ) incentive programme in England. They critiqued the narrow range of costs and outcomes considered within previous economic evaluations of P4P schemes, before proposing a new and more comprehensive framework, which they applied to the Advancing Quality (AQ) programme. Their findings suggest that the AQ is likely to have represented a cost-effective use of resources during the 18-month period of their study by generating approximately 5200 quality-adjusted life years and £4.4million of savings in reduced length of stay in the hospital. The AQ programme was shown to be cost effective within the study period. An important question remains whether the benefits and cost savings are sustainable. An impact evaluation of the AQ conducted by McDonald and colleagues (2014) found smaller mortality reductions in five clinical areas (acute myocardial infarction, heart failure, coronary artery bypass graft, pneumonia, and hip and knee replacement) in the long term (i.e. at 42 months) compared to mortality reductions at 12 months.

Some other reviews have looked at the cost effectiveness of P4P schemes but not comprehensively. Christianson et al., (2006) found two economic evaluations, which demonstrated that P4P was efficient. The methodology for these studies, however, was not assessed. The review by Van Herck et al (2010) also assessed the cost effectiveness of P4P schemes. They found mixed results and varying methodological quality from eight economic evaluations, which mirrored the studies included in the Emmert et al. (2012) review (see Appendix A6 for abstract of review).

In summary, there is little and mixed evidence regarding the cost-effectiveness of P4P schemes. Most of the studies were methodologically weak, apart from the recent study by Meacock et al., (2013), which sets the standard so far for cost-effectiveness studies of these schemes. It is important to question the cost effectiveness of these schemes even if they appear to be effective. It is important to know whether benefits observed are worth the investments. There might be other alternatives that could be implemented at lower costs producing similar or greater impact on outcomes or the same resources might be spent in other ways that may produce greater total health or health equity benefits.

2.4.2.5. Evidence on Sustainability of effect

The evidence regarding the long-term effects of P4P is even more scarce, and longer-term evaluations are needed to capture this. No reviews explicitly assessed the long-term effects of financial incentives and effects after removal of the incentives. However, there were a few primary studies assessing the sustainability of the effect of financial incentives in a few countries.

Researchers explored the removal of financial incentives in the QOF (UK) on some clinical quality indicators: influenza immunisation, lithium treatment monitoring, blood pressure monitoring, cholesterol concentration monitoring, and blood glucose monitoring. They found that all the indicators appeared to remain stable after the removal of incentive, apart from influenza immunisation which showed a statistically significant reduction (Kontopantelis et al., 2014).

Jha et al., (2012) assessed the long-term effect on the US Medicare Premier programme on patient outcomes. They compared data from hospitals implementing P4P and hospitals implementing public reporting alone on 30-day mortality of patients with acute myocardial infarction, congestive heart failure, or pneumonia, or who underwent

coronary-artery bypass grafting (CABG) between 2003 and 2009. They found similar rates at baseline for both premier and non-premier hospitals, similar decline in mortality rates in both, which remained similar after six years under the P4P scheme. Werner and colleagues (2011) also investigated the effects of the Premier programme for a period of five years. They found that even though performance (based on some of the indicators) in the intervention group improved within the first three years of the study, the performance of the hospitals in the control group caught up with and matched them in the last two years. The authors suggest the findings could be due to two things:

1. Participating hospitals could not improve their performance much more than they already had in the past two years.
2. It is also possible that the incentive programmes led non-participating hospitals to change their practices. For example, the hospitals in the control group assumed that the incentive programme might be extended to their hospitals therefore focused on improving their performance.

If true, this suggests that P4P might have knock on effects on non-participating hospitals or health facilities, which in turn might increase cost effectiveness.

2.4.2.6. Evidence on Unintended consequences

Two reviews and a few primary studies reported unintended and adverse effects of financial incentives on health care (Van Herck et al., 2010, Petersen et al., 2006). These include gaming, cheating, cherry picking, and neglect of non-incentivised services (Mannion and Braithwaite, 2012, Van Herck et al., 2010, Petersen et al., 2006).

A study conducted by Gravelle et al. (2008) showed evidence of gaming in the QOF. There is also evidence which shows that health service providers in P4P studies focus most of their attention on the monitored and evaluated health service(s), leading them to ignore/neglect other unevaluated but equally important aspects of health services (Cometto, 2008, Doran and Kontopantelis, 2013).

There is some evidence that P4P may worsen racial care disparities (Karve et al., 2008, Alshamsan et al., 2012). For example, the study by Karve and colleagues (2008) showed that under the Medicare P4P scheme, African American patients were less likely to receive evidence-based therapies compared to than white patients.

There is also some evidence that P4P may encourage health service providers to induce

demand for incentivised services over and above that which is clinically needed (Cometto, 2008, Canavan and Swai, 2008, Powell et al., 2012). This behaviour takes the focus off the patient, possibly exposing the patients to the risk of iatrogenic injury, which could undermine the patients' trust in the clinician (Cometto, 2008, Canavan and Swai, 2008, Powell et al., 2012). An example is the study conducted by Powell et al. (2012) on the QOF, which showed that P4P induced pressure on clinicians to achieve high performance scores on incentivised services. This led to actions to improve scores even when they were not necessarily in the patients' best interests, which took the focus off patient concerns and patient service, and made it difficult for patients to make informed decisions.

Lastly, evidence suggests that P4P may divert resources away from medically indigent communities or high-risk patients making health disparities worse (Friedberg et al., 2010). For example, a study exploring the impact of P4P on diabetes care in Taiwan found that P4P did indeed improve quality of care for enrolled patients. However, only a minority of the population were enrolled due to the physicians 'cherry picking' the healthiest patients because they were most likely to perform better on the selected measures compared to more complicated patients (with greater health needs) (Chang et al., 2012). This example is also similar to the issue of patients with multiple conditions, as most P4P programmes currently focus on single condition measures that might not necessarily reflect the complex nature of patient care with multiple conditions. Following the rigid guideline of single condition measures (in order to meet targets) might hinder the care of the other condition and might limit the freedom of the providers in deciding the best treatments for this set of patients (Chen et al., 2011, Silversmith, 2011).

2.5. Meta-analyses results

First, I present results of meta-analysis (including pooled estimates, heterogeneity tests, and forest plots) of all 36 included studies and explore evidence of publication bias in the form of a funnel plot. I then present results of meta-analyses by rigour of evaluation (RCTs, quasi-experimental studies, and studies with no adequate control groups).

The pooled effect of all included studies in each meta-analysis is illustrated in the forest plot presented. Each study's specific standardized mean difference (SMD) is plotted as a solid square symbol whose size is proportional to the weight of each study, with a horizontal line through each square representing the 95% confidence interval. The study with the smallest square symbol (widest 95% CI line) means that it is the least precise and carries the least weight in the meta-analysis (Rudnicka and Owen, 2012). A positive SMD value would indicate that P4P was superior to controls on improving quality of care or health outcomes (e.g. improved immunisation rates or reduction in hospital mortality), whilst a negative SMD value would represent superiority of the control group over P4P in improving quality of care (Higgins and Green, 2011).

The vertical line at SMD of 0.0 (also known as the null effect line) is the line of 'no effect', meaning that the effect of the incentive group is the same as that of the control group (no incentive). All the square symbols to the left hand side of the line of 'no effect' SMD <0 suggests that no incentives are favoured over incentives in these studies. Studies with block estimates to the right hand side of the 'no effect' line suggest that incentives are more effective than no incentives (favours incentive). However, the 95% CI lines for some of the studies overlaps the line of 'no effect', this is because the lower end of the 95% CI is <0 or the upper end is <0. Consequently, for these studies, the 95% CI includes the possibility of no benefit from incentives (for studies with blocks on the right side of the null effect line) or the possibility of benefit from incentives (for studies on the left side of the null effect line).

2.5.1. Meta-analysis results for all 36 evaluation studies

Figure 2.2 shows a pooled meta-analytical estimate (and a forest plot) of all 36 studies included in the meta-analysis. It also shows results of subgroup analyses by evaluation design (RCT, quasi-experimental, and no control groups) and by domain of performances (processes or outcomes). I now describe these results.

The pooled overall effect (meta-analytic estimate) of all included studies is illustrated on the forest plot (Figure 2.2) by the solid diamond, the centre of which is at 0.14 and the tips of the diamond represent the 95% CI limits, at 0.03 and 0.25, which indicates a very small but statistically significant ($p<0.0001$) overall positive effect of incentives. However, the measure of heterogeneity 1^2 was 99.9%, which suggests considerable

heterogeneity (>75%) according to Cochrane handbook of systematic reviews (Higgins and Green, 2011). This means that 99.9% of the variation between studies is due to true heterogeneity between the studies rather than chance alone. The funnel plot suggests that there is no evidence of publication bias, as the plot is roughly symmetrical (see Figure 2.3).

2.5.2. RCT studies (meta-analysis results)

The subgroup analysis for RCT studies in Figure 2.2 included the six RCTs included in the meta-analysis. The pooled effect indicates a much smaller effect (0.08 95% CI 0.01-0.15) compared to results of pooled estimates of all studies (in favour of P4P). The large study by Basinga et al., (2011) is dominant (accounting for about 75% of the overall weight). However, the results look similar and there is no heterogenity between these six studies: (I^2)=0.00, which suggests that variation between the pooled studies is likely due to chance. The funnel plot shown in Appendix A7 also suggests that there is no evidence of publication bias.

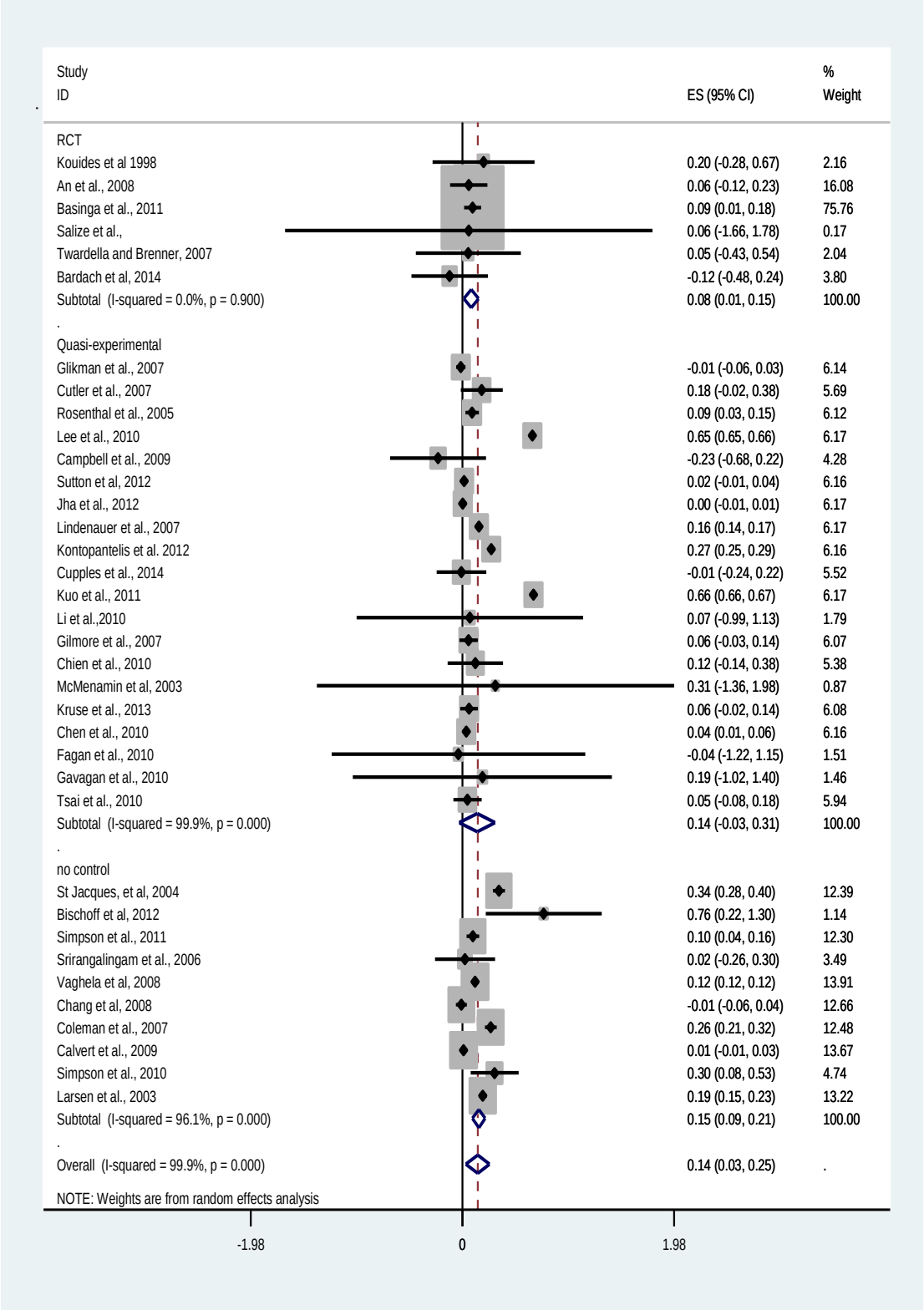


Figure 2.2 Forest plot with pooled estimate of all 36 studies (and subgroup analyses by evaluation design)

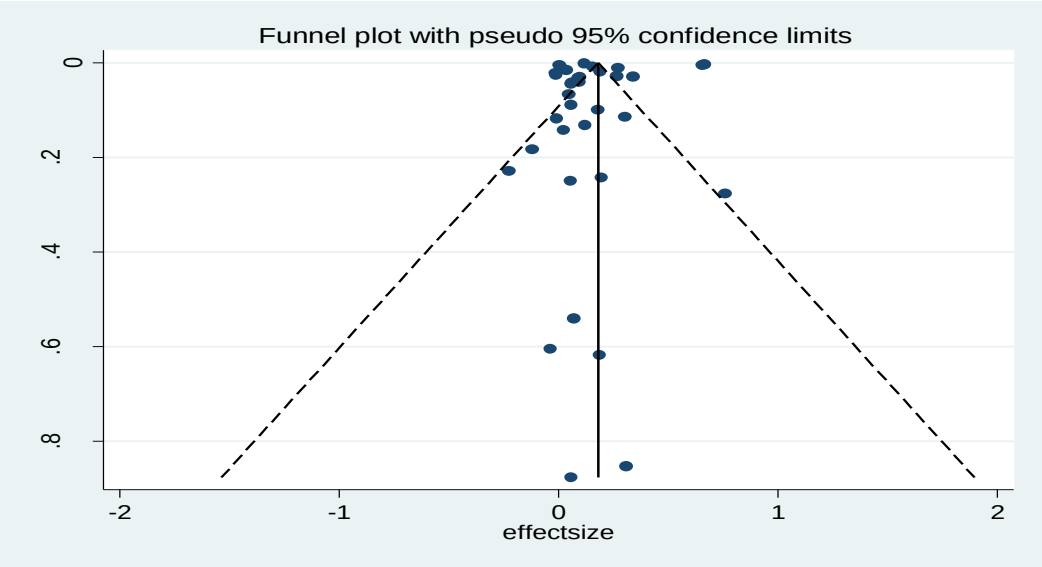


Figure 2.3 Funnel plot of all 36 pooled P4P studies

2.5.3. P4P Evaluation studies with no control groups (meta-analysis results)

Ten studies with no control group were included in the subgroup analyses presented in the forest plot shown in Figure 2.2. The pooled effect was 0.15 (95%CI 0.09-0.21), a higher effect estimate compared to RCTs and quasi-experimental studies. I^2 was 96.06%, which indicated substantial heterogeneity between pooled studies. There was no evidence of publication bias from the funnel plot presented in figure Appendix A9.

2.5.4. Quasi-experimental studies (meta-analysis results)

Twenty quasi-experimental studies were included in the subgroup analysis in the forest plot presented in Figure 2.2. This included non-randomised control studies, interrupted time series, and cross sectional studies. The overall pooled estimate shown in Figure 2.2 was 0.14, 95% (CI 0.03-0.31), an effect almost twice the size of the RCTs studies shown in section 2.5.2. Measure of heterogeneity (I^2) was 99.9%, which indicated substantial heterogeneity between pooled studies. The funnel plot presented in figure Appendix A8 exhibits asymmetry, which is likely as a result of heterogeneity and or poor methodological design (which leads to inflated effects in smaller studies), rather than publication bias (Lau et al., 2006, Sterne et al., 2011). Since there where different categories of quasi-experimental studies in P4P literature (pre-post design with control groups, cross sectional, interrupted time series (ITS), and longitudinal), I extended the subgroup analysis to reflect this. This is illustrated in Figure 2.4. Pre-post evaluations of P4P with control groups appeared to have the largest effect (0.22 95%CI -0.33, 0.47)

compared to cross sectional designs (0.03 95%CI 0.01, 0.05) or ITS designs (0.07 95%CI -0.40, 0.55). This might have been largely due to two studies (Lee et al., 2010 and Kuo et al., 2011), which appeared to have skewed the effect of P4P in pre-post studies with controls considerably to the right (thus appearing to have a larger pooled effect compared to other quasi-experimental designs).

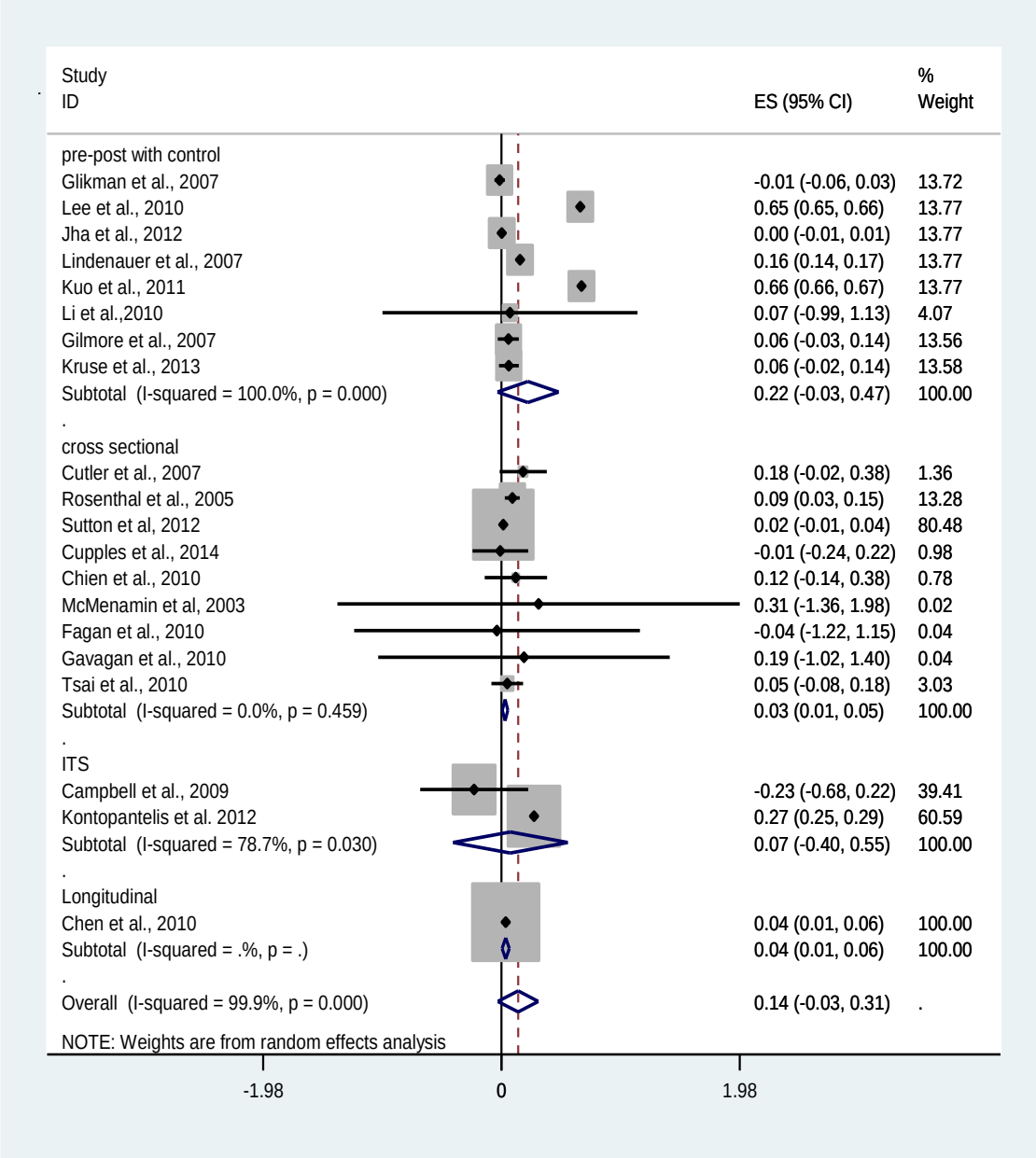


Figure 2.4 Forest plot showing subgroup analyses by quasi-experimental evaluation design

2.5.5. Subgroup analyses for domain of performance (process vs. outcomes)

Following the evidence from the narrative review that P4P is likely more effective for process measures as opposed to outcomes (section 2.4.2.2), I performed a subgroup

analyses on the studies included in the meta-analysis to investigate this statistically. This is illustrated in Figure 2.5. P4P studies focused on improving process measures (e.g. cancer screening or smoking cessation advice) appeared to have a bigger effect (0.18 95%CI 0.06, 0.31) compared with studies focused on improving outcomes such as smoking cessation or reduction in hospital mortality (0.00 95%CI -0.01, 0.01). P4P studies focused on intermediate outcomes such as blood pressure or cholesterol reduction also had a smaller effect (0.07 95%CI -0.01, 0.15) compared to studies focused on processes but a larger effect compared to studies focused on outcomes.

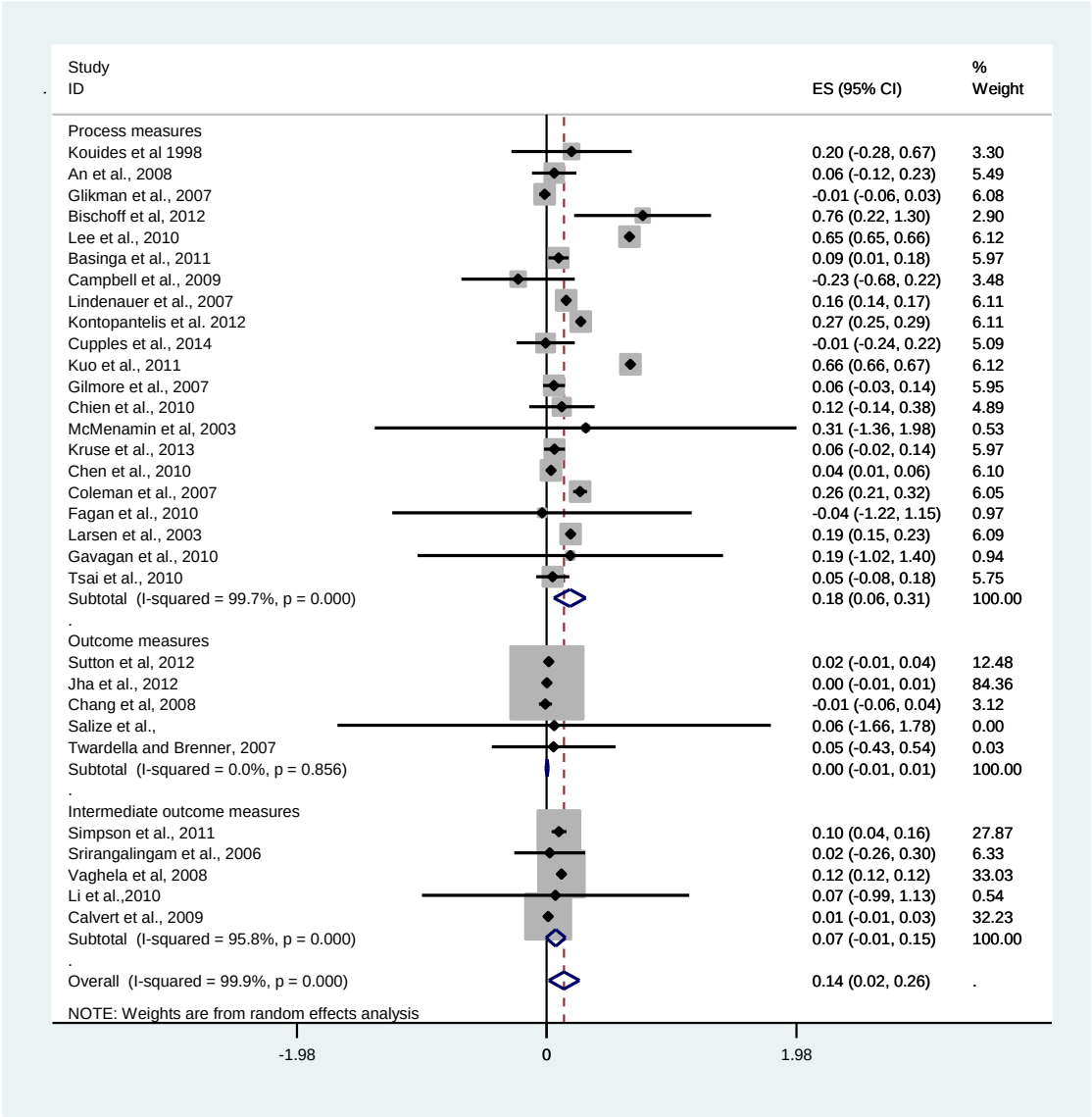


Figure 2.5 Forest plot showing subgroup analyses by domain of performance

2.6. Discussion

This aim of this review was to summarise evidence and identify and describe the shortcomings of the current evidence of the effects of P4P in healthcare. There was substantial evidence on the effectiveness of P4P but very limited evidence on cost-effectiveness. Guidelines suggested by Cohen (1998) indicate that P4P might have a very small effect (SMD= 0.14). In addition, there was evidence that P4P might be more effective in improving process measures (e.g. cancer screening coverage) compared to intermediate outcomes (e.g. blood pressure reduction) or final outcomes (e.g. smoking cessation or mortality reduction) (SMD process=0.18, intermediate outcomes=0.07, outcomes=0.00).

This review demonstrated that the evidence was, however, inconclusive due to substantial heterogeneity on the effect of P4P between and within schemes, discrepancies in the effects of P4P in studies with different evaluation design (as poorly evaluated P4P studies demonstrated a higher effect: SMD=0.15 compared to 0.08 for RCTs), and paucity of RCTs (and or high quality studies with adequate control groups) to make reliable and convincing conclusions. In the following subheadings, I briefly highlight the limitations of the review, before going on to discuss the shortcomings of the evidence: rigour of evaluation and heterogeneity.

Limitations

It might have been beneficial to assess the risk of bias (in detail) for all included studies in order to have a comparative subgroup analysis at an outcome level in the meta-analysis. In addition, the results of the review might have been further strengthened if all relevant primary studies were included in the meta-analyses. However, this was not possible due to poorly reported studies (and time constraints in contacting authors for more information). In the future, authors of P4P evaluation studies should endeavor to report more fully, providing raw data on sample size, and standard errors of estimates where possible.

2.6.1. Rigour of evaluation

There were very few RCTs that evaluated the effectiveness of P4P in healthcare. The majority of the studies were quasi-experimental and studies with no adequate control groups. The findings from this review suggest that estimates of the effects of P4P are

smaller in RCTs compared to quasi-experimental studies or evaluations without adequate control groups. This is in line with literature surrounding the effects of evaluation design in health care, which suggests that non-RCT studies are likely to over-estimate the effect of interventions (Shadish et al., 2002, Tilling et al., 2005). RCTs are likely to give less biased estimates than other evaluation designs (Booth and Tannock, 2014, Higgins and Green, 2011). Evaluations with no adequate control groups and to a lesser extent, quasi-experimental studies are prone to multiple biases and confounding factors such as additional funds or other quality improvement strategies (in the P4P context), which makes it difficult to disentangle the effects of P4P (Shadish et al., 2002). This could create ambiguity among researchers regarding the extent of attribution of success of P4P, as potential confounders may have influenced the observed effects.

It is understandable that the contexts in which P4P is implemented, a randomised controlled trial is not always feasible, especially in cases where the schemes have already been implemented without control groups. However, a rigorous evaluation is still possible using well conducted quasi-experimental designs with adequate controls, such as interrupted time series and controlled before and after studies (Tilling et al., 2005). This gives more credibility to the inferences made about cause and effect relationship, provided careful attention is paid to address potential confounders (Shadish et al., 2002, Higgins and Green, 2011).

Drawing from the findings, the important point to note is that rigorous evaluations are needed to generate more convincing evidence on the use of P4P in health care. The evidence of the effects of P4P will benefit from improved evaluation of future schemes using RCTs where possible and well-designed quasi-experimental studies (where RCTs are not feasible). In addition, exploration and analysis of current literature should take into consideration the influence of evaluation design on evidence of P4P.

2.6.2. Heterogeneity

The findings of this review provided evidence of substantial heterogeneity in the effects of P4P in health care, which is one of the reasons researchers struggle to make sense of the available evidence. The statistical heterogeneity observed in the findings of this review refers to the variation in the effects sizes of P4P in health care beyond that attributable to chance (Higgins and Green, 2011). Researchers have suggested that this

is mainly due to the differences in design features of the P4P schemes i.e. 'the problem of combining apples and oranges' (Van Herck et al., 2010, Witter et al., 2012, De Bruin et al., 2011)**.** These schemes vary in terms of incentivised activities (quality measures: process vs outcomes), basis for providing incentive, recipients of incentive, size of incentive, frequency of payment, and duration of the incentive. For example, P4P may have a different effect on process measures e.g. smoking cessation advice, as compared to patient outcomes e.g. smoking cessation rates (as demonstrated in section 2.4.2). In the same way, P4P schemes incentivizing health service providers with large sized incentives may have produced better effects than P4P schemes giving small sized incentives.

Others suggest that other sources of variation may also explain the observed heterogeneity (Eichler and Levine, 2009, Ssengooba et al., 2012). These include: (1) context in which the P4P scheme is implemented (health systems, increased funding, and complexity/degree of implementation of the scheme), (2) how well the programme has been piloted: use of baseline measurement, setting of targets, degree of preliminary work done, (3) evaluation design (as demonstrated in section 2.5), and (4) area of intervention e.g. prevention (immunizations and screenings), and disease management and outcomes (blood pressure monitoring and reduction) (Eichler and Levine, 2009, Ssengooba et al., 2012). Furthermore, P4P may work differently in contexts where preliminary work on what works in that context has been done, as opposed to contexts where no groundwork was done.

Emerging P4P literature is beginning to place more emphasis on careful consideration and selection of design choices, informed by the knowledge of 'what works' theoretically and empirically (Epstein, 2012, Eijkenaar, 2012, Scott et al., 2009). However, the evidence of what works is limited and has not been explored in a systematic way, despite the presence of theoretical and empirical evidence in this area. Therefore, in order to make sense of the available evidence and to explore 'what works' in P4P, it is important to examine sources of heterogeneity and the possible ways to address or investigate them.

2.6.3. Exploring heterogeneity

The heterogeneity observed in the findings of this review may be due to clinical difference between the schemes (for example, differences in elements of the P4P

intervention, contexts, or outcomes considered), or methodological differences (such as evaluation design) (Morton et al., 2004). The possible ways of systematically exploring this heterogeneity include a subgroup analyses and or a meta-regression, which both can be used to determine which elements might be contributing to the intervention effect and the extent of their influence on the intervention effect (Higgins and Green, 2011, Engels et al., 2000).

Subgroup analyses involve stratifying the studies into homogeneous categories, and fitting a separate effects estimate e.g., of the pooled standardised mean difference in each category to produce average effects for each category (Higgins and Green, 2011).

For example, pooled estimates of studies by evaluation designs (RCTs, quasi-experimental studies, and studies with no control groups) shown in section 2.5. Meta regression (an extension of the subgroup analyses) quantifies the relationship between the sizes of effect of the interventions and the categories or characteristics of the studies using weighted regression-based techniques. Thus, allowing multiple categories or characteristics to be investigated simultaneously (Morton et al., 2004).

In order to explore and investigate heterogeneity in the evidence of the effects of P4P using subgroup analyses and (or) meta-regression, a clear and reliable framework to categorise these studies into practical and coherent groups is required (for example, coherent categories based on designs, evaluation designs, and contexts of P4P). While there is clear and established literature for the basis of categorising studies by evaluation design, literature on categorisation of P4P designs or contexts is less clear and lacking structure (lacking a systematic approach). There is no reliable framework available in literature to categorise these studies into informed categories that would allow exploration of heterogeneity in a practical way. Therefore, a logical step in making sense of the evidence of P4P was to develop a reliable framework to categorise these schemes to facilitate exploration of heterogeneity.

2.7. What this chapter adds

This chapter reviewed and appraised the evidence surrounding the effects of P4P schemes in healthcare. Despite the popularity of financial incentives in healthcare, their effectiveness on improving quality of care are shown to be modest, fragmented (heterogeneous), and evaluations are of poor quality and often lacking adequate control groups. Improving the evidence of the effect of P4P requires more rigorous evaluations of future P4P schemes and exploring heterogeneity in the current evidence, while taking into consideration the evaluation designs of the P4P schemes. This forms the basis of the subsequent chapters of this thesis. In the next chapter, I focus on the development of a framework (typology) to systematically categorise P4P schemes in healthcare to help facilitate exploration of heterogeneity between the schemes, before going on in subsequent chapters to systematically explore heterogeneity by conducting a meta-regression analyses.

Chapter 3 Developing a framework to categorise P4P schemes (A P4P typology)

3.0. Introduction

The review and appraisal of evidence on effects of P4P in the previous chapter demonstrated the need for an informed, systematic, and reliable framework to categorise P4P schemes in order to systematically explore heterogeneity in P4P studies to allow researchers to make better sense of the available evidence to inform the use of P4P in health care. This is because heterogeneous evidence could be better explained by considering the difference in design, implementation, and contexts (Eichler and Levine, 2009, Meessen et al., 2006, Soeters et al., 2006, Epstein, 2012).

While it may be worthwhile to develop a common framework for contextual and implementation variables, it would be difficult for three reasons: (1) they vary extensively across P4P schemes, (2) they variables are less readily reported in evaluated P4P studies, as opposed to design features, and (3) they are setting specific and qualitative work or surveys is often required in that context to understand it (which is explored in the second part of this thesis).

Some researchers have reviewed and described common design features of P4P (e.g. what kind of incentive: fines or bonuses) based on the current empirical literature (Stockwell, 2010, Eijkenaar, 2013). However, consideration of the literature was not approached in a systematic way because both studies were informed only by empirical evidence, which has been shown to be substantially heterogeneous (see chapter two). For example, it seems logical that incentives could either be fines or bonuses (or both). However, some P4P schemes implementing fines or bonuses have resulted in improved quality, while other programmes have not had any desirable effect. Furthermore, literature suggests that P4P design features may interact with one another. For example, P4P schemes that have a certain combination of variables might be more effective than others (Stockwell, 2010, Eijkenaar, 2013). There are however no studies in literature combining design features in a sensible framework to possibly reflect this.

Despite the potential importance of design features of P4P, research that focuses on the impact of these design features on effectiveness of P4P is rare. There is, however, a range of theories and concepts in behavioural science and economics, which can be used to understand better how clinicians respond to incentives. This can contribute to the understanding of the impact of these design features on the effectiveness of P4P. However, this knowledge is unstructured and rarely explicitly used in the design of P4P schemes. Designers of P4P schemes rarely make clear the theoretical basis and justification for the designs of the schemes. Similarly, evaluations of some of these schemes do not relate the findings to the design features of the P4P schemes under scrutiny (Mehrotra et al., 2010, Doran, 2008).

A way to harness and structure this knowledge into a more reliable and practical framework to explore heterogeneity and 'what works' in P4P is through the development of a theoretically informed typology. This is a systematic form of classification (often having their roots in theory) that helps in description, identifying patterns, aids in reducing complexity in the empirical world, and theory testing (Bailey, 1994).

This chapter focuses on the development of a theoretical typology that categorizes P4P schemes based on their design features in order to attempt to sensibly describe, develop, and compare the results of evaluations of P4P schemes. This typology could aid in the description and comparison of P4P schemes. It would also aid in systematic exploration of the influence of the design features on the effectiveness of these schemes; whether certain design choices lead to more effective schemes or not (as theory suggests).

3.1. Aims and Objectives

The aim of this part of my research was to construct a P4P typology, which could be used to explore the heterogeneity in evaluations of P4P. In order to construct this typology, I set three objectives.

- First, to construct a typology based on design features of P4P schemes and relevant theories.

- Second, to test and refine the typology to ensure that it can be used to efficiently identify and categorise P4P design features on the basis of intervention descriptions.
- Third, to label the types in the typology to ensure that it is an efficient tool that would aid in systematic exploration of heterogeneity in results of evaluations of P4P schemes.

3.2. Methods

There are two common forms of typology namely, the ideal type and the constructed type (Bailey, 1994). For this research, I developed a constructed typology as opposed to the ideal type because the constructed type allows for categorisation, comparison and analysis of empirical cases (which was key in exploring heterogeneity in the effects of P4P). The ideal typology on the other hand is developed with only the most prominent features of the phenomenon (in their purest states), in which empirical cases might not exist, thereby making an ideal typology less relevant to cases in literature, and not helpful in exploring heterogeneity (Bailey, 1975; Bailey, 1994).

The constructed type is defined as "*a purposive, planned selection, abstraction, combination, and (sometimes) accentuation of a set of criteria with empirical referents that serves as a basis for comparison of empirical cases*" (McKinney, 1966).

The methods I used to develop the constructed typology (and the other two objectives of the study) involved six major steps (See Figure 3.1 for a summary of the steps taken). First, I identified nine potential design features to be included in the typology from the literature. Second, I identified and explored theories relevant to these design features. Third, I gave standard descriptions of the variables within the P4P design features and combined them in a multidimensional space, which resulted in 96 possible types. Fourth, to assess the functionality of the P4P typology, I piloted it on descriptions of P4P studies from a randomly selected review of the literature. After this, I reduced and refined the typology in order to make it more manageable and to meet the standard of what a good typology should be according to literature (Bailey, 1994). Finally, I labelled the types in the typology to ensure that it could be used efficiently in systematic exploration of heterogeneity in results of evaluations of P4P schemes (Bailey, 1994). Each step is described in more detail in the subsequent sections.

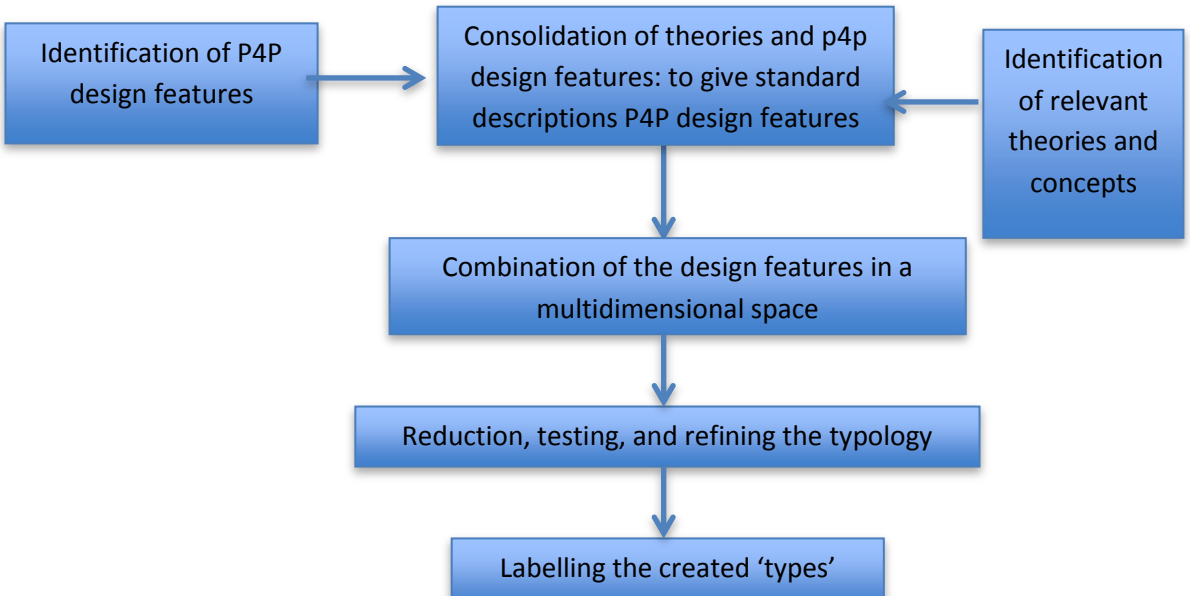


Figure 3.1 Flow chart of methods used to develop the P4P typology

3.2.1. Identification of potential design features

A few studies have identified and described design features in P4P literature (Stockwell, 2010, Eijkenaar, 2013). Both studies identified nine similar design features (patterns):

- Who receives the incentives (hospitals, individual clinical staff, teams)
- Type of payment (non-monetary or monetary)
- Type of incentives (fines or bonuses)
- Size of incentives
- Payment mechanism (absolute or tiered thresholds)
- Method of payment (is the incentive payment coupled/uncoupled from usual salary or budget)
- Performance measure (absolute or relative measure)
- The domain of performance measured (process, structures, and outcomes).
- The time lag between the measurement of performance and payment of the incentive

I used these nine design features as my starting point. Furthermore, to ensure comprehensiveness of the design features, I scanned 138 studies describing and evaluating P4P schemes identified in chapter two (see Appendix A4 and A5). There were no apparent design features that appeared to be left out. In the following sections, I

explore these design features in more detail with theory and empirical cases, where I also give standard descriptions and refining categories of P4P design features.

3.2.2. Identification and exploration of relevant theories to P4P design features

Potential relevant theories were identified from literature identified by a review using keyword searches (behavioural economics, behavioural theories, incentive theories, economic theories) on three databases (PubMed, PsycINFO, and EconLit) (see Appendix B1 for search output). I also found relevant theories from references in the identified P4P literature.

Since there were various economic theories that relate to how individuals respond to incentives. I purposively selected theories that had been used to describe or understand behaviour of clinicians or health service professionals in the context of P4P.

In the subsequent sections, the identified theories and concepts were consolidated and applied to the identified design features, to explore, explain, or predict human behaviour relative to design choices.

3.2.2.1. Who receives the incentives?

P4P schemes involve payment of incentives to individual health professionals, health institutions, clinical teams, or other organisations providing health care (Gross et al., 2008, Tahrani et al., 2008, Chee et al., 2007).

Trisolini (2011) suggests that P4P schemes with incentives targeted at groups (such as clinical teams, health institutions/facilities, and organisations) rather than individual health professionals are more likely to result in improvements in quality of care. This is because organizations are capable of setting up good management structures that could be strong enough to elicit a change in behaviour. For example, incentives paid to groups are used in different ways (purchase of equipment or hiring more staff), which could lead to improvements in quality and performance (Basinga et al., 2011, Vergeer and Chansa, 2008, Ssengooba et al., 2012). This argument is in line with the organisation theory, which proposes that payment of incentives to groups rather than individuals are more likely to bring desired effects because organisations are capable of promoting behaviour change in employees through a wide range of strategies e.g. better structures, improved supervision, enacting stricter guidelines and policies etc. (Stewart, 1998). However, this is dependent on the level of the managerial or organisational skills. So

paying incentives to the organisation as opposed to individuals is likely to result in higher impact provided managerial skills are good.

A counter argument against paying incentives to groups compared to paying directly to individual health care professionals is what is known as the 'free rider' theory (Kidwell and Benneth, 1993, Cooper et al., 1992, Shepperd, 1993). The free rider theory suggests that individuals are more likely to undersupply the service being incentivised when they share responsibility of providing that service because they might feel that the payment might be shared equally rather than based on individual contributions. Therefore there is less incentive to try to perform better because as an individual, one can 'get away with' not changing behaviour and still receive the incentive (free-riding). Furthermore, some researchers suggest that paying individual health professionals could create competition among the individual providers, which could heighten adverse consequences such as hoarding of knowledge and skills, thereby undermining the promotion of team based care, which is believed to be important to improving the quality of care (Town et al., 2004). This 'free rider' problem however, can be somewhat ameliorated if organisational structure and management is strong. For example, management can implement stricter policies/guidelines or create an opportunity for individual clinicians to earn part of the incentive received based on their contribution.

3.2.2.2. What type of incentive: Non-monetary or Monetary?

Financial incentives can be in the form of money or gifts, also known as monetary or non-monetary incentives. Monetary incentives are the most commonly used incentives in P4P programmes (Eijkenaar, 2013).

Non-monetary incentives (Justifiability and Evaluability Theories)

Two theories known as evaluability theory and justifiability theory support the use of non-monetary incentives as a more effective form of incentive to change behaviour compared to the use of monetary incentive (Jeffery, 2010, Hsee and Zhang, 2010). Evaluability theory suggests that some non-monetary incentives are more difficult to attach a monetary value to (Hsee and Zhang, 2010). For instance, an incentive of an all-expense paid holiday to Hawaii is likely to be considered a pleasurable experience. The things that come to mind are good weather, beaches, luxury, and room service. These positive attributes are difficult to 'put a price on' and thus may be ascribed a higher

value than the cash equivalent of the all-expense paid holiday. While justifiability theory proposes that when individuals are able to justify an award/incentive, there is a greater motivation to achieve the award or incentive (Jeffery, 2010). A number of non-monetary incentives are perceived as luxuries that individuals would not usually purchase with their own cash, primarily because they cannot justify the purchase. Behaviour change then becomes an effective way of acquiring something that someone could not normally justify purchasing with their own money. Non-monetary incentives can have both evaluability and justifiability. This means that they allow the earner/payee to justify the consumption of the incentive, thereby motivating the receiver to change behaviour and earn something that might have otherwise been more difficult to buy with money.

In 2007, Crifo and Diaye developed an economic model using both monetary and non-monetary incentives in which they gave the agents non-monetary incentives (an incentive of what matters to the agent was assumed, as it is important for the principal to be aware of the agents' preferences). They showed that non-monetary incentives are as important as monetary incentives in changing behaviour. They also found that there is the possibility of reward inflation occurring if agents are continually rewarded with money i.e. the agents can get adjusted to the incentives and might no longer be motivated to change behaviour by it. Crifo and Diaye (2007) also argued that non-monetary incentives are memorable, thereby creating a possibility of greater behavioural change compared to monetary incentives which might be combined with usual pay checks or salaries, making it less memorable. Despite this, non-monetary incentives are rarely used in P4P schemes.

Monetary incentives (Expectancy Theory)

Expectancy theory, proposed by Vroom (1964), suggests that: "*individuals act to maximize expected satisfaction with outcomes".* The theory assumes that individuals' motivation to work is dependent on two factors: (1) the expectancy about the relationship between effort and a particular outcome and (2) the valence (attractiveness) of the outcome. These two factors are believed to create the motivation that will lead to individuals changing their behaviour towards achieving the desired outcome. Vroom argued that money has valence because it is effective in acquiring things desired by individuals such as material goods of their choice. Therefore, money might be more effective in driving behavioural change compared to non-monetary incentives. This

might be particularly true for individuals whose salaries are barely sufficient. In such cases, money might be a more effective driver of behaviour change than non-monetary incentives.

It is also possible that giving money as incentives might be more effective in driving behaviour change compared to non-monetary incentives because the different agents might have different material goods that they desire and it might be almost impossible and challenging to determine what is of material good that is of value to every agent involved in the particular P4P scheme. A particular non-monetary incentive might be of value to one agent but might be of little value to another agent within the same P4P scheme (Furnham and Argyle, 1998). Furnham and Argyle further argue that money has symbolic value due to its perceived relationship to prestige, status, and other factors. Monetary incentives may have higher valence than non-monetary incentives, depending on the relative payment schedules.

3.2.2.3. Type of incentive: Fines or Bonuses (Loss Aversion Theory)

There are two forms of financial incentive used in P4P schemes: fines and bonuses. Kahneman and Tversky developed The Loss Aversion Theory in 1979 and it refers to the tendency for people to prefer to avoid losses compared to acquiring gains. Adam Smith said, "*Pain... is in almost all cases a more pungent sensation than the opposite and correspondent pleasure. The one almost always depresses us much more below the ordinary or what might be called the natural state of our happiness, than the other ever raises us above it*" (Smith, quoted in Maynard, 2012, p.8). From this perspective, fines are more likely to motivate behavioural change than bonuses. In addition, P4P schemes, which use fines, might be more sustainable compared to P4P programmes that only use bonuses because they could be less costly (Pope, 2011). This is supported by some experimental studies that have demonstrated that losses could be twice as effective as gains/bonuses in eliciting a positive behavioural response to increase performance (Tversky and Kahneman, 2004). The implication of this in P4P in health care is that practitioners will be more inclined to change behaviour or increase performance if they think they might lose something rather than get a bonus.

Despite theory suggesting that the use of fines compared to bonuses are more effective in implementing behaviour change, bonuses are still the most common form of incentives used in P4P programmes in healthcare. One of the reasons might be because fines can lead to a loss of intrinsic motivation in clinicians (Mehrotra et al., 2010,

Kinoti, 2011). This is because the use of fines might annoy clinicians who have altruistic purposes and they might feel they are not being appreciated for their job. In addition, it might be very difficult to implement the use of fines in certain contexts with weak health systems and poor governance because trade unions might object and create opposition leading to disruption in health services. For example, a country such as Nigeria where the union of doctors call strikes for several days because of delay in salary payments would most likely be against the use of fines (Hargreaves, 2002). The political challenge of using fines means that the few P4P schemes that implement fines usually include an opportunity to earn bonuses as well.

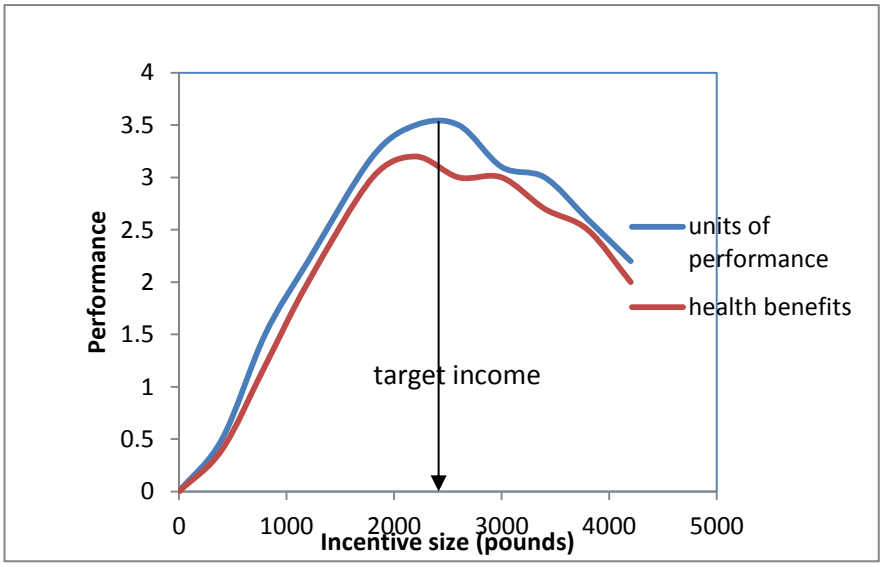
3.2.2.4. Size of incentives (The Target Income Hypothesis)

The most common form of description of size of incentive is the amount of money relative to the clinicians' salary, usual budget of the health institution or anticipated payment regarding the health service(s) in question, which ranges from 0.5% to 100%. Other P4P programmes simply report the size of incentive in absolute terms as the actual amount earned in the appropriate currency (Pope, 2011).

Hahn (2006) suggested that the effect of an incentive might be influenced by its size compared to the recipient's usual salary, budget, or anticipated payment (Hahn, 2006). Consequently, it is very important to specify appropriate reward levels as the incentives might be too small compared to the usual salary. This may produce little or no change even when the objectives are measured accurately and are fairly evaluated. Alternatively, the incentive might be too large resulting in paying more than necessary to bring about the desired behavioural change, thereby making it less likely to be cost effective.

As the size of the incentives (fines or bonuses) increase (everything else being equal), people may be motivated to work harder to reach the set targets. However, the relationship is likely to demonstrate diminishing marginal returns (see Figure 3.2). After a certain point, increasing the size of incentive will not bring about the required behaviour change and this will lead to a waste of resources (Mold et al., 2010).The size of incentive also raises the question of cost-effectiveness of P4P schemes, as money spent on the incentive might not be justified by the potential benefits in patient outcomes resulting from behaviour change.

There is some evidence that "*physicians have a desired income that they want to achieve whenever their actual income is below that income*" (Evans, 1974, p.162). This is commonly referred to as the Target Income Hypothesis and if this hypothesis is valid, it means that increasing the size of incentive would result in an increase in performance only until the clinicians reach their target income after which, increasing the size of incentive would reduce performance or not increase it any further (see Figure 3.2).



P4P interventions might not be cost effective beyond the point where increasing the size of incentives only gives a marginal increase in behaviour change and health outcomes

Figure 3.2 Illustration of physician target income relative to performance

Desquins and colleagues in 2009 further found that 80% of physicians would be willing to perform better to reach a target income, a finding supported by a number of other researchers (Folland et al., 1993, Rizzo and Blumenthal, 1994).

Though the Target Income Hypothesis provides some understanding as to how clinicians behaviour or performance might vary relative to size of incentive needed to achieve their target income, the size of incentive needed to reach this target is likely to differ according to context, as the target income would be related to the cost of living of that region.

Providers might judge the adequacy of the size of incentive to change behaviour based on the amount or difficulty of reaching the performance or targets that is required of them. This could mean that what constitutes an adequate incentive to improve performance or reach a certain target in a certain area of healthcare (e.g. immunisation rates) is likely to differ across contexts. For example, quarterly incentives of £100 to increase childhood immunisation rates to clinicians in developing countries (where

salaries are often insufficient) might be deemed sufficient, while clinicians in developed countries might not view the incentive as something worth doing. It is therefore important to have an idea of the average target income of the clinicians participating in the P4P programmes as this could contribute to the effectiveness of P4P schemes. The average target income could be determined through preliminary surveys before the implementation of the P4P schemes (Rizzo and Zeckhauser, 2003), which could help P4P scheme designers determine the appropriate size of incentive to ensure cost-effectiveness of these programmes.

The information that could help in meaningful categorisation of size of incentive in P4P schemes is limited and even if readily available, might be difficult to structure using absolute amounts. The closest way to subjectively capture the size of incentive is to use a function relative to the clinicians' usual salary/reimbursement. There are no set cut-offs in theory as to what size of incentive is adequate to change behaviour, I suggest arbitrary cut-offs guided by some empirical evidence.

The size of incentives in P4P schemes in healthcare tend to range from 0.5% to up to 100% increase in individual salary or institution budget. Studies indicate that most P4P initiatives with less than 5% increase/decrease in payment had no statistically significant effect on the performance indicator compared to P4P schemes with above 5% in salary or budget (Chen et al., 2011, Pope, 2011). For example, a P4P programme in the USA which rewarded hospitals with small incentives (less than 5% of usual reimbursement) showed no significant difference in 30-day mortality rates for conditions that were explicitly linked to incentives (acute myocardial infarction and Coronary artery bypass grafting) and among conditions not linked to incentives (congestive heart failure and pneumonia) (Jha et al., 2012). While the QOF programme in the UK with relatively larger incentives (up to 40% of usual reimbursement) showed significant reduction in 30-day mortality rates for pneumonia and non-significant reductions in myocardial infarction and heart failure (Sutton et al., 2012). Based on these examples and following the rationale of the target income hypothesis, for the purpose of developing the P4P typology, I proposed 3 categories of size of incentive namely: small (<5%), medium (5-10%), and large (>10%). I hypothesized that small incentives were likely to be the least effective in driving behaviour change, and the medium sized incentives and large incentives were likely to be more effective (although

large incentive sizes may not bring about the commensurate amount of behaviour change for the value of money) (see figure 3.2.).

3.2.2.5. Payment mechanism (absolute or tiered thresholds): The Goal Gradient Theory

There are two main kinds of payment mechanisms in P4P schemes. The first kind involves a payment for achieving an absolute target (e.g. 70% fully vaccinated children) and the second kind involves different payments triggered at tiered targets (e.g. 60%, 70%, and 80%) or a sliding scale.

Goal Gradient Theory (Hull, 1932) predicts a greater positive behavioural response if there are a series of stepped target thresholds to meet, for example paying increasing incentives for achieving a 65%, an 80%, and a 90% performance threshold rather than one target e.g., a payment for an absolute 80% increase in performance (Mehrotra et al., 2010). Therefore an incentive payment made for reaching an absolute threshold or a single target (only if you reach the target) might be less effective in changing behaviour compared to an incentive payment, which increases as you achieve higher thresholds. This might be because individuals in an incentive programme intensify their efforts as they sense that they are getting closer to their target goal (Heath et al., 1999). Furthermore, individuals are more likely to be motivated when the target goals appear to be 'realistic'. If the target goal is far from the baseline, it might be viewed as unrealistic or unachievable to the individuals who may see no reason to try to meet the target, as they are likely to fail. In addition to the risk of not getting any payment, this might also reflect the perceived cost to them of achieving it, when considered, the expected benefit might be too low.

There is also the risk of loss of interest or motivation when the target goal is achieved (this might contribute to an understanding of why some successful P4P programmes seem to reach a plateau or even dip after sometime) where there is just one target (Campbell et al., 2009). This suggests that having tiered targets or a sliding scale might challenge the clinicians to a continued effort in improving performance.

3.2.2.6. Method of payment (coupled or decoupled from usual reimbursement): Mental accounting theory

The method of payment in incentive programmes can be coupled or decoupled from salary or income. For example, increasing the usual salary of £2000 to £2080, compared

to making a separate payment of £80. Mental Accounting Theory states that individuals divide their current and future assets into separate, non-transferable portions and will assign different levels of utility of each asset in each group (Thaler, 1999). This predicts that people will value incentives more highly if not coupled with the usual salary (Jeffery, 2010). Applying this theory to P4P schemes means that it is likely that individuals would place more value on incentives not coupled with the usual salary compared to incentives coupled with salary (even though they might be the same amount). Decoupling the incentives from usual reimbursement might be administratively more burdensome. It could however be worth the additional cost, if it contributes to the success of the P4P programmes.

3.2.2.7. Performance measure, domain of performance, and time lag: Risk Aversion Theory

The last three identified design features; Performance measure, Domain of measurement and Timing of payment, share a common relevant theory known as the risk aversion theory, which is explained below.

Risk Aversion Theory predicts the behaviour of individuals when exposed to risk or uncertainty. An individual is less ready to accept an uncertain contract or agreement compared with another contract with a more certain consequence (Arrow, 1965). In P4P schemes in healthcare, when there are several elements of risk or uncertainty of not getting paid the anticipated or desired amount. This could reduce the impact of the scheme.

The riskiness of a scheme may be explained in terms of three design features:

a. The degree to which the target takes into account achievement in absolute terms or relative to how others perform (performance measure: absolute or relative measure)
b. The degree to which the person/organisation being incentivised can directly control the performance measured (domain of performance measured)
c. The confidence the agent has in being paid if they do achieve the relevant target. This might reflect reliability of measurement and (or) time lags between

performance and payment of incentive or belief in the administration[3] of the scheme (timing of payment and reliability of measurement).

These design features are explained below.

Performance measure (absolute and relative measures)

Absolute measure of performance is when an incentive is paid for quality improvement, not dependent on other providers (e.g. incentive paid per patient immunized). A relative measure on the other hand is when incentive is paid for attaining above a specified rank relative to other providers (e.g. incentives paid to clinicians for exceeding the bottom quartile immunisation rate). Relative performance measures create greater uncertainty for health service providers because their achievement depends also on how well others do. Hence providers may be less motivated to invest in improving performance. On the other hand, P4P schemes where absolute performance measures are used are more likely to be more effective. For example, paying health care workers per number of children they immunize against paying those above the median performance. This is because the clinicians or clinical teams are more certain of earning the incentive if they improve their performance.

Domain of performance (to what extent is it within the control of the provider)

The domain of performance measured is also important to the health service providers in assessing the risk involved in the P4P scheme. The domains of performance that could be measured include:

- Structure: this involves the use of resources to deliver care e.g. information technology (IT) including personnel, facilities, IT, and materials)
- Process: involves performing routine mechanical operations, specific tasks or recommended treatments e.g. periodic cholesterol screening, immunization.
- Intermediate outcomes: Intermediate outcomes are the steps or outcomes between the change in behaviour and the final health outcome (e.g. reduction in cholesterol levels, reduction in blood pressure). This means that the progress made towards achieving the intermediate outcome is likely to improve final outcomes (if evidence-based).

[3] In developing this typology, I considered only the time lag dimension because administration of the scheme is highly contextual and such information is not readily available in published studies. One might need to conduct in-depth interviews to explore the belief of health workers in the administration of each P4P scheme.

- Final outcome: these are effects on the quality and length of life and wellbeing of people (e.g. reduction in mortality and morbidity rates).

The structural, process and intermediate outcome domains of performance are often seen as more easily achievable or at least more under the control of the healthcare organization or clinician, compared with the health outcome measures which are influenced by a variety of other factors (including the patients) and so less under the control of the health services. Whilst some process and intermediate outcomes might partly depend on patient behavioural change as well (e.g. getting patients to show up for cancer screening in hard to reach areas), some outcome measures might be more dependent on external factors, and so might be viewed as more difficult to achieve compared to processes and intermediate outcomes. In addition, process and structural measures may be more sensitive to quality differences than outcome measures, because poor health outcomes do not always mean there is a quality problem (Oxman and Fretheim, 2009a). For example, if a clinician is to be incentivised based on a reduction in cardiovascular mortality rates, the positive efforts by the clinician may be confounded by other lifestyle choices of the patients e.g. exercise, diet and other factors outside their direct control for example, reduction of high blood pressure. Therefore, improvements in quality of care might not necessarily translate to improvements in patient outcomes.

For this reason, P4P interventions that focus on the final health outcome domain of performance might be perceived as higher risk (greater uncertainty in earning the incentive) and might not be as effective as P4P interventions that focus on structure and process domains of performance, because even though the healthcare professional might have changed behaviour, this might not necessarily reflect in the final outcomes as the patient might also need to change behaviour to obtain the desired final outcome.

Timing of payment (and frequency of payment)

Timing of incentive payment ranges from monthly to annually. When the time lag between the measurement of performance and payment of incentives is longer, it can create some uncertainty, particularly in countries with a track record of corruption, political instability and/or poor administrative infrastructure. This uncertainty in payment might reduce the motivation to improve performance. In addition, shorter time lags between payments may indicate smaller more frequent payments, which are more

likely to motivate a higher behavioural response in an individual compared to a one-time lump sum incentive (Thaler, 1985).

Furthermore, according to Price (1993), individuals often exhibit time preference (or time discounting) where *"happiness now is worth more to me than happiness next year"*. Consequently, individuals perceive incentives received soon after the behavioural change as having more value than incentives received in the future, a phenomenon called 'pure time preference'. Loewenstein and Prelec (1992) also suggest that time lag between measurement of performance and the receipt of the incentives could affect behavioural response. Individuals tend to ask themselves; is there anything that I could do now that will bring me immediate rewards instead of what I could do now that would reward me in a years' time? Consequently, P4P designs with minimal time lags between provision of care and receipt of incentive might be expected to produce greater behavioural response.

Arguably, in some P4P schemes, it may take months or even a year or more to collect and validate performance data. This means that payments based on performance might not be made until several months or a year after the delivery of care. People might be relatively motivated to change their behaviour even if the payment is a year away (after measurements of performance) for very large incentives (e.g. up to 40% increase), which implies that these design features might interact with each other to influence the impact of the scheme. This is another advantage of developing this typology, as each type (category) from the typology developed will be a unique combination of the dimensions of the design features of P4P.

There are no clear descriptions in theory or literature of what constitutes a 'long' time lag of payment. Previous studies have suggested that monthly, bi-monthly, or quarterly payments constitute shorter time lags, while payments after four months constitute a long time lag (Eijkenaar, 2013, Stockwell, 2010). An RCT conducted in the USA comparing annual payments of incentives to individual physicians totalling $5000 to quarterly payments totalling the same amount for quality improvements in treatments and outcomes of diabetes, cancer screening, and smoking, found that the difference in effect observed between the two arms were not as a result of the timing of payment, but rather likely as a result of difference in the timing of reporting of outcomes (Chung et

al., 2010). The quarterly payment group had to present reports every quarter to be approved for the payment of the incentive, which might have contributed to motivating the physicians in this group compared to submitting yearly reports. Therefore, based on these observations, for the purpose of categorisation in this typology, monthly to quarterly payments were considered as short time lags, whereas, payments made after four months were considered long time lags.

Reliability of measurement of performance

Similar to the timing of payment, the reliability of measurement of performance could also affect the confidence that the agent (health service provider) has in being paid if they do achieve the relevant target. Clinicians are likely to perceive the potential of earning the incentive as out of their control and uncertain if the tool for measuring performance is not reliable. Clinicians will most likely not change their behaviour even if the target is relatively easy to reach because they might think that the measurement tool might not accurately reflect behaviour change, thus, reducing motivation to change behaviour if it is unlikely to be measured correctly.

The reliability of measurement as perceived by the health service providers is important in exploring risk or uncertainty of not earning the incentive. However, it is difficult to judge this, as it depends on the views of the clinicians in the particular context, which are not commonly reported in P4P evaluations. It is, however, still important and should be explored in the implementation contexts when designing a scheme.

3.2.3. Combining the features in a multidimensional space

Following the exploration of the identified design features with theory, I proceeded to define a set of criteria for the variables in each design feature identified. This is presented in Appendix B2. This was done in preparation for the combination of these features in a multidimensional space to create the typology (see Appendix B3). This combination resulted in a typology of 108 possible types[4]. There were too many types, and therefore considered to serve little or no use as an analytical tool (overly simplified with too many categories). To deal with the large number of possible types, I followed

[4] Formula of estimating number of possible types: assuming all 9 variables are dichotomous= 9^2 (81) but since there is one variable with 3 categories (81 + 27 (adjusting for the 1 category left out)= 108 possible types.

Bailey (1994) suggestion that the typology could be reduced to ensure greater efficiency, manageability, and usability of the typology. The methods for reduction of this typology are discussed in the next section.

3.2.4. Reducing the Typology

I reduced the typology through (1) dichotomization of variables, which involved merging the categories within design features so that there are just two categories, (2) pragmatic reduction, which involved combining or compressing design features with the same underlying theory or concept, and (3) rescaling, which involved the removal of less relevant features from the typology. These methods were selected over other methods such as functional reduction (involving the removal of cells/types in which empirical cases cannot be found) and subtyping (involving the use of subtypes/versions within in type) in order to minimise loss of information, while retaining the typology's ability to be useful as a descriptive and analytical tool. Functional reduction and subtyping leads to loss of information and might create a typology in which some P4P schemes might not be categorised suitably (Bailey, 1994, Elman, 2005). The reduction processes are presented in more detail in the following sections.

3.2.4.1. Dichotomization of variables (reducing categories with multiple variables into two)

There were nine design features in the original typology and each one had two variables apart from the design feature 'size of incentive', which had three variables: small, medium, and large. According to Bailey (1973), one of the first steps to consider reducing a typology is to dichotomize the variables in the features of the typology, where there are limited features that do not have dichotomized variables because this reduces the loss of information associated with this form of reduction.

For the design feature of size, I collapsed the three categories: small, medium, and large, into two categories: small and large, by categorizing medium size of incentive as large because the target income theory (discussed in section 3.2.2.4) suggests that medium and large incentives are more likely to have similar effects compared to small and medium incentives. This resulted in a reduced typology of 81 types, which was still too large.

3.2.4.2. Pragmatic compression

According to Bailey (1994), it is permitted to collapse certain variables in a typology if they have the same underlying characteristic or theory, which is called a pragmatic compression. For this typology there were three design features that shared the same underlying theory- Risk Aversion theory. These features were timing of payment, domain of performance measured, and performance measure, which were subsequently collapsed into one conceptual variable called the 'the perceived risk of not earning the incentive' (see Table 3.1).

If a P4P scheme pays incentives one year after measurement of performance and the performance measure is a relative measure, it is more likely that the incentive payment will be considered as uncertain by agents (the recipients) even if they work hard to change performance. In the 'low risk' category, clinicians perceive the incentivised entity as a performance target that is easily achievable and there is little or no risk of not getting paid the incentives if the target is achieved. There is actually a guaranteed payment as long as they improve their performance. On the other hand, in the 'high risk' category, there is no guarantee of payment because they may not be among the target rank to be incentivised or/and that the incentivised entity will be achievable which introduces an element of risk (Arrow, 1965).

The three variables that might influence the risk or the uncertainty of payment as perceived by the recipients of the incentives is illustrated in this hypothetical case: "If I spend five more minutes with Mrs. Jones discussing the advantages of a mammogram I could increase my overall mammogram rate to X% (domain of performance: process measure, within clinicians control) which might put me in the 75^{th} percentile for my peer physicians (performance measure: relative measure) and possibly lead to an incentive at the end of the year (long time lag). This situation constitutes a high level of risk or uncertainty for the individual earning the incentive, as opposed to if the performance measure was absolute (e.g. payment for reaching a set target, which is not dependent on how others perform) and if the time lag was shorter (e.g. monthly payments), which constitute lower risk or higher guarantee of earning the incentive if targets are met.

Following this rationale, I set a dichotomous variable namely 'Perceived risk of not earning the incentive (Risk): low risk and high risk. Individuals who perceive the risk or uncertainty associated with earning the incentive as low are more likely to change

behaviour because there is a higher guarantee about earning the incentive compared to when individuals perceive the risk associated with earning the incentive as high.

After compressing the three design features into one variable, I defined a set of criteria for the two categories (high and low) for this new conceptual feature: Perceived Risk of not earning the incentive (RISK). In Table 3.1, I present features that are associated with a higher risk under the heading high risk and I present the factors that are associated with a lower risk of not earning the incentive under low risk.

Table 3.1 Collapsed design features to form a conceptual variable 'Risk'

Categories of new variable (Risk) / **Collapsed variables**	**Low risk**	**High risk**
Performance measure	Absolute: incentive is paid for quality improvement not dependent on other providers e.g. incentive paid per patient immunized	Relative: incentive is paid for attaining a specific rank relative to other providers e.g. incentives paid to clinicians or hospitals in top 2 performing quartiles
Domain of performance measured	Within clinicians control: incentive payments are based on process and structural outcomes e.g. number of children immunized, routine measurement of blood pressure of patients every month	Not within clinicians control: payment of incentives to health providers for health outcomes e.g. reduction in blood pressure of patients or reduction in mortality rates from a specific disease
Time lag	Short time lag: Payment of incentives immediately after measurement of performance) or four months or less.	Long time lag: Payment of incentives more than 4months after measurement of performance

To ensure that the typology was mutually exclusive (implying that no P4P schemes falls into more than one type) and to ensure that as many as possible P4P schemes could be categorized (despite poor reporting/missing some features in some evaluation studies), I set a decision rule that: P4P schemes are categorized as low risk if it has two or more of the characteristics in the new elements of the conceptual design feature: short time lag, domain of performance within clinicians' control, and absolute performance measure. While P4P schemes should be categorized under high risk if it has two or more of the characteristics in the new elements of the conceptual design feature: long time lag, domain of performance out of clinicians' control, and relative performance measure.

3.2.4.3. Rescaling

Rescaling was the third method used to reduce the number of types. This involved reducing the number of variables used in constructing the typology (Elman, 2005), to exclude redundant features, i.e. features that do not vary that much in empirical cases (and do not contribute much to the usefulness of the typology), which made loss of information minimal (Elman, 2005, Bailey, 1994).

In this typology, design features that did not vary significantly across the P4P studies were considered redundant. This does not mean they are not important in the consideration of designing P4P schemes. However, for the purpose of the development of the typology, it was likely the redundant features might not contribute significantly to the analytical and theory-testing functions of the typology (exploration of heterogeneity). Although they might be worth reintroducing as the typology evolves in the future.

Three design features were removed from the typology, which were kind of incentive (monetary and non-monetary), method of payment (coupled and decoupled), and mechanism of payment (absolute and tiered threshold). This was because in current P4P schemes the main form of incentive used was money (monetary incentive), payment usually is decoupled from usual payments, and the mechanisms of payment for a majority of the schemes were absolute payments.

The final features included in the typology therefore were:

- Who to incentivise (individuals or groups)
- Type of incentive (fines or bonuses)
- Size of incentives (small or large)
- Perceived Risk/uncertainty of payment (low or high)

This resulted in a typology of 16 possible types shown in Table 3.2.

Table 3.2 P4P Typology

Type	Who received the incentive	Type of incentive	Size of incentive	Perceived risk of not earning the incentive (RISK)
1	Groups	Fines	Large	Low
2	Groups	Bonuses	Large	Low
3	Groups	Fines	Small	Low
4	Groups	Bonuses	Small	Low
5	Groups	Fines	Large	High
6	Groups	Bonuses	Large	High
7	Individuals	Fines	Large	Low
8	Individuals	Bonuses	Large	Low
9	Groups	Bonuses	Small	High
10	Groups	Fines	Small	High
11	Individuals	Fines	Small	Low
12	Individuals	Bonuses	Small	Low
13	Individuals	Fines	Small	High
14	Individuals	Bonuses	Large	High
15	Individuals	Bonuses	Small	High
16	Individuals	Fines	Large	High

3.2.5. Piloting the typology

The next stage of developing a typology was to test it to assess its functionality. A typology is good and functional if it meets a set of pre-defined criteria of (1) relevance: all the core components considered, (2) manageability: not cumbersome with only a few types, (3) ease of use: to be sure all types of P4P programmes can be categorized easily, (4) mutual exclusivity: this requires that there be only one type for each P4P programme, and (5) Comprehensiveness: whether all the empirical P4P programmes be categorized (Bailey, 1994, Elman, 2005, Tiryakian, 1968).

Some of these criteria such as relevance have been demonstrated through the process of developing the typology, which involved careful and thorough consideration of relevant theories and literature applicable to design variables of P4P. Similarly, the manageability criterion has been achieved through reduction of the typology to a few types to facilitate its use in analyses (as described in the previous section). Other criterion such as ease of use was demonstrated by volunteers who used the typology to categorise P4P schemes in health care (evidence of this is presented in chapter 4).

In this section I tested the typology to assess whether categories were mutually exclusive (only one type for each P4P scheme) and if all identified empirical P4P schemes could be categorised (comprehensiveness). To do this, I applied the typology

to P4P studies that were included in a randomly selected review (Eijkenaar, 2012) from the set of previously identified reviews in chapter two. Data were extracted from the identified studies in a standardised way using a uniform template (see Table 3.4) to obtain information on design features from each study, such as who receives the incentive, size of incentive, performance measures, domain of performance measured, timing of payment etc. I identified 14 P4P schemes from the selected review.

I was able to identify the relevant design features and categorise 6 out of 14 P4P schemes identified from the review. This was because some of the P4P schemes did not fit in any category because certain combinations of the design variables that were initially not considered were used. There were two specific examples: (1) schemes that combined the use of fines and bonuses and (2) schemes that paid incentives to individuals as well as groups or schemes that individuals benefited from the group payments.

I also found that it was difficult to categorize large schemes that incentivised for multiple activities or quality measures that include both processed and outcomes. This is because judging whether the perceived risk of not earning the incentive is 'low' or 'high' is dependent on whether three design variables, which include whether quality measures are within the clinicians' control or not: processes are considered more within the control of clinicians compared to outcomes. Having a combination of both processes and outcomes was not considered earlier on in the process of developing the typology. An example was the quality and outcomes framework (UK), where large incentives (total bonuses of up to 40% of clinicians' salary) are paid to hospitals (groups) on an annual basis (long time lag) for a range of indicators spanning across process and intermediate outcome measures (management and treatment of hypertension and diabetes, reduction of cholesterol and blood pressure, and patient experience) (Eijkenaar, 2012). Following these issues, I therefore set out to refine the typology further, as illustrated in the next section.

3.3. Results (refining, retesting, and labelling the typology)

3.3.1. Refining the typology

To refine the typology to ensure its functionality, I defined stricter criteria for the design variables in the typology, a method suggested by Bailey (1994). This method of creating new variables for the newly identified design blends was chosen because it retains the size of the typology, as opposed to the method of expanding the typology to accommodate the newly identified design variables (which makes it less manageable) (Bailey, 1975; Bailey, 1994).

As illustrated in Table 3.3, the criteria for judgement of fines (under the variable kind of incentive) was redefined and expanded to include presence or absence of bonuses in the same scheme (may or may not have bonuses). The criteria for bonuses remained strictly just the opportunity to earn bonuses (and no penalty or fine of any kind). This follows the rationale that individuals are still likely to maintain their loss aversion attitude as long as there is an element of fine or penalty and whether there is the potential to earn bonuses or not is not likely to deter the risk averse behaviour. Instead, a potential to earn bonuses in an incentive scheme where fines are implemented is likely to further boost behaviour change.

I also redefined the criteria for categorization of payment of incentives under groups to include instances where individuals may or may not benefit from the group payments (see Table 3.3). This is because when incentives are paid to groups as opposed to individual clinicians, one of the ways a good management system could motivate behaviour change is to provide individuals an opportunity to earn from the incentives received by the group (among other things such as increased supervision and stricter guidelines, as argued in section 3.2.2.1).

Second, in order to address the problem of difficulty in deciding whether the domain of performance measured is mostly within the clinicians control or not (since a few P4P schemes tend to use a mixture of processes and outcomes), I set a rule to make such decisions based on the measures that are predominant. For example, P4P schemes with four outcome measures and 20 process measures will be categorized as mostly under the clinicians' control, since there are more processes than outcomes, as opposed to ten

outcome measures and two process measures, which will be categorised as mostly out of the clinicians' control (in section 3.2.4.2). In addition, in the unlikely case where there are equal number of processes and outcomes, the outcome measures are likely to outweigh the process measures.

Table 3.3 Criteria for categorisation of design variables in the P4P typology

Who received the incentive (Did Individuals or Groups receive the incentive)?	
Criteria for judging **Individuals**	• If the incentives are paid directly to individual health workers/clinicians/doctors only • If individual health worker/clinician/doctor's income is supplemented as a result of the incentive (e.g. reflected in the rise of personal income) only
Criteria for judging **Groups (including schemes where individuals and groups are paid bonuses)**	If the incentive is paid to a group or an organization in which individual clinicians may or may not benefit from the incentive directly Groups include any of the following • Hospital • Clinical team • General physician (GP) practice • NGO • Levels of government • Faith based organizations
Type of incentive (Was the incentive in the form of Fines or Bonuses)?	
Criteria for judging **Fines**	If the incentive is negative in the form of reduction in expected payments, penalty, punishment etc. In some cases, bonuses may or may not be paid as well
Criteria for judging **Bonuses**	If incentive is in the form of increase in payments, bonus, gifts etc. with NO fines levied
Size (Was the size of the incentive small or large)?	
Criteria for judging **Small**	If the incentive in the P4P programme is smaller than 5% of any one of the following: • Salary of individual clinician/health worker/doctor • Anticipated payments (to the health facility/hospital/clinical team) such as budgets (total budget or budget for the particular intervention in question), fee for service (FFS) and capitation
Criteria for judging **Large**	If the incentive in the P4P programme is 5% and above of any one of the following: • Salary of individual clinician/health worker/doctor • Anticipated payments (to the health facility/hospital/clinical team) such as budgets (total budget or budget for the particular intervention in question), fee for service (FFS) and capitation
Timing of payment after achieving targets (time lag): was it short or long?	
Criteria for judging **short**	If incentive payment (or penalty) is received not more than 4 months after measurement and confirmation of performance
Criteria for judging **long**	If incentive payment (or penalty) is received more than 4 months after measurement and confirmation of performance
Domain of performance measured (Was the domain of performance measure within	

clinicians control or out of clinicians' control)?	
Criteria for judging **within clinicians control**	If incentive payments to health service providers are **mostly/only** based on processes and structures e.g. number of children immunized, routine measurement of blood pressure of patients every month, number of referrals made, rate of cancer screening
Criteria for judging **out of clinicians control**	If incentive payments to health service providers depend on achieving a change in health outcomes e.g. reduction in mortality rates from a specific disease, blood pressure reduction, patient experience etc.
Performance measure (payment scale) Absolute or relative measure?	
Criteria for judging **Absolute measure**	If incentive is paid (fine levied) to the health service provider that based on their performance, not relative to how other health providers perform. For example, • Improvement in performance typically improvement from some baseline measure, using performance score/ performance points achieved • Achieving performance at/above a predetermined target • e.g. incentive paid per patient immunized, or 70% improvement from baseline
Criteria for judging **Relative measure**	If incentive payment is based on the performance of health service providers, relative to that of other providers. For example, • If bonuses are paid for to health service providers in a specific performance rank e.g. the providers above the top quartile of performance. • And/or • If fines are levied on health service providers in certain ranks usually the bottom ranks e.g. the providers below the lower quartile of performance
Risk: High risk or low risk? (based on judgements from Performance measure, Time lag, and Domain of performance measure	
Criteria for judging **High risk**	If the P4P programme has 2 or more of the following features • If incentive payment (or penalty) is made **after 4 months after measurement and confirmation of performance (long time lag)** • If the domain of performance measure was **mostly out of clinicians control** • If the perofmance measure (payment scale) is a **relative measure**
Criteria for judging **Low risk**	If the P4P programme has 2 or more of the following features • If incentive payment (or penalty) is made before or at 4 months after measurement and confirmation of performance **(short time lag)** • If the domain of performance measure was mostly **within the clinicians control** • If the performance measure (payment scale) is an **absolute measure**

3.3.2. Retesting the typology

After further refining the typology, I then retested the typology on the same P4P schemes identified from the review of Eijkenaar (2012) and on all descriptions of P4P schemes from evaluated studies identified from the review in chapter two (detailed results of these are shown in Appendix B4). In Table 3.4, I present the results from applying the typology to 14 P4P schemes identified from the review by Eijkenaar (2012).

In total, I used the typology to categorise 58 out of 73 P4P schemes (identified from previous searches in chapter two) into mutually exclusive types. The reason the rest of the schemes could not be categorised was that one or more of the design variables of P4P was not reported in sufficient detail (e.g. the Clalit scheme in Israel: Table 3.4). The least reported design variable was size of incentive. Studies often used vague terms such as ‘modest’ or ‘small’, without providing absolute amounts or sizes relative to the usual clinician income or hospital budget.

Table 3.4 Results of applying the typology to P4P schemes identified from the review by Eijkenaar et al. (2012)

P4P schemes	Who receives the incentive	Type of incentive	Size of incentive	Time lag	Performance measured	Domain measured	Risk	Type
Advancing quality (AQ) UK	Groups	Bonuses	Small	2/3months lag	Relative	Mostly within Physicians control (2 final outcomes and 26 processes)	High	8
Clalit Israel	Groups	Bonuses	Dependent on budget savings	Annually	Absolute	Mostly within Physicians control (10 processes and 8 intermediate outcomes)	Low	
Clinical Practice Improvement Pay (CPIP) Australia, Queensland	Groups	Bonuses	Large	Semi-annually, 3 month lag	Absolute	Within physicians control (12 structures and 7 processes)	Low	2
ERGOV Germany	Groups	Fines	Depend on other hospitals	4 month lag	Relative	Not completely within the physicians control (Final outcome)	High	
MACCABI Israel	Groups	Bonuses	Size not reported	Annually	Absolute	Mostly within Physicians control (12 processes and 5 intermediate outcomes)	Low	6
National Health Insurance P4P (NHI-P4P) Taiwan	Groups	Bonuses	Large	Monthly and annually	Absolute and relative measures	12 structures, 3 final outcomes, and 2 intermediate outcomes	High	6
Primary care P4P (PC-P4P) Netherlands Primary Care	Groups	Bonuses	Large	Annually	Relative	Within physicians control (31 processes)	High	4

P4P schemes	Who receives the incentive	Type of incentive	Size of incentive	Time lag	Performance measured	Domain measured	Risk	Type
Renewal Models (PCRM) Canada Ontario	Groups	Bonuses	Small	Annually	Absolute	Within physicians control (12 processes)	Low	2
Physician Integrated Network (PIN) Canada Manitoba	Groups	Bonuses	Maximum payment unknown	Immediately after performance measure	Absolute	Within physicians control (only processes)	Low	2
Practice Incentive Programme (PIP) Australia	Groups	Bonuses	Size not reported relative to income	Quarterly, semi-annually and annually	Absolute	Within physicians control (only structures and processes)	Low	
Performance management Programme (PMP) New Zealand	Groups	Bonuses	Small	Semi-annually and annually	Absolute	Within physicians control (8 processes)	Low	4
Programme of quality Improvement (PQI) Argentina	Groups	Bonuses	Large	Annually	Absolute	Mostly within physicians control (16 processes, 7 structures and 3 outcomes)	Low	2
Quality and Outcomes Framework (QOF) UK	Groups	Bonuses	Large	Annually	Absolute	Mostly within physicians control (85% processes)	Low	2

Following the categorisation of the P4P schemes identified from the reviews in chapter two, I found that descriptions of the schemes in evaluation studies of these schemes were often unsummarised and unstructured. This made it difficult to identify the design variables in the P4P scheme being evaluated. One of the few studies that summarised design variables of the evaluated scheme in such a way that could help readers use the typology more effectively was by Kirschner et al. (2013) illustrated in Table 3.5.

Table 3.5 Summary of design features presented in the P4P evaluation study by Kirschner et al. (2013)

Performance measurement
• Clinical care: diabetes (*n* = 9 indicators), COPD (*n* = 5 indicators), asthma (*n* = 4 indicators), cardiovascular risk management (*n* = 9 indicators), influenza vaccination (*n* = 2 indicators), cervical cancer screening (*n* = 1 indicator) • Practice management: infrastructure (*n* = 7 items), team (*n* = 8 items), information (*n* = 3 items), quality and safety (*n* = 4 items) • Patient experience: experience with GP (*n* = 16 items) and organization of care (*n* = 11 items)
Appraisal
• A benchmark with relative standards was set at the 25th percentile of group performance • For the appraisal, there was a series of tiered thresholds (seven levels) • Practices received feedback in the short term (4 months after data collection) • Valuing the quality level as well as the improvement of performance, weighing these levels as 3:1
Reimbursement
• A bonus of 5–10% of the practice income, not linked to the usual reimbursement • Bonus was paid in money and not in objects or services • Bonus to spend freely

3.3.3. Labeling the types on the P4P typology (preparing the typology for use in exploring heterogeneity)

Having demonstrated that the typology can be used to categorise several P4P schemes reported in the literature in health care, the next logical step was to use the typology in analysis to explore heterogeneity (as discussed in chapter two). For example, knowledge harnessed from theory suggests that incentives offered to groups, use of fines, payment of large incentives, and incentives that fall under 'low risk' have a higher chance of success as opposed to incentives offered to individuals, small incentives, and high-risk incentives. Thus, P4P schemes characterised with features that are likely to improve the chance of success or performance.

A way to make the P4P typology useful in exploring the heterogeneity between evaluation results is to label the types in the typology (Bailey, 1994). This involves using informed characteristics of the dimensions to assign labels (Kluge, 2000). One of the defining rules of labelling types in a typology is that one is allowed to label in such a way that is suited to the analysis one wants to carry out (Bailey, 1994, Kluge, 2000).

In this case, the main interest was to use the types to explore heterogeneity in P4P, which was needed to make better sense of the evidence (finding out what works and what does not). For this to be effective and to produce meaningful results with a certain level of confidence, the dimensions of the design features must vary sufficiently (Kluge, 2000). After using the typology to categorise evaluated P4P schemes identified from literature, I found that only four out 73 schemes used fines. For this reason, the design variable of type of incentive was excluded from the labelling process. This does not alter the original typology. Instead, it meant that in the analyses where I use the typology to explore heterogeneity (chapter 5), whether or not the incentives were fines or bonuses were not considered.

Therefore, to label the types in the P4P typology I considered three design features:

- Who receives the incentives: groups or individuals
- Size of incentive: small or large
- Perceived risk of not earning the incentive: high or low

As illustrated in Table 3.6, P4P schemes with all three design features were labeled as ‘type A’ (higher chance of success); P4P schemes with two out of the three design features were labeled as ‘type B’ (medium chance of success); and P4P schemes with one or none of the listed design features were labeled as ‘type C’ (low chance of success). These labels are the units of analyses (categories) used to systematically explore heterogeneous results of P4P in chapter five.

Table 3.6 Labelling the types in the P4P typology

Type	Who received the incentive	*Type of incentive	Size of incentive	Perceived risk of not earning the incentive (RISK)	Labels a- high chance of success b-medium chance of success c- low chance of success
1	Groups	Fines	Large	Low	A
2	Groups	Bonuses	Large	Low	A
3	Groups	Fines	Small	Low	B
4	Groups	Bonuses	Small	Low	B
5	Groups	Fines	Large	High	B
6	Groups	Bonuses	Large	High	B
7	Individuals	Fines	Large	Low	B
8	Individuals	Bonuses	Large	Low	B
9	Groups	Bonuses	Small	High	C
10	Groups	Fines	Small	High	C
11	Individuals	Fines	Small	Low	C
12	Individuals	Bonuses	Small	Low	C
13	Individuals	Fines	Small	High	C
14	Individuals	Bonuses	Large	High	C
15	Individuals	Bonuses	Small	High	C
16	Individuals	Fines	Large	High	C

***excluded from the labelling process**

3.4. Discussion

I developed a theoretical typology by merging and consolidating theories with designs variables potentially relevant to P4P to create a framework to help explore and possibly explain heterogeneous results of evaluations of P4P. The final typology consists of four key design variables namely: who receives the incentives, type of incentives, size of incentives, and perceived risk of not earning the incentive (a condensed variable consisting three design features: performance measure, time lag between the measurement of performance and payment of the incentive, and the domain of performance measured).

This P4P typology is helpful in clarifying the similarities and differences among the types of P4P schemes found in the literature. It is also helpful in categorizing the P4P schemes based on the design features. Without a typology or a similar framework to compare types, there remains the confusion and difficulty in exploring emerging literature surrounding P4P schemes in health care.

Adoption of this typology would also be helpful in facilitating effective communication between people who design P4P interventions, P4P implementers or adopters and P4P evaluators. It could also help provide structured information to P4P designers and developers, so that they understand the possible results of their design choices and possibly help guide their thinking towards design choices that might work in their context.

Yet another important use of this typology is to aid in interpreting the heterogeneous results of the evaluated P4P schemes and particularly, as a framework for the analysis of various theories relevant to the design of P4P schemes in health care. In order words, if what theory says about these design features is true to an extent, then we should be able to see significant association between these design choices and effectiveness. I explore these using empirical cases in literature in chapter five, where with the help of the typology, I explore how different designs variables and groups of variables (types) influence the impact of P4P in health care, while holding other factors constant.

Though this typology proves its potential usefulness in the description, categorisation and synthesis of evidence of P4P schemes, it is only a first attempt. It should be further tested and developed as more of these P4P schemes and their evaluations emerge to ensure its relevance. For example, design variables not included in this typology might be relevant in the near future and added on in a more extensive typology. Though adding more design variables to the current typology might make it a cumbersome framework to be used for analyses and exploration heterogeneity, a more extensive typology could be still be useful to describe current and future P4P schemes (a reporting template).

Finally, this typology of P4P design features provides only one element of the understanding of the variations in the effects of P4P. As noted and discussed in chapter two, other factors are likely to have a bearing on the impact of P4P as well (Kirschner et al., 2013, Canavan and Swai, 2008, Van Herck et al., 2010, Ssengooba et al., 2012). These include:

- The context in which the P4P scheme is implemented (health systems, increased funding, and complexity)

- How well the programme is being piloted: use of baseline measurement, setting of targets, degree of preliminary work done
- Rigour of evaluation (absence or presence of control groups)
- Clinical area of intervention. (Eichler and Levine, 2009, Ssengooba et al., 2012)

Some of these factors are considered and explored in subsequent chapters of this thesis. An example is the rigour of evaluation: evaluated programmes with inadequate control groups are likely to appear as effective compared to schemes evaluated with good control groups as a result of confounding factors (as demonstrated and discussed in chapter two). Therefore, in chapter five where I use the typology to explore heterogeneous results of P4P, the evaluation designs of the schemes were taken into consideration in the analyses.

In addition, as discussed earlier, unlike design variables, the contextual and implementation factors are harder to compare across different P4P because they are usually setting specific and non-generalizable i.e. contextual and implementation factors influencing effects of P4P in a LMIC is likely to be a non-issue in developing countries (e.g. distrust in the payment system). Usually, in-depth qualitative studies are usually required to capture this knowledge. Therefore, this was my focus in the second part of this thesis (chapter six-nine), where I explored the influence of contextual and implementation factors on the effectiveness of a Nigerian P4P scheme (a LMIC case study), to inform and make recommendations to improve the effectiveness of the scheme when implemented on a large scale across the nation.

Limitations of development and use of this typology

There were three main limitations.
First, to find a suitable trade-off between the typology being manageable and maintaining relevance, some of the design variables explored and discussed (such as method of payment and kind of incentive) were not included in the typology (used later in this research to explore heterogeneity). Thus the typology was not exhaustive, meaning that potentially important design variables might have been lost. Nonetheless, this typology can provide a foundation towards standardised categorizations of current P4P designs in literature, since it is the first typology of its kind to be developed. The

most important thing however, is that P4P designers and evaluators need to consider all discussed design variables in designing or evaluating these schemes, whether they are included in the typology or not.

Second, in order to have well rounded and practical categories, equal weights were assigned to the design variables included in the typology. This is a potential limitation because their relative importance is likely to vary i.e. size might be a lot more important in influencing P4P effects than who receives the incentive. I attempt to address this issue in chapter five, where I systematically and statistically explore the relative importance of the design variables in empirical cases (theory testing: one of the many functions of the typology).

The final limitation was the problem of poor reporting of evaluated P4P schemes. Some studies were incompletely reported with important design features or choices absent from the evaluations, despite the potential association between design features and effectiveness of the schemes. This meant that some P4P schemes could not be categorised. There is the need for a uniform way of reporting design variables of P4P schemes in evaluation studies if one is to be able to make sense of the evidence. The developed typology offers a way to improve upon this area, as it provides a standard and informed way to help describe P4P schemes.

Strengths of the process of developing this typology

There were two major strength of this study. First, I applied the typology and successfully categorised a number of P4P studies into mutually exclusive categories, which demonstrates that the typology is robust, has face validity, and that is potentially useful as a framework to systematically explore heterogeneity in P4P.

The second strength of this typology is that it was well informed through rigorous exploration of relevant theories and literature. Furthermore, the typology demonstrates strong content validity in that the process of development of the typology was transparent and decisions made were adequately justified and relevant to empirical cases in literature. In addition, to improve the credibility of the typology as a potential tool to categorise and describe P4P schemes in health care, measures of reliability, concurrent

validity, and pilot testing of the typology by other users are demonstrated in the chapter four.

3.5. What this chapter adds

The aim of this chapter was to explore and harness theories and literature in order to develop an informed framework (typology) to categorise P4P schemes based on their design features so that heterogeneity can be explored in a systematic way to make sense of the current evidence of P4P (exploring what works or what does not, and why). This typology builds on previous work of other researchers who have reviewed and described design features (using empirical literature) by adding on theoretical perspectives to explain and predict behaviours depending on design choices, and using this knowledge to create a practical framework (Stockwell, 2010, Eijkenaar, 2013). This P4P typology provides an important first step towards making synthesis of evidence easier by providing a quick and efficient way to categorise P4P schemes in an informed way in order to explore heterogeneity in the evaluations of P4P. In addition, the typology could help P4P developers' structure and inform design choices, and to establish a common language (reporting template) in which P4P designers, reviewers, implementers and policy makers can clearly specify the content of P4P designs in a standardised way so that other people can see what exactly is being done.

Given the potential usefulness of the typology in description, categorisation, and synthesis of evidence of P4P in health care, in the next chapter, using volunteer health service researchers, I assess the reliably and validity of the P4P typology, after which it is put up for critique and use in the public domain. The typology was then used to explore and explain heterogeneous results of P4P schemes in chapter five.

Chapter 4 Assessing the reliability of the Typology as a tool to categorise P4P schemes

4.1. Introduction

In previous chapters, I reviewed and appraised the available evidence surrounding the effectiveness of the use of P4P in healthcare, which highlighted the heterogeneity between the schemes and difficulty in interpretation of results and evidence synthesis.

Following that, I developed a typology as a framework for categorising P4P schemes in healthcare based on their design features, as a means to explore heterogeneity and establish a common language in which the contents of P4P schemes can be clearly specified. The typology consists of four items: who receives the incentives, the type of incentive, the size of incentive, and the perceived risk of not earning the incentive. Given the potential importance of the typology in categorising P4P schemes in healthcare and its potential to be a common language to describe P4P designs, it is essential to test whether this typology is easy to use, reliable, and valid as a categorisation tool for incentive schemes.

Reliability

A tool is reliable if it measures something in a reproducible and consistent fashion in the different conditions, in which it is likely to be used (Atkinson and Nevill, 1998, Streiner and Norman, 1989).

There are a variety of ways to estimate reliability. They include internal consistency reliability, test-retest reliability, and inter-rater reliability. The kind of reliability testing conducted on a tool is often dependent on the type of tool, its potential users, and conditions in which it is likely to be used (Charter and Feldt, 2002). Briefly outlined in subsequent paragraphs are some types of reliability tests.

Internal consistency reliability tests the consistency, with which the items on the tool measure the same thing, and the test-retest reliability is often conducted on tools or instruments used on individuals (as subjects) e.g. surveys, interview tools, and diagnostic tests (Trochim, 2006; Feder, 2008; Bowling and Ebrahim, 2005). Since all items on the P4P typology (tool of interest) assess different things, and the typology is

not designed for use on humans or diagnostic tests. Therefore, the internal consistency and test-retest reliability tests are not of relevance.

Inter-rater reliability on the other hand, assesses the degree to which the measuring instrument or tool produces similar results when used by different raters (users). In other words, it is the degree, to which the tool users/raters give consistent results for similar populations (Bowling and Ebrahim, 2005, Trochim, 2006b). The inter-rater reliability test is particularly useful to assess the reliability of tools in which there is the possibility of subjective judgements on items on the tool among users/raters e.g. categorisation tools (Feder, 2008).

For the developed P4P typology, I am particularly interested in how similarly the tool users/raters (health science researchers) will categorize P4P schemes. Therefore the appropriate reliability test of interest (in this study) is the inter-rater reliability test. There is a need for inter-rater reliability testing of the typology in order to assess and enhance consistency in its application, which will in turn support the uptake and use of the typology as a P4P categorization tool.

Validity

It is also important to assess the validity of the typology as a categorisation tool. Generally, a tool is considered valid when it measures or quantifies what is intended to measure (Streiner and Norman, 1989). A valid tool does what it is designed to do. There are different forms of evidence of validity of a tool or instrument. These include content validity, criterion-related validity, and construct validity.

Content Validity focuses on the theoretical or conceptual basis of a tool (Potvin, 2007). This is often demonstrated or established through a detailed description of the steps used to develop the instrument, which could include; review of literature and related theories, conducting focus groups or interviews, and expert consensus, depending on the type of tool developed (Trochim, 2006a). Content validity of the P4P typology has been demonstrated in the previous chapter (three) through a detailed description of the steps taken to develop the tool, which included review and analysis of literature and theory.

Criterion-related Validity: there are two types of criterion-related validity (concurrent and predictive validity) (Miller and Salkind, 2002).

- Concurrent validity is a measure of how well the results obtained from the tool in question correlates with a previously validated measure or a trusted criterion (Trochim, 2006a). The users of the tool might have high agreement, but in the wrong direction (wrong results). Therefore, concurrent validity is particularly important to the P4P typology because it is useful to know if the users of the typology are using it correctly. However, since this is a new typology and the first in literature, the concurrent validity could not be assessed due to lack of a previously validated measure (or gold standard).
- Predictive validity refers to the extent to which a tool can predict an outcome in the future (Miller and Salkind, 2002). In the case of the developed typology, I am interested in assessing whether theory can rightly predict the influence of design features on the impact of incentive schemes. In other words, will the 'types' in the typology, predict to an extent the degree of impact of incentive schemes. This type of validity is extremely important if the typology is being proposed as a predictive tool, which I explore in detail in the next chapter (5)[5] among other things.

Construct Validity centres on how well the tool measures the construct that is supposed to measure (Potvin, 2007). In other words, is it measuring what it is supposed to measure? Construct validity is demonstrated by comparing the tool to another test that measures a similar construct and it is mostly relevant to measurement tools (Miller and Salkind, 2002). Therefore, the evidence for construct validity is not applicable to this typology, since it is being proposed as a categorization tool.

Some evidence of validity, such as content validity, has been demonstrated in the previous chapter, while evidence of validity such as the construct validity is not relevant for the developed typology. Predictive validity will be explored in the subsequent chapter, leaving only the need to demonstrate evidence for concurrent validity for the P4P typology in this chapter.

The aim of this chapter is to assess the inter-rater reliability and ease of use of the typology as a categorisation for P4P schemes in health care. The subsequent sections of the chapter describe the methods, results, and discussion of the findings.

[5] In chapter 5, I employ statistical analyses to explore the relationship of each item on the tool with empirical outcomes of incentive schemes.

4.2. Methods

In summary, testing the inter-rater reliability and ease of use of the typology as a P4P categorisation tool involved having several raters/users apply the P4P typology to a sample of P4P studies identified from the review in chapter two. To assess inter-rater reliability, a statistical test that estimates agreement between the raters was used (McHugh, 2012). The inter-rater reliability and validity of all four items on the typology was assessed. This is the preferred method because it made it easier to identify sources of disagreement or confusion with the use of the typology (Lobbestael et al., 2011, Hartling et al., 2012, MacDermid et al., 2005, Oremus et al., 2012)

Finally, to assess the ease of use of the typology, a simple questionnaire was completed by the users, which incorporated understanding, ease of use, and time taken to apply the tool, and feedback/suggestions (Stewardson et al., 2013, Kastner et al., 2010).

In the following sections, I describe in detail the methods used to conduct this study, and the rationale for the methods chosen. This includes outlines of statistical tests for estimating inter-rater reliability and concurrent validity, sample size (number of raters and number of items rated), and selection of raters.

4.2.1. Statistical test for estimating inter rater reliability (kappa)

Cohen's kappa (or a variation of Cohen's kappa) is often used to estimate inter-rater reliability in categorisation tools in healthcare (Lobbestael et al., 2011, Hartling et al., 2012, MacDermid et al., 2005, Oremus et al., 2012) (see appendix C1 for a summary of similar studies consulted). Other commonly used statistical measures or indexes used in estimating inter-rater reliability include percentage of absolute agreement and intra-class correlation coefficient (ICC) (Ubersax, 2010).

Percentage absolute agreement is an index of inter-rater reliability in which the absolute percentage agreement between rater pairs is calculated. For example, for an item in a tool, if the raters agree six out of ten times, the tool/test has a 60% inter-rater reliability rate (Feder, 2008). ICC on the other hand is the fraction of the total variance within data that is explained by variance between two raters (Osborne, 2008).
Researchers have argued that kappa is preferred over percentage of absolute agreement and ICC because the latter does not take into account chance agreement or the agreement due to the raters guessing (McHugh, 2012, Cohen, 1960, Feder, 2008). In

addition, ICC might be a poor reflection of the amount of agreement between raters, resulting in extreme over or under estimation of the true magnitude of rater agreement. ICC is also very specific to the population sample and not directly comparable across population (McHugh, 2012, Osborne, 2008), making kappa the preferred option.

The kappa statistic gives a numerical rating of the degree to which two raters agree based on the difference between how much agreement is actually present, ("observed" agreement) and how much agreement would be expected by chance alone ("expected" agreement). The formula for calculating kappa=(Observed agreement – agreement expected by chance)/(100% – agreement expected by chance) (Viera and Garrett, 2005).

The kappa statistic varies between -1.00 and +1.00. A kappa value of +1.00 indicates a perfect agreement between raters and 0.00 kappa value indicates that the raters agreement is indistinguishable that expected by chance (Cohen, 1960). The Cohen's kappa is specific for estimating agreement between rater pairs and is not applicable for estimating agreement between multiple raters. Fleiss in 1971, however, developed an extension of the Cohen's kappa, which is known as Fleiss kappa that is commonly used to estimate agreement (inter- rater reliability) between more than two (multiple) raters (Fleiss, 1971, Ubersax, 2010).

Even though kappa is commonly used, it has its drawbacks. For instance, Ubersax (2010) argues that kappa does not really correct for chance agreement, because chance agreement is only relevant under the conditions when raters are independent. The author further argues that raters are not independent as long as they are rating the same cases. Therefore, the claim that kappa corrects for chance agreement is questionable[6]. Mchugh (2012) and Gwet (2002) further suggest that kappa estimates may be low even though there are high levels of agreement between raters and individual ratings are accurate, when the prevalence of an outcome is too low or too high, which can lead to high rate of agreement due to chance alone. When the probability of chance agreement is high, the estimated kappa is low because kappa values decrease as the probability of chance agreement increases among raters (see Figure 4.2.3) (McGinn et al., 2005).

[6] Ubersax argues that for kappa to adjust for chance agreement effectively requires and explicit model of how chance affects the decision of the raters. Kappa statistic does not do this; the claim of kappa adjusting for chance agreement follows the hypothesis that raters guess when not completely certain which is not very realistic.

Despite these draw backs, it is still often advised to use kappa because it definitely can verify that agreement between raters exceeds chance levels, having an advantage over percent absolute agreement and intra-class correlation (Ubersax, 2010, Cohen, 1960). This makes kappa the most appropriate test statistic in this study to assess the agreement between raters in order to estimate inter-rater reliability of the P4P typology.

4.2.1.1. Interpreting the kappa statistic (what level of agreement is 'good'/acceptable?)

How large kappa values should be to indicate good or an acceptable level of agreement between raters is the subject of debate. There are three common guidelines used to interpret kappa values (Altman, 1991, Landis and Koch, 1977, Fleiss, 1973) as shown in Table 4.1. The interpretations of kappa seen in table 4.1 are based on the value judgment of the authors. Kappa values are interpreted relative to the degree of chance agreement between raters. This is illustrated graphically Figure 4.1. This figure shows kappa values for two categories, 'yes' and 'no' by probability of a 'yes' and probability observer will be correct. The verbal categories of Landis and Koch are used for this example.

Table 4.1 Guidelines for interpreting kappa

Landis and Koch, 1977		**Altman, 1991**		**Fleiss, 1973**	
Kappa	Interpretation	Kappa	Interpretation	Kappa	Interpretation
0.81 – 1.00	excellent	0.81 – 1.00	very good	0.75 – 1.00	very good
0.61 – 0.80	substantial	0.61 – 0.80	good	0.41 – 0.75	fair to good
0.41 – 0.60	moderate	0.41 – 0.60	moderate	< 0.40	poor
0.21 – 0.40	fair	0.21 – 0.40	fair		
0.00 – 0.20	slight	< 0.20	poor		
< 0.00	poor				

Figure 4.1 shows that kappa is maximum when the probability of a true 'yes' is 0.5 (this is also when chance agreement equals 0.5). As this probability gets closer to zero or to

one (as chance agreement increases), the value of kappa reduces. This is an illustration of the high agreement, low kappa problem (see section 4.2.2.).

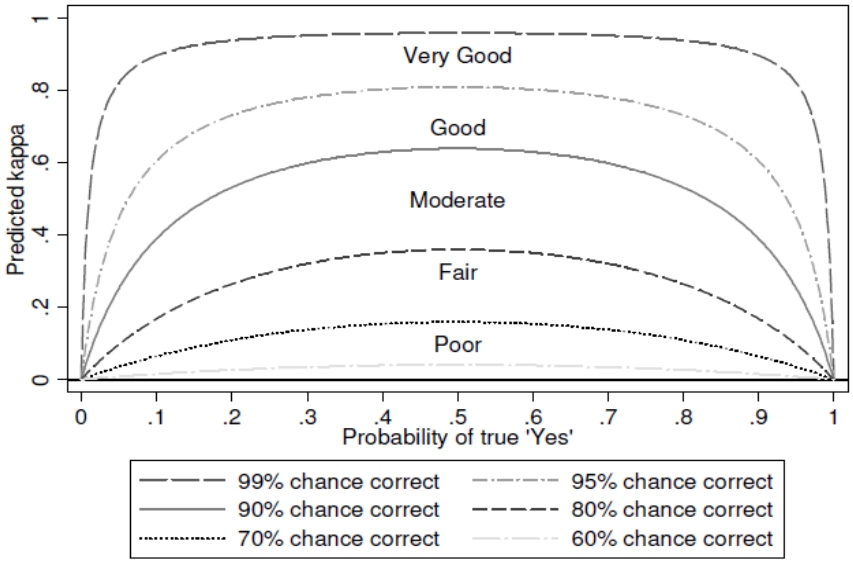


Figure 4.1 Predicted kappa for two categories, 'yes' and 'no', by probability of a 'yes' and probability observer will be correct (Source: Bland, 2008)

The lines represent the degree to which kappa corrects for chance agreement at various kappa values (tells you how much of the possible agreement is over and above chance). For example the line that represents 99% chance correct is interpreted as having kappa value of 0.8 (and above) which means that the observed agreement between the raters (with a kappa value of 0.8 and above) are 99% beyond chance or 99% of the agreement between raters are beyond/over chance or there is a 0.01 probability that agreement between raters is by chance (McGinn et al., 2005, Bland, 2008).

In most published studies assessing the inter-rater reliability of tools using the kappa statistic, the reason a particular guideline chosen to inform the choice of what kappa value represents good or acceptable agreement is unclear (Sim and Wright, 2005, Oremus et al., 2012, Lobbestael et al., 2011, MacDermid et al., 2005). McHugh, 2012 suggested that the available guidelines for interpreting kappa might be too lenient, as very low kappa values might be acceptable. For example, using the guideline by Altman (1991) and Landis and Koch (1977), kappa values as low as 0.41 are considered moderate level of agreement. However, since it is implied that kappa values are likely to be low despite high agreement between raters (Gwet, 2002, Ubersax, 2010), one might argue that using a lenient interpretation of kappa is justified.

Furthermore, studies suggest that choosing the acceptable level of agreement in order to consider a tool reliable depends on the clinical relevance of the tool or test (McHugh, 2012, McGinn et al., 2005). For example, one could argue that the kappa value deemed acceptable for clinical diagnostic tools for rare outcomes/diseases should be more lenient than for categorization tools. This takes into consideration that kappa values might still be low even though there is a good level of agreement between raters, which might minimize mishaps (or false negatives) in the clinical diagnosis of rare outcomes/diseases.

For the purpose of estimating the reliability of the P4P typology as a categorization tool in my research, the minimum acceptable level of kappa for each item on the typology was set at 0.6, which means that agreement observed between raters are 90% (and above) beyond chance or that the probability that agreement is by chance is ≤ 0.1. This decision is in line with the guideline for interpreting kappa values suggested by Altman (1991). This represents a suitable trade-off between having too many misclassifications of P4P schemes (making it an unreliable tool to classify P4P to aid exploration of heterogeneity) and adjusting for the possibility of having a low kappa value despite high agreement between raters (which avoids loss of a good and useful tool).

Having justified the use of kappa to assess inter-rater reliability and how to interpret the results, I move on to describe the sample size requirements in terms of (1) the number of items (in this case, reports or descriptions of P4P schemes) to be rated and (2) the number of raters needed.

4.2.2. Number of reports of P4P schemes

Literature concerning the number of subjects (P4P studies in this case) needed when estimating inter-rater reliability is quite scarce and only a few researchers have suggested a way to go about it. Most published studies testing the inter-rater reliability of tools in health care often use convenience samples, with no rationale for sample size (Hartling et al., 2012, Lobbestael et al., 2011, Oremus et al., 2012, MacDermid et al., 2005).

A few researchers suggest that choosing the sample size should be based on the probability of detecting a statistically significant kappa (the difference between the overall and chance agreement P_a-P_e) with a confidence interval of a desired width (or an

estimated relative error) (Sim and Wright, 2005, Gwet, 2010). A guide was proposed by Gwet (2010) to calculate sample size needed (see Table 4.2). However, the difference between the two agreement (overall and chance) probabilities is generally not known at the design stage. Gwet (2010) proposed a rule of thumb to assume the best case scenario that chance-agreement probability is zero, and use an anticipated value for Pa in place of P_a-P_e in Table 4.2 to obtain the absolute minimum sample size one should use. For example, if one anticipates that the raters will agree about 40% of the time, then one would use a sample size of 156 or 69 or 39, depending on the error margin chosen.

Table 4.2 Number of P4P reports needed to estimate Cohen's kappa

$p_a - p_e$	Relative Error		
	20%	30%	40%
0.1	2,500	1,111	625
0.2	625	278	156
0.3	278	123	69
0.4	156	69	39
0.5	100	44	25
0.6	69	31	17
0.7	51	23	13
0.8	39	17	10
0.9	31	14	8
1.0	25	11	6

Even though this method proposed by Gwet is for rater pairs, (it is not stated whether it applies to multiple raters). One might argue that if it can be applied, then the number of reports of P4P needed will reduce, because increasing the number of raters might increase statistical power. However, since there is no evidence to support this, it is safe to retain the number of studies estimated for rater pairs, as this will increase statistical power.

The method proposed by Gwet (2010) to estimate the adequate sample size requires the researcher to guess the value of kappa or agreement expected from the raters. However, since I have justified and specified what will be the minimum acceptable level of reliability of the developed P4P typology (kappa = 0.6), this was considered my 'guess/estimation'. This involved choosing a guessed kappa value and margin of error in which the lowest margin is 0.6 (see table 4.2): the possible kappa values and their percentage relative errors were 0.8 (20%), 0.9 (20-30%), and 1.0 (20-40%), highlighted in Table 4.2. It was however very unlikely that there will be perfect agreement between raters. Therefore, a kappa value of 0.9 (30% relative error) was selected based on a trade-off between precision and a reasonable number of P4P reports to avoiding burdening the raters. This meant the raters had to apply the typology on a minimum of 14 P4P reports (see table 4.2).

4.2.2.1. Selecting P4P studies to be rated

I selected 17 evaluations of P4P containing descriptions or reports of the designs of the P4P from a pool of previously identified evaluated studies in the literature review (chapter 2) of this thesis. Evaluations were chosen because other papers describing P4P schemes were usually large documents or web pages which were quite difficult to collate, and might have been overly ambitious to ask volunteer raters to categorize such. The selection of the evaluated P4P studies was not random due to the issue of incomplete reporting identified and discussed in the previous chapter. Instead, I selected studies that completely reported all the necessary design features to maximise the efficiency of the reliability test.

4.2.3. Number of raters

The literature concerning the number of raters required to investigate inter-rater reliability is mixed and limited. Some have suggested that inter-rater reliability can be investigated with two or more raters and that the number of raters does not affect the reliability (Ling, 2007; Saal, et al, 1980). Hallgren also suggests that inter-rater reliability is not affected by number of raters, instead it is more likely to be affected by having better raters and better guidelines (Hallgren, 2012). Walter and colleagues propose that using more than two raters could help improve statistical power (Walter et al., 1998) but this is still yet to be explored.

After consulting with a statistician (Professor Martin Bland) at the Department of Health Science-University of York, I decided to employ a sample of 15 raters (health science researchers with different research experience and expertise) based on the rationale of trying to replicate a real life scenario, which makes the results more generalizable. We concluded that 5 independent ratings for each of the 17 studies was a reasonable number, which was still within the range of minimum number of studies determined in the previous section. This means each rater will have up to 6 P4P studies to categorize.

4.2.4. Recruiting the raters (and ethics approval)

The Research Governance Committee of the Department of Health Science-University of York approved this study (see Appendix C2 for evidence of ethics approval).

The main rater population approached were health science graduate research students at the University of York. This was a convenience sample chosen due to limited time and resources. Apart from the population being easy to reach/access, it was also based on the rationale that it would be easier and more efficient to train those who are already at the University of York. I also approached health services researchers from the Nigerian Federal Ministry of Health.

The volunteer raters at the Department of Health Science, University of York were approached by an email (forwarded to the postgraduate students) by the health science postgraduate administrator (the other volunteer raters at the Nigerian Federal Ministry of Health were also contacted via email sent by me). The contents of the email included an information sheet explaining the research, and what is required of them should they choose to volunteer (see Appendix C3-C5 for a copy of the email, information sheet, and consent forms). The interested volunteers then contacted me via email.

In total, 12 participants agreed to be volunteer raters (nine postgraduate students from the University of York and three health service researchers from the Nigerian Federal Ministry of Health). This number of raters was short of three the pre-specified sample size of 15 raters. Therefore, in order to maintain the five independent ratings per study (see section 4.2.3), each rater was given between six to eight papers to rate.

4.2.5. Rater training

Inter-rater reliability studies on categorisation tools in healthcare often use raters that have some basic knowledge relevant to the tool (being tested) and even provide more training for the use of the guidelines/tool before inter-rater reliability is estimated (Sim and Wright, 2005, Oremus et al., 2012, Lobbestael et al., 2011, MacDermid et al., 2005). This is because the raters are likely to have different backgrounds and experience, which might affect agreement between the raters. The more heterogeneous the raters are, the less they are likely to agree (Graham et al., 2012). Rater agreement is affected by a number of other things, which include comprehensiveness and clarity of instructions in tool guideline, rater competence, errors in rater judgement, and rater bias (Berkman et al., 2013, Barclay and Harland, 1995, Myford and Wolfe, 2009, Hoyt and Kerns, 1999).

Inter-rater reliability is also affected by the complexity of the sample (e.g. P4P studies to categorize in the case of the typology). An example is the study by Berkman et al. (2013) examining inter-rater reliability of applying guidance for grading strength of evidence in systematic reviews, which found that inter-rater reliability of grading strength of evidence varied with the complexity of the evidence in systematic reviews. There was very good agreement between raters when the systematic reviews were straightforward but poor agreements when the raters had to use their subjective judgement to insinuate what the authors did. This issue is especially relevant to my research area because P4P studies are often poorly reported (ambiguous and sometimes incompletely reported). Therefore, one of the ways to improve rater agreement was to adopt a simpler and more straightforward complete reporting approach in evaluated P4P studies (a good approach would be to report the characteristics of the P4P program following the items in the typology, as described in the guideline). This however, could not be implemented in this study but was noted for future reference on reporting of P4P evaluations.

Other ways rater agreement might be improved is by improving the clarity and instruction on the tool guideline and training (extensive: exposing them to a variety of possible scenarios) of the raters (Woehr and Huffcutt, 1994, Cash et al., 2012, Gorman and Rentsch, 2009). Rater training gives the opportunity to improve on guidelines for use of the tool, as this will be informed by questions and feedback from the training session. Therefore, I developed a training manual to train the volunteer raters. The

method of development of the training manual, pilot testing the tool, and training sessions are discussed in the next section.

4.2.5.1. Training manual

I developed clear and concise decision rules (with examples where needed) to accompany the guidance for applying the tool to the P4P schemes as shown in Appendix C6. This was done by building upon and adding relevant examples to the decision rules developed for the typology in chapter three. This was then followed by the development of a training manual, which was based on the guidelines and decision rules for use of the typology. The training manual was considered to be instructive and comprehensive, with examples and relevant problem solving exercises.

4.2.5.2. Pilot testing

I tested the P4P typology on a subset of the raters (two Health Science PhD students), training them using the training manual and guidelines for use of the typology. The training for each researcher took about 40 minutes after which I asked them to apply the typology on six papers evaluating incentive schemes in healthcare. I gave them three weeks to complete the task and I estimated kappa for each item on the typology, using the STATA statistical package (version 12). Kappa for each item is shown in Table 4.3.

Table 4.3 Kappa values for each item on the typology (pilot test)

Item 1(who receives the incentive: individuals or groups)	**Item 2 (type of incentive: fines or bonuses)**	**Item 3 (size of incentive: small or large)**	**Item 4 (perceived risk of not earning the incentive: low or high)**
1	1	0.714	0.667

Kappa values indicate perfect agreement (kappa=1) between the raters for the first two items: who receives the incentive and type of incentive. Kappa value for size of incentive was 0.714 and kappa value for perceived risk was 0.667 indicating good agreement between raters.

In table 4.4, I present the results of applying the typology to the six evaluation studies by the two raters to better understand their sources of disagreement. Generally, the raters found the typology easy to use, but they disagreed on two occasions and they

indicated on one occasion that there was not enough information in the paper to make a decision.

Table 4.4 Ratings for individual studies by two raters

Study	Who received the incentive?		Type of incentive		Size of incentive		Perceived risk of not earning the incentive	
	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
Jha et al. (2012)	Groups	Groups	Fines	Fines	Small	Small	High	High
Ashworth et al. (2004)	Groups	Groups	Bonuses	Bonuses	Small	Small	High	Low
Basinga et al. (2011)	Groups	Groups	Bonuses	Bonuses	Unclear	Unclear	Low	Low
Cattaneo et al. (2001)	Groups	Groups	Fines	Fines	Small	Small	High	High
Harries et al. (2005)	Individuals	Individuals	Bonuses	Bonuses	Large	Large	Low	Low
Kirschner et al. (2013)	Groups	Groups	Bonuses	Bonuses	Small	Large	High	High

The first disagreement between the rater was in the 'perceived risk of not earning the incentive' category of the Ashworth paper. Both raters agreed that the timing of payment after measurement of performance was long, and that the domain of performance was within the clinicians' control, the raters disagreed on the performance measure (absolute or relative). The source of disagreement was traced to rater1 being confused with the text in the paper that was used to aid the judgement: "The average reward per general practitioner (GP) was calculated by dividing the total sum spent on incentive scheme payments in each primary care organisations (PCO) by the number of GP principals in that PCO". Rater 1 misunderstood how to judge the performance measure.

The second source of disagreement between the raters was in the category of size of the incentive in the rated paper: while 'rater 2' provided the right text that aided in making the decision (correct rating), 'rater 1' admitted to being carried away with words and did not read the whole paper before making a judgement.

The review of the sources of disagreements provided a number of recommendations to improve understanding and use of the typology:

- Clearer decision rules to judge when category is unclear
- Providing an easier template to fill the ratings
- Clear progression and clear distinction between the items on the typology.

I then applied these suggestions to the guideline and training manual for easier understanding. For example, I rearranged the items on the guideline and training manual in order of complexity. I also numbered each item on the tool for easier comprehension. Corrections were made to the guideline based on this pilot test as illustrated in Table 4.5 (updated guideline after pilot, see Appendix C6 for the original guideline). The corrections included clearer decision rules. In addition, I added examples to illustrate possible sources of confusion (when rating) in the training manual. I also stressed the importance of reading the papers thoroughly and taking time to do the task.

Table 4.5 Guidelines for use of the P4P typology

ITEM 1: Who received the incentive? Did Individuals or Groups receive the incentive?	
Criteria for judging **Individuals**	• If the incentives are paid directly to individual health workers/clinicians/doctors only • If individual health worker/clinician/doctor's income is supplemented as a result of the incentive (e.g. reflected in the rise of personal income) only
Criteria for judging **Groups (including schemes where individuals and groups are paid bonuses)**	If the incentive is paid to a group or an organization in which individual clinicians may or may not benefit from the incentive directly Groups include any of the following • Hospital • Clinical team • General physician (GP) practice • NGO • Levels of government • Faith based organizations
ITEM 2: Type of incentive Was the incentive in the form of Fines or Bonuses?	
Criteria for judging **Fines**	If the incentive is negative in the form of reduction in expected payments, penalty, punishment etc. In some cases, bonuses may or may not be paid.
Criteria for judging **Bonuses**	If incentive is in the form of increase in payments, bonus, gifts etc. with NO fines levied
ITEM 3: Size of the incentive Was the size of the incentive small or large?	
Criteria for	If the incentive in the P4P programme is smaller than 5% of any one of the

judging **Small**	following: • Salary of individual clinician/health worker/doctor • Anticipated payments (to the health facility/hospital/clinical team) such as budgets (total budget or budget for the particular intervention in question), fee for service (FFS) and capitation
Criteria for judging **Large**	If the incentive in the P4P programme is 5% and above of any one of the following: • Salary of individual clinician/health worker/doctor • Anticipated payments (to the health facility/hospital/clinical team) such as budgets (total budget or budget for the particular intervention in question), fee for service (FFS) and capitation
ITEM 4: Perceived Risk of not earning the incentive: High risk or low risk? (based on: Timing of payment after achieving targets (time lag), Domain of performance measure, and Performance measure (payment scale)	
Criteria for judging **High risk**	If the P4P programme has 2 or more of the following features • If incentive payment (or penalty) is made **after 4 months after measurement and confirmation of performance (long time lag)** • If the domain of performance measure was **mostly out of clinicians control** • If the perofmance measure (payment scale) is a **relative measure**
Criteria for judging **Low risk**	If the P4P programme has 2 or more of the following features • If incentive payment (or penalty) is made before or at 4 months after measurement and confirmation of performance **(short time lag)** • If the domain of performance measure was mostly **within the clinicians' control** • If the performance measure (payment scale) is an **absolute measure** **Note: It is possible that in some cases, you might still be able to judge the risk of the programme if one feature is missing/unclear. For example, if the time lag for payment is short and the domain of performance measure was mostly within the clinicians' control. We can judge from this information that the risk is low even when there is little or no information about the performance measure**
Timing of payment after achieving targets (time lag): was it short or long?	
Criteria for judging **short**	If incentive payment (or penalty) is received not more than 4 months after measurement and confirmation of performance
Criteria for judging **long**	If incentive payment (or penalty) is received more than 4 months after measurement and confirmation of performance
Domain of performance measured Was the domain of performance measured within clinicians' control or out of clinicians' control?	
Criteria for judging **within clinicians control**	If incentive payments to health service providers are **mostly/only** based on processes and structures e.g. number of children immunized, routine measurement of blood pressure of patients every month, number of

	referrals made, rate of cancer screening
Criteria for judging **out of clinicians control**	If incentive payments to health service providers depend on achieving a change in health outcomes e.g. reduction in mortality rates from a specific disease, blood pressure reduction, patient experience etc. **Note: sometimes, incentive programmes contain a mixture of processes and outcomes. However, one category out of the two is usually predominant. For example a programme with 6 process measures and 2 outcome measures. You will have to judge what category it falls into by deciding which category is predominant and for this example, the incentive programme falls within the clinicians control because the process measures are predominantly more than the outcome measures.** **Also, beware of the titles do not be carried away. For example, some authors report 'main outcomes of their study' make sure you read what this includes and it should not be confused with health outcomes.**
Performance measure (payment scale) Absolute or relative measure?	
Criteria for judging **Absolute measure**	If incentive is paid (fine levied) to the health service provider that based on their performance, not relative to how other health providers perform. For example, • Improvement in performance typically improvement from some baseline measure, using performance score/ performance points achieved • Achieving performance at/above a predetermined target • e.g. incentive paid per patient immunized, or 70% improvement from baseline
Criteria for judging **Relative measure**	If incentive payment is based on the performance of health service providers, relative to that of other providers. For example, • If bonuses are paid for to health service providers in a specific performance rank e.g. the providers above the top quartile of performance. • And/or • If fines are levied on health service providers in certain ranks usually the bottom ranks e.g. the providers below the lower quartile of performance

4.2.5.3. Training sessions

I set up a convenient time and place for the volunteer raters to be trained on how to use the typology to categorize P4P schemes (The raters that were not in York were trained over Skype using the same training manual).

Volunteer raters were asked to sign a consent form before the training session to demonstrate that their participation was entirely voluntary and they understood the

research. After which they were trained in an interactive training session and there were opportunities for questions and feedback. The training sessions lasted an average of one hour.

After the training, the volunteer raters were given six to eight academic papers (see Appendix C7 for full reference of studies used for this study), which described P4P schemes form different countries and were asked to apply the typology to each scheme.

The raters were asked to rate the studies independently. This was important because agreement between raters might be influenced by each other when they rate together (Swingler, 2001, Defloor and Schoonhoven, 2004). The raters were also asked to report the estimated time taken to apply the typology to each study and ease of use/difficulty level in understanding and using the typology by use of a simple questionnaire with three options of: easy, moderately difficult, and difficult.

The raters were given a uniform template to report the results of applying the typology to the papers. They were asked to report what portions of the paper helped them make their decision and how they came about the decision, which made it easier to identify sources of disagreements between raters (see Appendix C8 for a sample reporting template). I also collected information about the raters' qualifications, background, research expertise, years of experience, and their knowledge of incentive schemes in healthcare (see Appendix C9 for questionnaire used).

The volunteers were given a maximum of three weeks to complete this task, after which they were given a token of a £10 gift voucher as a thank you for their time upon completion of the task. Training all the volunteer raters and completion of the task by all rater lasted about two months.

The inter-rater reliability of each item on the typology was then estimated using the Fleiss' kappa. All analysis was done on STATA version 12.

4.3. Results

4.3.1. Rater characteristics

In total, 12 volunteer raters contacted me to express their interest in the study and were subsequently involved in the study. This number was three raters below the number I

wanted, to have a constant six rating for each of the 17 P4P studies. Therefore, I adjusted the number of studies given to the raters from five to six studies per rater. The rater population consisted of five PhD students, four Master's students, and three health service researchers (with a Master's degree being their highest qualification). Four of the raters had between zero to one year of research experience, seven raters had between two to four years of research experience, and one rater had over five years of research experience. Three of the raters had knowledge and experience of P4P schemes in healthcare.

4.3.2. Ease of use of the P4P typology

All the raters reported that the tool was easy to use. On average, the raters reported the time taken to apply the typology on one paper was an average of 20 minutes (See Appendix C 10). Furthermore, the raters seemed to understand the tool, which was reflected in their descriptions of how they applied the tool to each P4P study (see table 4.6 for an illustration from a rater). This suggests that the raters had applied the typology effectively, following the guideline for its use.

Table 4.6 An example of disagreement between raters

Study	**Who receives the incentive: individuals or groups**	**Type of incentive: fines or bonuses**	**Size of incentive: small or large**	**Time lag: short or long** **Domain of measurement: within clinicians control or out of clinicians control** **Performance measure: absolute or relative measure**	**Perceived risk of not earning the incentive: high or low risk**
Kirschner et al., 2013	**GROUPS** 'A **practice** with a quality score in the lowest group did not receive a bonus'	**BONUSES** 'A **bonus** was chosen instead of a possible more effective withhold'	**LARGE** 'A bonus of **5–10%** of the practice income'	**Short -** the payment was realized in relatively short time, **4 months** after the data collection 'For clinical care, the process indicators were incentivised' **Process measures-** **Within the clinicians control** **Relative** A practice with a quality score in the lowest group did not receive a bonus'. '…**relative instead of absolute** thresholds were chosen although these might provoke uncertainty and complexity that can negatively influence the effectiveness of a P4P programme	**Low risk** The P4P scheme had 2 'low risk' design features: short time lag and domain of performance within the clinicians control

4.3.3. Inter-rater reliability (kappa) of each item on the P4P typology

Kappa was estimated for each of the four items on the typology, shown in Table 4.7. The second column shows the kappa statistics, the third column shows the Z (test) statistics[7], and the fourth column shows the associated p-value.

Table 4.7 Kappa results for each item on the P4P typology

Items on the typology	Kappa	Z	Prob>Z
Item 1(who receives the incentive: individuals or groups)	0.95	12.40	0.00
Item 2 (type of incentive: fines or bonuses)	0.91	11.92	0.00
Item 3 (size of incentive: small or large)	0.72	9.33	0.00
Item 4 (perceived risk of not earning the incentive: low or high)	0.71	9.20	0.00

Kappa value for item 1 (who receives the incentive: individuals or groups) was 0.95, kappa for item 2 (type of incentive: fines or bonuses) was 0.91, both of which were considered almost perfect agreement between the raters. Kappa values for item 3 (size of incentive: small or large) and 4 (perceived risk of not earning the incentive: low or high) were 0.72 and 0.71 respectively, which were still considered good agreement among the raters. However, compared to the first two items, the kappa values indicate more sources of disagreements between the raters.

4.3.4. Sources of disagreement

Sources of disagreements between the raters were random and not specific to any particular rater. The source of disagreement for the first two items on the typology appeared to be as a result of human mistake. On the other hand, sources of disagreement in the third and fourth item (size of incentive and perceived risk of not earning the incentive) reflected differences in the subjective judgement between the raters. I illustrate in Table 4.8, an example of raters' responses to judging the size of incentive in a P4P study, which according to the typology guideline should be considered small if less than 5% of usual salary or budget and large if 5% or more than usual salary or budget.

[7] The test statistic testing the hypothesis that agreement (kappa) is beyond chance

Table 4.8 An example of source of disagreement between raters

Raters	**Quote/extract from study (An et al. 2008) used by rater**	**Raters response and judgment**
Rater 1	'Clinics that referred 50 smokers would receive a $5000 performance bonus. Clinics would also receive $25 for each referral beyond the initial 50'	"It is unclear because size was not reported relative to budget or salary. I consider it **small** in my judgment"
Rater 2	'Clinics that referred 50 smokers would receive a $5000 performance bonus. Clinics would also receive $25 for each referral beyond the initial 50'	"The study does not specify what percentage of the clinics budget the $5000 represents but I think it has the potential of being a **large** incentive".

Item 4 ('risk') consist of three design features (timing of payment, domain of performance, and performance measure), therefore, there is higher likelihood of disagreement between the raters because differences in judgement of just one of the design features led to different categorisations regarding the fourth item. Table 4.9 on the next page illustrates an example of sources of disagreement on item 4 (risk): both raters agreed on categories of domain of performance and performance measure, but one of the raters was unclear about the timing of payment and had indicated that he/she judged subjectively (the typology states that timing of payment should be considered short if payment is made anytime within four months of measurement of performance, while if payment is made after four months, it should be considered long). The lack of clarity as pointed out by the raters is suggestive of vagueness or lack of structure in reporting design features in the P4P papers.

Table 4.9 Sources of disagreement on judging item 4 ('risk') (Werner et al. 2011)

Rater 1	**Time lag: short or long** **Domain of measurement: within the clinicians control or out of clinicians control** **Performance measure: absolute or relative measure**	**Perceived risk of not earning the incentive: high or low risk**
	Unclear**:** The study does not specify the time lag between performance measure confirmation and payouts. It might have been a short time lag	Low risk
	Processes (within clinicians control); For two of the three clinical conditions we studied, Medicare's composite measures are based exclusively on process measures.	
	Partially relative; Two additional payment incentives were introduced in the fourth year (fiscal year 2007). Hospitals that attained a target performance level (defined as median performance two years previously) received an incentive. In addition, of the hospitals attaining that level, those that were in the top 20 percent in terms of improvement received another incentive.	
Rater 2	Long time lag: The first two years of the demonstration project (fiscal years 2004 and 2005), financial bonuses were distributed to the top 20 percent of hospitals.	High risk
	Processes (within clinicians control): Participating hospitals received higher payments for treating medicare patients with certain condition- acute myocardial infarction, heart failure, pneumonia, coronary artery bypass graft and knee and hip replacements.	
	Relative: Two additional payment incentives were introduced in the fourth year (fiscal year 2007). Hospitals that attained a target performance level (defined as median performance two years previously) received an incentive. In addition, of the hospitals attaining that level, those that were in the top 20 percent in terms of improvement received another incentive	

4.4. Discussion

This study assessed the inter-rater reliability and ease of use of what may be the first tool to categorise P4P schemes in health care, which is one of the strengths of this study. As a result, however, there were no studies to compare the findings.

Overall, all four items on the typology demonstrated good inter-rater reliability (kappa values were above the pre-set acceptable value of 0.6). This implies that if the typology is adopted as a P4P categorisation tool, misclassifications of P4P schemes will be minimised, which otherwise would have given wrong ideas, making the typology a futile tool to explore heterogeneity in P4P to aid evidence synthesis.

The results suggest that the raters found the typology was quick and easy to use. This, in addition to the training, and clear and concise guidelines, could have contributed to the good agreement between the raters.

The kappa value of item 3 (size of incentive K=0.72) and item 4 (perceived risk of not earning the incentive K=0.71) on the typology was moderately lower than that of item 1 (who receives the incentive K=0.95) and item 2 (type of incentive K=0.91). This might have been because item 1 (who receives the incentives: individuals or groups) and item 2 (type of incentive: fines or bonuses) were typically reported more clearly in the studies, and were easy to identify compared to item 3 (size of incentive) and item 4 (perceived risk of not earning the incentive).

In addition, whilst the findings demonstrated that the raters were clear on how to use the tool and applied it effectively to descriptions of P4P in the selected studies, the raters' interpretations of the authors' descriptions varied due to vague descriptions of the concerned categories in the scheme. Furthermore rater agreement might also have been reduced in item 4 (perceived risk of not earning the incentive: low or high) because it involved judgement of three design features (timing of payment, domain of performance, and performance measure) to decide whether the perceived risk of not earning the incentive is low or high, which means disagreement in one of the three design features often led to disagreement in categorisation of item 4.

The findings further show that the sources of disagreement were likely to be not inherent to the typology, as the findings that the disagreements between raters were random. Arguably, there might be better agreement with more experienced raters conversant with P4P in healthcare. However, this was not explored in this study, as the sample size was restricted to health science research students and early researchers.

The ease of use of the typology and the inter-rater reliability of the third and fourth item can be greatly improved with clearer and better reporting of P4P designs in evaluation studies. The size of incentive is often reported in absolute terms (e.g. £500); and there is often no way to make comparisons with usual earnings of the clinician. For example, an annual incentive of £500 to a clinician earning less than £5000 per year is likely to be considered as large (above 5% of salary) as opposed to a clinical earning £20,000 annually. In the same way, the findings suggest that vague and unstructured

reporting of some design features in the evaluation studies made it difficult for the raters to use the typology. To effectively categorise P4P schemes using this typology, P4P description in evaluation studies needs to be clearer and more structured, which can be achieved by using the typology to describe the design features of these schemes.

Limitations

A limitation of this study is that most of the raters were students at the same University, department, with similar research background or expertise, which might have an influence on reliability. The more similar the rater population is, the more they are likely to agree. Though if the typology is adopted, these are the same sort of people who are likely to be applying the tool.

Another limitation was that the papers assigned to the raters were not randomly selected, which was due to the problem of poor, vague, and incomplete reporting of descriptions and evaluations of incentive schemes in healthcare. Instead, the papers assigned to the raters were selected based on how well described/how completely described the incentive schemes were. This was unlikely to reflect a real life situation where poorly reported papers exist. Inter-rater reliability might have been lower if the papers were randomly selected to include poorly reported and well-reported papers.

4.5. What this chapter adds

This developed typology demonstrated preliminary evidence of reliability when used by health service researchers to categorise P4P schemes in health care. The raters also appeared to understand the typology and found it quick and easy to use. The findings of this study suggests that the typology is ready for use by other researchers, as a simple and effective tool to categorise well reported P4P schemes in health care, which will improve evidence synthesis and aid interpretation of results of incentive schemes in health care. In the next chapter, I use the typology to explore heterogeneity and synthesise evidence on how much influence the four items on the typology (who receives the incentive, type of incentive, size of incentive, and perceived risk of not earning the incentive) have on the effectiveness of evaluated P4P schemes.

Chapter 5 Exploring the heterogeneity of the results of evaluations of P4P in health care

5.0. Introduction

The findings from the review of literature (chapter two) showed the difficulty in synthesising evidence of the effectiveness of P4P schemes in healthcare, which is due to heterogeneity between the results of evaluation of these schemes. This might be explained by variation in design features, contexts, implementation factors, and evaluation design between the schemes. Some researchers have considered some of these variations in these schemes through narrative summaries, with subjective and inconclusive findings (Van Herck et al., 2010, Eijkenaar, 2012). There are however no studies that explore heterogeneity in a structured, quantitative, and systematic way.

In chapter two, I considered some ways to explore heterogeneity in P4P schemes, such as subgroup analyses (a form of meta-analyses for a subset of studies) and meta-regression (an extension of subgroup analyses that quantifies the relationship between the size of effect of the interventions and the categories or characteristics of the studies using weighted regression based technique, thus allowing multiple characteristics to be investigated simultaneously) (Higgins and Green, 2011, Engels et al., 2000). These analyses often require a reliable, structured, and informed framework, which is lacking in P4P literature.

In chapter three I developed a framework (P4P typology) based on some design features of P4P schemes (who receives the incentive, type of incentive, size of incentive, and perceived risk of not earning the incentive), to describe, categorise, and help think about P4P schemes targeted at health service providers. After demonstrating the reliability and potential of the P4P typology as a P4P categorisation tool in chapter four, I then used the typology to categorize evaluated P4P schemes into three theoretical ‘types’ or categories: type A (high chance of effectiveness) B (medium chance of effectiveness) and C (low chance of effectiveness), based on knowledge harnessed from theoretical literature. In this chapter, I describe and present the results of how I used this P4P

typology to explore heterogeneity in the results of P4P evaluations in a quantitative and scientific way.

5.1. Aims and objectives

The aim of this study was to systematically explore heterogeneity in P4P schemes in healthcare using the developed typology.

The objectives were:

- To explore the relationship between the effectiveness of P4P schemes and the design features included in the typology.
- To investigate whether certain 'types' of P4P schemes result in better performance

5.2. Methods

Meta-regression analyses and multilevel logistic regression analyses were performed to explore heterogeneity in results of evaluations of P4P schemes. This involved identification of relevant studies and data extraction, model specification, and sensitivity analyses, which are described in the following sections.

5.2.1. Identification of studies and data extraction

I selected evaluated P4P schemes in healthcare from primary studies identified from the literature review in chapter two, following the same inclusion and exclusion criteria applied in chapter two. In total, 96 evaluation studies were included in this study (see Figure 2.1).

As in chapter 2, data were extracted in a standardized way. A uniform template was used for all studies. In addition to the outcomes and effect sizes extracted in chapter 2 (see Appendix D1), information was extracted from each study using the new typology on design features (who receives the incentive, type of incentive, size of incentive, performance measures, timing of payment, and domain of performance measured), evaluation design (whether there was an adequate control group or not), and country. I also extracted information on sample size and raw numbers of events, from studies that reported them (see Appendix D4).

5.2.1.1. Data features

There were two notable data features. First, the measures of effect extracted were reported in different forms, which included: odds ratio (19%), percentage points (15%),

mean differences (13%), and percentages (45%). These estimates had to be converted to a standardised measure so they could be included in the analyses. This required additional data such as absolute differences (percentages or numbers), sample size, standard deviations or standard errors or variance (Borenstein et al., 2009). I discuss this in detail in the subsequent section.

Second, some of the P4P schemes incentivised a wide range of activities or outcomes, which sometimes were evaluated by more than one study (some of which presented results for more than one incentivised outcome). For example there were four evaluations of the USA premier P4P programme (Jha et al., 2012, Werner et al., 2013, Lindenauer et al., 2007, Glickman et al., 2007), assessing over 15 different outcomes including hospital mortality (for conditions such as pneumonia and acute myocardial infection), prescribing conduct, and smoking cessation interventions. A few of the evaluation studies reported a composite measure of all outcomes considered in the study. For example, the evaluation by Lindenauer et al. (2007) assessing 14 outcome measures on prescribing conduct and appropiate care of heart disease and pnemonia, also provided a composite outcome measure for all 14 outcomes. However, many just reported each outcome separately. This resulted in a non-uniform multilevel/clustered structure to the data, which needed to be taken into account in my analysis (see Figure 5.1).

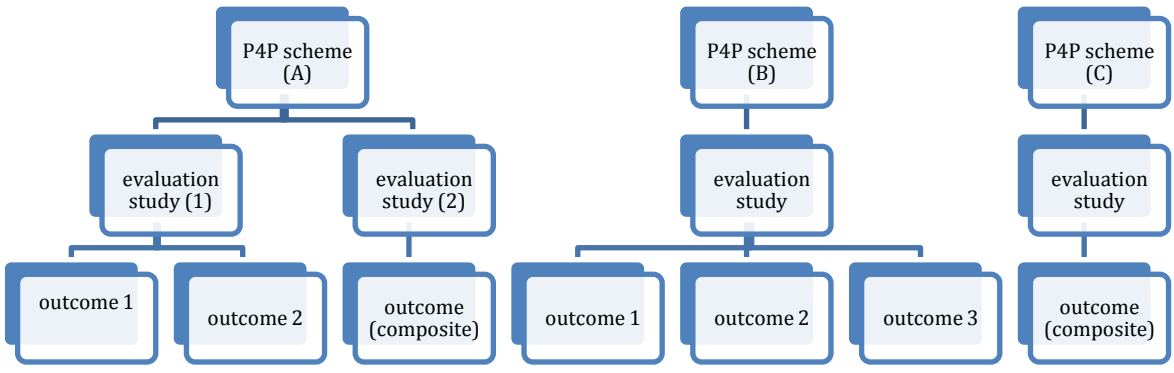


Figure 5.1 Illustration of the multilevel structure of the data

In total, I extracted information on effect sizes of 270 outcomes from 96 evaluation studies, which were from 68 P4P schemes (see Appendix D1). Six P4P schemes were evaluated by more than one study, ranging from two to 22 studies per P4P scheme

(Table 5.1). The number of outcomes per P4P scheme evaluation ranged from one to 77 (see Figure 5.2).

Table 5.1 Distribution of multiple evaluations OF P4P schemes

P4P schemes evaluated by more than one study	Number of evaluations
Quality and Outcomes Framework (QOF), UK	22
Premier, USA	4
California P4P, USA	2
Hudson P4P, USA	2
Payment incentive programme, Australia	2
National Health Insurance P4P, Taiwan	5

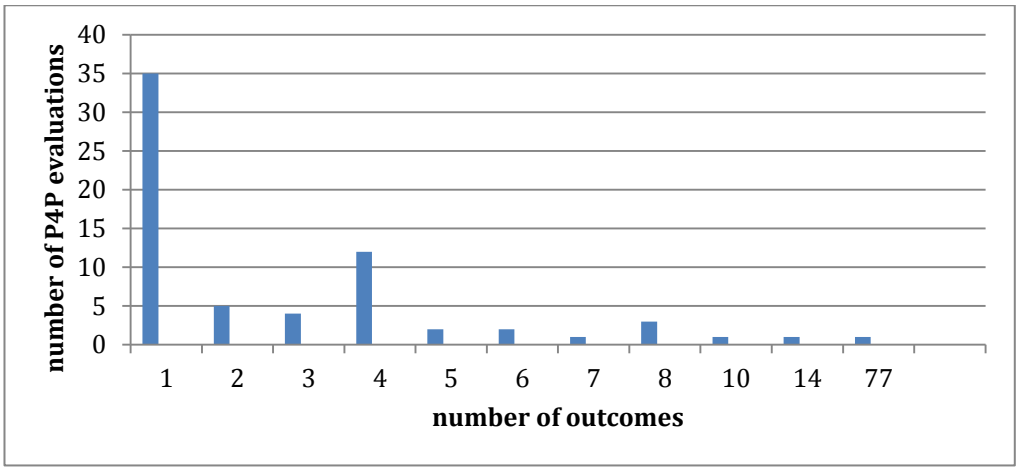


Figure 5.2 Distribution of number of outcomes per P4P evaluations

In the next section, I describe the regression models employed in this study and discuss the consequences of the data features: multiple formats of estimate of effect and multilevel data (multiple outcomes within schemes), and how I addressed them.

5.2.2. Regression models

I used a meta-regression and multilevel logistic regression model to explore heterogeneity in results of P4P evaluations. The following sections lay out the rationales for the selection of the models and exact steps taken to perform both analyses, which included specification of variables.

5.2.2.1. Meta-regression

Meta regression was performed in STATA 12 using the 'metareg' command by Roger Hardboard (originally written by Steven Sharp in 1998). This command performs

random-effects regression using aggregate level data. This extends the variance-weighted least squares regression by estimating an extra additive component of variance, which requires specification of the standard errors within each study (Harbord and Higgins, 2008). The regression coefficient obtained describes how the outcome variable (in this case P4P effect estimates: standardized mean difference) changes with a change in explanatory variables (P4P design features) (Borenstein et al., 2009, Morton et al., 2004).

Meta-regression is similar to a simple regression model (e.g. ordinary least square regression). The meta-regression however, was a more robust choice for two reasons. First, it is a weighted regression where larger studies (with smaller standard errors) have more influence than smaller studies, as meta-regression estimates appropriate weights to the studies using the precision of their respective effect estimate (Borenstein et al., 2009). Second, meta-regression considers the residual heterogeneity among intervention effects not modeled by the explanatory variables, using the random-effects meta-regression model (Thompson and Sharp, 1999).

The data features highlighted in the previous section (multiple effect estimates measure and multiple outcomes per evaluation study) were taken into consideration in the meta-regression analyses. I explain how I dealt with these issues in the sections below.

Converting effect estimates to a standardised measure (multiple effect estimates measure)

I converted different formats of effect estimates to a standardized measure in order to be able to include as many studies as possible in the analysis. Using formulae from Borenstein et al. (2009), I converted the different formats of effect estimates (e.g. odds ratio and percentage points) to standardized mean difference (a standardized measure), with associated standard errors (SE) (see Table 5.2).

Unfortunately, in this study out of 96 primary studies identified to be included in the analysis, only 36 studies reported enough information to compute the standardized mean difference, unavoidably resulting in loss of information[8] (see Appendix D3 for list of included studies). I considered avoiding this loss of information by fitting a separate meta-regression model for each reported format of the outcome effect. However, each

[8] Authors were not contacted due to time constraints

model would have had insufficient number of studies (small sample size<10 studies) to confidently carry out the desired exploratory analyses (Borenstein et al., 2009) (see section 5.2.1.1.). In addition, even if this were possible, the results would have been difficult to merge and interpret as a whole.

Table 5.2 Formulae used in converting effect estimates to standardised mean difference

Formulae obtained from Bland (2000) and Borenstein et.al. (2008)

$$\text{Standardised mean difference } (d) = logOddsRatio \times \frac{\sqrt{3}}{\pi}$$

$$\text{Variance of converted odds ratio } (V_d) = V_{LogOddsRatio} \text{ X } (3 \div \boldsymbol{\pi}^2)$$

Where $\boldsymbol{\pi}$ is the mathematical constant (approximately 3.14159), $V_{LogOddsRatio}$ is log of the variance of the odds ratio.

$$\text{Standardised mean difference (d)} = \frac{Mean\ difference}{Standard\ deviation\ (SD)}$$

Where SD = $\sqrt{N}$ * $\frac{\text{(95\% upper limit confidence interval (CI) - 95\% lower limit CI)}}{3.92}$

Variance of standardised mean difference (V_d) =SE^2

Standard error (SE)= $\frac{\text{(95\% upper limit confidence interval (CI) - 95\% lower limit CI)}}{3.92}$

Another approach used to address the problem of loss of information and still statistically explore the data was to convert the effect estimates of all the studies to binary data (e.g. was P4P effective or not), which was analysed using a logistic regression model. This eliminated the problem of differences in effect estimate measures, and allowed incorporation of all (96) relevant P4P evaluation studies. However, because effect sizes were not needed to transform the outcomes to binary, the magnitude of the effect size was lost (there was no differentiation between evaluations with large or small effect estimates). Therefore, the results from this analysis were interpreted in light of this limitation (Spitznagel, 2008, Greene and Hilbe, 2008). I examined the extent to which the results from both analyses were similar to increase confidence in the findings (this logistic regression analyses is described in section 5.2.2.2).

Adjusting for multiple outcomes within evaluation studies

Second, it was necessary to account for the clustered nature of the data (multiple outcomes within a study) in the meta-regression, because if unaccounted for, it might lead to incorrect estimates of the standard errors, and so, the precision of the summary of effect (Snijders and Bosker, 2012, Borenstein et al., 2009). The clustering observed in this dataset means that the outcome variables within each cluster at each level (outcomes within studies or outcome within programmes) might be dependent or correlated with one another and likely to have similar results (Luke, 2004). Correlations between the outcomes in these studies are likely for a number of reasons. Outcomes measured from the same population of clinicians, outcomes evaluated using the same study design, and outcomes focused on the same clinical area or domain of performance, are all likely to be correlated because the outcomes are not independent of one another. Therefore, ignoring the clustering nature of the data (treating each outcome as independent) is likely to underestimate the standard errors associated with the coefficients because sample size or amount of independent information is exaggerated (as the analysis assumes that each observation is independent of one another), which results to misleading conclusions (Van den Noortgate et al., 2005, Snijders and Bosker, 2012).

There were a number of ways to deal with this. The first way considered was to select only one outcome per P4P scheme. However, this would have resulted in a substantial loss of information. The second way considered was to compute a summary measure for each P4P evaluation study with multiple outcomes, which takes into account the correlation among the outcomes (Borenstein et al., 2009). To do this, I used formulae shown in Table 5.3 to calculate the summary effect and its associated variance to estimate its standard error for multiple outcomes within each study (see Appendix D4 and D5).

Table 5.3 Formulae for computing summary effect of multiple outcomes within scheme

Formulae from Borenstein et al. (2009)

- Summary effect for two outcomes in a study $\sum Y = \frac{1}{2}(Y1 + Y2)$

Or

- Summary effect for more the two outcomes in a study $\bar{Y} = \frac{1}{m}\sum_{j}^{m} Yj$

Where Y1, Y2 etc. are effect sizes from different outcomes, m=number of studies and $Y_{j=}$ sum of the effect sizes in the study.

- Variance for mean of combined outcomes within a study= V_Y * Variance inflation factor (VIF)

- $V_y = \frac{1}{M}v\ (1+ (m-1)\ r)$

- VIF= 1+(m-1)r

Where m= number of outcomes, V= sum of all the variances of the combined outcomes, and r = the correlation coefficient that describes the extent to which the outcomes in the study co-vary.

The problem, however, with using the formulae in Table 5.3 was that the correlation (r) between the outcomes were not reported in these studies, and so there was no data on the likely degree of clustering. Therefore, I had to make assumptions on plausible correlations between multiple outcomes (within P4P schemes) based on their similarities, as suggested by Borenstein et al. (2009). For instance, P4P schemes with multiple outcomes within the same clinical area (e.g. diabetes related outcomes such as LDL tests and HbA1c outcomes) are likely to be more highly correlated, compared to outcomes (within the same P4P scheme) in different clinical areas (e.g. childhood vaccinations and reduction in blood pressure). A correlation of one (perfect correlation) or zero (outcomes are completely independent) is unrealistic and unlikely in this case because multiple outcomes within P4P studies are likely to be correlated to a certain degree i.e. same programme, same method of evaluation, same population etc. (as discussed in earlier in section 5.2.2). Therefore, I assumed that 0.5 was a plausible correlation between multiple outcomes in a P4P scheme. The assumption of correlation of 0.5 was based on taking a conservative position of a midway value between zero and one. However, based on the argument that correlations between outcomes within a scheme are likely to be higher if they were around the same clinical area (e.g. diabetes

outcomes such as eye tests, and hba1c levels), I assumed a higher correlation of 0.75 for such outcomes.

Since the correlation values used in the above estimations were assumptions, I conducted a sensitivity analysis to see the extent to which results obtained varied for lower values of correlation of 0.5 for multiple outcomes in similar clinical areas and 0.25 for multiple outcomes in different clinical areas within the P4P scheme, as shown in Table 5.4 (see Appendix D4 for summary measures).

Table 5.4 Correlation (r) values used in the estimation of a summary measure of effect for multiple outcomes within P4P schemes

Multiple outcomes within a P4P scheme	**Main analysis**	**Sensitivity analysis**
Outcomes around a similar clinical area (e.g. LDL tests and HbA1c)	0.75	0.5
Outcomes around different clinical areas (e.g. immunizations and blood pressure reduction)	0.5	0.25

5.2.2.2. Logistic regression model (multilevel)

A logistic regression analysis was performed using the outcomes transformed to whether P4P was effective or not (as discussed in the previous section). The logistic regression model allows exploration of relationship between this binary outcome variable and explanatory variables (P4P design features) (Hosmer and Lemeshow, 2000, Kleinbaum and Klein, 2010). It was however, still necessary to consider the multilevel nature of the data, because results from an ordinary logistic regression does not take into account the effects of clustering (Hox, 2010).

The way I dealt with the clustering/multilevel nature of the data was to specify a mixed effects multilevel logistic regression model using the 'Xtmelogit' command on STATA 12, which analyses clustered data (balanced or unbalanced)[9] by accounting for intra-

[9] Unbalanced multilevel data like the one in this study is where the clustering is not equal e.g., there are some P4P schemes with just one evaluation study (with one outcome), whereas there are other P4P schemes with multiple evaluation and multiple outcomes.

class correlation (by correcting standard errors) and correcting denominator degrees for number of clusters (giving an estimate of the variability caused by each level of cluster) (Gelman, 2006). The mixed effects multilevel model consists of fixed effect and random effect elements. The fixed effect part of the model presents estimates that are interpreted similarly to a logistic regression (with robust standard errors) and the random effects part of the model accounts for the clustering effect or the amount of unexplained variation at each level (Snijders and Bosker, 2012, Hox, 2010).

The multilevel logistic regression model was the most robust method of dealing with a three-level data compared to other methods, such as aggregation (which combines the outcomes at lower levels of the data to eliminate the multilevel structure and estimation of clustered robust standard errors), which produces less biased standard errors associated with regression estimates by inflating the standard errors (Snijders and Bosker, 2012). This is because it is often difficult to create a decision rule for aggregation of binary outcome variables, aggregation also leads to loss of statistical power due to loss of information and diminished sample size. Estimation of clustered robust standard errors is mostly useful for two-level data and may lead to misleading conclusions in three level data such as the one in this study (Stata Library, 2014, Moerbeek, 2004, Van den Noortgate et al., 2005). In the next section, I describe how I determined whether P4P was effective or not based on the effect estimates.

Outcome variable specification for logistic regression model

Extracted data of effects estimates revealed four main categories: statistically significant[10] positive effect, statistically non-significant positive effect, statistically significant negative effect, and statistically non-significant negative effect. Outcomes that had statistically significant positive effect were classified under outcomes that P4P were seen as effective, and outcomes that had either statistically significant negative effect or no statistically significant difference between intervention and control group were considered as outcomes that P4P were not effective (see Table 5.5). However, failure to find a statistically significant difference (or positive effect) between the intervention and control group might not necessarily mean there was no effect, because there might have been a small sample size (low power) to detect a statistically significant effect (Borenstein et al., 2009). Therefore, in order to examine how my

[10] Statistically significant here means that probability of the event happening by chance is less than 0.05 or 5%.

decision of outcome variable specification affects the results and conclusions, I performed another analysis in which outcomes with non-statistically significant positive effect were also considered as outcomes where P4P was effective, in addition to outcomes with statistically significant positive effects.

Table 5.5 Outcome variable specification for multilevel logistic regression model

Outcome variable	**P4P was effective**	**P4P was not effective**
Effect of P4P	Statistically significant positive effect	• Statistically non-significant positive outcomes* • Statistically significant negative outcomes • Statistically non-significant negative outcomes

***considered as effective in a sensitivity analyses**

5.2.3. Model specification

In this section, I summarise the statistical models used to explore heterogeneity in the results of evaluations of P4P schemes. This includes outline of the outcome and explanatory variables in each model, and variants of the models.

As discussed in previous sections, two models: meta-regression (model A) and multilevel logistic regression (model B) were considered. The outcome variable for model A was effect estimates (in the form of standardised mean difference) and the outcome variable for model B was binary (whether P4P was effective or not).

5.2.3.1. Explanatory variables for Meta regression and multilevel logistic regression models

In both models, two sets of explanatory (categorical) variables were considered. The first sets of explanatory variables (variant 1) are three design features included on the P4P typology namely: who receives the incentives (individuals or groups), size of incentive (large or small), and perceived risk of not earning the incentive ('risk': low or high). The second sets of explanatory variables (variant 2) are the 'types' of P4P from the typology developed in chapter three. There are three types, namely type A, type B, and type C, which depict the prospects of the effectiveness of P4P schemes. Theory suggests that paying groups instead of individuals, paying large incentives instead of small, and low risk schemes instead of high risk ones are likely to yield better results in

incentive schemes. Following this rationale, I categorised schemes that had all three features present (payment to groups, large incentive size, and low risk incentive) as type A (labelled as P4P schemes with high chances of effectiveness); if two of these features were present, the schemes were categorised as type B (P4P schemes with medium chance of effectiveness); if one or less than one of the design features was present, the incentive scheme was categorised as type C (P4P schemes with low chance of effectiveness). In addition, to the above-specified explanatory variables, I generated an additional exploratory variable (whether there was an adequate control group or not), which was included in both models as an adjuster to reflect the rigour of evaluation (see Table 5.6 for all variables). This follows the argument put forward and explored in chapter two that evaluations of incentive schemes with inadequate control groups are likely to over-estimate the effect of the incentive scheme.

5.2.4. Statistical analyses

Two sets of analyses were performed in each model. First, a univariate analyses for each explanatory variable and second, a multivariate analyses with each set of explanatory variables in the model simultaneously (see Table 5.6). Multivariate analyses were performed because they present a more robust approach to exploring the data, when it is expected that more than one explanatory variable will influence the outcomes of the schemes simultaneously. In addition, multivariate models allowed the exploration of relationships between the explanatory variables (Flom, 2014, Hair et al., 2006).

Table 5.6 Summary of statistical models

	Model A (meta-regression) **Outcome variable: effect estimate (standardised mean difference)**			**Model B (multilevel logistic regression)** **Outcome variable: Binary (whether P4P is effective or not)**
Variant 1 Univariate and multivariate analyses performed	**Explanatory Variables**			
	Who receives the incentive	Groups	*Individual	
	Size of incentive	Large	*Small	
	Perceived risk	Low	*High	
	Evaluation design	No control group (before and after studies)	*Adequate control group (quasi-experimental and RCTs)	
Variant 2 Univariate and multivariate analyses performed	**Explanatory variables**			
	Type of P4P scheme	Type A High chance of effectiveness	Type B Medium chance of effectiveness	*Type C Low chance of effectiveness
	Evaluation design	No control group (before and after studies)	*Adequate control group (quasi-experimental and RCTs)	

*reference category is the group to which the other categories are compared. So, if the estimate of regression in model A for a univariate analyses of who receives the incentive is 4.5, it means than P4P schemes in which groups are paid the incentive are 4.5 times likely to be effective compared to P4P schemes that pay incentives to individuals.

5.2.4.1. Model modifications

There were three slight modifications in the analyses described in Table 5.6. First, in the multivariate analyses of model A (variant 1), 40 studies were required to confidently perform meta-regression analyses on four explanatory variables (following the sample size rule of thumb of 10 studies per explanatory variable) (Hosmer and Lemeshow, 2000). However, there were only 36 studies included in this model and inclusion of evaluation design as an adjuster might have resulted in unrealistic conclusions (Borenstein et al., 2009). Therefore, in order to have meaningful estimates from the analyses, I excluded the evaluation design as an adjuster in model the multivariate analyses of model A (variant 1). Though, the variable was still explored in the univariate analysis of variant 1 and both analyses in Model A (variant 2). Also, a sensitivity analysis that would exclude P4P evaluations without an adequate control group from the model was considered, but was faced with the same problem of inadequate sample size because out of the 36 studies, 10 had no control group, leaving the sample size less than the desired number required per explanatory variable.

Second, as seen in Table 5.6, each explanatory variable is a dichotomous variable. Though, for two of these variables: size of incentive and evaluation design, I originally wanted to explore three categories each:

- Size of incentive: small (<5% of clinician salary), medium (5-10% of clinician salary), and large (>10% of clinician salary),
- Evaluation design: before and after studies with no control group, quasi-experimental studies (controlled trials, interrupted time series, pre-post designs with control groups), and RCTs.

However, this was not done because of inadequate sample size (96 studies), following the guideline of 20 studies per dichotomous explanatory variable. A sample size of over 120 studies would have been required to detect reasonable size effect with reasonable power in multilevel regression models (Harrell, 2010). So, in all the analyses, the medium sized incentives were categorised as large and the RCTs and quasi-experimental studies were grouped together, thereby making both variables dichotomous as seen in table 5.5. I however explored the frequency distribution of the effects of P4P in the original three categories in the two variables (results of which are shown in section 5.3.1).

Third, non-dichotomous variables were not allowed in meta-regression (model A). Type A and type B schemes were merged to form one variable, which was compared to type C schemes in the analyses.

5.3. Results

This section presents the results of the statistical analyses, first in the form of a descriptive summary and then statistical outputs from the regression models (univariate and multivariate) used to explore the heterogeneity.

5.3.1. Descriptive statistics

The descriptive statistics presented in this section describe the distribution of the variables in relation to the outcome variable (of whether P4P was effective or not, as specified in the multilevel logistic regression model (see section 5.2.2.2).

There were 270 outcomes (from 96 studies) across different areas of health care, which

included cancer screening, tuberculosis screening, curative care, disease management, patient satisfaction, smoking cessation advice, hospital mortality, vaccinations, diabetes management, cost containment, change in use of prescription budget, and utilization of health services etc. These outcomes were from 68 incentive programmes, which had different scopes and sizes, and were implemented in different countries and contexts (ownership and funding), with varying levels of complexity.

Illustrated in Table 5.7 are the results from descriptive analysis, which shows that the majority (70%) of the identified outcomes were statistically significantly positive. A majority of these significantly positive outcomes also had favourable design features (according to theory): payment made to groups (81.7%), large and medium size incentives (60.2% and 23.8%), and 'low' perceived risk of not earning the incentive (78.2%). In addition, a majority of the positive outcomes were categorised as type A (57.3%) and type B (27.5%), which were more likely to have statistically significant positive outcomes compared to type C (15.2%). In the same way, a majority of the statistically significantly positive outcomes were from studies that had either an adequate control group (77.5%) or a randomised controlled trial (16.2%), which supports the argument that lack of good control groups might be partly responsible for the effectiveness observed in some of these schemes.

These results mirror theoretical arguments from chapter four, that certain design features are associated with the impact of the scheme. These findings are, however, are not definite. The next section presents more robust evidence of relationships and measures of association between the outcome variables and the explanatory variables from the regression analyses.

Table 5.7 Characteristics of included studies

	Recipient of incentive		Size of incentive			Risk		Evaluation design			Type		
	Individuals	**Groups**	**Small**	**Medium**	**Large**	**Low**	**High**	**No control**	**Quasi-experimental**	**RCT**	**A**	**B**	**C**
Outcomes with statistically significant positive effect (N=190) **n (%)**	33 (17.3)	156 (81.7)	29 (16)	43 (23.8)	109 (60.2)	147 (78.2)	41 (21.8)	93 (48.9)	86 (45.3)	11 (5.8)	102 (57.3)	49 (27.5)	27 (15.2)
Other outcomes (N=80) • **no statistically significant effect** • **statistically significant negative effect** • **negative effect** **n (%)**	12 (15)	68 (85)	21 (27)	12 (15.4)	45 (57.6)	79 (75.2)	26 (24.8)	5 (6.3)	62 (77.5)	13 (16.2)	37 (47.5)	26 (33.3)	15 (19.2)
Total number of outcomes (270) **n (%)**	45 (16.7)	224 (83.3)	50 (19.3)	55 (21.2)	154 (59.5)	226 (77.1)	67 (22.9)	98 (36.3)	148 (55.2)	24 (8.5)	139 (54.8)	75 (29)	42 (16.2)

***type A-schemes with high chance of success, type B-schemes with medium chance of success, type C- schemes with low chance of success *number of studies= 96**

5.3.2. Meta-regression

Table 5.8 shows the coefficients (standardized mean difference) obtained from the meta-regression model. The third column shows the coefficients for each of the dependent variables put in the model individually (univariate model), while the fourth column shows results of all the dependent variables put in the model simultaneously (multivariate model).

Table 5.8 Meta-regression coefficients for Model A (Outcome variable: P4P effect estimate)

	Explanatory variables (Number of studies=36)	**SMD (univariate model) (95% CI)**	**SMD (Multivariate model) (95% CI)**
Variant 1 (First set of explanatory variable)	**Who receives the incentive: payment to groups compared to payment to individuals**	0.002 (-0.186, 0.190) P= 0.981	-0.009 (-0.201, 0.182) P= 0.922
	Size of incentive: large incentive compared to small incentive	0.101 (-0.068, 0.270) P=0.222	0.116 (-0.077, 0.309) P=0.229
	Perceived risk of not earning the incentive (Risk): low risk compared to high risk	0.009 (-0.146, 0.163) P=0.910	0.002 (-0.205, 0.138) P= 0.693
Variant 2 (Second set of explanatory variables)	**Type: type A and B (with high/medium chance of effectiveness) compared to Type C (with low chance of effectiveness)**	0.103 (-0.197, 0.121) P=0.222	0.098 (-0.077, 0.270) P=0.263
		-0.038 (-0.197, 0.121) P=0.629	**Evaluation**: -0.019 (-0.179, 0.142) P=0.815

Table 5.8 shows that in the univariate and multivariate model, the standardized mean difference (SMD) for the variable 'who receives the incentive' were 0.002 (-0.186, 0.190) and -0.009 (-0.201, 0.182) respectively. In the same way, the SMD for the

perceived risk of not earning the incentive (Risk) in both models were 0.009 [-0.146, 0.163] and 0.002[-0.205, 0.138] respectively. These estimates (all of which were not statistically significant, P= 0.69-0.98) were very small and insignificant in magnitude, following guidelines for interpreting these estimates by Cohen (1988). Therefore, it is safe to say that there was no evidence in the meta-regression model that who receives the incentive or the risk of incentive predicts the outcome variable (estimate of effect of P4P).

On the other hand, the regression coefficients from the univariate and multivariate meta-regression models for the variable 'size of incentive' were 0.101 (-0.068, 0.270) P=0.222 and 0.116 (-0.077, 0.309) P=0.229 respectively. Similarly, the coefficients for the variable 'type' for both univariate and multivariate meta-regression model were 0.103 (-0.197, 0.121) P=0.222 and 0.098 (-0.077, 0.270) P=0.263 respectively. Though these estimates were small in magnitude and not statistically significant (using the 5% significance cut-off), the magnitudes of effect were considerably higher (with lower p-values) compared to the variables 'who receives the incentive' and perceived risk of not earning the incentive. In addition, the small sample size may have led to low power to detect a statistically significant estimate. Therefore, with a considerable degree of uncertainty (p-value: 0.222), one can assume that the positive estimates suggest that payment of large sized incentive (above 5% of clinicians salary) and (or) P4P schemes that were classified as type A or B[11] (with high/medium chance of effectiveness) might bring about a bigger effect size than P4P schemes paying small incentives and (or) classified as type C schemes (with low chance of effectiveness).

5.3.3. Multilevel logistic regression results

This section illustrates results from the logistic regression shown in Table 5.9. A test of the full model multilevel logistic regression against an ordinary logistic regression model (likelihood ratio test) was statistically significant (LR test: Chi2 (2) 24.43 prob>chi2= 0.0000), which indicates that the multilevel model was a better fit to the data than the ordinary logistic regression model.

[11] Type A and B has 2 or more of the following design features: payment to groups, payment of large size incentive, and low perceived risk of not earning the incentive, whereas type C has one or less of the design features.

Table 5.9 Regression coefficients for multilevel logistic regression (Model B)

	Explanatory variables (Number of studies=96)	**Odds Ratio (univariate model) (95% CI)**	**Odds Ratio multivariate model) (95% CI)**
Variant 1 (First set of explanatory variable)	**Who receives the incentive: payment to individuals compared to payment to groups**	1.32 (0.32-5.54) P=0.703	2.01 (0.62-6.56) P=0.369
	Size of incentive: large incentive compared to small incentive	4.33 (1.02-18.31) P=0.047	3.38 (1.07-10.64) P=0.037
	Perceived risk of not earning the incentive (Risk): low risk compared to high risk	2.90 (0.78-10.83) P=0.113	0.61 (0.22-1.75) P=0.369
	Evaluation design: No adequate control group compared to RCTs or quasi-experimental studies	23.34 (6.28-86.73) P<0.0001	24.16 (6.31-92.78) P<0.0001
Variant 2 (Second set of explanatory variables)	**Type: type A and B (with high/medium chance of effectiveness) compared to Type C (with low chance of effectiveness)**	3.04 (1.04-11.76) P=0.042	1.81 (0.63-4.14) P=0.225
			Evaluation: 18.49 (4.94-69.19) p<0.0001

- ***Coding for this variable was inverted, making payment to individuals the reference group to aid interpretation of odds ratio (easier to interpret OR >1).**
- **An Odds ratio (OR) of >1 indicates the other group has a higher probability of being successful compared to the reference group**
- **See Appendix D7 multifactorial model**

In Table 5.9, the regression coefficients in the logistic regression model are presented as odds ratio (OR), which is interpreted as If OR (X) >1, the odds of the having a statistically significant positive outcome is X times more than the reference group. Size of incentive (univariate model) shows the OR to be 4.33 (95% CI: 1.02- 18.31) P=0.047, meaning that the odds of having a significant positive outcome in P4P

schemes that paid large incentives was 4.33 times than schemes were small incentives were paid. In the same way, in the univariate model the OR for the variable 'Risk' was 2.90 (95%CI: 0.78-10.83), which means that the odds of having a statistically significant positive outcome in P4P schemes in which the perceived risk of not earning the incentive was low was 2.9 times than P4P schemes where perceived risk of incentive is high. Similarly, the OR of the evaluation design was 23.34 (95%CI: 6.28-86.73), meaning studies with no control groups were 23.34 times more likely to have a significant positive outcome compared to studies with an adequate control group (such as RCTs and quasi-experimental studies). Likewise, the OR of 'type of scheme' in the univariate logistic regression model was 3.04 (95%CI: 1.04- 11.76) P=0.042 means that the odds of type A and B incentive schemes (labelled high and medium chance of effectiveness) was 3.04 times more than type C incentives (labelled low chance of effectiveness).

Table 5.10 Random effects parameters of the multilevel logistic regression model

Random-effects parameters	**Estimate**	**Standard error**	**95% CI**
P4P scheme sd(_cons)	3.51e-08	0.62	0.00-0.00
P4P study sd(_cons)	1.83	0.45	1.12-2.96

sd(_cons): standard deviation at each level

The results presented in Table 5.10 show the random effects parameters of the multilevel logistic regression model, which gives an indication of variation that is unaccounted for in the model through the standard deviations at each level. At the programme level, there appears to be no evidence of unexplained variation, but there is some evidence of some unexplained variation at the study level. A standard deviation of 1.83 at the study level indicates that outcomes in an evaluated P4P study which is one standard deviation above the mean have odds of being effective 6.2 times higher than comparable outcomes in an average evaluated P4P study [exp (1.83) = 6.20].

5.3.4. Sensitivity analyses

Two major assumptions were made in this study. The first one was regarding decisions about selecting plausible correlation sizes to estimate a summary measure and variance for multiple outcomes within evaluations of P4P (see section 5.2.2.1). A correlation size of 0.75 was chosen for multiple outcomes in the same clinical area, while 0.5 was

selected for multiple outcomes in different clinical areas. I then performed the same analysis using calculations of slightly lower correlation values of 0.5 for multiple outcomes in the same clinical area, and 0.25 for multiple outcomes in different clinical areas. The results presented in Table 5.11, show the estimates of regression did not differ, though the 95% confidence intervals differed slightly, but did not lead to different conclusions (compared to those shown in Table 5.8).

Table 5.11 Results for change in correlation values to account for multiple outcomes within schemes in the meta-regression model

	Explanatory variables (Number of studies=36)	**SMD (univariate model) [95% CI]**	**SMD (Multivariate model) [95% CI]**
Variant 1 (First set of explanatory variable)	**Who receives the incentive: payment to groups compared to payment to individuals**	0.002 (-0.184, 0.193) P= 0.989	-0.009 (-0.200, 0.184) P= 0.925
	Size of incentive: large incentive compared to small incentive	0.101 (-0.064, 0.272) P=0.220	0.116 (-0.077, 0.309) P=0.229
	Perceived risk of not earning the incentive (Risk): low risk compared to high risk	0.009 (-0.146, 0.163) P=0.930	0.002 (-0.202, 0.139) P= 0.693
Variant 2 (Second set of explanatory variables)	**Type: type A and B (with high/medium chance of effectiveness) compared to Type C (with low chance of effectiveness)**	0.102 (-0.199, 0.122) P=0.224	0.099 (-0.074, 0.276) P=0.288 **Evaluation**: -0.020 (-0.173, 0.142) P=0.834

Outcome variable: P4P effect estimate (standardized mean difference)

The second assumption was regarding the multilevel logistic regression model (model B), where I had to convert the outcome variable (estimate of effects) to a binary variable (whether P4P was effective of not). I classified outcomes that had statistically significant positive estimates as effective, while others (effect estimates that were statistically non-significant positive, statistically significant negative, and statistically non-significant negative) were classified as not effective. However, because failure to find a statistical significance estimate does not always mean the intervention is not

effective, as it might be due to low power (small sample size), I repeated the analysis, and this time classifying outcomes with statistically non-significant positive effects were classified as effective. The findings from the sensitivity analysis presented in Table 5.12 shows that there were no material changes to the results of the original model (compared to those shown in Table 5.9).

Table 5.12 Results for change in categorisation of Binary outcomes in the multilevel logistic regression model

	Explanatory variables (Number of studies=96)	**OR (univariate model) (95% CI)**	**OR multivariate model) (95% CI)**
Variant 1 (First set of explanatory variable)	**Who receives the incentive: payment to groups compared to payment to individuals**	1.25 (0.31-5.89) P=0.756	1.98 (0.72-6.88) P=0.350
	Size of incentive: large incentive compared to small incentive	4.24 (1.02-17.66) P=0.049	3.36 (1.09-10.88) P=0.039
	Perceived risk of not earning the incentive (Risk): low risk compared to high risk	2.95 (0.78-9.86) P=0.113	0.68 (0.22-1.94) P=0.369
	Evaluation design: No adequate control group compared to RCTs or quasi-experimental studies	23.22 (6.28-85.73) P<0.0001	24.09 (6.31-90.76) P<0.0001
Variant 2 (Second set of explanatory variables)	**Type: type A and B (with high/medium chance of effectiveness) compared to Type C (with low chance of effectiveness)**	3.01 (1.04-11.22) P=0.042	1.9` (0.76-4.87) P=0.229 **Evaluation:** 18.49 (4.43-68.22) P<0.0001

Outcome variable: P4P effect estimate (standardized mean difference)

5.4. Discussion

This study presents results of an exploration of heterogeneity between incentive schemes in healthcare, showing how design features influence effectiveness of these schemes. The results show that there was evidence that some of the design features explored appeared to be significant predictors of the results of P4P (whether it was effective or not). In particular, the size of incentive appeared to be the most significant predictor (OR= 4.33 95%CI 1.02, 18.31). Furthermore, the findings suggest that P4P schemes evaluated without adequate control groups were likely to appear more effective compared to schemes evaluated with adequate controls. Finally, P4P schemes classified as type A or B using the developed typology (from chapter three) were more likely to be effective compared to type C schemes.

In this section, first I outline the limitations and strengths of the study, after which I discuss the results of the study further.

Study limitations

This study had three main limitations. First, due to poor quality of reporting of details (of effect estimates, study sample sizes, and standard errors) in the evaluated P4P studies, 60 out of 96 studies were excluded from the meta-regression model, which led to loss of information. This problem was somewhat mitigated by adapting the outcomes variables of all 96 studies to binary (whether or not there was a statistically significant positive effect), thus allowing inclusion of all evaluated P4P studies in a multilevel logistic regression model. The problem with this, however, was that unlike the meta-regression model, the logistic regression model does not take into consideration each study's sample size (each is treated as a data point of equal weight), which is likely to affect precision of estimates. While this could have been accounted for by adjusting for the sample size in the multilevel model, this was difficult due to poorly reported studies. In addition, there is often the problem of selective reporting (of statistically significant effects) in evaluation studies multiple outcomes (Saini et al., 2014), which further creates uncertainty around the findings. However, since both the meta-regression and the multilevel regression analyses revealed similar trends, some confidence can be placed in the findings.

Second, the limited sample size combined with inclusion of multiple covariates (explanatory variables) simultaneously in the analyses could have led to low power to detect small effect sizes and the failure to find a statistically significant estimate (type 2 error) especially in the meta-regression models (Borenstein et al., 2009). While researchers generally agree on this problem, there is no consensus on how to deal with it. Borenstein et al. (2009) suggest that failure to obtain a statistically significant effect for a covariate should never be interpreted as evidence that there is no relationship between the covariate and the effect size. Therefore, in this study, interpretation of results or measures of association between the covariates and effect size were interpreted with caution (as a trade-off between statistical significance and magnitude of effect), not following the usual simplistic approach of a 5% statistical significance cut off, as a p-value greater than 0.05 might mean that there was not a enough power to detect a statistically significant relationship, as opposed to no relationship between the covariates and the effect size.

Third, it might have been worthwhile to explore the individual effects of the design features (timing of payment, domain of performance, and performance measure) that made up the compressed variable ('risk') in the typology. Nonetheless, these design features were considered to be captured to a certain extent in the quantitative exploration of the variable 'risk' in this study.

Fourth, studies with poor methodological quality or poor evaluation designs were not excluded from the statistical analyses due to the limited number of studies. I however, included the method of evaluation design in the regression models, which adjusts to some extent for the possible bias, caused by lack of adequate control groups.

Strengths

This study had three major strengths. The first strength of this study is that it is the first study to systematically and statistically explore heterogeneity in P4P schemes in healthcare.

Second, this study improves on previous narrative reviews because it gives insight to previously unexplored areas such as the proposed relative importance of these design features, and testing the predictive validity of the novel typology tool for categorising incentive schemes in healthcare.

Third, heterogeneity was explored using two methods of regression (meta-regression, and logistic regression), which presents a more robust way of exploring heterogeneity in the dataset.

Therefore, the findings of this study, which are discussed in the following paragraphs, represent an advance in understanding some of the factors (design features) that influence the impact of P4P schemes.

The findings of this study build on previous evidence in four major ways.
First, there was evidence (weak) that some of the design features explored appeared to be significant predictors of whether P4P was effective in the multilevel logistic regression model (larger sample). Though in the meta-regression model, there was no statistically significant evidence of relationship between the effect estimates and the explanatory variables. This was likely as a result of low power due to small sample size (36), as discussed earlier. In addition, because the overall effect of P4P was small, the interactions/relationships between the outcome variable and the explanatory variables in the meta-regression are much smaller, which require large sample sizes to detect it.

The odds of the P4P schemes with large incentives (>5%) having a significant positive outcome was 4.33 (1.02- 18.31) P=0.042 and the odds of the P4P schemes with low perceived risk of not earning the incentive having a significant positive outcome was 2.90 (0.78-10.83) P=0.113 (not statistically significant at the 5% level).

These findings regarding the ‘perceived risk of not earning the incentive’ are consistent with results from the review of reviews by Eijkenaar (2013), that found that incentive schemes were reported as more effective where there was the ‘perceived risk of not earning the incentive was low’, which were characterised by the following design features: smaller time lag between verification of performance and incentive payment, absolute measures compared to relative measure, and process domains of performance compared to outcome domains. However, evidence from reviews regarding the influence size of incentive was inconclusive, even though a few primary studies had reported a dose-response kind of relationship of the size of incentive with the impact of the scheme (Van Herck et al., 2010, Eijkenaar et al., 2013).

On the other hand, the variable ‘who receives the incentive’ did not appear to have any association with the outcome variable of whether P4P was effective or not: OR= 1.32

(0.32- 5.54) P=0.703. This contradicts theory, which predicts that paying groups the incentive could bring about better performance than paying individuals because organisations are capable of promoting behaviour change in employees through a wide range of strategies e.g. better structures, improved supervision, enacting stricter guidelines and policies etc. (Stewart, 1998). A possible reason that the design feature 'who receives the incentive' did not predict the outcome variable might be because in most P4P schemes where incentives were paid to groups, individual clinicians benefitted as well. So, it might be difficult to distinguish the singular effect to payment to groups alone and payment to individuals alone. This reflects the conclusions from two narrative reviews, which found that incentives aimed at the individual provider level and/or team level generally reported positive results compared to very large groups (Van Herck et al., 2010, Eijkenaar et al., 2013).

An interesting observation was that some of the results in the univariate model were significantly different to the multivariate model, which is often an indication that some the variables are related (Hair et al., 2006, Flom, 2014). For example, in Table 5.9, the OR of the variable 'who receives the incentive' increases in the multivariate model, while the p-value reduces, which suggests that this variable is not closely related to the others in the model. On the other hand, the OR of the 'size on incentive' (p-value decreased) and the 'perceived risk of not earning the incentive' (p-value increased) decrease in the multivariate model, which suggest that these two variables might be closely related and their effects on the outcome variable are not independent of one another.

The relationship between the size of incentive and the perceived risk of not earning the incentive is logical. It is possible that large incentive sizes might offset the perceived risk of not earning the incentive. For example in situations where the incentive sizes are large (> 10% of salaries or usual budget), clinicians are likely more willing to take on a higher risk to earn the incentive. In other words, even though the perceived risk of not incentives might be high, the promise of a very large reward might still be enough to drive behaviour change. Furthermore, the perceived risk of not earning the incentive is likely not only affected by the design features (timing of payment, domain of performance, and performance measure), but also by context. For example, the perceived risk of not earning the incentive is likely to be much higher in contexts where

there does exist distrust in the payment system. Hence, the relative weight of the 'perceived risk of not earning the incentive' might not be fully captured in this variable.

The second major finding was that in all three design features explored, the size of incentive had the highest magnitude of effect in both models, which suggests that the size of the incentive may be one of the most important design features to consider for the effectiveness of P4P. The findings suggest that P4P schemes paying incentive of 5% and above of (clinician's salary or hospital/group budget) are likely to be more effective compared to schemes paying small incentives of less than 5% of usual salary or budget. These findings reflect evidence from studies, which suggests that payment of incentive of 5% and above (of usual salary) results in a more effective P4P scheme (Chen et al., 2011, Pope, 2011). Furthermore, theoretical literature suggests that large incentives might drive higher performance because of its potential to supplement clinician income and help reach what is known as the target income (Desquins et al., 2009, Evans, 1974) i.e. the larger the size of incentive, the higher the potential of reaching their target income, and the more the clinicians are willing to change behaviour and or improve performance. However, careful attention needs to be paid to the size of incentive because increasing size of incentive or paying large amounts might not necessarily bring about change in behaviour if the clinicians are already close or at their desired 'target income' (Desquins et al., 2009, Evans, 1974). Therefore, it might be worthwhile to explore the average target income of clinicians' participating in P4P schemes on a case-by-case basis, in order to explore the required/adequate size of incentive to motivate behaviour change and improve performance, while maximising returns.

The third major finding in this study is that P4P schemes evaluated without an adequate control group appeared more likely to a positive or greater effect (OR=24.16 95%CI 6.31, 92.78). The findings are in line with the body of literature that suggests that interventions evaluated with inadequate control groups (no RCT or no quasi-experimental studies) are likely to over-estimate effect of incentive schemes in healthcare (Ireland et al., 2011, Shadish et al., 2002, Tilling et al., 2005). RCTs and quasi-experimental studies are likely to provide less biased estimates of the effect of interventions compared to studies with no adequate control groups. This is because adequate controls reduce biases due for example to additional funds or other quality improvement strategies (in the P4P context), which makes it difficult to disentangle the

effects of P4P. Studies with adequate control groups on the other hand, are able minimise the risk of bias, and the effect is often singularly as result of the intervention (Booth and Tannock, 2014, Higgins and Green, 2011).

Another factor surrounding evaluation of the incentives schemes that might influence the outcomes of the schemes, which was not explored in this study is the length of follow up. For example, the USA premier programme evaluated after two years of implementation demonstrated some positive results, while the same programme evaluated after five years demonstrated no significant positive impact when compared to non-incentive control groups (Jha et al., 2012, Lindenauer et al., 2007). These findings demonstrates the need for more rigorous evaluations, at different time points of incentive schemes in healthcare in order to have a clearer idea on the impact of the scheme on clinical effectiveness and sustainability of effect, and possibly, what happens after removal of the incentive.

The fourth main finding in this study is that there was evidence that P4P schemes classified as type A or B using the developed typology (from chapter three) were more likely to be effective compared to type C schemes [OR= 3.04 (1.04- 11.76) P=0.042]. This suggests that the developed typology is useful in helping to explain sources of variation in outcome of the P4P schemes, and thus, could be a useful tool in predicting the outcomes of P4P schemes.

The findings from this chapter that type A and B schemes (high/medium chance of effectiveness) were likely to be more effective than type C schemes (low chance of effectiveness) are consistent with the findings from theoretical exploration in chapter four. Type A and B schemes have two or three of the following features: incentives are paid to groups, large size of incentive, and low perceived risk of not earning the incentives, while type C schemes have one or less than one of the ‘desired’ design features i.e. payment is made to individuals as opposed to groups, small incentives as opposed to large, and perceived risk of not earning the incentive is high as opposed to low.

Whist it is possible that the reason type A and B schemes appear to be more effective than type C schemes might be largely due to the effect of the large size of incentive (as it appeared to be the most significant predictor), these findings are in line with theory.

According to theory, paying incentives to groups instead of individuals is likely to lead to a more effective scheme because groups (or organizations) are probably able to set up good management structures to induce behaviour change (Trisolini, 2011). In the same way, large size of incentive will be more effective in improving performance compared to small incentive size because clinicians are more likely to change behaviour to achieve larger sizes of incentive because of the need to earn a desired income (Desquins et al., 2009, Evans, 1974). Finally, theory suggests that incentive schemes in healthcare are likely to work better if 'perceived risk of not earning the incentive is low'. I.e. if two or more of the following features are present in the scheme: absolute performance measure instead of relative performance measure, short time lag between measurement of performance and payment of incentive, and domain of performance within the clinicians control (processes) instead of outside the clinicians control (outcomes).

P4P schemes with minimal time lags between meeting the target and receipt of the incentive are more likely to show a higher positive behavioural response, because individuals are likely to do things that will bring about immediate rewards instead of reward in a years' time, as individuals tend to place more value on incentives received soon after the behavioural change compared to the incentives received in the future (also known as pure time reference) (Price, 1993, Loewenstein and Prelec, 1992). In the same way, improvements in processes are often seen as more easily achievable or at least under the control of the healthcare organization or clinician, compared with the health outcome measures, which are influenced by a variety of other factors less under the control of the health services. Therefore, outcome measures are likely to appear less effective than process measures, because even though the healthcare professional changes behaviour, it might not necessarily reflect in the final outcomes, as the patient usually needs to change behaviour as well to obtain the desired final outcome. For example, schemes incentivising clinicians for patient smoking cessation are less likely to be successful compared to schemes incentivising smoking cessation advice to patients because in the first scenario, the clinicians are aware that even if they provide the necessary smoking cessation advise, counselling and appropriate help to quit smoking to patients, it is mostly still up to the patient to quit smoking and it is likely that all their 'hard work' will not be recognised and rewarded. Therefore, it might not be worth the effort. Similarly, relative performance measures create greater uncertainty for individuals because their achievement depends on how well others do. Hence they may

be less motivated to invest to improve performance compared to absolute performance measures that are not dependent on how others perform.

Other sources of variation in P4P schemes

Exploration of three key design variables of P4P suggests that (1) size of incentive has the highest influence the results of P4P and (2) P4P schemes with combination of certain design features (such as payment of large incentive, payment of incentive to groups, and low risk of not earning the incentive) increases the schemes chances of effectiveness. However, it is important to note that even a genuine difference between categories is not necessarily due to the essence of their classification (Borenstein et al., 2009). This might be explained by some variation that was unaccounted for at the study level (variation other than design features of P4P that influences results) in the multilevel logistic regression model (as the findings suggest). Study parameters such as setting, area of care, and target population of the health service(s) in question might influence the results of P4P. For example, an evaluation of a P4P scheme with outcomes relating to the management of asthma (mostly affecting children) and diabetes (mostly affecting middle age and the elderly) management might produce different results because they affect different populations. Furthermore, researchers hypothesize that a combination of design features, contexts, and implementation factors are all likely to influence the impact of P4P schemes. It is possible that P4P schemes with design features that increase the chance of effectiveness are likely to be well planned and implemented, which in turn means that it is possible that in most of the schemes where design features are favourable, the contextual and implementation factors are likely favourable as well. However, this might not necessarily be the case. For example, the 'risk of not earning the incentive' (characterised in the P4P typology by certain design features such as timing of payment, domain of performance, and performance measure) might not have been fully captured by these design features, as other factors that might be context or implementation related could also influence the recipients risk of not earning the incentive (e.g. clinicians working in settings where there is low trust in the payment system might have higher level of risk or uncertainty associated with earning incentive, even if the design features suggest otherwise). Other examples of contextual and implementation factors that could P4P schemes include length of programme, clinician awareness of the incentive programme, hospital/health facility preparedness, healthcare system, provider/patient characteristics, dimension of care (prevention or

treatment or disease management), ownership of the scheme, communication within the programme, and degree of clinician inclusion in planning and implementing the incentive scheme (Eijkenaar et al., 2013, Van Herck et al., 2010, Stockwell, 2010, Pierce et al., 2007, Ssengooba et al., 2012).

It is important that these contextual and implementation factors are investigated on a case-by-case basis because what might be a contextual issue in one country might not be an issue in another.

5.5. What this chapter adds

Emerging literature suggest that variation in design features of P4P might explain heterogeneous evaluation results and that certain design features might contribute to success of failure of P4P schemes. A few researchers have attempted to explore this narratively, with subjective and inconclusive results of what works and what does not (Stockwell, 2010, Eijkenaar, 2013). This study presents the first systematic and quantitative exploration of heterogeneity in P4P schemes in healthcare (using the theoretical typology developed in chapter three); the findings of which represent early strides towards the understanding of how design features influence the impact of P4P schemes. The findings suggest that P4P schemes with a combination of certain design features are likely to be more successful than others. These design features include payment of large incentive (>5% of salary or usual budget), payment of incentive to 'groups' (hospitals, clinical groups etc., where individual clinicians may benefit from the incentive), and payment of incentive in which the risk is perceived as low from the perspective of the recipient (meaning the recipients think are likely to get the incentive if they change their behaviour). This study also suggests the potential usefulness of the P4P typology in predicting outcomes, in addition to it being a tool to help categorise and think about incentive schemes in healthcare. These findings are valuable in informing design choices for developing and implementing incentive schemes in different contexts, as P4P scheme developers in health care can pay careful attention to designing schemes using the typology as a guide for choosing suitable design features that are likely to enhance effectiveness. Though there is the need to produce more rigorous and better reporting of evaluations of upcoming P4P schemes to aid further test the hypothesis of design choices affecting the impact if the scheme. In addition, there is

the need to explore other sources of variation in P4P such as contextual and implementation that may influence P4P results, which serve as the basis for the second part of this thesis (where I explore the influence of contextual and implementation factors on a P4P scheme in Nigeria).

Chapter 6 The Nigerian Health System

6.0. Introduction

In the first part of this thesis, I examined the effectiveness of P4P schemes to improve quality of health care by way of a narrative review and a meta-analysis. Evidence from the literature review on the impact of incentive schemes were mixed and emerging literature suggested that effectiveness was linked with the design features of the scheme, contexts, and implementation. The pooled estimate of the meta-analysis revealed that the incentive schemes in health care has a statistically significant small positive effect on improving quality of health care, with substantial heterogeneity between the pooled studies.

I progressed by exploring this heterogeneity by investigating the influence of design features on the effectiveness of the schemes. First, I developed and tested a theoretical typology to categorise P4P schemes based on their design features. After demonstrating the reliability of the P4P typology, I then used regression models to explore to investigate the influence of P4P design features on the impact or effectiveness of the P4P schemes (with the help of the P4P typology). This exploratory analyses revealed that while certain 'types' (characterised by payment to groups, payment of large incentive sizes, and low risk of earning the incentive) of P4P schemes are likely to have better chances of effectiveness, there remains still heterogeneity which might be explained by other factors related to context and implementation. These factors were not captured in the typology because they are often setting specific and acquiring knowledge about such factors often requires preliminary qualitative work that is often not carried out or reported in evaluated P4P schemes.

In this part of the thesis, I explore the influence of contextual and implementation factors on the impact of a new P4P scheme in Nigeria. This was introduced in 2011 as a strategy to improve maternal and child health outcomes by incentivising health service providers to improve quality and increase utilisation of basic maternal and child health service in primary health care facilities. It was implemented in three out of 36 States: Ondo, Nassarawa, and Adamawa.

The choice of this case study was driven by a rationale to contribute and improve a newly introduced P4P scheme (in its early stages of implementation) in a low and middle-income country (LMIC), where the evidence base of P4P is sparse. As a result, the main focus of the study was to harness the findings of my research to improve the effectiveness of the scheme (a formative evaluation) (Øvretveit, 1998).

This part of the thesis is comprised of four chapters. Chapters 6 and 7 focus on the strategic and operational context of the P4P scheme in Nigeria, both of which are important for the providing the background to the formative evaluation of the scheme. Chapter 8 focuses on the rationale and methods employed to conduct the formative evaluation. Finally, chapter 9 presents and discusses the findings of the study, considers the strengths and weakness of the study, and draws out the implications for policy and research of P4P in Nigeria.

In this chapter, I provide an overview of the Nigerian health system. I describe the extent of the issues in the health system, such as poor maternal and child health indicators, poor quality of healthcare, low utilisation of care, and significant health inequalities. I then, move on to discuss previous health reforms implemented to tackle the health care issues and their impact on maternal and child health outcomes (drawing out core underlying challenges of the health system). Finally, I discuss the introduction of P4P to the Nigerian health system and consider how this approach attempts to address the health system challenges. This chapter is largely informed by government documents, country reports, and project documents from international organisations such as The World Bank and the World Health Organisation (WHO)[12].

[12] Data from these organisations often supplements the lack of reliable data in the government bodies in Nigeria.

6.1. Country overview and organization of the Nigerian healthcare system

Nigeria has a population of almost 169 million, with a growth rate of 6.5%, fertility rate of 5.13 children born per woman, with 68% of the population living on less than $1.25 per day in 2012 (The World Bank, 2013). Nigeria has three tiers of government (the Federal, State, and Local Governments) consisting of 36 States (and the Federal Capital Territory), which has 774 local government areas (LGA) (National Bureau of Statistics, 2011). These three tiers of government share the responsibilities of providing public health services in Nigeria (see Figure 6.1).

The Federal Government handles health affairs through the Federal Ministry of Health (FMOH), which is funded by 50% of the total annual country budget allocated to health expenditures. The FMOH is responsible for health policy reforms, provision of technical support to the health system as a whole, and the provision of health services through 71 tertiary and teaching hospitals across the nation (Scott-Emuakpor, 2010, Abimbola et al., 2012).

The State Government through the State Ministry of Health (SMOH) is responsible for providing health services through the State hospitals, which is funded by 25% of the total annual country budget allocated to health expenditures. These State hospitals include general and specialist hospitals, which are responsible for treatment and management of complicated diseases. The SMOH also provides technical support to the next level of government which is the local government area(s) within each State (Abimbola et al., 2012, WHO, 2013, National Bureau of Statistics, 2011).

The Local Government Areas (LGAs) through the primary healthcare department are responsible for primary healthcare (PHC) services, which is funded by 25% of the total annual country budget allocated to health expenditures (Abimbola et al., 2012, WHO, 2013, National Bureau of Statistics, 2011). The PHC facilities serve as the first point of contact for basic healthcare services, preventive care, community health hygiene, and sanitation. There are 22,000 PHC facilities in Nigeria within 5 km of 71% of most households, (Abimbola et al., 2012, Abdulraheem et al., 2012) , which reflect a uniform and sufficient distribution according to the World Health Organisation (WHO, 2013).

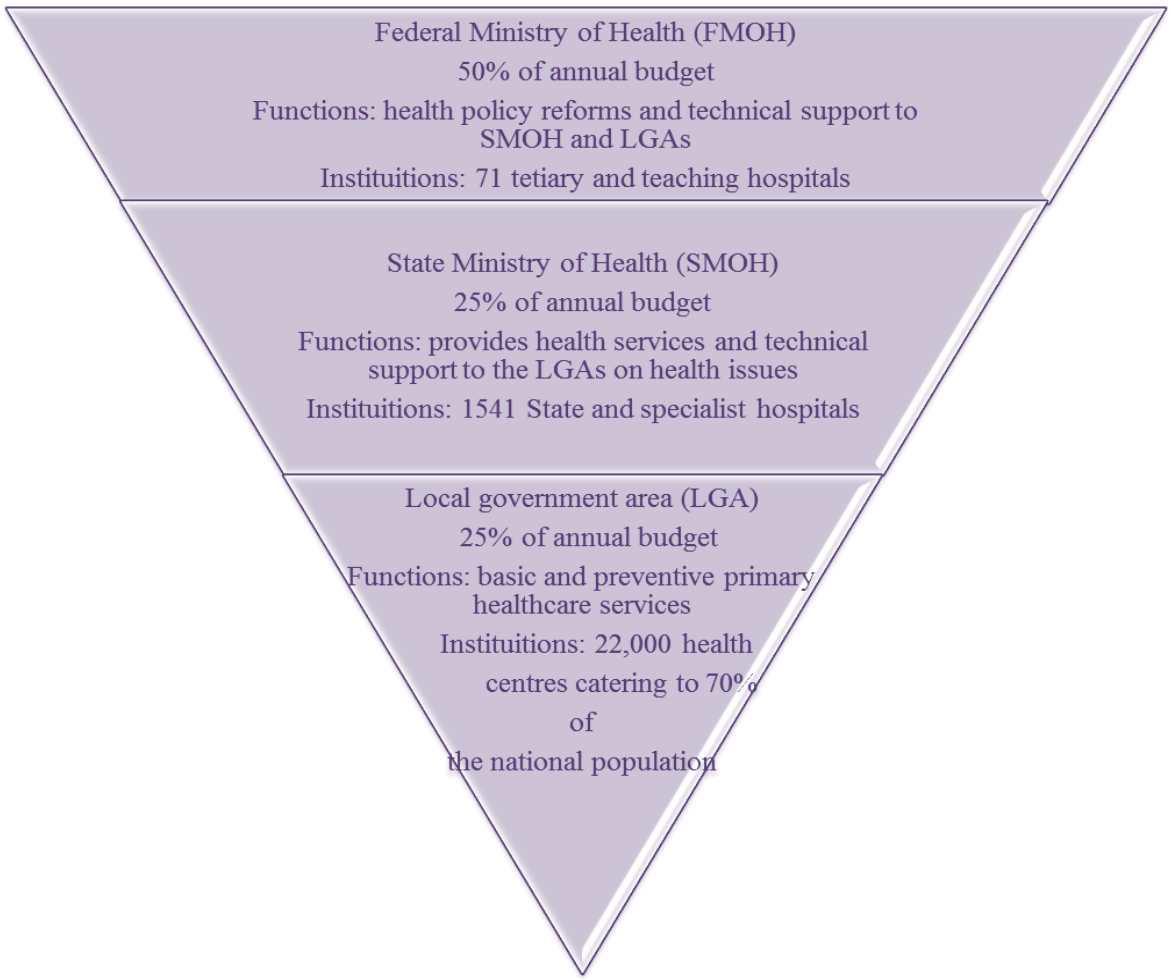


Figure 6.1 Structure and function(s) of the Nigerian healthcare system

6.2. Health system challenges

The Nigerian health care system is characterised by poor health outcomes especially in the areas of maternal and child health, despite evidence of cost effective interventions (Abimbola et al., 2012). There are also significant inequalities between urban and rural areas. In addition, public health facilities are underutilised and the quality of health services delivered is poor. I discuss the challenges below.

6.2.1. Health status overview (child and maternal health)

Over one million children (under five) die annually in Nigeria (second highest in the world after India), a mortality of 130 per 1000 live births. In addition Nigeria has the fifth highest maternal mortality rate (630 per 100,000 live births) in the world (The World Bank, 2013). About 70% of the causes of under-five deaths and maternal deaths are preventable and/or easily treatable (Hogan et al., 2014, Ebeigbe, 2013, Khan et al., 2006). These include conditions such as pneumonia, diarrhoea, malaria, measles, HIV/AIDS, pre-term birth, birth asphyxia, and infections for child deaths; and conditions such as anaemia, HIV, ectopic pregnancies, haemorrhage, sepsis, abortions, obstructed labour, and hypertensive disorders for maternal deaths (see Figure 6.2).

Studies have shown that access to antenatal care (ANC), skilled birth attendant (SBA), and postnatal care (PNC) can reduce maternal and neonatal mortality by up to 60% (Smaill and Hofmeyr, 2002, Hogan et al., 2014, Duley et al., 2003, Yakoob and Bhutta, 2011). Similarly, access to basic health care for treatments of ailments such as malaria, diarrhoea, and access to vaccines can reduce child deaths by up to 80% (Jones et al., 2003, Bhutta et al., 2003, Eisele et al., 2012).

In the Nigerian health system, the PHC facilities are responsible for providing services to treat a majority of the causes of maternal and child deaths. These PHC facilities are uniformly distributed across Nigeria (WHO, 2013). However, there has been no decrease in maternal or child deaths (Abimbola et al., 2012, Abdulraheem et al., 2012). This might be as a result of low utilisation and (or) poor quality of health services in the Nigerian PHC facilities, which are discussed in the next sections.

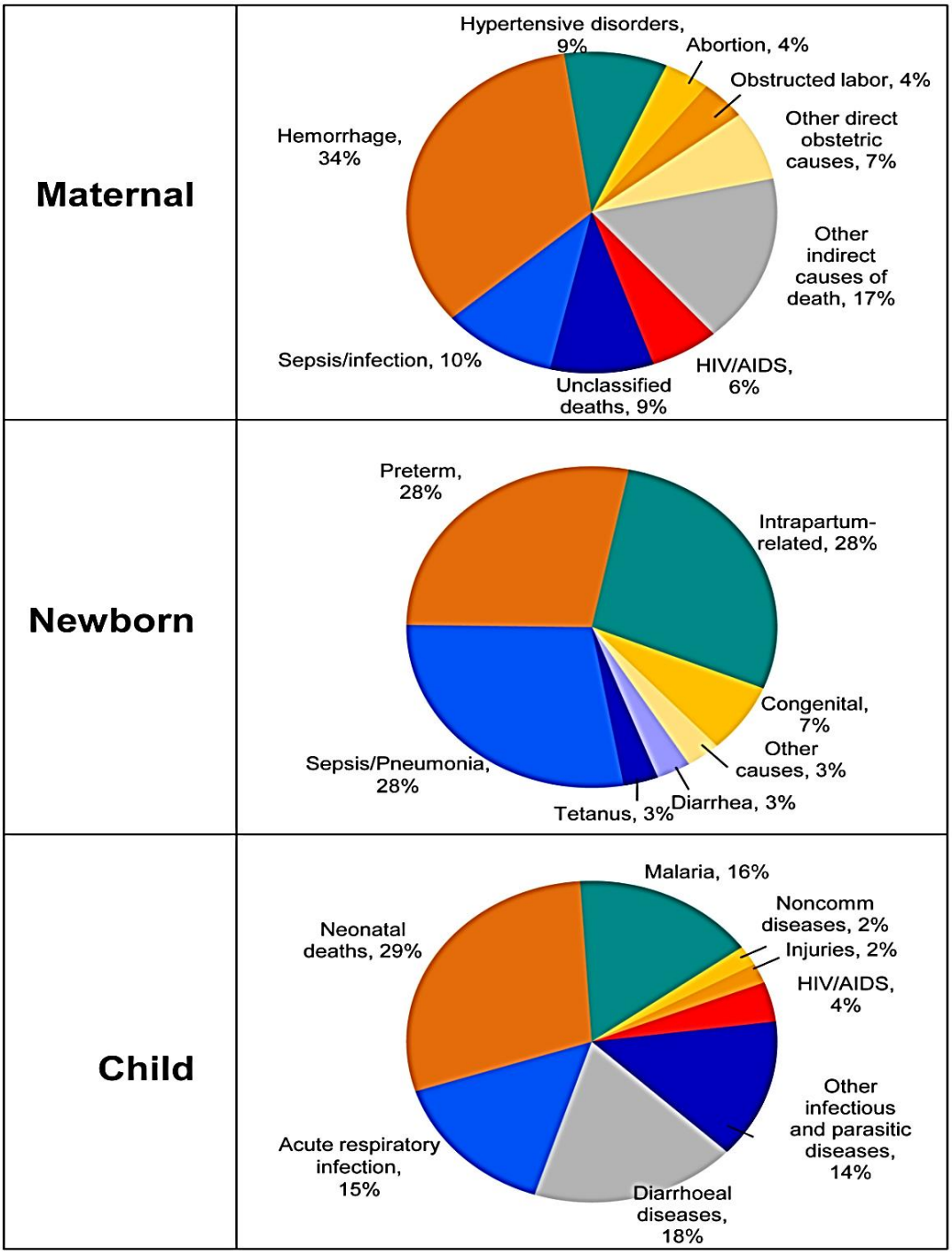


Figure 6.2 Sources of maternal and child deaths in Nigeria (Hogan et al., 2014).

6.2.2. Distribution of utilisation of maternal and child health services in Nigeria

Evidence from Demographic and Health Surveys from 1990-2008 indicate that there was only a limited increase over time in the use of maternal and child health services. Figure 6.3 shows that since 2003, there have been modest improvements in childhood immunizations and treatment of pneumonia, whilst use of other services such as diarrhoeal disease treatment, antenatal care, and skilled birth attendance has slightly declined (National Population Commission, 2009, National Bureau of Statistics, 2011, Nigeria count down to 2015, 2012).

About two thirds of the children aged 12-23 months had received BCG (tuberculosis vaccine) by 12 months, but only 43% received the recommended three doses of DPT (diphtheria and pertussis (whooping cough) and tetanus toxoid). About 46% received the third dose of polio vaccine, 49% coverage for measles and 40% coverage for the yellow fever vaccine was achieved. This suggests that about half of the population of who require these maternal and child health services receive them. Furthermore, evidence from the National Bureau of Statistics showed that around 58% of women (aged between 15-49 years) with a live birth(s) received ANC at least once by a skilled personnel and 39% of pregnant women (aged 15-49 years) births were attended by skilled personnel.

In the subsequent sections, I consider some of the reasons for low utilisation of maternal and public health services, such as health inequalities, cost of health care, and quality of health services.

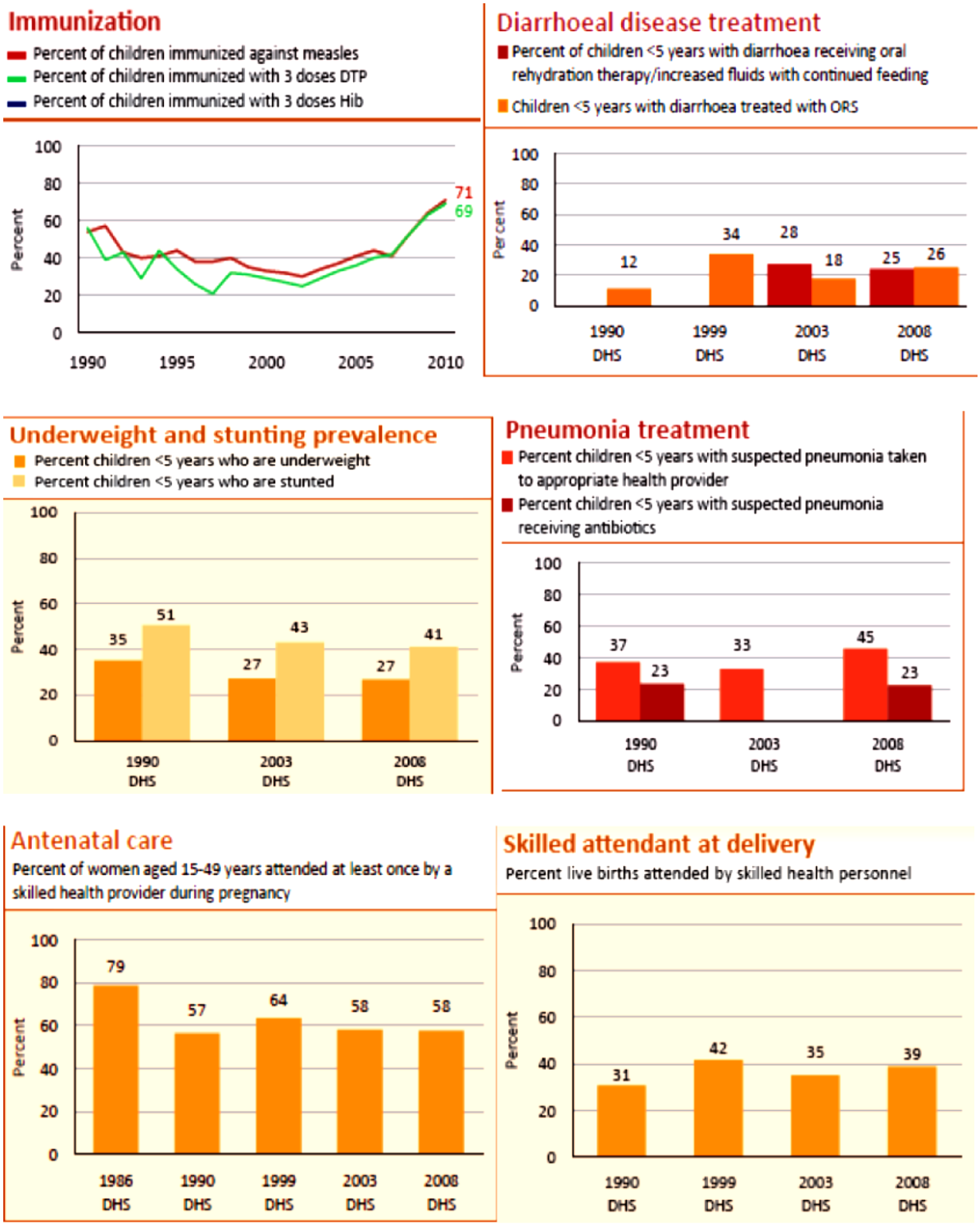


Figure 6.3 Utilisation of maternal and child health services in Nigeria (Source: Nigeria countdown to 2015, 2012)

6.2.3. Health inequalities

The country level statistics presented in the previous section hide variations in health outcomes and indicators in Nigeria. There are significant disparities in health outcomes between rural and urban, despite similarities in government expenditure in healthcare (National Bureau of Statistics, 2011, Federal Republic of Nigeria, 2010). For example, under-five mortality in rural areas exceeds 180 per 1,000, while it is less than 110 per 1,000 in urban areas (see Figure 6.4). Also, the infant mortality rate in urban areas is 68 per 1000 births compared to 110 per 1000 in the rural areas (National Bureau of Statistics, 2011). In the same way, utilization of maternal and child health services are often lower in rural areas where the poorest of Nigerians reside (National Bureau of Statistics, 2011).

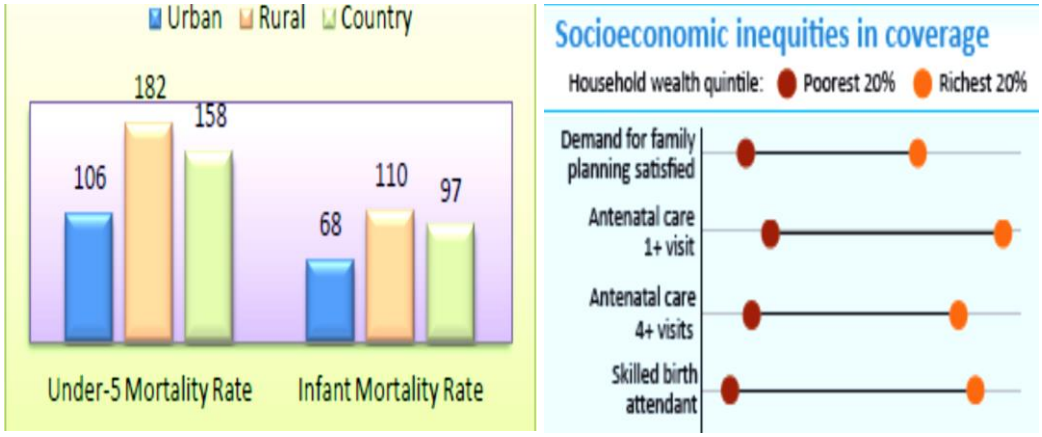


Figure 6.4 Inequalities in health outcomes and health service utilisation (National Bureau of Statistics, 2011)

Health service researchers have argued that the poor and unequal utilisation of the maternal and child health services in Nigeria stems from two key factors: cost of health care and perceived quality of care by the public (Gustafsson-Wright and van der Gaag, 2008, Fapohunda and Orobaton, 2014, Abdulraheem et al., 2012), which I now discuss.

6.2.4. Cost of health care

Nigeria's budget for healthcare in 2010 was around 6% of the total annual budget of the country, which amounted to approximately US $5 per capita government expenditure on health (compared to a total of US $60 per capita expenditure on health). This government expenditure on health falls short of the World Health Organization's recommendation of 15 percent or (US $14 per capita government expenditure on health) for developing countries (Federal Republic of Nigeria, 2010, Xu et al., 2010).

Furthermore, due to the decentralized nature of the health care system in Nigeria (see section 6.1), PHC (catering to about 70% of the population) is allocated less than 1.5% of the total country budget, most of which goes towards payment of salaries of the health workers in the PHC facilities (Abimbola et al., 2012, National Bureau of Statistics, 2011). Consequently, over 80% of healthcare expenditure (mostly going towards the purchase of drugs and payment of government set service fees) in Nigeria comes from out-of-pocket spending. This means individuals or households have to pay for health care/services, which is often a limiting factor especially for the poor or rural dwellers in accessing and utilising basic health services, (Riman and Akpan, 2012, Emmanuel et al., 2013, Abdulraheem et al., 2012, Lawanson et al., 2012).

6.2.5. Current state of maternal and child health services (quality of healthcare)

Most of the PHC facilities in Nigeria lack basic equipment, essential drugs, and proper infrastructure, due to insufficient funding, and lack of strategic planning and allocation of resources (Abdulraheem et al., 2012). While the Federal and State health institutions tend to have good drug and equipment supply, the reverse is often the case in the PHC facilities (Abimbola et al., 2012).

About half of the PHC facilities in Nigeria have less than half of the essential drugs (to treat diarrhoea, malaria, anaemia, sepsis etc.), as defined by the WHO. The PHCs also lack basic equipment such as weighing scales, centrifuges, laboratory equipment, and stethoscopes (Akinwale, 2010, Abdulraheem et al., 2012). In addition, most of the health facilities lack basic infrastructure such as running water, 'flush toilets' and adequate beds (Akinwale, 2010). This is because there are often barely any funds left after payment of salaries of the health workers(Akinwale, 2010, Abdulraheem et al., 2012).

Furthermore, health workers assigned to these PHC facilities are often poorly motivated due to delays in low salary payments (sometimes lasting for months), exacerbated by the poor state of the PHC facilities (Akinwale, 2010, Akinyemi and Atilola, 2013). This in turn leads to high health worker absenteeism at the PHC facilities (despite a minimum of three trained health workers including nurses, skilled birth attendants, community health workers, and lab technicians attached to each PHC) (NPHCDA, 2012). Most health workers engage in other activities to supplement their income, which includes selling essential drugs at high prices to the patients (to make a profit), so

making it unaffordable for most users (Akwataghibe et al., 2013). Consequently, the combination of perceived low quality of healthcare by the users and the 'high' cost of health care contributes to the poor utilisation of the PHC facilities (Emmanuel et al., 2013). In addition, some patients rely on 'traditional healers' or quacks who sell counterfeit drugs (cheaper than the original brands) (Garuba et al., 2009, Akinyandenu, 2013).

Having described the main causes of poor utilisation of care and poor maternal and child health outcomes, I now summarise and discuss the health reforms that have been adopted in recent years (before the introduction of P4P in 2011) to improve the health outcomes.

6.3 Past Health Reforms

The FMOH has implemented several healthcare reforms since 2000 to attempt to improve maternal and child health outcomes. This aimed to meet the health related Millennium Development Goals (MDGs) of two-thirds reduction in under-five mortality rate and three-quarters reduction of maternal mortality ratio by 2015. These reforms include the National Health Insurance Scheme (NHIS), National Immunization Coverage Scheme (NICS), Midwives Service Scheme (MSS), and most recently the Nigerian P4P scheme in 2009 (Welcome, 2011, Wagstaff et al., 2006, Haddon, 2013, Uneke et al., 2013). In this section, I summarise these reforms, highlighting their objectives and the approaches implemented to improve utilisation of basic maternal and child health services.

6.3.1. National Health Insurance Scheme (NHIS)

The NHIS was set up in 1999 operating as public-private partnership with the aim of providing accessible, affordable and quality healthcare for all Nigerians (National Health Insurance Scheme, 2012). The main approach of the NHIS is to bear some of the financial risk of incurring healthcare expenses, thereby reducing the cost of healthcare (service and drug fees) in a bid to improve equitable access to health services (National Health Insurance Scheme, 2012, Oyekale and Eluwa, 2009).

The NHIS is funded primarily by contributions from members of the scheme based on their income. For those employed formally, premiums are up to 15% of an individual's

basic salary (the employer contributes 10% and 5% comes from the employee), which covers their spouse and up to four children. Members of the NHIS employed informally (mostly in the rural areas) pay a monthly flat rate payment not related to income, which is meant to cover basic health care (Ilesanmi et al., 2014, Oyekale and Eluwa, 2009).

However, only 3% of the Nigerian population was insured in 2012, most of whom were formally employed individuals residing in urban areas. This suggests that the scheme did not reach the poor and those in rural areas where health outcomes are the worst, due to low willingness to pay premiums as a result of the poverty levels in most households (Oyekale and Eluwa, 2009, Ilesanmi et al., 2014, Ebeigbe, 2013, National Health Insurance Scheme, 2012). This suggests that to improve the uptake of NHIS among the poor and rural dwellers, a pro-poor approach is needed. An example of this approach is the National Health Insurance Scheme implemented by the Rwanda Government in 2000, in which the government identified and paid premiums for the poorest who could not afford it. This resulted in 97% coverage by 2010 and a tripled increase in health service utilizations (Dhillon et al., 2012, Nyandekwe et al., 2014).

6.3.2. National Immunization Coverage Scheme (NICS)

The NICS was launched in 2001 to improve child health outcomes. The goal of the NICS was to reduce the occurrence of the deadly diseases (e.g. tuberculosis, poliomyelitis, diphtheria, whopping cough, tetanus, and measles) through an increase in routine immunisation coverage and the provision of vaccines and supplies (Abanida, 2012, The Complete Laws of Nigeria, 1997).

The NICS is a simple input based financing scheme where the FMOH supports the SMOH and the LGAs in their immunization services by supplying them with vaccines, needles, syringes, cold chain equipment and other logistics required for routine immunizations (Abanida, 2012). Since the scheme was introduced, routine immunization of children increased from about 40% in 2001 to 68% in 2010.

6.3.3. Midwives Service Scheme (MSS)

The National Primary Health Care Development Agency (NPHCDA) under the 2009 Appropriation Act implemented the midwives service scheme (MSS) in 2009, which was specifically aimed at increasing coverage of Skilled Birth Attendance (SBA) in

rural areas to reduce maternal, new-born and child mortality (Abimbola et al., 2012, Federal Government of Nigeria, 2011).

The MSS is a collaborative effort between the three tiers of government in Nigeria, which provided monetary incentives for midwives willing to relocate to rural health facilities. Usual monthly salaries (average of $350) of midwives mobilised to selected facilities were supplemented with $100 by the State government and an addition $150 by the federal government (Abimbola et al., 2012, Federal Government of Nigeria, 2011).

The PHC facilities were selected using rigorous criteria, which included hard-to-reach areas or among underserved populations, availability of potable water supply and basic minimum equipment and laboratory diagnostic facilities for maternal and child health conditions (Abimbola et al., 2012, Federal Government of Nigeria, 2011).

As of July 2010, 2,622 midwives had been deployed to 652 PHCs in rural areas. Preliminary results (outcomes in 2010 compared to previous year) show some increase in maternal health services such as family planning visits, pregnant women with new ANC visits and those with at least four ANC visits, facility-based deliveries, and the number of women receiving two or more doses of tetanus vaccine (see Figure 6.5).

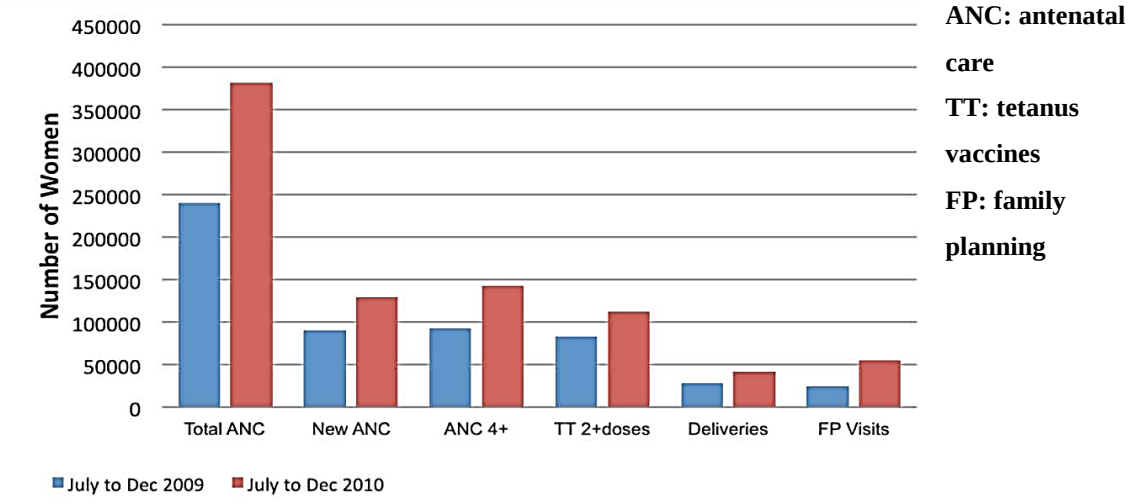
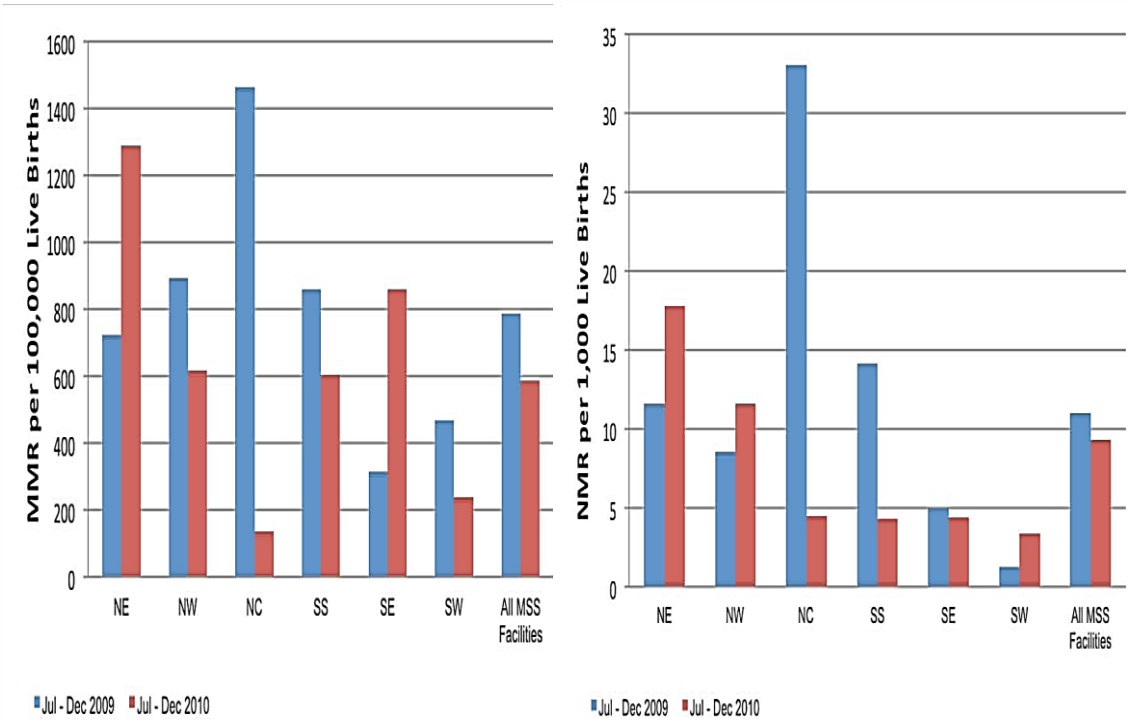


Figure 6.5 Change in utilisation of maternal health services (Abimbola et al., 2012)

There has also been an overall corresponding reduction in maternal and neonatal deaths in the MSS health facilities from 2009 to 2010. Maternal mortality rate (MMR) in 2010 was 572 compared to 789 per 100,000 live births for the same period in 2009 and neonatal mortality rate (NMR) in the same period in 2010 was 9.3 per 1,000 compared to 10.97 per 1,000 live births for the same period in 2009 (see figure 6.6).



NE: north east, NW: northwest, NC: north central, SS: south south, SE: south east, SW: south west

Figure 6.6 Change in Maternal mortality rates (MMR) and Neonatal mortality rates (NMR) (Abimbola et al., 2012)

However, the overall decreases in maternal and neonatal deaths were not uniform across the six regions in Nigeria as seen in Figure 6.6. The MMR actually increased in the North East and South East region, while NMR increased in all except two regions (North central and South-South) that showed dramatic reduction when compared to the previous year. The lack of improvement in MMR and/or NMR in specific zones may have been due to an increase in the proportion of high-risk deliveries in the midwives service scheme (MSS) PHC facilities (Abimbola et al., 2012). Findings from process evaluation of the MSS suggest that the overall limited improvements of utilisation of care and decrease in maternal and child deaths was mainly due to the persistent

challenge of inadequate government spending at the PHC facilities (see section 6.2.5). This often resulted in lack of essential drugs and lifesaving equipment like antibiotics for sepsis treatment and vacuum extractor for obstructed labour (Okoli et al., 2012, Abimbola et al., 2012).

6.4. Evidence of the impact of past health reforms on maternal and child health outcomes

There are no published formal evaluations of the above mentioned these reforms apart from the interim evaluation of the MSS described earlier. This is because of the weak monitoring and evaluation capacity in Nigeria (Igbokwe-Ibeto, 2012, Society for monitoring and evaluation Nigeria, 2012). The impact of the reforms is therefore assessed by the progress of meeting the MDG targets of reduction of maternal mortality ratio by three-quarters and reduction of child mortality by two-thirds by 2015 (Wagstaff et al., 2006, Nigeria count down to 2015, 2012).

Overall, child mortality (under-five) has reduced from 213 deaths per 1,000 live births in 1990 to 143 deaths per 1,000 live births in 2010. Maternal deaths have also decreased from 1100 deaths per 100,000 live births in 1990 to 630 deaths per 100,000 live births in 2010 (see Figure 6.7) (Nigeria count down to 2015, 2012).

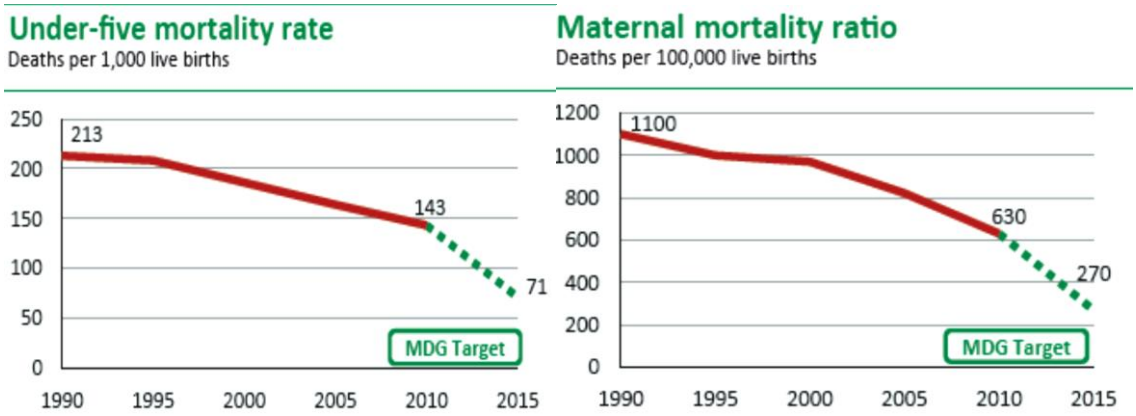


Figure 6.7 Reduction of Child and Maternal mortality rates in Nigeria (Nigeria count down to 2015, 2012)

There has been a gradual reduction in maternal and child deaths over the years in Nigeria. However, the rate of decline to meet the MDG targets is small compared to other African countries such as Rwanda (see Figure 6.8) where under 5 morality rates is very close to meeting the MDG target of reduction by two-thirds (from 151 to 55 per 1000 births), and maternal morality ratio has decreased from 1400 to 320 per live births

(past the MDG targets) (see figure 6.8). The improvements in maternal and child health outcomes in Rwanda might be attributable to the pro-poor approach that has been taken by the government to insure over 90% of the population, and other quality improvement strategies such as P4P in order to strengthen its health system and ensure access to basic maternal and child health services (Dhillon et al., 2012, Basinga et al., 2011).

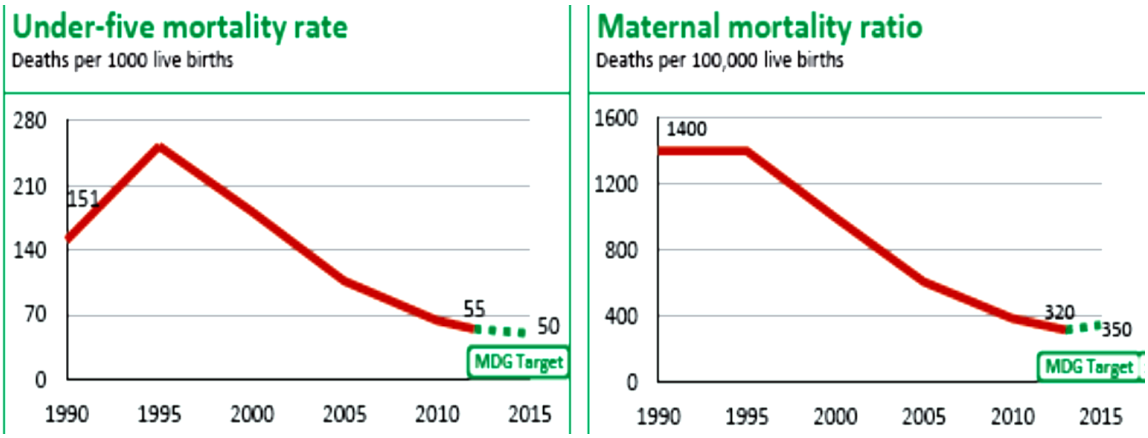


Figure 6.8 Reduction in child and maternal mortality rates in Rwanda

6.5. Failure of past health reforms to meet MDG health targets in Nigeria

In the preceding sections, I have identified some of the reasons why these reforms have not sufficiently increased utilisation of health care and health outcomes to correspond closely to the MDG targets. These include inadequate government spending on primary health care, low willingness to pay insurance premiums especially by the poor and rural dwellers, poor quality of healthcare (lack of infrastructure and equipment), and low health worker motivation. However, central to the persisting challenges in the Nigerian health system are core underlying factors that these reforms have failed to effectively address such as misappropriation of funds, poor governance, and lack of transparency and accountability (Welcome, 2011, Haddon, 2013, Uneke et al., 2013), which I now outline.

6.5.1. Misappropriation of funds (lack of transparency) in the Nigerian health system

The problem of insufficient government spending on healthcare identified in preceding sections is made worse by misappropriation of funds and resource leakage. In 2008

Nigeria was ranked 121 out of 180 countries on the corruption perception index (Transparency International, 2008). In addition, there is substantial evidence that suggests that the annual health budgets drawn do not correlate with actual health expenditures (Nair, 2012, Mohammed, 2013, NPHCDA, 2012, Umukoro, 2012, Eno, 2010). This corruption takes on different forms in the health system, such as overpayments for supplies (ignoring competitive bidding) and payment of salaries to 'ghost' health workers (non-existent individuals put on the payroll in PHC facilities) (McCoy et al., 2008, Health Systems, 2012, Umukoro, 2012). Thus, reducing healthcare funding, translating into drug and equipment shortages, poor infrastructure, and delayed salary payments to health workers at primary health care facilities.

6.5.2. Poor governance and Lack of accountability

The issue of poor governance and lack of accountability in the health system is one of the disadvantages of the decentralized nature of the health care system in Nigeria. Whilst there are clear responsibilities between the different levels of government in the health system, there is often duplication of roles and responsibilities. This leads to weaknesses in coordination, performance tracking, supervision, and monitoring, generally resulting in poor performance in the healthcare services delivery (Khemani, 2004, Abimbola et al., 2012).

Furthermore, there are no formal mechanisms by which health service users in the community can hold the providers (PHC facilities) accountable for provision and access of quality health services. There is also lack of accountability on expenditure and or performance of the PHCs to the Local government (Lawanson et al., 2012, Partnerships for Transforming Health Systems, 2009). The consequences of this at the PHC facilities include poor record keeping and poor transmission of information through the health systems leading to poor allocation of resources and lack of feedback to the PHC facilities (from the Local and State Government, service users) (Partnerships for Transforming Health Systems, 2009, NPHCDA, 2012).

As a result of the multiple underlying challenges outlined, researchers and policy makers have argued that in order to accelerate the improvements of health outcomes, there is a need for a multifaceted pro-poor approach that has the potential of addressing the core persisting problems in the Nigerian health care system (Ilesanmi et al., 2014, Oyekale and Eluwa, 2009, Okoli et al., 2012, Ebeigbe, 2013). This led to the

introduction of P4P in Nigeria (NPHCDA, 2012, Nair, 2011). In the next section, I summarise the main P4P scheme, the focus of this research (a detailed description follows in chapter seven). I also outline how it aims to address the underlying challenges of the Nigerian health system using a multifaceted approach.

6.6. P4P as a strategy to improve the Nigerian healthcare system

There have been a few P4P schemes introduced to the Nigerian health system (NPHCDA, 2012). In this thesis I focus on 'The Nigerian P4P Scheme', which is a large scale scheme implemented by a FMOH Parastatal: National Primary Health Care Development Agency (NPHCDA).

I initially considered focusing on a different P4P scheme implemented by an international non-governmental organisation: 'Save The Children'. This P4P scheme also aimed to improve uptake of maternal and child health services by incentivising health service providers in Nigeria. However, due to terrorist attacks in States (Katsina, Kano, Kaduna, and Zamfara) where the scheme was being implemented, the project was put on hold. Hence the focus of this research shifted to the Nigerian P4P scheme, which was implemented in three States (Ondo, Nasarawa, and Adamawa). It was considered that since both schemes were similar, and implemented in similar contexts, lessons learnt from this research could be applied to the 'Save The Children' P4P scheme if implementation resumes.

The Nigerian P4P scheme was introduced as a pilot scheme that incentivises health facilities in rural areas (pro-poor approach) in three selected States in Nigeria for verified health services (output based) and quality structures such as hygiene and general management. The Nigerian P4P scheme was implemented in December 2011 (to continue until 2018) (NPHCDA, 2012).

The proposal for the introduction of P4P in Nigeria was led by a senior health specialist in 2009 at The World Bank funded by a loan of 150 million US dollars. This represents an influx of additional funds (to a previously underfunded sector), with up to 100 million USD going towards payment of incentives in PHC facilities in the three selected

States over a period of six years, while the remainder of the funds go towards technical support and incentives to LGAs and SMOH (NPHCDA, 2012, Nair, 2011).

The World Bank's rationale for the introduction of this P4P scheme in the Nigerian context was twofold. First, the agency theory that offers the use of incentives as one of the strategies to align the interests of both the principal (purchaser of health care: NPHCDA) and the agent (healthcare provider: PHC facilities) (described in detail in chapter one). Second, the evidence of the effectiveness of P4P schemes in improving quality and utilisation of maternal and child health services in similar LMICs such as Rwanda and Tanzania (Nair, 2012, NPHCDA, 2012). However, as demonstrated in the review of reviews in chapter two, the evidence of effectiveness of P4P schemes in different countries is mixed (including Rwanda and Tanzania), and it is likely dependent on designs (size of incentive, who receives the incentive etc.), contexts (implications of which are discussed in the next chapter), and evaluation design (Van Herck et al., 2010, Witter et al., 2012, Canavan et al., 2008). In addition, P4P could also have unintended consequences, such as falsification of records and neglect of incentivised activities (Gravelle et al., 2008, Doran et al., 2011). Therefore, the implementation of P4P in Nigeria should be with careful consideration. It is important to note that whilst P4P has the potential to improve quality of care, it needs to be implemented using optimal design features and contextual conditions and evaluated with adequate control groups.

Outlined below are the means by which the incentives could be used to align the interests of the PHC facilities (health service providers) and that of the NPHCDA (purchaser). This would address the underlying core challenges that limit utilisation and quality of maternal and child health services at PHC facilities.

- Increased funding/influx of new funds, which could reduce cost of health care through subsidised user fees and drug fees using part of the incentives earned for health services provided (pay for service).
- Increased health worker motivation through monthly bonuses to supplement salaries, leading to reduction in health worker absenteeism at duty posts.
- Improvement in quality of health services delivered: the PHC facilities have immediate direct access to the incentives earned, thereby increasing resources to purchase essential drugs, equipment, and improve infrastructure, as opposed to

usual government funding that may be embezzled before reaching the PHC facilities.

- Improvement in proper allocation and management of resources, achieved through the autonomy given to the PHC facilities in utilising the incentives earned (the Local or State Government have no control on how the incentives are used, this would be determined by the health facility manager). This should reduce corruption/embezzlement and allow the funds to be used in ways that the health facility managers see fit to improve the quality of health services (e.g. purchase of TV sets or provision of alternative power supply to ensure proper running of the health facilities).
- Strengthening accountability, transparency, and good governance simultaneously, through incentives paid to the LGA and SMOH for independent verification of health services delivered and reported by the PHC for incentive payment. This would curtail falsification of data (corruption associated with P4P). Verification of spending and audit trails at the PHC facilities and monthly supervision and monitoring of the quality of care provided by the PHC facilities (through community validation of facility performance and feedback) would strengthen accountability.

6.7. Summary

In this chapter, I have described the elements of the Nigerian health system, the challenges, previous health reforms, and the context surrounding the introduction of the Nigerian P4P scheme. I have also outlined the potential of the P4P scheme to address the core challenges that previous reforms have not been effective at addressing. However, P4P or the use of incentives to improve quality and efficiency of healthcare is not a panacea, as evidence from previous chapters suggests a range of impact from no effect to very effective schemes. Furthermore, findings from previous chapters suggests that important to the effectiveness of P4P schemes are design choices, contexts and implementation. In the next chapter, I describe and review the design features of the Nigerian P4P scheme on the PHC operational level. I also review and discuss the preliminary results of the P4P scheme taking into account context and implementation, before going on to carry out a formative evaluation to explore the influence of context and implementation on P4P.

Chapter 7 Overview of the Nigerian P4P Scheme

In the previous chapter, I have explained the P4P scheme in the Nigerian health system was developed to improve the quality and utilisation of basic health services. The rationale for using the P4P approach is to motivate the health workers either by supplementing their income through bonuses, and or using the incentive to improve infrastructure, drugs, and equipment, if targets are achieved.

In this chapter, I describe the design, implementation, and preliminary findings of the Nigerian P4P scheme.

This chapter is largely informed by project documents (implementation manual and reports) and informal interviews with some of the scheme implementers and designers at the World bank and NPHCDA.

I commence by describing the aim of the scheme and the implementation context, detailing: phases of the programme, funding, and the project sites. I then describe the design features using the typology developed in chapter three. Finally, I conclude the chapter by reviewing the preliminary results of the incentive scheme.

The incentive scheme described in this chapter is referred to as the Nigerian P4P scheme throughout the thesis.

7.1 Aim of the Nigerian P4P scheme

The main aim of the Nigerian P4P scheme is to increase the delivery and utilisation of high impact maternal and child health services and to improve the quality of primary care at selected health facilities in the participating States (NPHCDA, 2012).

7.2. Phases of the Nigerian P4P Scheme

The P4P programme spans across eight years (July 2010- June 2018), and consists of three phases: pre-implementation, pre-pilot, and pilot (Nair, 2011).

7.2.1. Pre-Implementation phase

The pre-implementation phase from July 2010 to November 2011 included the evolution of the programme from conception to first implementation. In this stage, public health specialists from the World Bank proposed the design and implementation of the P4P scheme to gain approval for funding from the World Bank.

The designers proposed the scheme be implemented in primary health care (PHC) facilities in three out of the thirty-six States, representing three out of the six geo-political zones in Nigeria. The States were Adamawa, Ondo and Nassarawa (see Figure 7.1). These States were chosen based on the broad principles of the Country Partnership Strategy (CPS), which include strong governance capability and commitment, greater health needs (priority), willingness to use P4P approaches, geo-political representation, and filling gaps in donor support (i.e. targeting States where international donor support is minimal) (Nair, 2011). The scheme designers also proposed that three States (Ogun, Benue, and Taraba) serve as controls for evaluation purposes (see Figure 7.1) (Nair, 2012). The evaluation design is described in section 7.2.3.

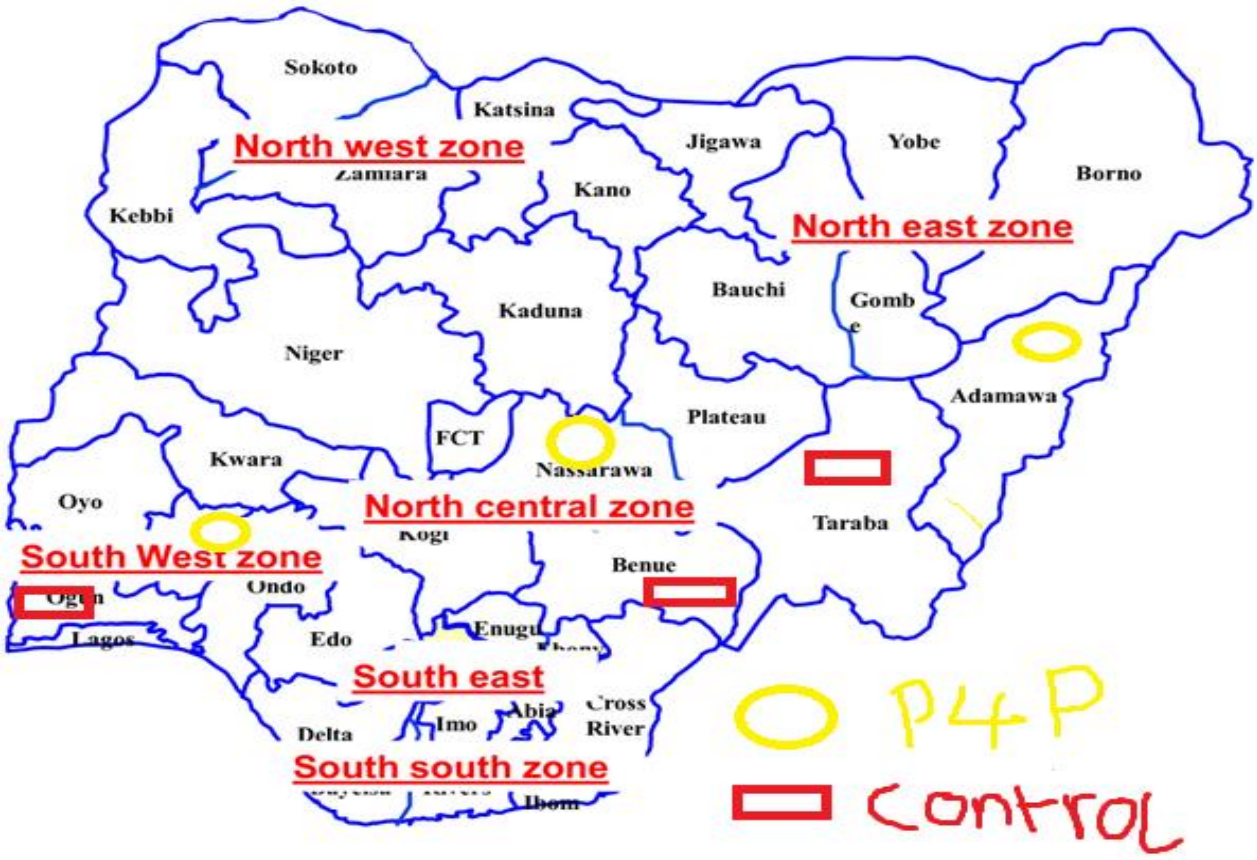


Figure 7.1 Geographical Map of Nigeria indicating the three P4P and control States

The programme design was informed primarily by lessons learnt from similar projects implemented in similar countries by the World Bank. Some of the lessons learnt and reflected in this P4P scheme included:

- A strong capacity building and technical assistance component is necessary for successful implementation of the programme, since it is a new approach in the Nigerian healthcare system.
- Engagement of the Ministry of Health (MOH) at the Federal and State levels is essential for programme ownership and coordination.
- Stringent monitoring of results and evaluation, which is essential to ensure progress towards the aim of the programme (Nair, 2011, NPHCDA, 2012).

7.2.2. Pre-pilot phase

The pre-pilot phase was an experimental phase in which the P4P scheme was rolled out on a small scale in the selected States after the approval of the project by The World Bank. The pre-pilot spanned 36 months (from December 2011 to July 2014) and was implemented in PHC facilities in one Local Government Area (LGA) in each of the States: Adamawa State (Fufore, population: 240,160); Nassarawa State (Wamba, population: 90,454); and Ondo State (Ondo East, 85,323) (see Figure 7.2) (Nair, 2012).

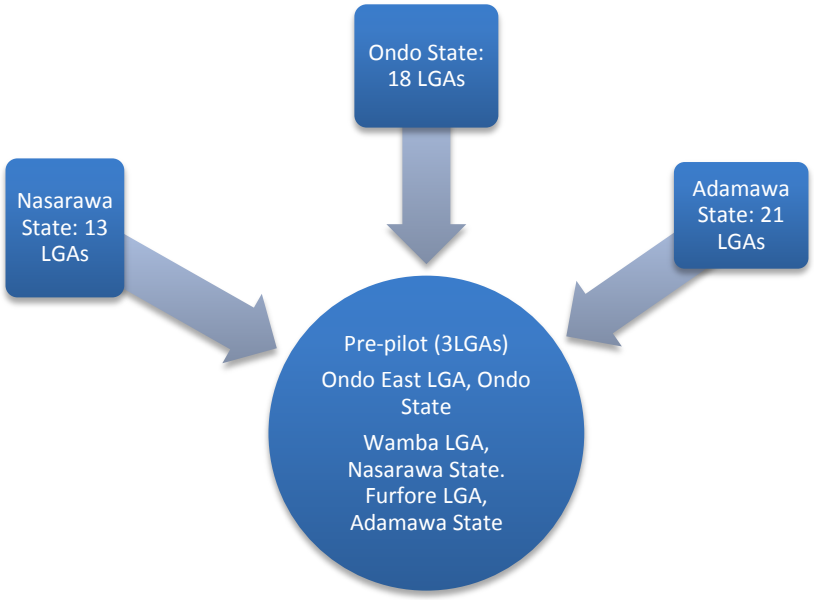


Figure 7.2 Nigerian P4P pre-pilot implementation design

The LGAs for the pre-pilot in Adamawa and Nassarawa were selected at random. In Ondo State, the selection of LGA was politically influenced. The State 'policy makers' argued that the worst LGA (in terms of health indicators and services) should be chosen

(with a rationale that such a LGA will benefit from the pre-pilot) (Nair, 2012). This decision is an example of potential issues the scheme evaluators might face regarding the level of control the State government has over the P4P scheme. However, it has the potential of providing very valuable lessons on context and implementation of the P4P schemes through exploration and comparisons with the randomly selected LGAs in the other States.

The pre-pilots were conducted to inform the implementation of the pilot. This included assessing risk, building capacity, building State specific models of P4P mechanisms, setting up suitable systems, and conducting a formative evaluation (which is the focus of the second part of this thesis). The NPHCDA collected baseline and monthly data relevant to the incentivised services in the all the health facilities in each LGA throughout the pre-pilot phase to monitor preliminary progress (NPHCDA, 2012).

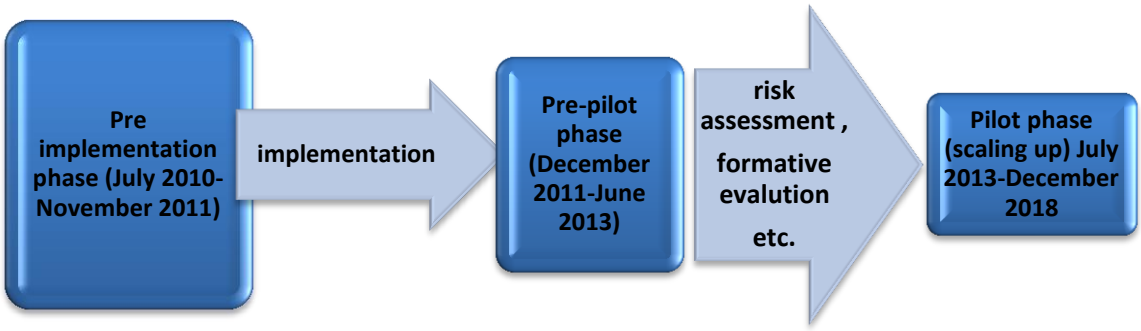


Figure 7.3 Timeline of the Nigerian P4P scheme

7.2.3. Pilot phase of the Nigerian P4P

The pilot phase commenced in July 2014 and will last until 2018. In this phase, the implementation of the scheme has been extended from PHCs in one LGA (in each participating State) to PHCs in all LGAs (in each participating State) after lessons on feasibility, capacity, context, and implementation had been learnt (NPHCDA, 2012). The findings of the second part of this thesis informed this phase by making recommendations informing design, implementation, and context towards a more effective programme.

7.2.3.1. Planned evaluation for the Pilot phase of the Nigerian P4P in 2018

The effectiveness of the Nigerian P4P programme in 2018 will be assessed using a randomised three-armed trial. The LGAs in each State were randomly assigned to P4P or decentralized facility financing (DFF) (providing similar levels of funds but not based on performance). Health facilities in both groups (P4P and DFF) are then compared to heath facilities in non-P4P control States (carefully matched in terms of health and socioeconomic indicators: see Figure 7.1) where there is a 'do nothing approach' or conducting business as usual (NPHCDA, 2012).

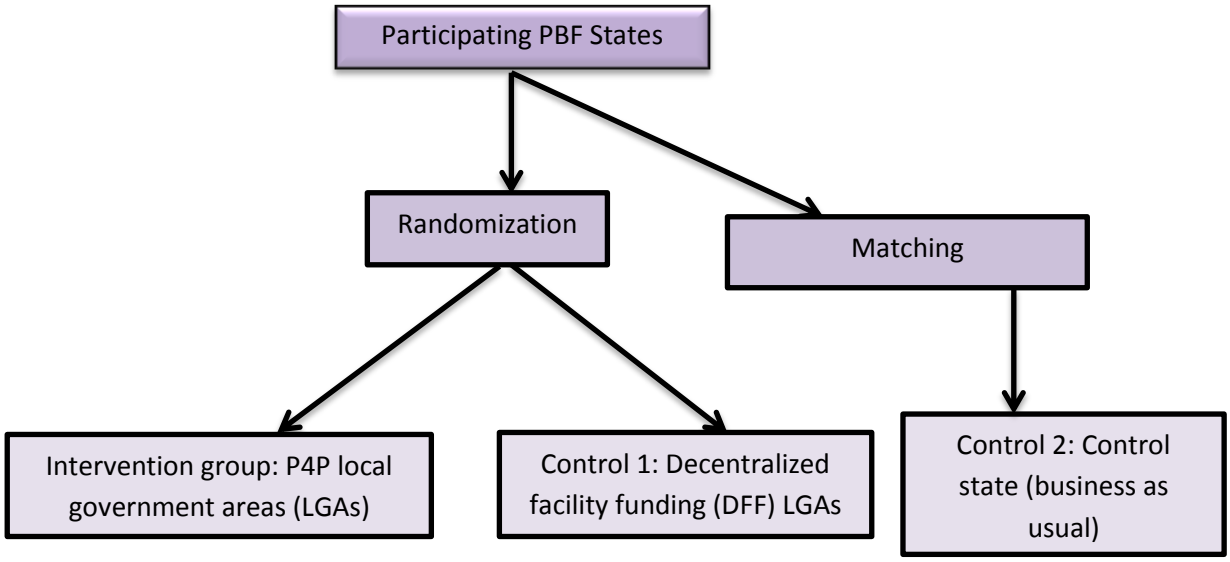


Figure 7.4 Impact evaluation design of the Nigerian P4P pilot (Source: NPHCDA, 2012)

7.3. Approval to conduct research on the Nigerian P4P pre-pilot

The World Bank and NPHCDA granted me approval to conduct research on the pre-pilot in December 2012 after about a year of correspondence, which was facilitated by my supervisor (Professor Trevor Sheldon) and a member of my Thesis Advisory Panel (Professor Alan Maynard). The approval was granted on the premise that the findings of my research would provide recommendations to inform and improve the implementation of the Nigerian P4P scheme in addition to lending some of my research skills to the project while in Nigeria to collect data.

I was then integrated with the rest of the team of researchers and specialists working on the project. For my data collection (a total of about six months), I was based in the

health financing unit of the NPHCDA, which gave me the opportunity to be actively involved in the project through verification exercises, workshops, sites visits, stakeholder meetings, and informal interviews with key Senior Health Specialists at The World Bank such as the head P4P scheme designer, the P4P project manager, and some of the P4P project supervisors. This opportunity to familiarise with and immerse myself in the project improved my understanding and allowed me to make useful observation which partly informed the area of enquiry in my research. This is discussed in detail in section 7.5 where I review the preliminary results of the pre-pilot. In the next section, I review the design features of the Nigerian P4P scheme.

7.4. Design Features of the Nigerian P4P scheme

I now describe the design features of the Nigerian P4P scheme using the P4P typology developed in chapter three of this thesis. The design of the Nigerian P4P scheme described in this section is informed by the project implementation manual (NPHCDA, 2012).

7.4.1. Who receives the incentives? (And timing of payment)

Incentives are paid to the health facility on a quarterly basis. However, the scheme implementers recommend that up to 50% of the incentives earned can be used as bonuses to the health workers (on a monthly basis) and the other 50% for operational cost (maintenance, repair, drugs, consumables, outreach and other quality enhancement measures). Autonomy is given to the health facility on how they allocate and utilise the money (decided by the health facility manager).

7.4.2. Type of incentives: Fines or Bonuses

The Nigerian P4P scheme pays only bonuses (uncoupled from usual salaries) for the achievement of targets. Arguably, theory suggests that fines might be a better incentive to change behaviour, as individuals are usually loss averse (as discussed in chapter 3). However, fines might be more difficult to implement, especially in contexts characterised by frequent industrial actions due to delay in salary payment (sometimes up to 6 months), such as Nigeria (Akinyemi and Atilola, 2013). As a result, part of the rationale behind payment of bonuses is that they would serve as a source of motivation for the health workers to do their jobs while they wait for their salaries to be paid.

7.4.3. Performance measure and domain of performance

This section is discussed in two parts, considering that both the health facility and individual health workers have potential to earn incentives. First, I describe how the health facility earns the incentive (what they have to do and what is measured) and second, I describe what domain of performance is measured for the individual health workers to earn bonuses.

7.4.3.1. Domain of performance for Health facilities

The PHC facilities are rewarded with cash based on the quantity and quality of certain health services they provide (absolute measures). Thus, the incentive is fixed based on each health facility's performance, and in no way dependent on how other health facilities perform. The two areas in which the health facilities are assessed in order to earn incentives are made up of structures and process, which to a certain extent are within the control of the health workers. There are twenty quantity incentivised health services, collectively known as the minimum package activities (MPAs) (see Table 7.1) and 12 quality indicators ranging from availability of essential drugs (to treat minor ailments) to health facility hygiene standards as outlined in Table 7.2 (see Appendix E1 and E2 for a detailed summary of all the indicators).

Table 7.1 Incentivised health services

Minimum Package Activities (MPAs) incentivised at the PHC facilities
1. New outpatient consultation
2. New outpatient consultation of an indigent patient
3. Minor Surgery
4. Referred patient arrived at the General Hospital
5. Completely Vaccinated Child
6. Growth monitoring visit Child
7. 2 - 5 Tetanus Vaccination of Pregnant Woman
8. Postnatal consultation
9. First ANC consultation before four months pregnancy
10. ANC standard visit (2-4)
11. Second dose of Sulfadoxine Pyrimethamine (SP) provided to a pregnant woman
12. Normal delivery
13. Family Planning (FP): total of new and existing users of modern FP methods
14. FP: implants and Intrauterine device (IUDs)
15. Voluntary Testing and Counselling (VCT)/ Prevention of Mother to Child Transmission (PMTCT) test
16. PMTCT: HIV+ mothers and children born to are treated according to protocol
17. Sexually Transmitted Diseases (STDs) treated
18. New Alcohol and acid fast bacilli (AAFB) + Pulmonary Tuberculosis (PTB) patient
19. PTB patient completed treatment and cured
20. Insecticide Treated Nets (ITN) Distributed

Table 7.2 Incentivised quality indicators

No	Service	Points	Weight_%
1	General Management	11	4.4%
2	Business Plan	9	3.6%
3	Finance	10	4.0%
4	Indigent Committee	7	2.8%
5	Hygiene	25	10.0%
6	OPD (Outpatient department)	34	13.7%
7	Family Planning	22	8.8%
8	Laboratory	10	4.0%
9	Inpatient Wards	10	4.0%
10	Essential Drugs Management	20	8.0%
11	Tracer Drugs	30	12.0%
12	Maternity	21	8.4%

The quantity and quality of health services provided are recorded by the health facilities and submitted on a monthly basis to the NPHCDA. They are then verified on a quarterly basis by representatives from the SMOH and LGAs, which are then, counter verified by independent consultants hired by the NPHCDA on a quarterly basis.

The quality measures are verified by inspections done by the consultants, while the verification of the MPAs (quantity measures) involves members of the community. Consultants train members (about five) of the community to go into the community and ask randomly selected individuals from the health facility records (about 10-15 for each health service) if they have received the services recorded by the health facility. The results verified are then calculated as follows: (number verified/sample size asked)*unit price of health service. For example, if the health facility reports that they had 100 obstetric deliveries in the last quarter and out of the 10 randomly selected individuals from the records asked if they delivered in the facility, only 9 confirmed that they delivered in the facility, then the amount verified for which incentives to be paid is calculated thus: (9/10)*100= 90 deliveries. Verification is done to detect falsification of reports, which is not uncommon in incentive schemes (Gravelle et al., 2008, Van Herck et al., 2010). The penalty for evidence of falsification (e.g. substantial discrepancies) in the Nigerian P4P scheme is termination of the P4P contracts of such facilities.

7.4.3.2. Domain of performance for Health workers

The size of bonuses earned by individuals is based on an assessment (carried out by the health facility manager) using five weighted criteria: professional awareness, team

spirit, technical competency, willingness and aptitude for personal development, and availability at work (see Table 7.3). For example, a health worker who arrived frequently late will score two points (25% of eight points) in the domain of timeliness, part of the professional awareness criteria.

While all five criteria used to assess performance to allocate bonuses to the individual health workers are important, and might contribute partly to the performance of the health facility (e.g. coming to work every day, will most likely translate to seeing more patients). It is important to note there is an apparent disconnect between what the health workers are required to do to increase quantity and quality of health services (for the health facility to earn the incentive) and what the health workers are required to do to earn bonuses (out of that earned by the health facility). This might create an issue about whether or not health workers' contribution to the health facilities performance is adequately measured. For example certain core components in which health workers increase utilisation of health care are not adequately reflected in the individual assessment form**,** such as health promotion and outreach (Abdulraheem et al., 2012, Roodenbeke et al., 2011).

Table 7.3 Individual evaluation tool for health workers in the Nigerian P4P scheme

	25%	**50%**	**100%**
Criteria 1: Professional awareness (total of 20 points)			
Timeliness (8 points)	Arrived frequently late *(at the least four times past month)*	Arrived sometimes late *(1 to 3 times per month)*	Was always on time
Availability (8 points)	Has been frequently absent from his/her service without any clear motive (*at the least four times past month)*	Has been a few times absent from his service without clear motive *(1 to 3 times per month)*	Was never absent from his/her service without known and valid motive
Uniform (4 points)	Did not wear a uniform during working hours *(even once per month)*	Neglected uniform *(dirty or torn or not ironed)*	Uniform always worn and proper (*washed ; ironed and not torn*)
Criteria 2: Team spirit (total of 30 points)			
Interpersonal Relationship (8 points)	Frequently in conflict with colleagues *(reported more than once to his/her superior during the past month)*	Sometimes in conflict with colleagues *(reported once to his/her superior)*	Never in conflict with colleagues
Collaborative spirit (8 points)	Frequently refused to assist colleagues when asked (more than once per month)	Sometimes refused to assist colleagues (even once)	Never refused to assist colleagues
Dedication (8 points)	Frequently left work unfinished without somebody taking over under the argument that official working hours were up *(more than 3 times past month)*	Sometimes left work unfinished without somebody taking over using the argument that official working hours were up (1 to 3 times per month)	Never left work unfinished without somebody taking over
Initiative (6 points)	Has never done any additional work	Has always awaited a command from higher up to carry out additional work	Has at least once done additional work without supervisor asking him/her to do so
Criteria 3: Technical Competency and flexibility during work (total of 40 points)			
Organization (10 points)	Never has a daily work schedule *(assessed during internal work supervision)*	Not always has a daily work schedule (*at least once during internal supervision*)	Always has a daily work schedule
Quality of work (14 points)	Never adheres to specific work related	Not always adheres to work	Always adheres to specific work

	25%	**50%**	**100%**
	norms and standards (assessed during internal supervision)	related norms and standards (*found at least once during internal supervision*)	related norms and standards
Quantity of work (16 points)	Never finishes his/her daily work based on his/her own daily work schedule *(assessed during internal supervision)*	Not always finishes his/her work based on his/her own daily work schedule (*found at least once during internal supervision*)	Always finishes his/her work according to his/her daily work schedule
Criteria 4: Willingness and aptitude for personal development (total of 10 points)			
Takes into account advice and recommendations from previous internal and external supervisory visits	Never takes care of such recommendations *(concluded during internal and external supervisory visits)*	Not always takes care of such recommendations (*if this happens once or more)*	Always takes into account recommendations of internal and external supervisory visits
Total 100 points			
Criteria 5: Participation to Results and the Past Monthly Performance Score			
Participation to Results and the past month's performance score (quantity and quality) through presence during working days during the past month, not taking into account reasons for absence such as vacation, leave, sickness, absence through disciplinary action, formal trainings etc.	Percentage of days performed = (P) Number of official working days = (N) Number of days actually worked = (n)	P = (n/N) * 100	
Final result of Individual performance (X) = total points from criteria 1-4 * P% (criteria 5) **Bonus earned= 'X'% * Bonus originally due.** **Sample calculation of bonus due to health worker in a health facility (with 4 other health workers) that earned 250,000 naira in the past quarter** The first step is to divide the incentive the health facility earned into two, as only half of it can be used as bonuses to the staff: 250,000/2= 125,000 naira the other 125,000 naira goes towards the improvement of the health facility. The second step is to divide the bonuses by the total number of health workers: 125,000/ 5 health workers= 25,000 naira per health worker per quarter, translating to 6,500naira (25,000/4) per month (bonus originally due). The third step is to multiply the bonus originally due by the final result of Individual performance (X %). If 'X'= 70, then bonus due to that health worker= (70/100) *6500= 4, 375 for that month			

7.4.4. Size of incentives and payment mechanism

Like most incentive programmes, the size of the incentive is not calculated or reported relative to the health facility's usual income/budget or in the case of the individual health worker, size reported relative to usual salaries. This is understandable in the Nigerian context, as these PHC facilities do not have assigned budgets. The health system funding operates in such a way that money is not allocated to the health facilities, rather an elaborate process of procurement is used to obtain what is needed in the health facility from the LGA (which would either be approved or not depending on available funds).

The potential size of incentive that can be earned by the health facility is however described in absolute terms. Each of the incentivised service carries a unit price (see Appendix E3 for unit prices of all the health services) e.g. normal delivery in the facility = 1000 Naira (so if there were 90 deliveries in the past quarter, the health facilities would expect to get 90,000 Naira for that service). This is calculated for each of the incentivised activities and quality indicators and summed to estimate the total amount of money expected by the health facility. However, the maximum size of incentive that can be earned per quarter in each State is capped at 150 Naira ($1) per capita (Adamawa LGA Fufore: 240,160, Nassarawa LGA Wamba: 90,454, Ondo: 85,323). Meaning at the end of each quarter, on the average, each P4P facility has the potential to earn a maximum of $16,000 in in Fufore LGA (240,160/15 health facilities); in $8200 in Wamba LGA (90,454/11 health facilities); and $8500 in Ondo East LGA (85,323/10 health facilities).

7.4.5. Review of design features of the Nigerian P4P scheme

The design features of the Nigerian P4P schemes summarised indicates that it falls into a 'type A' category of incentive schemes using the P4P typology developed and tested in previous chapters. Evidence from theory (chapter three) and empirical findings (chapter five) suggests that these type A schemes (characterised by large, quarterly, monetary incentives, paid to groups for health services and domains of quality within the health workers control) are likely to be more effective than schemes characterised with other design choices such as payment of small sized incentives paid on a yearly basis, dependent on how others perform (relative measure) for domains of performance out of the clinicians control (e.g. mortality reduction) (see Table 7.4).

Table 7.4 Key design features in the Nigerian P4P scheme

Core design features	Category: 'Type A'
Who receives the incentives	Incentive paid to Groups (health facilities) but individual health workers have the opportunity to earn part of it as bonuses
Type of incentive	Bonuses
Type of payment	Monetary (cash)
Size of incentive	Large (up to 100% of performance budget can be earned)
Payment mechanism	Absolute targets (pay per increase in incentivised activity or quality measure e.g. availability of drugs at the health facility)
Performance measure	Absolute measure (pay per activity)
Domain of performance measured	Within clinicians control (Processes e.g. health service delivery such as ANC and hygiene/cleanliness of the health facility)
Timing of payment	Quarterly: health facility, Monthly: health workers

I have explored and discussed in detail, the influence of design features on effectiveness of P4P in previous chapters, which I now summarise in the context of Nigerian P4P scheme.

First, paying health facilities the incentive could be advantageous because the health facility managers might be able to effect a higher behaviour change in individual health workers and performance in health facilities by implementing good management structures such as supervision, resource management, and motivating health workers (Trisolini, 2011). Similarly, the large size incentive that the health facilities could potentially earn in the Nigerian P4P scheme represents an influx of new funds, which if used effectively, could go a long way in improving problems such as lack of drugs or infrastructure and manpower to improve health service delivery (Abdulraheem et al., 2012, Akinwale, 2010). The large size of incentive also has the potential to supplement the salaries of the health workers, which will be particularly beneficial in the Nigerian context because it has the potential to help the health workers focus on the health facility (thus improving health service delivery), rather than supplementing their income in other ways that take their attention away from the health facility (Abdulraheem et al., 2012). Lastly, in the Nigerian P4P scheme, quarterly payments for health services under the control of the health workers might be translated by the health workers as guaranteed incentive or lower uncertainty of earning the incentive (all things being equal), as opposed to design features, such as yearly payments that are dependent on ranks (based on the performance of others) for outcomes such as mortality. This might be translated as 'unguaranteed incentive (higher uncertainty of not earning the incentive) even if they improve performance or change behaviour. Therefore, the health

workers participating in the Nigerian P4P scheme are more likely to change behaviour in a situation where they view the incentive as more certain because individuals are generally risk averse and would rather invest their time and resources in other activities that give them a higher guarantee of returns on their efforts (Arrow, 1965). However, the uncertainty in earning the incentive might also be affected by contextual factors such as trust in the payment system, which I discuss in the next chapter.

In summary, the Nigerian P4P scheme demonstrates the potential to be effective based on its design features (based on findings from the typology and statistical exploration in chapter 3 and 5). However, as resonated throughout this thesis, variables other than design of the scheme, such as contextual and implementation factors might also influence the effectiveness of these schemes. Therefore, the main focus of the formative evaluation of the Nigerian P4P scheme was exploration of the influence of contextual and implementation factors. In the next section, I review and discuss the preliminary results of the Nigerian P4P pre-pilot, before moving on to explore contextual and implementation factors in the Nigerian P4P pre-pilot in subsequent chapters.

7.5. Early findings of the Nigerian P4P scheme

In this section, I describe how I estimated the effect of the P4P scheme (change in utilisation of health services) in the health facilities where the pre-pilots were being implemented. The aim of this was to provide a rich detail to enable the exploration of contextual and implementation factors in the Nigerian P4P scheme.

7.5.1. Method of estimating change in utilisation of health services

I calculated the change/difference (from December 2011 to December 2012) in utilisation of maternal and child health services in each health facility in all three pre-pilot LGAs using monthly performance data collected and made available by the NPHCDA on the P4P website https://nphcda.thenewtechs.com/data.html). The analysis was done using SPSS statistical package (version 19). The maternal and child health incentivised services were tetanus vaccination for pregnant women, postnatal care (PNC), antenatal care (ANC), malaria treatment for children and pregnant women (Sulfadoxine Pyrimethamine: SP), voluntary counselling and testing (VCT), normal delivery at the health facility, Prevention of Mother to Child Transmission of HIV (PMTCHIV), and complete childhood immunisations.

This method (before and after) used to estimate the impact of the pre-pilot after one year was the only feasible approach due to lack of reliable data on the incentivised services before the implementation of the pre-pilot, a long standing problem with the Nigerian health system (Uneke et al., 2010). This is likely to have introduced some bias because this method does not take into consideration confounders or the variability of the baseline data like other evaluation designs such as RCTs and quasi-experimental studies (Shadish et al., 2002, Stigler, 1997). For example, a previously high performing health facility with little improvements after one year would be indistinguishable from a previously low performer with little improvements. In the same way, improvements noted in certain health facilities might not necessarily be due to the introduction P4P because the health facility might have been steadily improving over the years.

Whilst these are limitations of the analysis that could not be adequately addressed, it was considered to be acceptable for this research because the baseline data showed similar levels of performance in the health services in the health facilities in the three States (see appendix E4). In addition, the effect of P4P in the Nigerian pre-pilots was not the main focus of this research. Rather, the results were meant to provide richer detail to enable the exploration of contextual and implementation factors in the Nigerian P4P pre-pilot, which is the main focus of this part of my research. The next section presents the results of the data analysis of P4P from December 2011 to December 2012.

7.5.2. Results

There was an overall improvement in utilisation of most of the incentivised maternal and child health services in PHCs in the three States as shown in Figures 7.5-7.7). However, the difference/change in utilisation of maternal and child health services varied considerably between health facilities in each LGA. In Adamawa State (see Figure 7.5), Gurin health facility and Chigari health facility showed very good improvement compared to Karlahi health facility (the health services were on the average worse off after P4P) and St. Mary health facility (very little improvement in the health services. Similarly, in Nassarawa State (see Figure 7.6), performance improved considerably in Wamba health facility and Zali health facility compared to Yahsi Madaki and Kwarra health facilities.

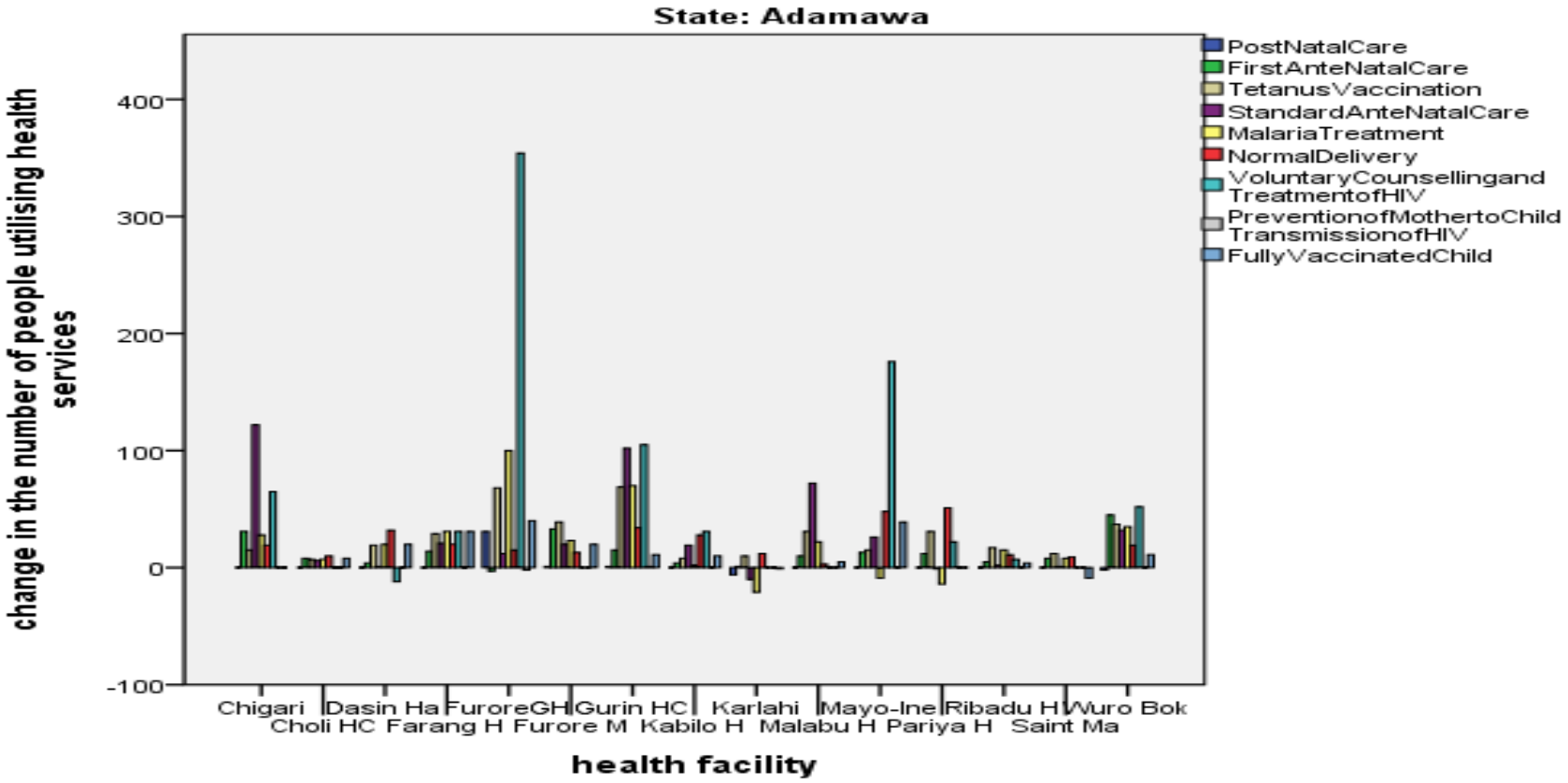


Figure 7.5 Change in utilisation of incentivised maternal and child health services in Adamawa State (Fufore LGA) from December 2011-December 2012

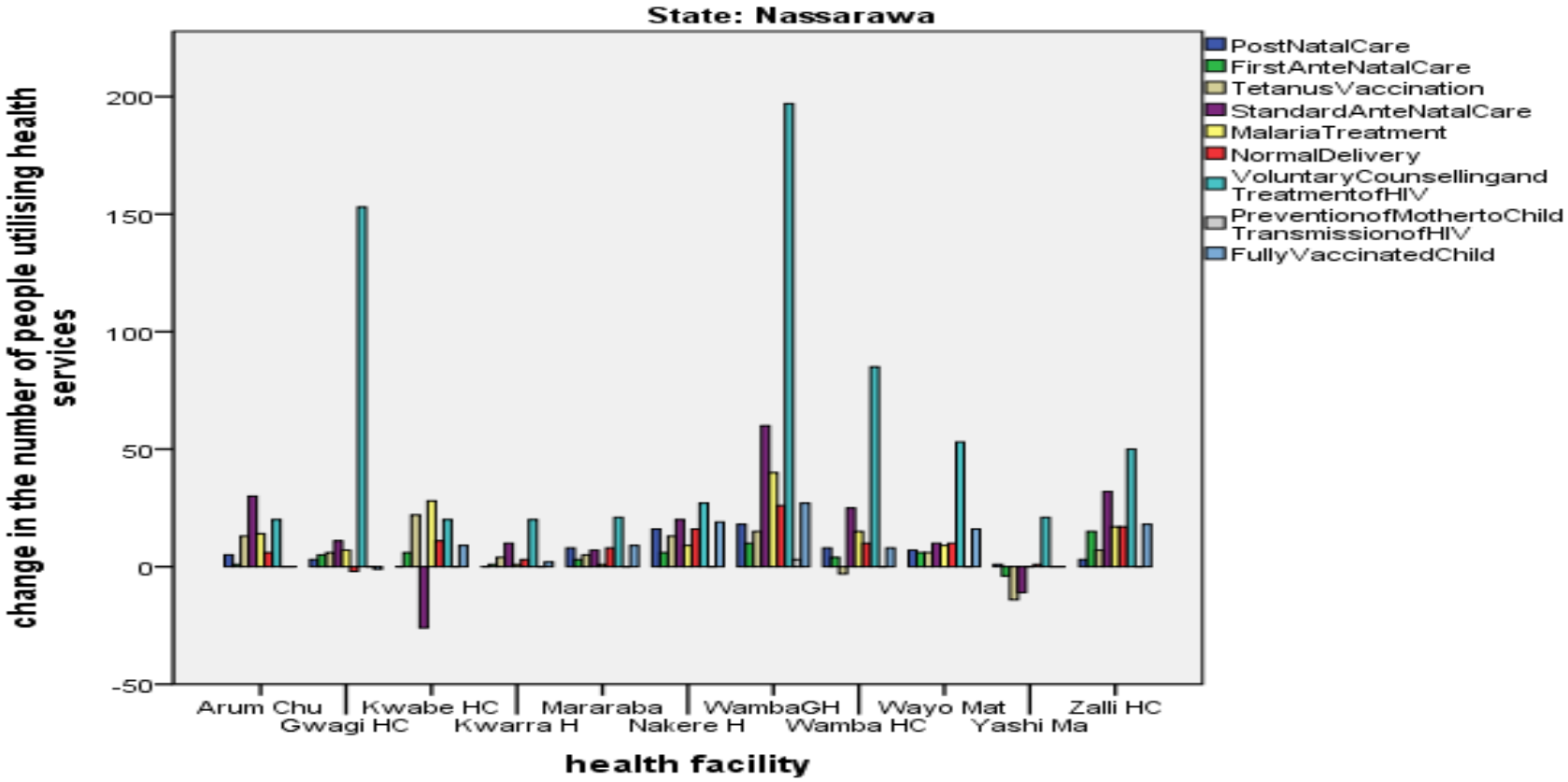


Figure 7.6 Change in utilisation of incentivised maternal and child health services in Nassarawa State (Wamba LGA) from December 2011-December 2012

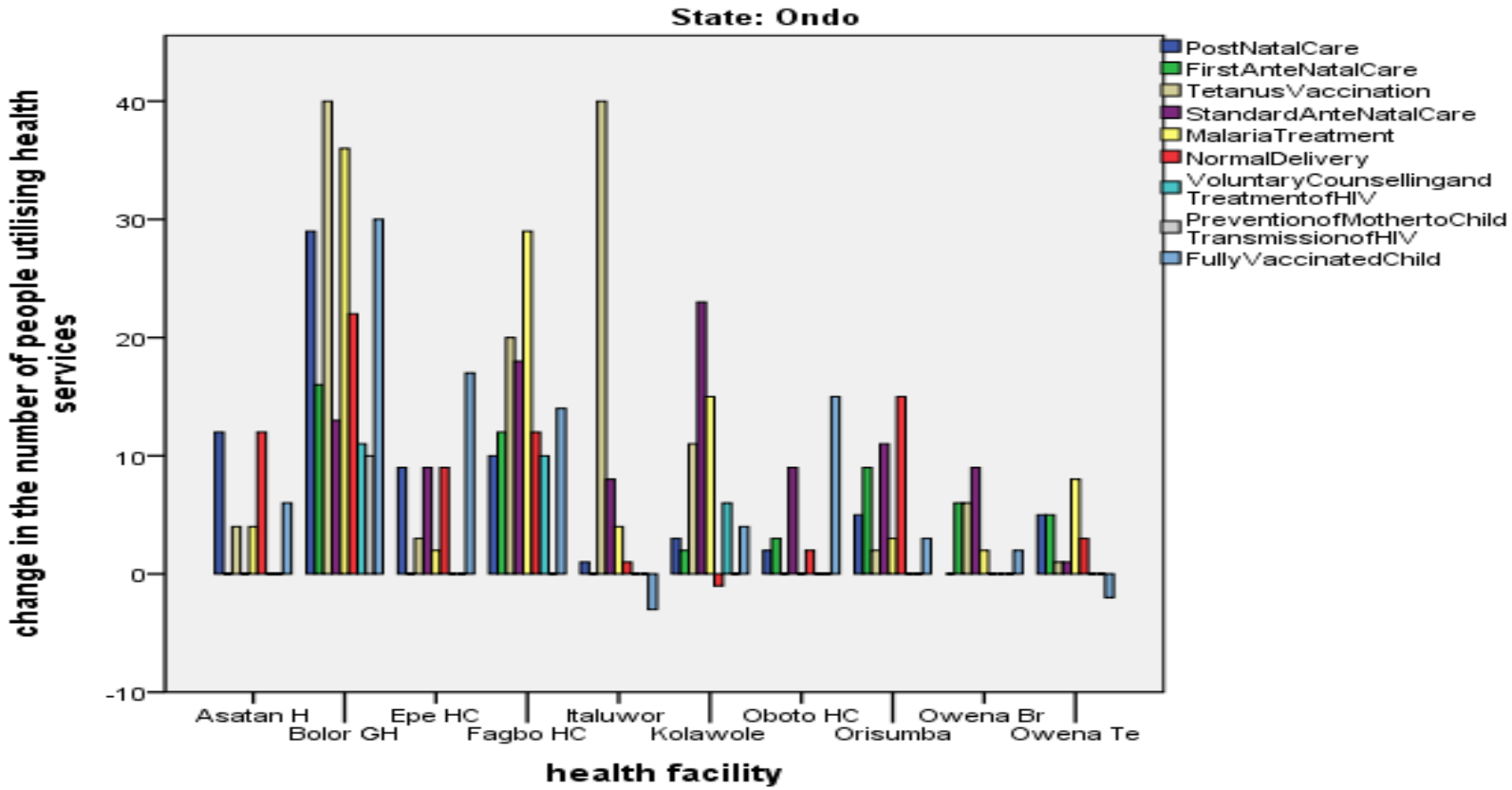


Figure 7.7 Change in utilisation of incentivised maternal and child health services in Ondo State (Ondo east LGA) from December 2011-December 2012

In the same way, the results varied between the three States. Most of the health services in each State showed improvement in coverage/utilisation, ranging from small/modest to significant changes. On the average, Adamawa State showed the best improvement, followed by Nassarawa and then Ondo state as seen in Figure 7.8.

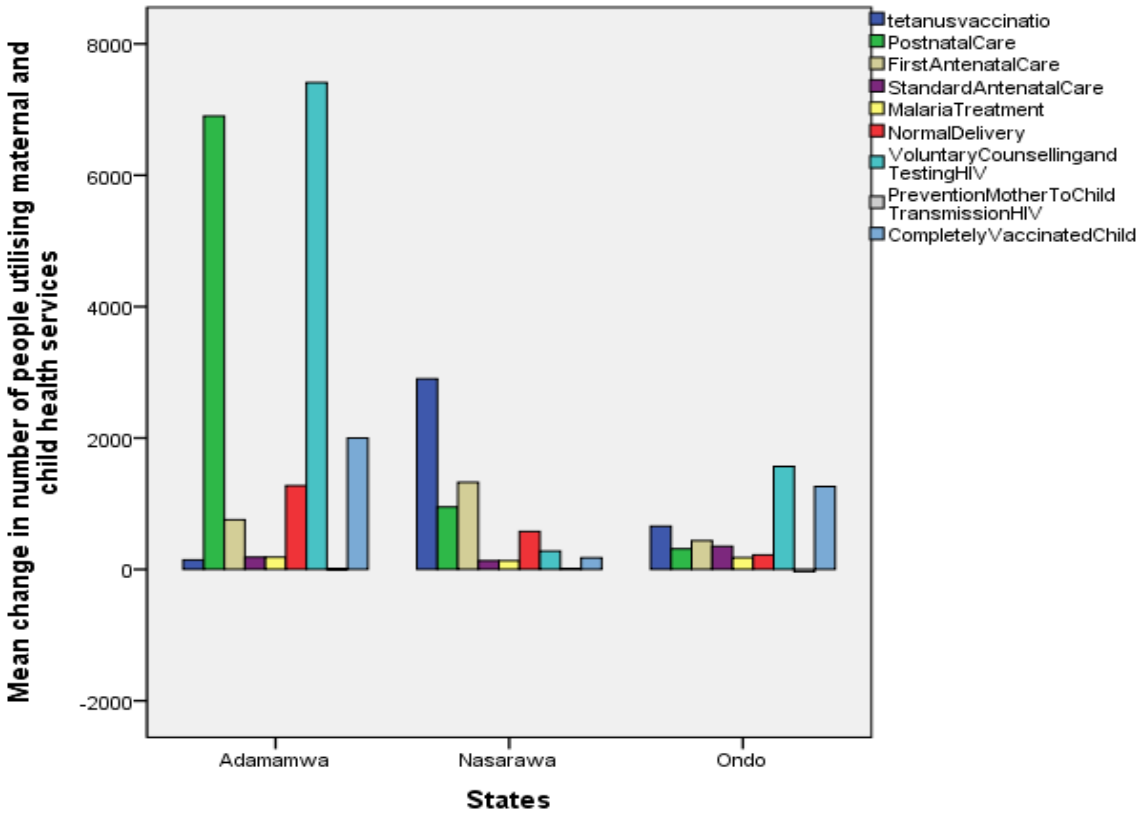


Figure 7.8 Mean change in coverage of maternal and child health services from December 2011 to December 2012 in Adamawa, Nassarawa, and Ondo States

7.6. Discussion of the early results of the Nigerian P4P scheme

The preliminary results show that despite similar design features and method by which performance has been measured across all three States, change in performance varied considerably across States and across the health facilities within the LGAs. This suggests that other sources of variation beyond design features, such as contextual and implementation factors might explain the results.

Previous studies and literature suggests that there are a number of contextual and implementation factors, which could lead to heterogeneous results of incentive scheme. These include uncertainty in earning the incentive, management of the scheme, health worker awareness and understanding of the scheme, communication (flow of

information), infrastructure (equipment/readiness to meet targets), staff skill mix, patient characteristics, and the health care system (Kirschner et al., 2013, Canavan and Swai, 2008, Van Herck et al., 2010, Ssengooba et al., 2012).

In the Nigerian P4P scheme, contextual factors such as staff skill mix, patient characteristics, and characteristics of health care system are unlikely to explain results because, the degree to which they vary in the chosen sites for the implementation of the P4P scheme was minimal. The three sites had similar structure of organisation and funding of the healthcare system (as described in chapter six) i.e. decentralised health system. In the same way, the three sites chosen for the pre-pilots were all rural Local Government Areas (LGA), which have similar patient characteristics, catering to the same primary health care needs, and a similar skill mix in all the health facilities comprised mostly of community health workers (CHEWs) and nurses (NPHCDA, 2012).

Through preliminary visits to the research sites and informal conversations about the P4P scheme with some scheme implementers and health workers (see section 7.3), and exploration of literature, I was able to identify potential contextual and implementation factors that could have influenced the outcome of the Nigerian P4P pre-pilot. This informed the area of enquiry/research questions for the formative evaluation. The potential factors identified were delay in incentive payment, communication between the scheme implementers and the health workers, and scepticism in the measurement of individual performance (all of which are likely to create uncertainty in the health workers regarding earning the incentive). I also identified factors such as management of the scheme (at the health facility level), health worker awareness and understanding of the scheme, and infrastructure as factors with potential to explain or help interpret the results of the Nigerian P4P pre-pilot. I now discuss how these factors might influence the results of the schemes (supported by relevant literature), thus providing a rationale for exploring these factors in the Nigerian context in subsequent chapters.

7.6.1. Risk (uncertainty of earning the incentive)

Arrow (1965) argued that individuals are generally risk averse, meaning that they are less ready to accept an uncertain contract or agreement rather than another contract with a more certain consequence. In the case of incentive schemes, this means that the extent to which an incentive results in behaviour change is likely to be affected by the degree

to which the clinicians are confident they will receive the incentive if they change their behaviour or improve performance.

The element of perceived 'risk" or uncertainty of earning the incentive spans across both design features and contextual and implementation factors. I have analysed and discussed in detail in chapter 3 and 5, the design elements associated with perceived risk of not earning the incentive. I argued that recipients of the incentives have a low perceived risk when incentive schemes are characterised with certain design features, which include:

- Short time lag between measurement of performance and receipt of incentive, as opposed to long time lags (between five months and one year)
- The use of absolute instead of relative performance measures (and so not dependent on how others perform)
- The use of process and structural domains of performance, which are more under the providers control, rather than clinical or health outcomes.

In this discussion, I focus on contextual and implementation factors (rather than P4P design factors) that might contribute to the uncertainty or risk of not earning the incentive.

I identified three potential factors in the Nigerian P4P scheme that were likely to have affected trust in the payment system, thereby increasing the risk of not earning the incentive. They include delay in payment, communication, and individual performance measure used.

Delay in promised payment has been shown to reduce health worker motivation and affect performance in some P4P schemes in implemented in developing countries like Uganda (Ssengooba et al., 2012). In the Nigerian context, there is existing distrust in the payment system brought on by delay in payment of salaries, lack of transparency, and poor governance (Hargreaves, 2002, Garuba et al., 2009, Okafor, 2009, Akinyemi and Atilola, 2013). As a result, some health workers are likely to interpret the delay in payment as deception or not getting paid what was promised, thereby reducing performance results.

In the same way, communication of the changes and the reasons for changes in the scheme (for example, reasons for delayed payment) between the stakeholders might be important in the Nigerian context because the likely distrust brought on by the delay in payment could be exacerbated by misinformation or speculation, which could lead to

poorer outcomes. On the other hand, effective communication has the potential to enhance transparency, which boosts the trust of the health workers in the payment of the incentive because behaviour of health workers are to less likely be affected in cases where a genuine reason for the delay in payment has been communicated effectively by the scheme implementers to the health workers.

Similarly, the assessment upon which the sharing of bonuses to individual health workers (as described in Table 7.3), is based on elements such as clean uniforms, attendance, initiative etc. which are different from the services the health workers have to render in order for the health facility to earn the incentive. It is possible that some health workers might feel that the measure by which the bonuses are being shared to them does not adequately reflect the actual contribution to the health facilities performance. This in turn might lead to uncertainty in the amount of money expected by the health workers. For example, health workers who participate more in outreach and home visits, thereby drawing more patients to utilise the services at the health facility, might feel they deserve more bonuses than other health workers who do not participate in outreaches, but come to work every day, wearing clean uniforms etc. (and doing other things on the individual assessment tool upon which the bonuses are shared). This ambiguity might affect performance results because such health workers might feel like they are not getting what is due them, and therefore, the behaviour change is not worth it.

This aspect of 'risk' is particularly relevant in the Nigerian context because of the existing distrust of the health workers in the payment system, characterized by frequent union strikes (Hargreaves, 2002, Garuba et al., 2009, Okafor, 2009, Akinyemi and Atilola, 2013). Therefore, the health workers are more likely to be uncertain and these additional factors identified in the scheme are likely to further increase their level of uncertainty, which may have negative consequences on the results of the scheme. For instance, if health workers are doubtful about earning the incentive even though they improve performance, they are not likely to change. Rather, they might invest their time and efforts in other things likely to bring them returns on their investments such as doing additional jobs (farming, trading etc.) to supplement their income (Akwataghibe et al., 2013).

7.6.2. The role of health facility managers

Proper management of the scheme at the health facility level has been shown to improve the outcomes of P4P schemes in some LMICs such as Tanzania (Canavan and Swai, 2008). Managers of the scheme at each health facility might influence the impact of the Nigerian P4P scheme because they might handle the scheme differently, in terms of how the manager motivates the health workers and the strategies for improvement implemented (Bredenkamp et al., 2014). This might include different levels of supervision, and monitoring, transparency, communication etc. For example, health facilities where there is constant and appropriate supervision from the health facility manager are likely to produce better results compared to ones with minimal supervision.

7.6.3. Health worker understanding the P4P scheme

Hillman (1998) found that the failure of financial incentives to improve physicians compliance with cancer-screening guidelines in Medicaid health maintenance organization in the USA was associated with low level of awareness and understanding among the participating physicians. Young et al. (2007) also found a link between the impact of P4P and health worker awareness and understanding of the scheme in a number of P4P schemes in the USA. They found that low levels of understanding among clinicians about the way the programme works (the indicators used, payment mechanisms, use of the incentives, individual performance measures, and financial incentive payment specifics) were linked to the negative impacts observed in the P4P programme. This suggests that when health workers understand the P4P scheme, they are more likely to change behaviour to improve performance.

A number of factors might affect the level of understanding of the Nigerian P4P scheme. These include level and effectiveness of training for the scheme, technical support/assistant, qualifications, literacy levels, and personal interest (NPHCDA, 2012). Differences in these factors might affect the way they understand the scheme, which might in turn influence the degree of behaviour change or performance, leading to heterogeneous results. If some health workers lack adequate understanding about the terms and conditions of the scheme, such as the potential incentives to be earned, it is possible the incentives might not reach or have any impact on them. This might reduce the effectiveness of the scheme because the individual health workers might not put as much effort into changing behaviour the way they would if they were fully knowledgeable about the P4P scheme.

7.6.4. Infrastructure (readiness to implement the P4P programme)

Studies that have examined the attitudes and perceptions of healthcare professionals towards P4P schemes found that some healthcare professionals felt that their clinics lacked the infrastructure that was needed to provide the level of quality required, which reduced the impact of P4P in these practices (Locke and Srinivasan, 2008, McDonald and Roland, 2009, Kaczorowski et al., 2011, Canavan and Swai, 2008).

The potential influence of infrastructure on the outcomes of the P4P scheme in Nigeria is very important, because one of the main issues weakening the health system is lack of appropriate infrastructure (Akinwale, 2010). Health facilities in the Nigerian P4P schemes have different infrastructure and equipment because of the influence of the different State governments and international donors. For instance, in Ondo State, the State government provides free drugs for maternal and child health related cases, which is not the case in Adamawa and Nassarawa State. In the same way, the distribution of equipment by donors is not controlled and unbalanced in the health facilities (NPHCDA, 2012).

Specifically, if these infrastructural differences are related to the health services and quality indicators incentivised by the Nigerian P4P scheme, it is likely that it will result in differences in performance results. For example, for an incentivised health service, such as childhood immunisations, health facilities that lack fridges or power supply to maintain the cold chain storage of the vaccines are likely to have low performance results compared to health facilities with appropriate infrastructure.

7.7. What this chapter adds (rationale for exploration of contextual and implementation factors in the Nigerian P4P scheme)

The importance of contextual and implementation factors in P4P schemes have been hypothesized and proposed in this chapter. These factors were considered to be relevant to the Nigerian context and may help explain the varied results of the Nigerian P4P scheme. The potential factors identified were uncertainty in payment of the incentive and attainment of target, administration and management of the scheme, health worker awareness, and infrastructure.

These factors are likely to affect attitude or behaviour of the health workers in the pre-pilot sites in different ways. This in turn may explain or help interpret the preliminary results, thereby providing valuable insight on appropriate contexts and better implementation strategies for better performance results. There is however no systematic evidence of the influence of these factors in the Nigerian context because the scheme is among the first of its kind to be introduced to Nigerian healthcare system. There are a few studies exploring some contextual and implementation factors in incentive schemes in developed countries (Locke and Srinivasan, 2008, McDonald and Roland, 2009, Kaczorowski et al., 2011, Stockwell, 2010), but research in this area is sparse in LMICs, with only two identified studies exploring this area in Tanzania and Uganda (Canavan and Swai, 2008, Ssengooba et al., 2012).

A study to investigate the effect of these contextual and implementation in the Nigerian P4P scheme, will contribute to this area in three main ways.

- The findings will help interpret the variations in preliminary results of the Nigerian P4P pre-pilot.
- Findings from the study will inform implementation and provide recommendations for scaling up of the Nigerian P4P pilot across the whole country.
- The study will enrich the database of evidence of the influence of implementation and contextual factors in incentive schemes especially in low and middle-income settings.

In the subsequent chapters, I empirically explore and investigate in detail, how these factors influence the early results of the Nigerian P4P pre-pilot, providing vital recommendations to the scheme implementers based on the findings.

Chapter 8 Methods of the formative evaluation of the Nigerian P4P scheme

The second part of the thesis focuses on evaluating a P4P scheme in a low and middle-income country (LMIC), using Nigerian as a case study. In chapter six, I described the context of the P4P scheme (Nigerian P4P scheme). In chapter seven, I described the design of the Nigerian P4P scheme and reviewed its preliminary results. The results showed considerable variation between the three States (Adamawa, Nassarawa, and Ondo) and also between the health facilities within States, despite the implementation sites having the same design features. This was followed by an exploration of literature (informed by observation) of other relevant factors beyond design features, such as contextual and implementation factors that may influence the outcomes of P4P scheme and explain the variation in results. The key potentially relevant factors considered were: (1) Uncertainty in earning the incentive in terms of delay in payment, communication, and the assessment tool; (2) Health worker understanding of the P4P scheme; (3) The role of health facility managers in the scheme; and (4) The role of infrastructure in the P4P scheme. These factors are likely to elicit different views and experiences in health workers in health facilities participating in the pre-pilot, which may affect their attitude or behaviour. Thereby providing a rationale to explore the views and experiences of the health workers to capture the influence of contextual and implementation factors on the Nigerian P4P pre-pilot. This represents a formative evaluation, which takes place before or during the implementation of an intervention with the aim of improving the intervention's performance. I.e. it is focused on trying to understand which factors influence the effectiveness of the intervention (Øvretveit, 1998).

In this chapter, I describe the aims and objectives of the research. I also describe and justify the methods used to explore the views and experiences of health workers participating in the Nigerian P4P scheme. This included the study design and setting, details of data collection and analysis, and methods of demonstrating rigour in the research.

8.1. Aims and objectives

This aim of this study was to investigate the influence of contextual and implementation factors with a view to interpret the results of the Nigerian P4P pre-pilot and to inform the development on the scheme.

Objectives

1. To investigate the views and experiences of health workers participating in the Nigerian P4P scheme on:
    - Uncertainty in earning the incentive in terms of: delay in payment, communication, and the assessment tool.
    - Understanding of the P4P scheme.
    - The role of health facility managers in the scheme.
    - The role of infrastructure in the P4P scheme.
2. To examine whether and how these responses vary within and across professional roles, health facilities, and States.

8.2. Study design: a qualitative approach

This was a cross-sectional qualitative study of health workers in the Nigerian P4P pre-pilot scheme. The approach was influenced by a pragmatic paradigm, which guided my methods about gaining the knowledge to answer my research questions, as well as conducting and reporting the research (Creswell, 2009). The pragmatic paradigm was best suited to this research because unlike others such as postpositivisim or constructivism, it is not aligned to a particular philosophical approach. Rather, the pragmatic paradigm allows exploration of research questions by whatever methods are most appropriate or "what works" where there can be singular or multiple realities. In addition, it focuses on exploration of problem oriented research questions in real life situations, which corresponds with the aim of this research (Creswell and Plano Clark, 2011).

I chose a qualitative approach to answer the research questions in this study for two main reasons. First, qualitative approaches can explore areas of human behaviours, beliefs, attitudes, and experiences, which cannot directly be answered or explored by quantitative approaches (Britten, 1995, Creswell, 1998). Unlike quantitative approaches such as surveys, using a qualitative approach takes an in-depth approach to the exploration of theories or topics, conveying an intensity and richness in detail to

understand the topic of interest more thoroughly (Carter and Henderson, 2005). A qualitative approach thus enabled me to gather rich and detailed data concerning the views and experiences of the health workers in the Nigerian P4P scheme on contextual and implementation factors that might affect the outcomes of the scheme.

Second, a qualitative approach has been useful in similar studies to investigate the influence of contextual and implementation factors on P4P schemes. A formative evaluation of a P4P scheme in Tanzania conducted by Canavan and Swai (2008) employed a combination of in-depth interviews, observation, and focus group discussions to provide policy recommendations on design, context and implementation. Ssengooba et al. (2012) also used a combination of qualitative approaches (observation and in-depth interviews) to explore why a large-scale P4P scheme in Uganda was not effective. Finally, Felt-Lisk et al. (2007) employed the use of in-depth interviews to explore and compare the views of physicians on the implementation of P4P schemes in California and England. Thus, a further advantage of using a qualitative approach is to facilitate comparison of the findings with these studies.

8.3. Method of data collection (semi-structured face to face interviews)

Semi structured interviews were used to collect the data. This approach presents three key strengths. First, this method allows researcher to directly intervene in the interview process (Carter and Henderson, 2005, Bowling, 2014b). For example, in situations where the participant does not understand or has difficulty answering a question or provides only a brief response, the interviewer can prompt or probe to encourage the participant to consider the question further and to talk at length or expand on areas, which they feel they are important. Common prompts include “is there any other thing you would like to add” and “can you tell me more about that” (Bowling, 2014b). Second, the flexibility in semi-structured interviews gives room for emergent themes or topics to be captured during the interview process, as the researcher can adjust the interview questions or use prompts to explore emergent themes as the interview proceeds (Carter and Henderson, 2005, Bowling, 2014b). Finally, semi-structured interviews allow consistency. The researcher can ask participants the same broad questions on a particular issue or area (Carter and Henderson, 2005), which provides reliable comparable data to adress the research question.

Other ways of gathering data using a qualitative approach were considered for this study. They include observation, document analysis, focus groups, and unstructured in-depth interviews (see Appendix F1) (Barbour, 2013).

Observation was impractical for this doctoral thesis because of the large amount of time and resources required to collect and analyse data across different health facilities in all three States (Gobo, 2011, Creswell, 1998). In addition, whilst observation can provide richly detailed data, it is difficult to observe 'why' participants do what they do or feel how they feel (reasons for their attitudes or behaviours), especially if it is inherent to the participants (Kawulich, 2005). Similarly, document analysis was not employed in this study because there were no publicly available documents relevant to the research questions. This is a new area of inquiry in the Nigerian context and it was possible to engage the key participants directly, which is regarded as a superior method of collecting data (Bowling, 2014b).

The limitation of focus group discussions is the issue of maintaining confidentiality and anonymity within the group, as participants might be hesitant in expressing their views (Carter and Henderson, 2005, Bowling, 2014b, Barbour, 2013). Hence, it was not a suitable method for exploring sensitive areas in this study, such as uncertainty in payment and role of management, both of which may involve corruption and transparency issues. Furthermore, focus group discussions can be dominated by the voice or one of two participants, and unlikely to achieve the 'in-depthness' obtained from interviews (Finch and Lewis, 2004, Bowling, 2014b).

Finally, unstructured interviews were not appropriate in this study because I had specific and focused areas of enquiry (Bowling, 2014b). Data produced from unstructured interviews, whilst richly detailed is often non-comparable to other participants (Carter and Henderson, 2005, Bowling, 2014b), which was needed to achieve the objectives of this study.

8.4. Ethics approval

Ethics approval was sought and obtained from The Research Governance Committee of the Department of Health Science-University of York (see Appendix F2). The study was governed by the principles of informed voluntary consent, confidentiality, and anonymity (see Appendix F3 and F4 for information sheet and consent forms). Potential

participants (health workers) were provided with information to make sure that they understood the research. They were also assured of confidentiality and anonymity if they were to be interviewed. Finally, participants who expressed interest in being a part of the research (to be interviewed) provided voluntary written consent.

8.5. Data collection

Semi structured face-to-face interviews were undertaken from June to October 2013 (two years after the start of the implementation of the Nigerian P4P pre-scheme). In this section, I describe and justify the approaches used to investigate and explore the views and experiences of the health workers participating in the Nigerian P4P scheme.

8.5.1. Developing the interview questions

According to Erlandson et al. (1993), the key to obtaining rich reliable data from interviews is by asking 'good' questions that reflect the research question(s), which are often informed by preliminary observation of the context and relevant literature. Good interview questions according to Creswell (1998), should be well-informed, non-leading, and unambiguous. The following paragraphs describe the process and features of the interview questions and how they were developed to ensure the credibility of this study.

First, I conducted preliminary visits to the research sites and I had informal conversations about the P4P scheme with some stakeholders (scheme implementers and health workers), which helped me to develop an understanding of the context and areas of interest. This allowed me to focus my research questions, by helping to identify likely relevant issues and topics (guided and supported by literature) in the Nigerian context (see chapter seven), which guided me in developing meaningful semi-structured questions.

The areas/topics I set out to explore in this study were:

- Uncertainty or risk of not earning the incentive (delay in payment, communication, and assessment tool)
- Health worker understanding of the scheme
- The role of the health facility manager in the P4P scheme
- The role of infrastructure

I then developed questions within each topic, asking about the participants' views and experiences. I ensured that these interview questions were focused and asked in a neutral or non-leading manner, which increased credibility of the study by ensuring that my personal opinions about the topics were not expressed (Creswell, 1998). For example, questions regarding the measurement tool were phrased as "what are your thoughts about the measurement tool"? As opposed to "do you think the measurement tool is bad"?

I also made sure the questions were not ambiguous by having just one idea per question and using simple language for easier understanding, which helped the participants understand the questions, thereby producing more reliable and credible data (Britten, 1995).

Finally, because these were questions for semi structured interviews, I had a number of probes and prompts to be used in cases where participants give shallow/brief responses to questions where I particularly wanted rich detail or when following up on a response. Some of the probes I used included:

- Could you give me some examples…
- Why do you feel that….
- You mentioned earlier that… could you please tell me more about that.

Examples of prompts in the interview questions included:

- Q: What has your experience been with the delay in payment?
- Prompt: has that affected the performance of the facility at all?

8.5.2. Piloting the interview questions

Following the development of the interview schedule, I piloted it on five health workers (a subset of the population of interest) in Wamba health facility in Wamba LGA, Nassarawa State to further refine the interview questions, look out for emergent themes (ideas that come from looking at the data) to inform questions, and to improve my interviewing skills. Table 8.1 presents the interview questions piloted. I ended the questions in each area/topic by asking the participants if they would like to add anything else before moving on to the next question.

Table 8.1 Preliminary interview questions for health workers in the Nigerian P4P pre-pilot

Topic	Interview questions
Uncertainty and risk of not earning the incentive	• I've heard that payments have been delayed in the past; can you please share your experiences with the delay in payments of incentives. • Are explanations for the delays in payments communicated to you? • Now, let us move on to talk about how individual performance is measured for payment of bonuses. What are your thoughts about the way the bonuses are shared to the health workers?
Health worker understanding of the scheme	• I would like get an idea of your understanding of the scheme; can you please tell me how this scheme works in this health facility?
The role of the health facility manager in the P4P scheme	• Can you tell me about some of the approaches that have been used in this health facility to improve performance? • How is the incentive earned used in this health facility]? • [For the health facility manager]: How do you decide how to utilise the incentives earned?
The role of infrastructure	• Have you faced any infrastructural challenges in delivering the health services required to earn the incentive? • If yes, what are they?

Conducting the pilot interviews helped me to familiarise myself with the interview schedule, which allowed the interview subsequently to feel more natural in the other sites, helping the participants feel more comfortable and hopefully encourage honest answers (Shenton, 2004). In addition, the pilot interviews were transcribed verbatim and were reviewed by my supervisor (Professor Trevor Sheldon) and Dr Cath Jackson (member of my Thesis Advisory Panel) who has extensive experience with interviews.

Based on lessons learnt from the pilot interviews and expert scrutiny, I made a number of changes to subsequent interviews, which are outlined below.

- I used a simpler information sheet different from the one approved by the ethical committee. It contained all the relevant information but with less technical terms (see Appendix F5 for amended information sheets and evidence of ethics approval for the changes made). The original information sheet was still submitted to the NPHCDA and PHC director at the LGA, but the simpler version were given to the health workers.

- Interviewing style: I became more comfortable with long silences and avoided jumping too quickly from one topic to another.
- I added more prompts and edited the interview questions to follow a coherent structure, with related question grouped together to ensure a smooth flow of the interview process (see Appendix F6 for amended interview questions).
- I reframed several sensitive questions after I noticed participants' discomfort in answering, perhaps because they did not want to be perceived in a negative light. For example, a question regarding the effect of the delay in payment was originally framed as "does the delay in payment affect you?" and the answers from all the participants in the pilot was that it does not affect them. However, after I rephrased the question as "does the delay in payment affect the health workers?" there were more open and detailed responses (see Appendix F6 for amended interview questions).
- Finally, the pilot interview gave me an insight to potential additional emergent themes to explore what motivated the health workers to improve performance. Based on this, in the subsequent interviews, I asked participants to share their views and experiences on factors that motivated or demotivated them (see Appendix F6).

8.6. Setting

The Nigerian P4P pre-pilot was implemented in one Local Government Area (LGA) each of the three States: Fufore LGA in Adamawa State, Wamba LGA in Nassarawa State, and Ondo East LGA in Ondo State. There were 15 health facilities in Fufore LGA, with an average of six health workers in each centre; 11 health facilities in Wamba LGA with an average of five health workers in each centre, and ten health facilities in Ondo East LGA with an average of four health workers in each health facility (see Figure 8.1). The health workers in each health facility in each LGA had a mix of a health facility manager, nurses and community health extension workers (CHEWs). The study population was approximately 36 health facilities and 175 health workers, from the three LGAs of the three States (see Appendix F7 for detailed characteristics of the health facilities).

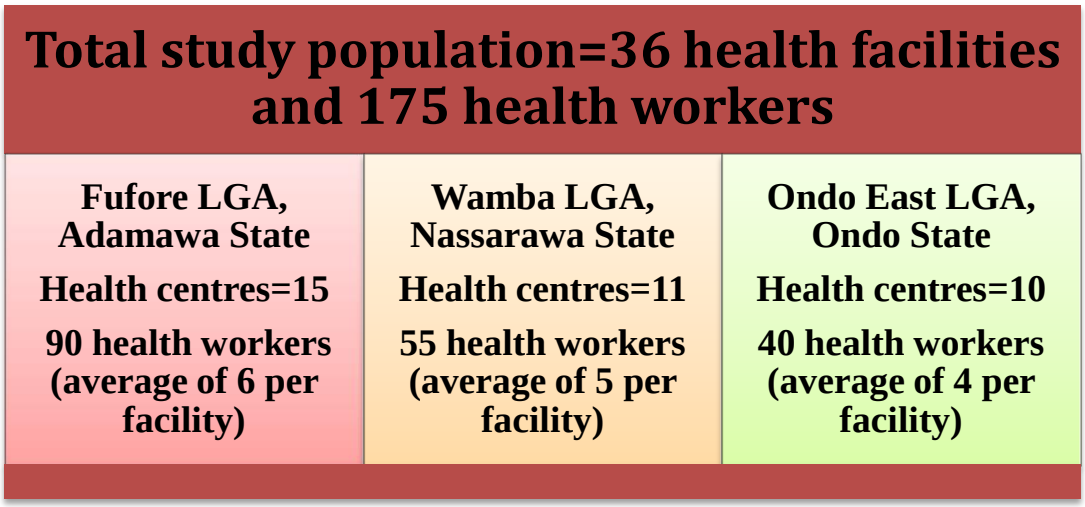


Figure 8.1 Overview of the Nigerian P4P pre-pilot

8.6.1. Sample size

Qualitative research unlike some quantitative methods has no strict rules as to the correct sample because one is not making statistical inferences about estimates (Bowling, 2014b). Researchers have suggested different sample sizes for different qualitative approaches. Morse (1994) suggested: 30 -50 interviews for an ethnographic approach, 30-50 for grounded theory approach and at least a sample size of 5 for a phenomenological approach. Creswell (1998) suggested 20-30 interviews for grounded theory approach and 5-25 for a phenomenology approach. Guest and his colleagues (2006) suggested that 15 is the smallest sample size acceptable for all forms of qualitative research. Alder (2012) suggested graduate students should aim for a sample size of loosely around 30 because it is a medium sized subject pool which permeates beyond a very small sample size without the problem of endless data gathering with limited time.

Others determine sample size by saturation, which according to Ritchie et al. (2003) is the point of diminishing return to the qualitative sample when an increase in sample size does not necessarily lead to more information. However, one usually does not know the number of interviews it will take to reach saturation, and it might require a large sample size that exceeds the researchers' resources (Mason, 2010). Therefore it was an impractical approach for this study.

Having established that there are no clear-cut rules to appropriate sample size, Baker and Edwards (2012) suggest that the decisions about sample size should be made with a

number of considerations: the resources available, the time frame of the study, and whether the sample is large enough to reflect the variation within the target population.

Therefore, based on available resources and time frame of the study, I aimed to interview a sample of 30-45 participants with an approximately equal skill mix (health facility managers, nurses, and CHEWs), reflecting the variation within the target population.

The sample size of 30-45 health workers represented an average of 10-15 participants in each LGA from the three States (Adamawa, Ondo, and Nassarawa) with an approximately equal mix of a maximum of five health facility managers, five nurses, and five CHEWs in each state. However, I could only visit the two participating LGAs in two States (Ondo State and Nassarawa State). I could not visit the third State (Adamawa) due to terrorist attacks and safety issues. The sample size was therefore adjusted to 15-20 health workers per State to retain the overall sample size of 30-45 participants.

8.6.2. Sampling strategy

This study considered one sampling strategy to select the health facilities of interest and the health workers from each of the selected health facilities. A 'maximum variation' type of purposive sampling was used to select the health facilities and health workers within each selected health facility. This method of sampling allows the researcher to select units or cases that maximise diversity to aid exploration of variations (Palys and Fraser, 2008). With respect to health workers, it was considered that health workers across a range of qualifications might have different views and experiences (Palys and Fraser, 2008). For example, nurses might have different views or experiences with delay in payment compared to community health extension workers (CHEWs), which might affect their performance in different ways.

Reports from the NPHCDA suggest that there were at least three health workers (the health facility manager, a nurse, and a CHEW) in each health facility (NPHCDA, 2012). Based on the assumption that all the health workers approached would be willing to participate in the research, I then estimated that I would need to interview health workers from at least five to six health facilities to achieve my target sample size of 15-20 in each State. In order for the five to six selected health facilities to reflect maximum

variation in the performance/results of the P4P scheme, I aimed to select the two best performing, two worst performing, and two average performing health facilities (based on performance data presented in the previous chapter) in Wamba and Ondo East LGAs.

8.6.3. Selecting the health facilities

In order to purposively select the health facilities, I calculated the change in coverage of maternal and child health services (antenatal care, postnatal care, normal deliveries, VCT, tetanus vaccination, completed childhood immunization) from December 2011 to December 2012 for each health facility in each of the local government areas (outlined in chapter seven).

I ranked each of the activities from the lowest to the highest, after which an aggregate of the ranks (numbers) were collected for each health facility. This was used to rank the health facilities in terms of absolute change in coverage of the selected activities (see Table 8.2) for ranks of the facilities for each State based on change in performance).

Table 8.2 Ranks of the health facilities in each State

Rank (according to performance)	**Nassarawa (Wamba LGA)**	**Ondo (Ondo East LAG)**
Top performers	*Wamba GH*	*Bolorunduro GH*
	Zalli	*Fagbo*
	Nakere	Orisumbare
	Wamba	Kolawole
Average performers	Wayo Matti	*Asatan*
	Kwabe	*Epe*
	Arum Chugbu	*Owena Bridge*
	Mararaba Gongon	
Worst performers	*Gwagi*	*Oboto*
	Kwarra	*Owena Tepo*
	Yashi Madaki	*Italuworo*

(*Selected health facilities in italics*)

As seen in Table 8.2, the top two performers in in Nassarawa State were Wamba GH and Zali. However, Zali health facility was inaccessible at the time of data collection

because of bad road networks combined with the bad weather (rainy season). Therefore, I substituted it with Nakare health facility (the next best performer). In the same way, I could only visit one average performing health facility Arum Chugbu in Nassarawa because of the poor road networks. Finally, I chose Kwarra and Gwagi (instead of Yashi Madaki) as the two worst performers because Yashi Madaki was also inaccessible at the time of data collection. In total, I collected data from six health facilities in Nassarawa State: Wamba (pilot), Wamba GH, Nakare, Arum Chugbu, Gwagi, and Kwarra. The substitutions of some health facilities made in Nassarawa as a result of unforeseen consequences were considered to be acceptable, as they were still in the desired performance range, thus achieving the variation desired in the sample.

In Ondo State, the health workers who were interviewed were selected from seven health facilities (see Table 8.2). The purposive (maximum variation) sampling method could not be strictly implemented for two reasons:

- Fewer numbers of health workers per health facility compared to Nassarawa States
- High levels of health worker absenteeism

In order to maintain the maximum variation purposive sampling of the health facilities and to get close enough to the desired sample size (with a roughly equal mix of Health facility managers, nurses, and CHEWs), I interviewed health workers from seven health facilities: top two best performing health facilities (Bolorunduro and Fagbo), all three average performing health facilities (Asatan, Epe, and Owena Bridge), and all three worst performing health facilities (Owena Tepo, Oboto, and Italuworo) (see Table 8.2).

8.6.4. Identification and approaching potential participants

I sought and was granted permission to visit and approach participants from the selected health facilities from the PHC director and the State Primary Health Care Development Agency (SPHCDA) in each LGA.

I was introduced by a representative of PHC director in each of the LGAs to the manager at each health facility (otherwise known as the ‘in charge’) as an independent researcher from the University of York. The managers of each of the health facilities then introduced me to all the health workers after which I introduced my research, read out the contents of the information sheet (see Appendix F3 for information sheet) and

provided a copy for each person. Some of the health facilities had to be visited twice to introduce my research to all the members of staff, as they work in shifts.
Appointment forms (see Appendix F8), and envelopes were distributed and left with the health workers for five days to allow sufficient time to think about if they wanted to participate in the research. Those who wished to be a part of the interview indicated this by picking an appropriate day and time for an interview in their appointment forms (with their names and telephone numbers) in a sealed envelope. Those who did not wish to participate were also instructed to leave the appointment form blank but also handed this back to me in a sealed envelope. After all the health workers had handed in their appointment forms, I telephoned the interested participants to set up and confirm an interview.

8.6.5. Interview sessions

Data collection took place over six months: a month preliminary visit in March 2013 to observe the context, and interviews from June to October 2013.
The interviews were conducted in English language and an interpreter was available. Participants were interviewed one-to-one in a private place that was free from distractions. This was chosen by the participant, usually an unoccupied room at the health facility, based on the premise that this familiarity may help the participants to relax, therefore resulting in productive interview (Shenton, 2004).

The interviews began with me providing a summary of the research using the information sheet and ensuring the participant understood the research before progressing. I then read out the consent form to the participant, emphasising confidentiality, anonymity, and the right of the participant to withdraw from the study at any point, without an explanation. I also emphasised my status as an independent researcher, not affiliated with the P4P scheme implementers (The World Bank or the NPHCDA). Next, both the participant and I signed the consent forms (two copies: one kept by the participant and I kept the other copy). Assuring the participants of confidentiality, anonymity, and my status as an independent researcher is said to increase the likelihood of the participants' honesty and openness about their experiences, views, and feelings (Krefting, 1991).To further try to encourage the participants to be honest, I stated that there were no right or wrong answers to the questions that would be asked.

To establish rapport, the first questions asked were general questions such as participants' role and qualifications, and their general experience at the health facility, before moving on to the potentially more sensitive topics. Also, from time to time, during some of the interviews, I reflected back on what other participants had said (using them as probes) to gain a better understanding of their views and experiences.

The interview sessions were audio recorded, with the permission of the participants. I also listened carefully, took notes and impressions of participants during and after the interviews (which were used in the data analyses in section 8.7). I used the interview questions as a checklist/topic guide. I also used probes and prompts in cases where the participants did not understand the question, needed further clarification, and to follow up from previous responses (to drill deeper). I was sensitive to cues, body languages, and silences, knowing when to wait or prompt. For example, in Ondo State, I sensed hesitation from one of the participants, after which I reminded her of the assurance of confidentiality and anonymity and the right to withdraw from the research at any time. The participant decided to withdraw from the study, which might have been beneficial to this study because a hesitant participant may give dishonest and biased data/contribution, thereby reducing the credibility of the study (Shenton, 2004).

I then ended each interview by thanking the participant for their time and contribution. Each interview lasted an average of 50 minutes. The longest interview lasted about one hour 25 minutes and the shortest interview was about 30 minutes.

8.7. Data Analysis

To analyse the data generated from the semi-structured interview, I used the framework approach developed by Ritchie and Spencer (1994), which is an approach considered appropriate for health policy research, and for this study for a number of reasons. First, the matrix output proves an efficient way to organise, manage, and become familiar with the data, which is practical and feasible in this study to explore the variation in the views and experiences of the health workers interviewed (Gale et al., 2013, Ritchie and Spencer, 1994). Second, the framework approach is adaptable for both pre-set themes (inductive approaches), and emergent themes (deductive approaches) **(**Smith and Firth, 2011**),** which was appropriate for this study in which emergent themes were anticipated. Third, in the framework analysis, the stages by which the results have been obtained

from the data are clear, visible and systematic, which enhances the rigour of the analytical processes of the study through an effective and transparent audit trail (Gale et al., 2013). Finally, framework analysis significantly facilitates comparison of data across the matrix (Gale et al., 2013), which was important in exploring and comparing the variations in views and experiences of the health workers to see if they explained the heterogeneous results of the Nigerian P4P pre-pilot.

Other approaches to qualitative data analyses are often associated with a specific discipline, theoretical or philosophical ideas, which shape the process of analysis (Gale et al., 2013). Some examples are: Discourse analysis, which is associated with different aspects of language use in social interactions; Phenomenology, which involves experience and meaning; and Grounded theory that develops theory through data analysis (Strauss, 1987). The Framework approach, however, is not associated with a particular philosophical or theoretical approach. Rather it is flexible tool that can be adapted for use in qualitative approaches with pre-set or emergent themes (Gale et al., 2013).

The framework approach is part of the family of data analysis approaches known as thematic analysis. This seeks to identify similarities and differences in qualitative data and explore relationships between themes in order to draw explanatory conclusions clustered around the theme (Ritchie and Lewis, 2003, Gale et al., 2013). The defining feature of framework analysis is the matrix output consisting of rows (cases: interviewee or groups of interviewees), columns (themes) and 'cells' of summarized data, which allows a comprehensive and robust analyses of the themes across the dataset (theme based approach), while maintaining the connection of the participants views to other aspects of the account (case based approach) (Ritchie and Spencer, 1994). The framework approach consists of five stages: Familiarization with data, Identification of thematic framework, Indexing/coding, Charting, and Mapping and interpretation (Ritchie and Spencer, 1994). The data were managed using Microsoft Word. I now describe the stages of analyses.

8.7.1. Stage 1: Familiarization with the data

Familiarization with the data involved transcription, immersion in the dataset, and preliminary interpretation of text to facilitate coding. In this stage, I transcribed

(verbatim) all 36 interviews which helped me familiarise myself with the data. I was able to remember almost each participant and revisit notes taken and impressions formed during the fieldwork as I transcribed each interview.

After initial transcription, I read each transcript several times, further making notes and writing down impressions of possible emergent themes and/categories (see Figure 8.2 for a sample of transcribed interview). Each transcript was then labelled and stored in a password-protected folder.

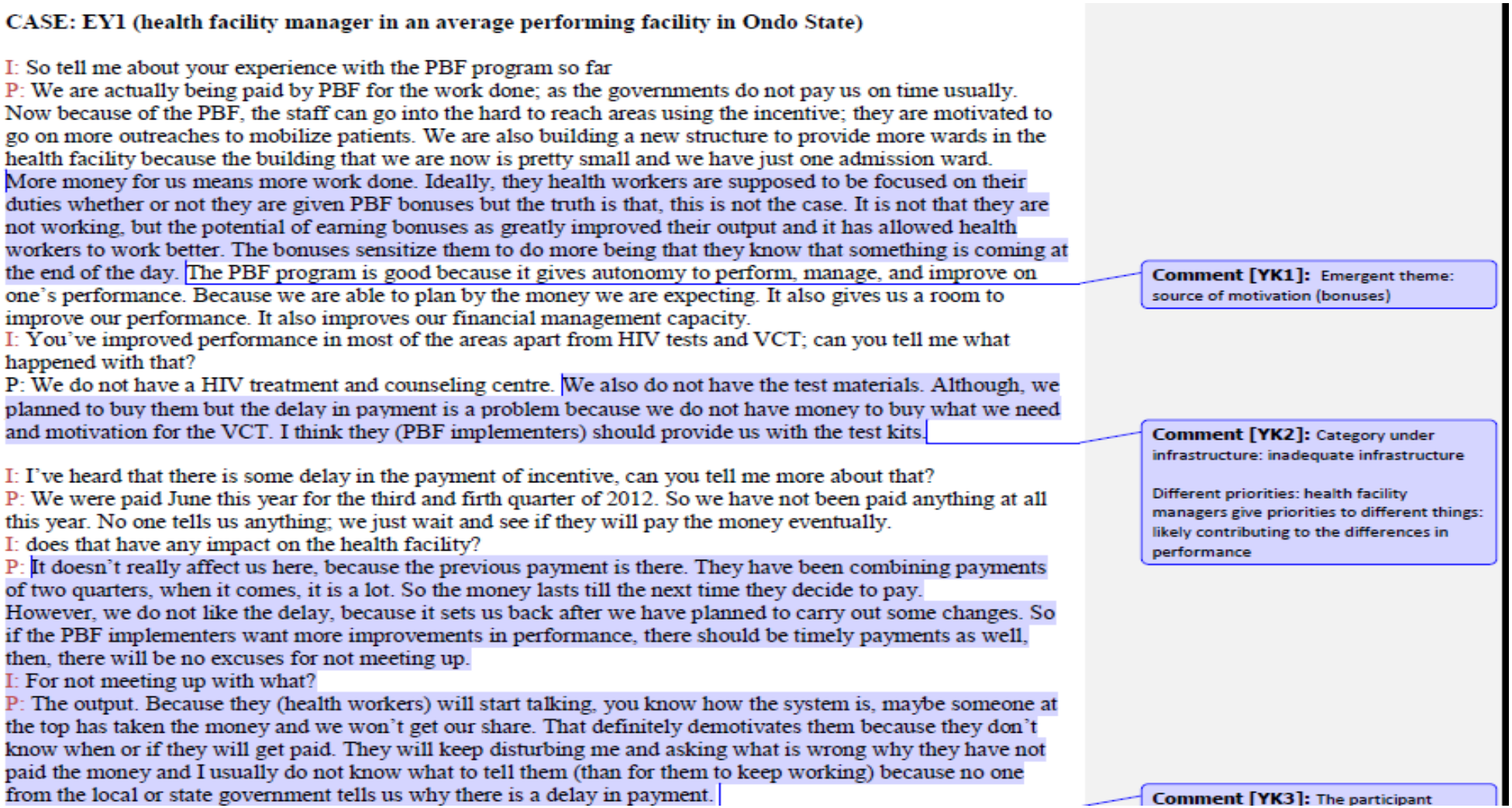
8.7.2. Stage 2: Identification of the thematic framework

The second step taken in the data analysis was to develop the thematic framework, which involved identifying and refining initial and emergent themes. This was carried out simultaneously with the third stage of analysis (indexing/coding), but for the sake of clarity, I first outline how I identified the thematic framework in this section, before explaining the process of indexing in the next section.

I started out with initial themes selected based on previous literature and evidence (from chapter seven), a deductive approach (Gale et al., 2013).
The initial themes were:

- Uncertainty of earning the incentive in terms of: delay in payment, incomplete payment, communication, and the assessment tool.
- Health worker understanding of the P4P scheme.
- Management and administration of the P4P scheme (the role of health facility managers)
- The role of infrastructure in the P4P scheme

After familiarising myself with the data, I looked to see if other themes (dominant and frequent pattern that was relevant to the research question) emerged from the data (an inductive approach to analysis) (Burnard et al., 2008). Using this combined approach of deductive and inductive analyses ensured that all the relevant experiences or views of the participants on the area of interest were analysed, producing findings that reflect the participants' views and experiences and not mine (the researcher) (Mays and Pope, 2000).



CASE: EY1 (health facility manager in an average performing facility in Ondo State)

I: So tell me about your experience with the PBF program so far
P: We are actually being paid by PBF for the work done; as the governments do not pay us on time usually. Now because of the PBF, the staff can go into the hard to reach areas using the incentive; they are motivated to go on more outreaches to mobilize patients. We are also building a new structure to provide more wards in the health facility because the building that we are now is pretty small and we have just one admission ward. More money for us means more work done. Ideally, they health workers are supposed to be focused on their duties whether or not they are given PBF bonuses but the truth is that, this is not the case. It is not that they are not working, but the potential of earning bonuses as greatly improved their output and it has allowed health workers to work better. The bonuses sensitize them to do more being that they know that something is coming at the end of the day. The PBF program is good because it gives autonomy to perform, manage, and improve on one's performance. Because we are able to plan by the money we are expecting. It also gives us a room to improve our performance. It also improves our financial management capacity.
I: You've improved performance in most of the areas apart from HIV tests and VCT; can you tell me what happened with that?
P: We do not have a HIV treatment and counseling centre. We also do not have the test materials. Although, we planned to buy them but the delay in payment is a problem because we do not have money to buy what we need and motivation for the VCT. I think they (PBF implementers) should provide us with the test kits.

I: I've heard that there is some delay in the payment of incentive, can you tell me more about that?
P: We were paid June this year for the third and firth quarter of 2012. So we have not been paid anything at all this year. No one tells us anything; we just wait and see if they will pay the money eventually.
I: does that have any impact on the health facility?
P: It doesn't really affect us here, because the previous payment is there. They have been combining payments of two quarters, when it comes, it is a lot. So the money lasts till the next time they decide to pay. However, we do not like the delay, because it sets us back after we have planned to carry out some changes. So if the PBF implementers want more improvements in performance, there should be timely payments as well, then, there will be no excuses for not meeting up.
I: For not meeting up with what?
P: The output. Because they (health workers) will start talking, you know how the system is, maybe someone at the top has taken the money and we won't get our share. That definitely demotivates them because they don't know when or if they will get paid. They will keep disturbing me and asking what is wrong why they have not paid the money and I usually do not know what to tell them (than for them to keep working) because no one from the local or state government tells us why there is a delay in payment.

Comment [YK1]: Emergent theme: source of motivation (bonuses)

Comment [YK2]: Category under infrastructure: inadequate infrastructure

Different priorities: health facility managers give priorities to different things: likely contributing to the differences in performance

Comment [YK3]: The participant

Figure 8.2 Sample of transcribed interview with initial impressions of categories

Following the familiarisation with the data, an emergent theme centred on 'motivation', in which participants talked about other contextual and implementation factors (that influenced their performance in the scheme), some of which were not inherent to the P4P scheme. These factors included infrastructure, bad roads/mobility issues, and lack of manpower. Based on this, to avoid repetition in the findings, I incorporated the original initial theme 'the role of infrastructure in the P4P scheme' under this emergent theme, as it was one of the categories discussed by the participants under motivation. Therefore, the final thematic framework consisted of four themes (three initial themes and one emergent theme).

8.7.3. Stage 3: Indexing (coding)

Indexing/coding involved applying the thematic frame work to the data using labels or codes that correspond to different themes. In this stage, I reread the transcripts several times to develop textual codes or categories, which summarised the participants' views within each theme while retaining links to original data (see Table 8.3). I also highlighted quotes in transcripts that illustrated the view or experience being described.

These categories identified were refined several times to accommodate all the relevant data provided by the participants. The categories and themes were then used to form a 'coding index' (see Table 8.4), which was applied to the whole dataset as a means of systematically organising the data in preparation for the next stage of analysis (charting).

Table 8.3 An example of development of categories for under Theme: motivation

Cases	Quotes from interview transcript	Category
EY1	*"We are actually being paid by P4P for the work done; as the governments do not pay us on time usually. Now because of the P4P, the staff can go into the hard to reach areas using the incentive; they are motivated to go on more outreaches to mobilize patients. We are also building a new structure to provide more wards in the health facility because the building that we are now is pretty small and we have just one admission ward.* *More money for us means more work done. Ideally, they health workers are supposed to be focused on their duties whether or not they are given P4P bonuses but the truth is that, this is not the case. It is not that they are not working, but the potential of earning bonuses as greatly improved their output and it has allowed health workers to work better. The bonuses sensitize them to do more being that they know that something is coming at the end of the day".*	Bonuses (money)
EX1	*"Well, we let them realize that the bonuses are not our salary. We do this work for altruistic purposes. If the government is not paying us, they will still pay you later on. The bonuses are minor. How much is the bonus compared to the salaries? If the bonuses stop, I think I will still continue because a foundation has been laid. I have realized that the more you see patients, the more you understand, and the more you have knowledge. Without seeing patients and practicing, you cannot progress.* *One health worker even told me that she has improved on her skills because we now attend to more patients and we can put our knowledge to work. We are very happy about that. That makes me happy even more than the money. I feel more exposed to many new cases. I now feel like I my doing my job. Even if the bonus goes away, the way we work now will be sustained. P4P can go, it is a programme. We have had many programmes before but I know that this one will have a longer lasting impact".*	Knowledge
NX1	*"The environment is better for we the health workers to stay at our duty posts now. We are now enjoying the place; out of the money we got, we bought essentials like generator, fridge, some equipment, and TV. There is nothing that other health facilities are enjoying in the city that we don't have. So we enjoy the place better now".*	Equipment and Structural improvement

Table 8.4 Final coding index

Themes	Subthemes	Categories/codes
Uncertainty of earning the incentive	Delay in payment	• Delay in payment reduces motivation and or performance • Delay in payment reduces performance but not motivation
	Incomplete payment	• Negative effect on performance • No effect on health facility performance
	Individual assessment tool	• Assessment tool is fair and should not be changed • Assessment tool is fair but should be improved to further reflect individual contribution • Assessment tool is biased and should be improved to further reflect individual contribution
	Communication	• Reasons for changes communicated through 'hearsay' • Reasons for changes not communicated effectively
Health worker understanding of the P4P scheme		• Good working knowledge of the programme • Aware of changes in the programme • Average working knowledge of the programme • Unaware about changes in the programme
Management and administration of the P4P scheme (the role of health facility managers)		• Hiring more staff • Gifts to patients • Outreaches and home visits • Equipment and structural improvement • Improved supervision • Health workers nicer and more welcoming to patients • Free/subsidized services and or drugs • Effective use of feedback • Feedback given but not implemented
Motivation	Motivating factors improving performance	• Bonuses (money) • Knowledge (education and experience) • Infrastructural improvement • Positive thoughts towards peer reporting
	Demotivating factors decreasing performance	• Structural challenges (insufficient infrastructure to meet targets) • Competition from 'quacks' /other health facilities • Mobility (bad terrain/roads) • Man power • Negative thoughts towards peer reporting

8.7.4. Stage 4: Charting

Charting involved grouping themes and subthemes, elaboration of themes, and comparisons of themes across the participants. After coding the data, I entered the summarised data into a framework matrix in order to easily look across the dataset to identify patterns and connections within and between the themes. The framework matrix combined both the theme based (looking down) and case based approach (looking across) for the whole dataset. Table 8.5 shows a sample of the framework matrix. Each row represents a participant (case), columns represent themes, and cells contain the summarised data (code) for each case within the corresponding theme or subtheme.

8.7.5. Stage 5: Mapping and interpretation

Mapping and interpretation involved searching for patterns and associations within the data, and linking the interpretation of the themes with literature to construct an explanation or meaning. In this final stage of analysis, I created descriptive and explanatory accounts of the data. The explanatory accounts began with reflection on the original data and on the previous analytical stages. This was to ensure the views and experiences of the health workers were accurately reflected, and to minimise misinterpretation of the data (Gale et al., 2013). I then identified patterns in the data through general comparisons of the individual participants and participants' clusters: looking for related themes and searching for explanation/causality (with the help of existing knowledge). Specifically, I manipulated the framework matrix (produced from the subsequent section, Table 8.2) for each participants cluster to facilitate comparisons. The participants' clusters explored were professional qualification (health facility managers, nurses, CHEWSs and JCHEWs), performance (top performers, average performers, and low performers), and States (Ondo and Nassarawa).

In the final stages of the analysis, I used visual charts and maps to help illustrate and make sense of relationship between themes and categories (see Figure 8.3 and 8.4 for sample charts). The findings were then compared and contrasted with established literature and the theoretical perspectives relating to incentive schemes in health care to create a comprehensive explanatory account.

Table 8.5 A sample of the framework matrix

Case	Uncertainty of earning the incentive (Theme)	Delay in payment Incomplete payment Individual assessment tool Communication	Health worker understanding of the P4P scheme	Management and administration of the P4P scheme(role of the health facility manager)	Motivation	Motivating factors improving performance Demotivating factors decreasing performance
NX1a	Delay in payment reduces motivation and or performance		• Good working knowledge of the programme • Aware of changes in the programme	• Free/subsidized services and or drugs • Changes in professionalism and manners • Equipment and structural improvement • Health workers nicer and more welcoming to patients • Effective use of feedback		• Bonuses • Positive thoughts towards peer reporting
	Negative effect on health facility performance					• Lack of man power • Competition
	Assessment tool is biased and should be improved to further reflect individual contribution					
	Reasons for changes not communicated effectively					
NX1b	Delay in payment reduces motivation and or performance		• Good working knowledge of the programme • Aware of changes in the programme	• Equipment and structural improvement • Free/subsidized services and or drugs • Hiring more staff • Gifts to patients • Health workers nicer and more welcoming to patients		• Bonuses • Knowledge • Positive thoughts towards peer reporting
	Negative effect on health facility performance					• Lack of man power • Infrastructural challenges
	Assessment tool is biased and should be improved to further reflect individual contribution					
	Reasons for changes not communicated effectively					
NX1c	Delay in payment reduces motivation and or		• Good working	• Free/subsidized services and or drugs		• Knowledge

Case	Uncertainty of earning the incentive (Theme)	Delay in payment Incomplete payment Individual assessment tool Communication	Health worker understanding of the P4P scheme	Management and administration of the P4P scheme(role of the health facility manager)	Motivation	Motivating factors improving performance Demotivating factors decreasing performance
	performance Negative effect on health facility performance Assessment tool is biased and should be improved to further reflect individual contribution Reasons for changes not communicated effectively		• knowledge of the programme • Aware of changes in the programme	• Hiring more staff • Gifts to patients • Outreaches and home visits • Equipment and structural improvement • Health workers nicer and more welcoming to patients • Effective use of feedback		• Lack of man power
NX2a	Delay in payment reduces performance but not motivation No effect on health facility performance Assessment tool is fine but should be improved to further reflect individual contribution Reasons for changes not communicated effectively		• Average working knowledge of the programme • Aware of changes in the programme	• Equipment and structural changes • Health workers nicer and more welcoming to patients • Effective use of feedback		• Positive thoughts towards peer reporting
NX2b	Delay in payment reduces motivation and or performance Negative effect on health facility performance Assessment tool is fine but should be improved to further reflect individual contribution		• Average working knowledge of the programme • Aware of changes in the programme	• Health workers nicer and more welcoming to patients • Effective use of feedback		• Bonuses • Lack of man power

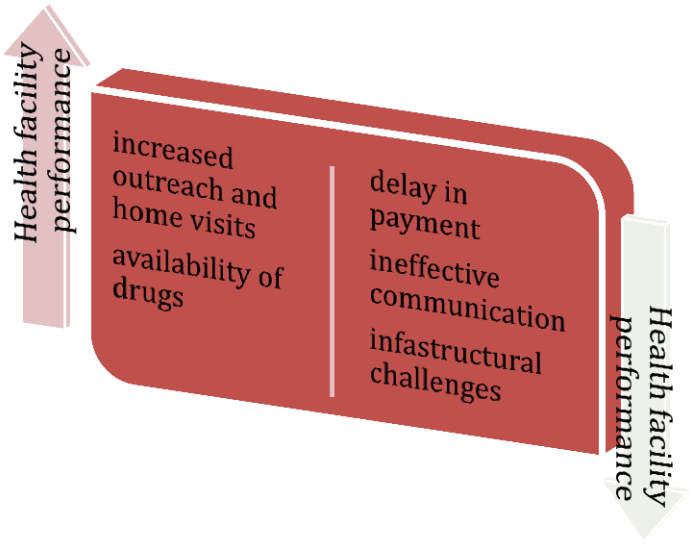


Figure 8.3 Sample chart of the relationship between general performance and contextual and implementation

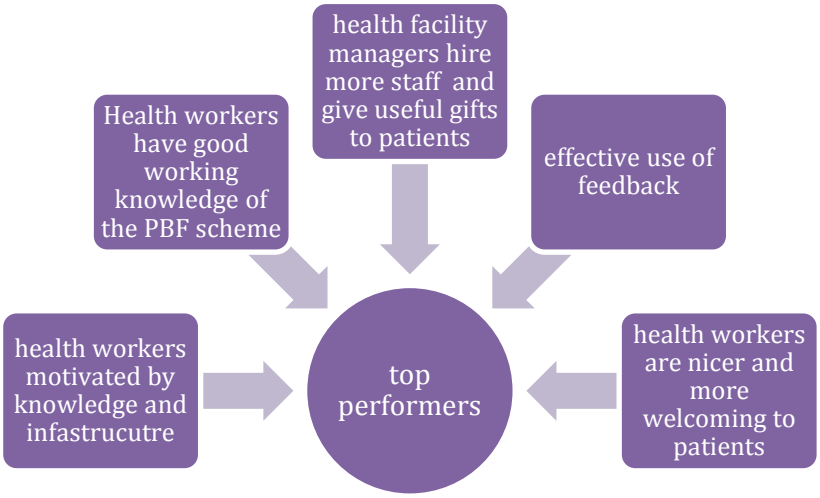


Figure 8.4 Sample chart illustrating the contextual and implementation factors linked to top performing health facilities

8.8. Trustworthiness of the research

The final section of this chapter describes the strategies I employed to ensure trustworthiness of the research. Trustworthiness refers to the methods employed in the research process to improve quality and rigour of the research, which is comparable to reliability and validity in quantitative studies (Shenton, 2004, Murphy et al., 1998).

There is considerable debate associated with terminologies used in assessing the quality/rigour of qualitative research but the techniques suggested by researchers are similar. They include clarity of research questions, suitability of the qualitative enquiry

to the research questions, consideration of other plausible methods, transparency, reflexivity, detailed description and justification of methods and analysis, relevance (potential generalizability beyond groups or settings studies) (Murphy et al., 1998, Mays and Pope, 2000, Dixon-Woods et al., 2004, Lincon and Guba, 1985).

In the subsequent subsections, I describe the techniques I employed to ensure the trustworthiness of this research under the four criteria proposed by Lincon and Guba (1985), the most common criteria for judging trustworthiness in health services research. They are credibility, dependability, conformability, and transferability.

8.8.1. Credibility

A credible study is one in which the findings are as close to reality as possible (Shenton, 2004, Guba, 1981). Credibility is threatened by participants responding based on social desirability rather than social experience (Krefting, 1991). This can be improved by the researchers being familiar with context (prolonged engagement) and strategies to promote openness in participants (Seale and Silverman, 1997, Shenton, 2004). I also sought to achieve credibility in a number of other ways, namely adoption of established research methods, triangulation, negative/deviant case analysis (elements of the data that appear to contradict patterns or explanations that are emerging from data analysis), peer/expert scrutiny of the research, and consistency of findings with previous work (Shenton, 2004, Porter, 2007), as described in the following sections.

Familiarisation with context and promoting open responses in participants

I familiarized myself with the contexts where the data were to be collected, through preliminary observations and informal chats with potential participants and P4P scheme implementers (captured in chapter seven). This was useful in informing the research questions, shaping the interview questions, and gaining the trust of the potential participants, which can encourage open and honest, producing data close to reality of the respondent (Krefting, 1991, Shenton, 2004). Other tactics used to encourage open and honest responses (see section 8.6.1 and 8.6.5) were as follows. I sought to establish rapport during interview sessions, I interviewed the participants in places where they felt comfortable, I emphasised anonymity and confidentiality, and used iterative questioning in which I reframed questions in certain ways to help illicit more personal responses.

Established research method

The adoption of well-established and justified research approach from literature and comparable studies in section 8.2, demonstrates credibility by reflecting the use of appropriate methods for answering the research question (Shenton, 2004, Porter, 2007).

Triangulation

Triangulation involves analyzing the research question from multiple perspectives, which enhances credibility if the findings from the different perspective arrive at the same conclusion (Rolfe, 2006). There are different types of triangulation methods, which include data, environmental, investigator, theory, and methodological (Shenton, 2004). The methods of triangulation used in this study were data and methodological triangulation.

Data triangulation uses one method (e.g. in-depth interviews) to collect information from different sources or groups of people and/or groups of people in different sites/places (Flick, 2014). In this study, I collected data from different groups of health workers in different health centers (see section 8.6.2) to compare multiple perspectives. Although Mays and Pope (2000) argue that data triangulation may not necessarily be a way to 'validate' findings or conclusion because the different groups might have opposing views and experiences. Either way, data triangulation ensures comprehensiveness in the data collection process (Mays and Pope, 2000).

Methodological triangulation involves the use of other qualitative and/or quantitative methods to explore the same research question for consistency in the findings (Shenton, 2004). This qualitative research was conducted to help make sense of the variation in the preliminary results of the Nigerian P4P pre-pilot. The qualitative findings (in chapter nine) explain and coincide with the preliminary evaluation results of the Nigerian P4P pre-pilot, which increases confidence in the findings.

Environmental triangulation was not applicable because environment was not a potential influencing factor on data collection (Thurmond, 2001). Similarly, investigator and theory triangulation (which involves the use of several different investigators in the analysis process and multiple perspectives, usually professionals outside the field of study to interpret a single set of data) were not practical in this study because this study

is part of a doctoral thesis and there are large amounts of data/transcripts involved, which would have been time consuming (Flick, 2014).

Deviant/negative cases

I thoroughly examined the data generated from the interviews during analysis, which included identification and discussing cases that contradicted the dominant findings (deviant cases) in the findings of this study (chapter nine). Explanation of deviant cases enhances credibility by reflecting the thoroughness of the analysis, indicating the findings presented and discussed reflects those that are relevant to the research question, and not based on assumptions or preferences of the researcher (Krefting, 1991, Shenton, 2004).

Expert scrutiny of the study

The rigour of this research was monitored and thoroughly scrutinised by my PhD supervisor and my Thesis Advisory Panel, consisting of my supervisor and two experienced health service researchers (including one with expert experience in qualitative research). Credibility was enhanced through their feedback and suggestions, which brought a fresh perspective to the study, pointing out issues in methodology and analysis, as discussed in section 8.5.2.

Consistency of findings with previous research

Krefting (1991) proposed that relating the findings of the study to previous research is also an important factor in enhancing credibility because the confidence in the findings is increased by the degree or level of consistency of the findings of the research with an established or existing body of knowledge. This might not be possible in situations where the research is a completely new area of enquiry. In addition, consistency with previous work may not necessarily reflect the credibility of the findings but reflect similar biases in the studies. Therefore, in this study, it was important to demonstrate the rigour of the research to increase confidence in the findings, before comparing it to previous research in the next chapter.

8.8.2. Confirmability

In a conventional sense, confirmability refers to the degree to which the findings of the research could be confirmed or corroborated by others (Shenton, 2004, Lincon and

Guba, 1985). Researchers have, however, argued that the concept of confirmability is almost impossible to achieve in qualitative research, because the researcher brings a unique perspective to the research, and is part of the research process, through their assumptions, background, qualifications etc. (Mantzoukas, 2005, Ortlipp, 2008). Miles and Huberman (1994) suggest that a key technique for confirmability is the extent to which the researcher states his or her own predispositions or assumptions and considers how this affects the research process (a process known as reflexivity). A reflexive study should include steps taken to ensure the findings of the research are the result of the experiences and ideas of the informants, rather than the characteristics and preferences of the researcher (Ortlipp, 2008, Meyrick, 2006, Mantzoukas, 2005).

To establish confirmability, this study was conducted and reported reflexively. First, I acknowledged reasons and rationales underpinning decisions made and methods adopted (the reasons for favouring one approach over others) and I explained the strengths and weaknesses in the methods actually employed throughout the study. Second, I considered the strengths and limitations associated with my background, qualifications, and experience in the discussion of the findings in chapter nine.

8.8.3. Dependability

Dependability is concerned with the extent of the repeatability of the process of the research (Miles and Huberman, 1994). A dependable study will, therefore, clearly document every step of the research in detail, describing the changes that occurred in settings and their potential effect on the research, thereby enabling a future researcher to repeat the work (Shenton, 2004). In addition an in-depth transparent description of the research allows the reader to assess the extent to which robust research practices have been followed and to develop a thorough understanding of the methods and their effectiveness (Meyrick, 2006, Porter, 2007).

In this study, I attempted to achieve dependability by:

- Stating the aims and objectives of the study clearly in section 8.1.
- Describing in detail the research design and its implementation, describing what was planned and executed (section 8.6).
- Providing the operational detail of data gathering and what was done in the field including the changes that occurred (section 8.6.3 and 8.6.4).

- Presenting the study as a reflexive one (acknowledging how my background and experiences might have influenced the research, in chapter nine)

8.8.4. Transferability

Transferability in qualitative research refers to the degree to which the findings of the research can be applied or generalised to other contexts or settings with similar populations, parameters, and characteristics (Miles and Huberman, 1994). It has however been argued that the researcher conducting the study knows only the context of their research (not the one the reader wants to apply it to) and he or she cannot make transferability inferences, but that other researchers are able to relate or transfer the findings of the research to their contexts if it is similar to that described in the study (Firestone, 1993, Lincon and Guba, 1985). Therefore, transferability can be enhanced through a detailed description of the research context and assumptions central to the research, which helps the readers to determine how far they can be confident in transferring the findings to their own context (Shenton, 2004).

Transferability in this study is enhanced by detailed description of the context of this study in the two previous chapters. In addition, the methods used are written up clearly, detailing the context in which the data were collected, data collection methods, length of data collection, time period over which data were collected, the number and description of sites and participants as seen in section 8.6.

8.9. Summary

In summary, in this chapter, I have provided details of the methods I used to explore health workers' views and experiences on potentially relevant contextual and implementation factors in the Nigerian P4P scheme. I also focused on the strategies employed to increase the trustworthiness of the study, namely reflexivity, detailed description of methods and context, and deviant case analysis. In the next chapter, I report and discuss the findings, providing participants' quotes from multiple perspectives and providing policy and research recommendations (based on the findings) to The World Bank and The NPHDCA on the Nigerian P4P scheme.

Chapter 9 Views and experiences of health workers in the Nigerian P4P scheme

In this chapter, I present the views and experiences of 36 health workers from 14 health facilities who took part. I start by providing an overview of the participants. I then provide a comprehensive report of the findings, supported by participant quotes. Finally, I discuss the findings of the study, incorporating the strengths and weakness of the study and drawing out recommendations to improve the Nigerian P4P scheme and research in the field.

9.1. Overview of Participants

All 37 participants approached gave informed consent to be interviewed for the research. However, one participant in Ondo State dropped out 15 minutes into the interview. The data from this participant were not included in analysis, leaving a sample size of 36. The target sample size of 30-45 participants was therefore achieved.

Table 9.1 presents the characteristics of the participants. In total, I interviewed 36 participants from 14 health facilities (six in Nassarawa and eight in Ondo), comprising of 13 health facility managers (six in Nassarawa and seven in Ondo), eight nurses (six in Nassarawa and two in Ondo), ten CHEWs and lab technicians (six in Nassarawa and four in Ondo), and five Junior CHEWs (three in Nassarawa and two in Ondo).

Table 9.1 Overview of participants

Health worker qualification	Top performers		Average performers		Worst performers		Total
	Ondo	Nassarawa	Ondo	Nassarawa	Ondo	Nassarawa	
Health facility managers	2	3	2	1	3	2	13
Nurses	1	4	1	1	0	1	8
Community health extension workers (CHEWs) and Lab technicians	1	4	0	1	3	1	10
Junior CHEWs	0	1	1	0	1	2	5
Total	4	12	4	3	7	6	36

9.2. Findings

I present the findings under each theme in two parts. First, to portray a full account of the participants, I present views and experiences of all participants on contextual and implementation factors that influence the outcomes of the Nigerian P4P scheme, which are supported by participant quotes. The selection of participant quotes[13] was guided by a number of aims: giving a voice to all the participants while ensuring diversity across participant clusters. Some quotes were used to typify a theme while others were chosen because they provided a well-articulated point. In some cases, I present quite lengthy quotes to give a full understanding of the account and a richer context.

Second, I describe and use 'frequency' tables to illustrate the variation in views and experiences of 'participant clusters', which were purposively selected (see chapter 8) to facilitate comparison between participants to examine the extent to which they may possibly explain the variation in performance results in the Nigerian P4P pre-pilot. The groups compared were: health Facilities (worst performers vs. average vs. top), States (Ondo vs. Nassarawa), and health worker qualification (health facility managers vs. other ranks).

While the use of numbers or frequencies may be considered to be controversial in qualitative research, some researchers have argued that the use of counts or display of numeric data is an important part of interpreting data because patterns and deviations from patterns emerge with greater clarity (Maxwell, 2010, Sandelowski, 2001). Therefore, I present a tabular description of the findings, focusing mostly on the variation. In describing the findings, I use terms such as 'few' (below 25%), 'some' (25-50%), 'many' (51-75%), 'most' (above 75%) instead of the actual percentages, so as not to 'over-count' (a disadvantage of using numerical data in qualitative research), which can detract the reader from the primary focus of understanding the views and experiences of the participants (Maxwell, 2010, Sandelowski, 2001).

The findings are presented under four (three pre-set and one emergent) themes: uncertainty of earning the incentive, health worker understanding of the P4P scheme, management and administration of the scheme (the role of health facility managers), and motivation.

[13] To avoid repetition, participant quotes are only presented in the first part of the findings, as they still apply to the second part of the findings. In cases where there was considerable variation in the responses between participant clusters, the quotes selected were chosen to reflect this.

9.2.1. Theme 1: Uncertainty of earning the incentive

Most of the participants expressed distrust and uncertainty of earning the incentive, which was evident across four areas (subthemes), namely: delay in payment, incomplete payment, communication, and individual assessment tool. The first three subthemes were inter-connected and are discussed together in the following section.

9.2.1.1. Subthemes: Delay in payment, Incomplete payment, and Communication

Most of the participants thought that delay in payment, incomplete payment, and ineffective communication (about reasons for delay in payment and other changes in the programme) triggered uncertainty and distrust in the P4P payment system, which had a negative impact on their behaviour and hindered potential improvements in the health facilities. This was mostly because planned improvement strategies that required funds, such as transportation to hard-to-reach areas or purchase of essential equipment were restricted due to delay in payment of the incentive and incomplete payment of the incentive. In the same way, most participants expressed that their motivation to keep up with the required quality and performance targets was reduced as a result of lack of communication about reasons for delay in payment or change in unit prices, as they felt they were being cheated for the money they worked for. Table 9.2 illustrates some examples of these views.

An alternative (deviant) view offered by a small number of participants was that while the delay in payment reduced performance of the health facility, it did not affect their motivation because they felt that bonuses should not affect the way they did their jobs, some of whom had chosen that line of work for altruistic purposes. Similarly, regarding incomplete payment a few participants thought that it did not affect the health facility's performance because they had saved from previous payments, which accommodated the subsequent reduction in payment. Finally, a few of the participants said they had received explanations for the changes in the programme, which turned out to be rumours and speculations (a result of ineffective communication). Table 9.2 presents examples of these alternative views.

Table 9.2 Health workers' views and experiences with uncertainty of earning the incentive

Dominant views		
Delay in payment	**Incomplete payment**	**Communication**
EX1 (Health facility manager in a top performing facility in Ondo State) "*When the payments were not forthcoming, initially the people were complaining but later, they were frustrated and were insinuating that maybe there are some sharp practices (corruption) going on somewhere, which is not impossible. So it is affecting trust and when the health workers don't trust that they will get paid, they will be lax about carrying out their duties. The P4P programme needs to be implemented efficiently and consistently for maximum impact*". **NY1 (Health facility manager in an average performing facility in Nassarawa State)** "*…They (health workers) started saying that I (the in-charge) have received the money and I have spent it instead of sharing it with them. But I told them no, it is not like that, keep working the money will come. But they said they will not work extra hard and not get the money. So they stopped working… and when they money finally came it was small and they were sad, saying look at what we could have gained. So it really affected us, you can see the fluctuation in the results because of that…*"	**NY4 (JCHEW, average performing facility in Nassarawa State)** "*When we first started, the bonus was OK and the members of staff were happy. But they have reduced the money and now the staffs are complaining that they are being overworked and not getting what is due them"* **EZ1 (Health facility manager, low preforming facility in Ondo State)** "*Some health workers have started complaining that the work is too much and the money is not enough; they are not putting in as much effort as they used to before the money was reduced*	**EX1 (Health facility manager, top performing facility in Ondo State)** *"The programme is a very good programme but if there is way it could be fine-tuned so that it would be much more efficiently run. There is need for more transparency and communication. The communication gap should be filled"* **NZ3(CHEW, worst performing health facility, Nassarawa State)** "*the lack of information affected some health workers because they thought they were being deceived because we didn't get any information regarding the money, so it made people more relaxed, thereby affecting performance*"
Deviant/negative cases		
Delay in payment	**Incomplete payment**	**Communication**
NZ3 (lab technician at a low performing health facility in Nassarawa State) *"I have human sympathy and that is why I am doing this job. So not receiving the bonuses on time will not make me relent in my efforts; it doesn't affect my motivation as long as I have the necessary kits. I can only speak for myself; I have that heart in which my primary aim is the welfare of the patient. I am not trying to praise or flatter myself". The delay however affects other things such as buying of some test kits and other laboratory equipment, which reduces our performance here"*	**NX1 (health facility manager, top performing facility in Nassarawa State)** *"The reduction in payment didn't affect us here because we save some of the money, we do not spend everything at once, because we anticipate that things like this will happen, we are in Nigeria, it is just the way the system is. So we plan accordingly. In the end, when they reduced the money, we still had money saved up, so we worked with that".*	**EY2 (Nurse, average preforming health facility in Ondo State)** "*Our in charge said the people at the NPHCDA want to 'spoil' us with money that they are not supposed to give us that money and that is why we have not received out payment. But we know we are working extra hard and we deserve the money*".

9.2.1.2. Subtheme: Individual assessment tool

The views and experiences of the participants regarding the subtheme: 'individual assessment tool' were more diverse in terms of its perceived influence on health worker behaviour in the P4P scheme. There were three distinct views.

Some of the participants thought that it was a fair way of distributing the bonuses because all the criteria they were being judged on contributed to improving performance of the health facility. They also recognised that assessment by their individual contribution (e.g. number of deliveries they assisted with) might be unfair to some because they work shifts and some shifts might be beneficial. For example, there might be more opportunities to go on outreach and home visits patients during the day compared to night shifts.

Other participants thought that whilst the method of assessment was good, it should be improved to include individual contribution to health facility earn the incentive. These health workers also thought that if the bonuses were shared based on their direct input (not captured in the assessment form) such as, the number of deliveries they take, home visits and outreaches; they would try harder because they would have the potential to earn more.

Finally a few participants discussed that the way the bonuses was shared was unfair because health facility managers used ranks to share the bonuses, which did not recognise their own individual contribution in helping the health facility earn the incentive. This made them feel that the allocation of the bonuses was not performance based, thereby discouraging high performance of some health workers.

Participant quotes on the views and experiences of the regarding the assessment tool are presented in Table 9.3.

Table 9.3 Health workers' views and experiences regarding the individual assessment tool

Assessment tool is fine and should not be changed	Assessment tool is fine but should be improved to further reflect individual contribution	Assessment tool (rank) is biased and should be improved to further reflect individual contribution
EY1 (Health facility manager in an average performing facility in Ondo State) "*I like the assessment form because flow of patients varies by the duty that you are. While it may benefit others, it wouldn't be fair on some. The way I see it, the assessment form makes it fair".* **NY1 (Nurse, top performing facility in Nassarawa State)** "*We work according to our qualification: e.g. I'm the only midwife here so I handle deliveries and I let the CHEW handle the outpatient department (OPD). So it is teamwork and all of us are doing our part. So, I don't think it will be a good idea to share the bonuses based on our direct contribution because we have different skills and some things attract more money than others. For example, we receive more money for deliveries than for growth monitoring (OPD). Both are equally important health work and different people do it but to base the allocation of bonuses on that is unfair".*	**EX1 (Health facility manager, top performing facility in Ondo State)** "*…there was a time when the staff went for outreach for immunization and I divided the health workers into two to go to different wards (villages). At the end of the day, when they got back from the outreach, I reviewed the patients both groups have seen. It turned out that one group had seen a lot more patients than the other. Apparently the group with fewer patients had sat down in one place and didn't really bother to move about. Then I told them that when the P4P money comes, this will be taken into consideration and I will not pay them as much as the other group. They were not happy with what I had told them but because they didn't want their bonuses reduced, the next day they went back to that same ward and went to see a lot more patients. So if P4P includes outreach for basis of sharing bonus, I know the work will boost more…*" **EZ1 (Health facility manager, low performing health facility in Ondo State)** *"I think it will be better to assess performance based on the individuals direct input, because you will know that it is based on exactly the amount of work that you do that you are being paid for, which will make you want to work more. They (health workers) will be more focused, they will have a clear target on how much they want to earn and they will work to earn it. All the other items on the individual evaluation form are still important and we can share the money according to a combination of the two. Doing it this way will make the health workers do more work because they will be focused and will have a target"*	**NX2 (Nurse, top performing health facility in Nassarawa State)** "*Somebody who is not doing any work in a higher rank will receive more money compared to me that is working in the maternity section and the labour room, on my feet most of the day working hard. I receive less, just because I am in a lower rank, and you call that performance based? It is not. I would prefer if the bonuses were shared based on the work that I do (direct input) because that will mean it is performance based. For example, if I take 10 deliveries, I expect to be paid based on that because I have worked for it. I want something that will reflect how much work I have done but if the bonuses are shared by rank, it does not reflect the amount of work done because a lot of them in the higher ranks just seat in their offices, doing nothing"* **EX1 (Health facility manager, top performing facility in Ondo State)** *"You see, the individual evaluation form is quite vague, and to tell the truth, as the health facility manager, I don't look at all the points, I just share the bonuses according to ranks when it comes. If the programme implementers can come up something like this: you have performed this in outreaches and you should be give this as bonus, it will be a great idea because it will encourage people to work harder and it is more transparent and it reflects what P4P Is about, because sometimes, a volunteer health worker does more work than a nurse on the same duty or even 2 nurses on the same duty and one does more work than the other and just because they are on the same rank, they get the same bonuses. The one who worked harder will be inclined not to work as hard in the coming quarter since the bonuses will still stay the same"*

9.2.1.3. Variations between participant clusters (Theme: Uncertainty of earning the incentive)

For this theme 'uncertainty of earning the incentive', comparisons between participant clusters revealed similar views and experiences across all subthemes (delay in payment, incomplete payment, individual assessment tool and communication) as seen in Table 9.3. The only noticeable difference was between participants in Ondo state where most health workers thought that the incomplete payment had negative consequences compared to Nassarawa State where only some (about half of the participants) thought incomplete payment had negative consequences on the results of the scheme (as illustrated in the participant quotes presented in table 9.2). In addition, whilst there were diverse views about the subtheme: individual assessment tool, participant responses between participant clusters were fairly similar (See table 9.4).

Table 9.4 Comparison between participant clusters (Theme: Uncertainty of earning the incentive)

Subthemes	Categories	States		Performance			Health workers			
		Ondo n=14(%)	Nassarawa n=22 (%)	Top n=16 (%)	Average n=7 (%)	Worst n=13(%)	Health facility managers n=13 (%)	Nurses n=8 (%)	CHEWs and Lab technicians n=10 (%)	JCHEWs n=5 (%)
Delay in payment	Delay in payment reduces motivation and performance	9 (64)	21 (95)	15 (94)	5(71)	10 (77)	9 (69)	7 (88)	9 (90)	5 (100)
	Delay in payment reduces performance but not motivation	4 (28)	1 (5)	1 (6)	2 (29)	2 (15)	2 (17)	1 (12)	1 (10)	0 (0)
Incomplete payment	No effect on health facility performance	2 (14)	10 (48)	6 (38)	2 (29)	4 (31)	4 (31)	3 (30)	3 (30)	2 (40)
	Negative effect on health facility performance	11 (79)	12 (52)	10 (62)	4 (57)	9 (69)	7 (54)	6 (60)	7 (70)	3 (60)
Individual assessment tool	Assessment tool is fine and should not be changed	6 (43)	10 (45)	7 (44)	3 (43)	6 (46)	3 (23)	2 (25)	7 (70)	4 (90)
	Assessment tool is fine but should be improved to further reflect individual contribution	4 (29)	6 (27)	3 (19)	4 (57)	4 (31)	5 (38)	2 (25)	2 (20)	1 (10)
	Assessment tool (rank) is biased and should be improved to further reflect individual contribution	1 (7)	4 (18)	5 (31)	0 (0)	3 (23)	3 (23)	1 (13)	1 (10)	0 (0)
Communication	Reasons for change communicated through 'hearsay'	2 (14)	1 (4)	0 (0)	0 (0)	3 (23)	0 (0)	1 (13)	1 (10)	1 (20)
	Reasons for changes not communicated effectively	11 (79)	21 (96)	15 (94)	7 (100)	10 (77)	13 (100)	7 (87)	9 (90)	5 (80)

Numbers (%) presented in each column for each category within themes (or sub-themes) are representative of participants within that cluster. Comparisons are made across rows for each category within themes (or sub-themes) to explore variation in responses within participant clusters: States, performance, and health workers

9.2.2. Theme 2: Health worker understanding of the P4P scheme

The P4P programme was well understood by some participants and not by others. Similarly, some participants were aware of the changes in the programme and expected payment, while others were not. Participants who appeared to be more knowledgeable about the schemes often expressed a higher level of enthusiasm, motivation, and behaviour change compared to participants with poor understanding of the scheme. Specifically, the more knowledgeable participants thought that introduction of the P4P scheme provided them the opportunity to learn more about and improve on their jobs, thereby, providing them with the motivation to work harder. Participants' quotes in Table 9.5 illustrate these views.

Table 9.5 View and experiences regarding health worker understanding of the P4P scheme

	Q: Tell me about the P4P programme?	**Q: So what has changed since the beginning of the programme?**
NX1 (health facility manager, top preforming facility in Nassarawa state)	"*P4P is a good programme; we have the opportunity to earn money based on the number of services we render (lists out all the services and unit prices). The more we do, the more we get and we are given autonomy in the way we use the money. I feel P4P allows funds to get to where it is needed the most… we share 50% as bonus for the health workers, 25% for drugs and the other 25% for other things needed. The autonomy itself is something to us and we can decide ourselves what we need to do it"*	*"…I have changed a lot since P4P, I manage the clients better, record keeping is better, we do exactly what needs to be done, and I have learnt a lot more because of the P4P programme…."*
EZ3 (CHEW in a low performing facility in Ondo state)	*".. hmm because of P4P, we have more drugs and we have renovated the health facility…. I know we get bonuses from the programme and it is from the World Bank, I'm not sure how the bonuses are shared, by seniority I think. All I know is that the in-charge brings money and gives me that this is from P4P. I do not know much about the programme all I know is that I get bonuses twice a year".*	*"I have not changed anything about myself. It is what I was doing before I am doing now, because I was doing the work before. It is my job whether there is P4P or not"*

9.2.2.1. Variations between participant clusters: Health worker understanding of the P4P scheme

The findings as illustrated in Table 9.6 reflect noticeable variation in the understanding and level of awareness of changes in the programme. About half of the participants in Nassarawa State had a good working knowledge of the programme compared to the few participants who had a good working knowledge of the programme in Ondo State. Similarly, many participants in the top performing facilities had a good working knowledge of the programme compared to the low performing facilities where there were very few health workers with a good working knowledge of the programme. In the same way, many of the health facility managers and nurses were aware of the changes in the scheme compared to participants with lower qualifications (CHEWS and JCHEWs).

Table 9.6 Comparison between participant clusters (Theme: Health worker understanding of the P4P scheme)

Subthemes	Categories	States		Performance			Health workers			
		Ondo n=14 (%)	Nassarawa n=22 (%)	Top n=16 (%)	Average n=7 (%)	Worst n=13 (%)	Health facility managers n=13 (%)	Nurses n=8 (%)	CHEWs and Lab technicians n=10 (%)	JCHEWs n=5 (%)
Understanding	Good working knowledge of the programme	2 (14)	10 (45)	9 (56)	0 (0)	3 (23)	6 (46)	1 (22)	4 (40)	1 (20)
	Average working knowledge of the programme	12 (86)	12 (55)	8 (44)	6 (86)	10 (77)	7 (54)	7 (88)	6 (60)	4 (80)
Level of awareness of changes in the scheme	Aware of expected payment	5 (36)	15 (68)	11 (69)	4 (57)	5 (38)	10 (77)	6 (75)	4 (40)	0 (0)
	Unaware of expected payment	9 (64)	7 (32)	5 (31)	3 (43)	8 (62)	3 (23)	2 (25)	6 (60)	5 (100)

Numbers (%) presented in each column for each category within themes (or sub-themes) are representative of participants within that cluster. Comparisons are made across rows for each category within themes (or sub-themes) to explore variation in responses within participant clusters: States, performance, and health workers.

9.2.3. Theme 3: Management and administration of the P4P scheme (role of the health facility manager)

The participants expressed several views on and experiences of the influence of management and administration of the P4P scheme at their various facilities through diverse strategies the health facility manager had implemented to improve performance under the P4P scheme. These included practical gifts to patients, hiring additional staff, improved supervision, change in health workers attitude towards patients', equipment and structural improvement, effective use of feedback, outreach, home visits, and free or subsidized drugs and health services.

A few health facility managers gave gifts to patients (e.g. sanitary towels and soaps for pregnant women who deliver at the health facility), which they thought encouraged other patients to come to the health facility, thereby increasing utilisation of health services. Similarly a few health facility managers who hired additional staff said it had helped reduce the workload in the health facility, which led to a more efficient system and increased performance. In the same way, a few of the health workers thought that improved supervision played a role in improving performance because that meant that they could not inefficient, since the supervisory visits were usually impromptu.

Some health workers thought that their changed attitude towards patients' made the patients more comfortable in the health facility, which encouraged utilisation of health services. Some health workers thought that presence of equipment and structural improvement in the health facility brought about increased utilisation because the patients perceived 'a higher quality of care'. Also, some health workers thought that performance feedback from consultants helped them to properly implement changes needed to improve performance. Although a few participants had a contrary experience that they did not receive feedback or that they had difficulty in implementing changes to improve performance based on the feedback received.

Finally, most health workers thought that increase in number of outreach and home visits led to increase in performance because they were now able to attend to patients in hard to reach areas of the community. In the same way, most of the health workers thought that the free or subsidized drugs and health services improved utilisation because members of the community could now afford healthcare **(**see Table 9.7 for participants' quotes, some of which reflect more than one view or experience**).**

Table 9.7 Health workers' views and experiences regarding Management and administration of the P4P scheme (role of the health facility manager)

• **Useful gifts to patients** • **Hiring additional staff**	• **Improved attitude towards patients** • **Equipment and structural improvements**	• **Free or subsidized drugs and health services** • **Improved supervision** • **Outreach and home visits**	• **Effective use of feedback**
NX1 (health facility manage, top performing facility in Nassarawa State "*…With P4P we understand that we have money to do what we need to do in the health facility, we have problem of manpower but we have subcontracted, this fat man writing something outside is one of the workers we hired using P4P money and we hired another attendant so that we can do the work more effectively… we also give small gifts such as sanitary towels and soap for the women who come to deliver at the health facility*"	**EX1a (Health facility manager, top performing facility in Ondo State)** "*We started going for outreaches, we try as much as possible to shorten the waiting time, improve patient satisfaction, we have improved attitude towards the patients, and we also improved the outlook of the hospital and make our patients more comfortable, we have running water around all the time"*	**NY3 (CHEW, average performing facility in Nassarawa State)** "*The supervision of this programme is the driving force for the change. The LGA, state and Federal government sends people to monitor and supervise us almost every week. So we must always be alert and working. We have also increased the mobilization of pregnant women; we campaign for ANC within our catchment area population. We reduced the cost of our drugs and now people can afford our treatments"* **EX1b (Health facility manager, top performing health facility in Ondo state)** "*Well the environment is better for the health workers to stay at their duty posts now. We are now enjoying the place. I used some of the money to buy essentials like generator, fridge, TV. There is nothing that other health facilities are enjoying in the city that we don't have. So we enjoy the place better now and the health workers are motivated to come to work. I even employed a new health worker from the money earned from P4P"*	**NZ1 (Health facility manager, low performing facility in Nassarawa State)** *"We have weekly meetings where I use the feedback form to discuss our past performance and how to improve on where we have performed low. For example, we saw that we got zero in waste management and HIV tests. So now we have started working on our incinerator and we have purchased some HIV kits now so we can improve our performance and earn more money"* **Deviant view on effective use of feedback** **EZ1 (Health facility manager, low performing facility in Ondo State)** *We receive the feedback form, we get zero in some areas and we know we have to improve but there are just some things we don't know how to go about right now. For example, we do not get a lot of pregnant women coming to deliver in this facility, maybe because the building is small or because they just prefer to go to the private clinics. Even with the outreaches we do, they just won't come, so I don't know what else to do to improve that area".* **NZ4 (JCHEW, low performing health facility in Nassarawa State)** "*I don't get any feedback. I don't know if the in-charge gets it and doesn't share it with us"*

9.2.3.1. Variation in participant clusters (Theme: Management and administration of the P4P scheme/role of the health facility manager)

Findings as illustrated in Table 9.8 indicate that view that outreach, home visits, and free or subsidised drugs and health services improved health facility performance had similar patterns across the participant clusters. Other strategies cited by the participants, however, varied considerably across some participant clusters. These included improved health worker attitude towards patients, equipment and structural and effective use of feedback. Most of the participants who cited effective use of feedback were from Nassarawa State, while most of the participant who said they didn't get or use their feedback were from low performing facilities in Ondo State. Similarly, most of the participants who thought or said that improvement in performance was due to their improved behaviour/attitude towards patients or and equipment and structural improvement were from top performing facilities or Nassarawa State. Finally, improvement strategies such as hiring additional staff, giving useful gifts to patients, was less commonly used and specific only to participants in top performing facilities in Nassarawa State.

Table 9.8. Comparison between participant clusters (Theme: Management and administration of the P4P scheme/role of the health facility manager)

	States		**Performance**			**Health workers**			
	Ondo n=14 (%)	Nassarawa n=22 (%)	Top n=16 (%)	Average n=7 (%)	Worst n=13 (%)	Health facility managers n=13 (%)	Nurses n=8 (%)	CHEWs and Lab technicians n=10 (%)	JCHEWs n=5 (%)
Hiring more staff	1 (7)	3 (14)	4 (25)	0 (0)	0 (0)	2 (15)	2 (25)	0 (0)	0 (0)
Gifts to patients	0 (0)	6 (27)	5 (31)	1 (14)	0 (0)	2 (15)	2 (25)	2 (20)	0 (0)
Outreach and home visits	10 (71)	15 (68)	10 (63)	6 (86)	9 (69)	10 (77)	5 (63)	5 (50)	5 (100)
Equipment and structural improvement	3 (20)	18 (82)	12 (75)	5 (71)	4 (31)	6 (46)	5 (63)	8 (80)	2 (40)
Improved supervision	1 (7)	3 (14)	3 (19)	1 (14)	0 (0)	2 (15)	2 (25)	0 (0)	0 (0)
Health workers nicer and more welcoming to patients	4 (29)	19 (86)	14 (88)	6 (86)	3 (23)	8 (61)	6 (75)	7 (70)	2 (40)
Effective use of feedback	5 (36)	14 (67)	9 (56)	5 (71)	5 (38)	8 (62)	5 (63)	4 (40)	2 (40)
Feedback given but not implemented	5 (36)	1 (5)	1 (6)	0 (0)	5 (38)	2 (15)	0 (0)	1 (10)	3 (60)
Free/subsidized services and or drugs	11 (79)	20 (91)	13(81)	5 (71)	13 (100)	10 (78)	6 (75)	10 (100)	5 (100)

Numbers (%) presented in each column for each category within themes (or sub-themes) are representative of participants within that cluster. Comparisons are made across rows for each category within themes (or sub-themes) to explore variation in responses within participant clusters: States, performance, and health workers.

9.2.4. Theme 4: Motivation

The participants expressed their views and experiences about motivation within two sub-themes: sources of motivation to improve performance under the P4P programme and demotivating factors that reduced their performance.

9.2.4.1. Subtheme: Motivating factors improving performance

There were diverse views about the sources of motivation, which included bonuses, and knowledge and experience. Some participants felt that their main source of motivation was the knowledge and skills they had acquired due to exposure to new clinical cases since the introduction of the P4P scheme. Therefore, this encouraged them to want to see more patients and to improve the quality of health services they provide, as this would improve their skills and knowledge.

A few participants also felt that the availability of drugs and equipment at the health facility improved the motivation. This was because prior to P4P, they felt that going to the health facility was futile since there were no drugs or equipment to treat or attend to patients (see Table 9.9 for participant quotes).

Furthermore, most of the participants were positive towards peer reporting and thought that it would help improve performance. They suggested that if peer reporting was introduced to the scheme in the future, it would make them work harder because they would want to be the best performing facility, since other health facilities would be aware of their performance. That said, there were contrasting views from a few participants who thought that peer reporting might not necessarily bring about improvement in performance because it might promote unhealthy competition, as illustrated in Table 9.9.

Table 9.9 Participant quotes on motivating factors improving performance

Bonuses	Knowledge and experience	Availability of drugs and Infrastructural improvement	Positive thoughts toward peer reporting
EY1 (Health facility manager, average performing facility in Ondo State) "*We really appreciate the money; we would do even a lot more if we get more money"* **NZ1 (Health facility manager, low performing facility in Nassarawa State)** "*The truth is that the bonuses encourage the staff a lot. When I start delegating duties to the health workers at first, they are a bit reluctant but once I tell them that their performance on the task will determine part of their bonus, you will see them putting their best into it"*	**EX1 (Health facility manager, top performing facility in Ondo State)** *"One health worker told me that she has improved on her skills because we now attend to more patients and we can put our knowledge to work. We are very happy about that. That makes me happy even more than the money. I feel more exposed to many new cases. I now feel like I my doing my job"* **NX1(Health facility manager, top performing facility in Nassarawa State)** *"…The bonus is minor because how much is it really relating to the salary? I feel a foundation has been laid because the more we see patients, the more we understand, the more we gain more knowledge, like before P4P, we sit down, no patients coming and we are not really practicing; we do not add to our knowledge (we don't know anything) one of the health workers even told me that she now likes this health facility, that before she could not perform a certain procedure well but now she knows it well. So for us, the money is minimal, we feel more exposed to more cases and we are gaining more knowledge…"*	**EY2 (Nurse, average performing facility in Ondo State)** *"A lot has changed, in the sense that before P4P, we were short of drugs and other equipment, but since P4P, the facility can afford to buy those things now. No shortage of drugs now. The patients are happy now that they can come and they will not hear some story about how we don't have drugs in the health facility and this has caused a very rapid great change in the health workers. There has been a massive improvement in punctuality and coming to work.: before P4P, usually the health workers just tell themselves; if there are no drugs in the health facility, why bother come anyway and what are we coming here to do but now, they have no excuse for not coming to work"* **NX2 (Nurse, top performing facility in Nassarawa State)** *"The health workers are coming to work now because we have all the equipment and drugs we need that we didn't have before P4P"*	**NY3 (CHEW, average performing facility in Nassarawa State)** "*It will get us to improve. At least if we cannot be first, it will aim to be 2nd or third. It will really help us improve our services. Even if we are first, we will still improve so we do not come last"* **NX2 (Nurse, top performing health facility in Nassarawa State)** "*We will feel great if the results are published and it will motivate us and create healthy competition among the health facilities"* **Negative thoughts toward peer reporting** **EY1 (Health facility manager, average performing facility in Ondo State)** *"Hmmm, it might encourage sharing ideas within the different health facilities from strategies used, but there are some conditions or situations that cannot be applied to other health facilities. So there are bound to be differences in performance. For instance, this health facility is centrally located and if you look around, the people in the community can easily go to the city or nearest GH to get treatment, instead of coming here. They have closer options. Compared to when you visit other health facilities (hard to reach areas). The community basically have no choice but to go there"* **EX2 (Nurse, top performing facility in Ondo State)** *"I will liken that to competition. Because when there is competition, everyone will be striving to be first. I don't think it will work because some might not care if they are first or not as long as they are still performing and still getting their money. So I cannot really say. I prefer the feedback we are getting to competition"*

9.2.4.2. Subtheme: Demotivating factors decreasing performance

Participants also had diverse opinions about challenges that demotivated them and decreased the performance of the health facility. These included mobility problems, inadequate manpower, competition, and insufficient infrastructure. Some participants expressed instances where the health workers could not go on home visits due to the poor road networks and lack of means of transportation, which often deterred their motivation and overall performance. In the same way, some participants shared experiences of when they had to turn patients away from the health facility due to inadequate infrastructure (see Table 9.10 for participant quotes).

Table 9.10 Participants' views and experiences with demotivating factors decreasing performance

Mobility	Man power	Competition	Infrastructural challenges
EY1 (Health facility manager, average performing facility in Ondo State) *"Our performance is really affected by the terrible roads which make transportation very difficult, so we can't do many outreaches or home visits as much as we would like"* **EX1 (top preforming health facility in Ondo state)** *"The main challenge is transport. Even though we have two means of transportation (a bike and a tricycle) the roads are very bad. We have to repair the motorbike every time it goes out. The terrain is bad. For instance, one of our patients wanted to deliver and before she could get to us, she delivered on the road"*	**NX2 (Nurse, top preforming health facility in Nassarawa State)** *"I will say this is still caused by the problem of lack of manpower as the patients complain that we waste a lot of time before attending to them when they come for anc. This is simply because we don't have enough health workers to juggle the duties"* **NY1 (Health facility manager, average performing facility in Nassarawa State)** *"The number of health workers here is not enough to be on 24 hour duty, attending to patients, like the P4P wants us to do".*	**EZ4 (JCHEW, low performing health facility in Ondo State)** ***"****Some of the patients prefer to go to the 'quacks' we don't know why, we've tried health promotion but some people just prefer their old ways****"*** **NZ1 (Health facility manager, low performing facility in Nassarawa State)** *"One of the major challenges we have is that the general hospital is just a mile away and the patients prefer to go there. When some patients come here they will tell me it is because they don't have money and that is why they are here but once they have money, the go to the general hospital and this reduces the number of patients that make use of our services*	**EZ1a (Health facility manager, low performing facility in Ondo State)** *"We don't have some of the kits or reagents for some of the tests that P4P want us to do, which makes our performance low in those areas; so we are trying to save some money from past bonuses to buy some things and improve our performance"* **EZ1b(Health facility manager, low performing facility in Ondo State)** *"We have just two rooms in this place, so we can only attend to a few people at a time and we have to send people away when the ward is full; in the end it reduces the amount of bonuses we receive"*

9.2.4.3. Variation between participant clusters (Theme: Motivation)

The view that money and acquisition of knowledge was a motivating factor had a similar pattern across the participant clusters (see Table 9.11). On the other hand, the views and experiences of health workers on infrastructural improvement as an important source of motivation was less common among the participants and was mostly specific to a few participants in Ondo State. In addition, a number of participants mostly in Ondo State and in low performing health facilities cited structural challenges as a major source of demotivation, while lack of manpower was cited by participants mostly in Nassarawa State. This suggests a link between motivation and the role of the health facility manager in the scheme described in the previous section, where some health facility managers in Nassarawa State recognised the challenge of lack of manpower, and hired additional staff to help improve performance of the health facility. Whereas very few participants in Ondo State stated cited improved infrastructure, which appeared to be a major source of motivation for them. Other less common challenges such as competition had similar patterns across participant clusters, while mostly participants in Ondo State cited issues with mobility and bad roads.

Table 9.11 Comparison between participant clusters (Theme: Motivation)

		States		**Performance**			**Health workers**			
		Ondo n=14 (%)	Nassarawa n=22 (%)	Top n=16 (%)	Average n=7 (%)	Worst n=13 (%)	Health facility managers n=13 (%)	Nurses n=8 (%)	CHEWs and Lab technicians n=10 (%)	JCHEWs n=5 (%)
Motivating factors improving performance	Bonuses (money)	6 (43)	10 (45)	8 (50)	3 (43)	5 (38)	7 (54)	2 (30)	4 (40)	3 (60)
	Knowledge (education and experience)	6 (43)	10 (45)	8 (50)	3 (43)	2 (15)	8 (62)	0 (0)	3 (30)	2 (40)
	Availability of drugs and Infrastructural improvement	4 (29)	1 (5)	1 (6)	2 (29)	2 (15)	4 (31)	1 (13)	0 (0)	0 (0)
	Positive thoughts towards peer reporting	3 (21)	16 (73)	10 (63)	4 (57)	5 (38)	6 (46)	4 (50)	7 (70)	2 (40)
Demotivating factors decreasing performance	Infrastructural challenges	10 (71)	6 (27)	4 (25)	3 (43)	8 (62)	6 (46)	2 (30)	2 (20)	3 (60)
	Competition	3 (21)	2 (9)	1 (6)	3 (43)	2 (15)	3 (23)	0 (0)	1 (10)	1 (20)
	Mobility	7 (50)	3 (14)	4 (25)	4 (57)	3 (23)	5 (38)	1 (13)	3 (30)	1 (20)
	Man power	1 (7)	7 (32)	2 (13)	2 (29)	4 (31)	2 (15)	1 (13)	3 (30)	2 (40)
	Negative thoughts towards peer reporting	2 (14)	1 (5)	1 (6)	2 (29)	0 (0)	2 (15)	1 (13)	0 (0)	0 (0)

Numbers (%) presented in each column for each category within themes (or sub-themes) are representative of participants within that cluster. Comparisons are made across rows for each category within themes (or sub-themes) to explore variation in responses within participant clusters: States, performance, and health workers.

9.3. Discussion

The Nigerian P4P scheme was introduced in Nigeria to improve access and quality of health services (with a major focus on maternal and child health) in primary healthcare centres (PHC) in three States (Adamawa, Nassarawa, Ondo). First, as pre-pilots in PHCs in one local government are (LGA) each in the three states from December 2012-July 2014, after which the scheme was scaled up to cover all LGAs in each of the three States.

This qualitative study aimed to explore the views and experiences of health workers on influence of contextual and implementation factors on the Nigerian P4P scheme and to consider the extent to which they explain the varied results of the pre-pilot (in the form of a formative evaluation). This is a relevant and significant research focus because previous literature has rarely considered the influence of contextual and implementation factors on the outcomes of incentive schemes especially in low and middle income countries (Van Herck et al., 2010). In addition, because the scheme is a relatively new approach in the Nigerian health system, a formative evaluation was critical to inform the scaling up of the P4P scheme for improved outcomes or effectiveness

In this section, first, I discuss each theme contextualising them in literature and theory where relevant. I then discuss the strengths and limitations of the study, before highlighting the practical implications of my study and providing recommendations to the NPHCDA to improve the Nigerian P4P programme.

9.3.1. Key findings

There were four key themes that captured the views and experiences of health workers in the Nigerian P4P scheme on the influence of contextual and implementation factors on the scheme. They were uncertainty of earning the incentive, health worker understanding of the scheme, management and administration/role of the health facility manager, and motivation, I now discuss each in detail.

Theme 1: Uncertainty of earning the incentive

The findings of this study suggest that factors such as delay in payment, ineffective communication, incomplete incentive payment, and scepticism in the division of bonuses (individual assessment tool) generally led to distrust and uncertainty in payment, possibly leading to decreased health worker motivation and health facility

performance in the Nigerian P4P scheme. This research is the first to directly explore the influence of uncertainty in payment in P4P schemes in the Nigerian context. My findings are consistent with that of Stockwell (2010) who found that clinicians' uncertainty in earning incentive constituted a risk to the effectiveness of the clinical practice improvement payment (CPIP) incentive programme in Australia. More recently, Ssengooba and colleagues (2012) found that one of the implementation factors that contributed to the failure of a World Bank funded incentive scheme in Uganda was delay in payment.

The study findings are also consistent with economic theory that suggests that individuals are 'risk averse' i.e. they tend to go with a less risky alternative (Arrow, 1965). In the Nigerian context, the delay in payment creates a longer time lag between measurement of performance and payment of incentives and some of the health workers appeared to feel that they were not guaranteed to receive their bonuses. Therefore, due to the associated 'risk' of not earning the incentive, the health workers talked about focusing their efforts on other things that were likely to bring immediate rewards as opposed to some promise of payment that they believed they were unlikely to get.

The effect of the uncertainty in payment in reducing performance may have been expounded in the Nigerian P4P scheme because it lies within a context where transparency has not been a strong feature and corruption is widespread within the system (Hargreaves, 2002, Garuba et al., 2009, Okafor, 2009). In addition, delay in payment of the incentive translates to a lack of funding for most of the health facilities, which ultimately leads to reduction in performance or quality of health service delivery. For example, the health workers stated that the main reasons for improvement in quality of care and performance were structural changes, availability of drugs (Free/subsidized), availability of equipment, outreaches, and home visits; all of which require funds. These facilities do not receive any government budget or funding to run themselves (Welcome, 2011, Asuzu, 2005). Therefore, if the incentives are not being paid on time or the money promised is not paid, it consequently reduces health worker motivation and performance of the health facility, which was evident in the health workers' accounts.

These findings suggest that health workers in the Nigerian P4P pre-pilot were bothered by difficulties in accessing the incentive that they thought they were entitled to. These

findings however, did not seem to explain the variation in results of the Nigerian P4P pre-pilot, as the views and experiences of the different participant clusters were similar. It appeared that the distrust in the payment system and the uncertainty of not earning the incentive was a general characteristic across participants in the different States, facilities, and professional qualifications. This however, does not mean the level of uncertainty in payment is not an important contextual and implementation factor. Rather, the findings suggests that the Nigerian P4P pre-pilot generally might have been more effective if there was higher level of trust in the payment system or a higher degree of certainty in earning the incentive. Therefore, it is likely the P4P scheme would work better if the overall trust of the scheme was improved by minimising delay in payment of incentives, improving communication between the scheme implementers and the health workers, and reviewing the individual assessment tool used to measure individual performance and allocate bonuses to the health workers.

Theme 2: Health worker understanding of the P4P scheme

In exploring the levels of health worker awareness and understanding about the Nigerian P4P scheme, it appeared that some of the participants were more knowledgeable about the scheme than others. The findings also suggested that there was a link between health worker understanding and motivation. Health workers who understood the scheme appeared to be more highly motivated and were more willing to do more to improve performance compared to those with less understanding of the scheme.

Patterns of variations in this theme emerged between participant clusters. The health workers in Ondo State and low performing health facilities who had a weaker understanding of the P4P scheme were not clear on how much they would receive as a facility at the end of the quarter (after verification of performance), how often to expect payment, and the basis of sharing performance bonus to individual health. This might partly explain the preliminary results of the scheme, which suggest that health facilities in Nassarawa State performed better than health facilities in Ondo State. It might also explain one of the factors that contributed to difference in performance between the top and worst performing facilities.

Another important pattern emerging within this theme was that understanding of the P4P scheme varied by health workers' qualifications. It is perhaps not surprising that

most of the health facility managers in the participating health facilities understood the scheme better because they are the ones being trained directly by the P4P consultants, and they in turn train health workers in their facilities. This suggests that the efficiency of training of the health workers may have varied across health facilities and that the efficiency of training of health facility managers possibly varied across States as well.

So far, this study is the first to explore health worker understanding of a P4P scheme in a LMIC and in Nigeria in particular. However, the link between understanding of the scheme and motivation or performance is consistent with a recently conducted review by Eijkenaar (2013) who found that P4P schemes in which health service providers were not knowledgeable about the schemes were mostly ineffective. Similarly, other studies have found an association between clinician motivation and understanding on the incentive programme in USA and Australia (Young et al., 2005, Stockwell, 2010).

An explanation of the link between understanding of the scheme and motivation might be because the health workers with better understanding of the programme also thought the P4P scheme reminded them of what they should be doing as health workers. In addition, these health workers also saw their participation in the P4P scheme as a platform for acquisition of knowledge (through quality guidelines provided by the scheme). The findings are consistent with the work of Leshabari et al. (2008) and Luoma et al. (2005), which both showed that acquisition of knowledge was an incentive to improve motivation in health workers in LMICs.

In summary, based on the reflections on the findings, there is need to engage with key stakeholders such as the health facility managers and the P4P consultants in each State on how to improve understanding of the health workers of the P4P scheme in order to improve the effectiveness of the scheme.

Theme 3: Management and administration of the P4P scheme (role of the health facility manager)

The findings of this study suggest that the health facility managers had implemented various strategies to improve performance and quality of care. This included outreach, use of feedback, home visits, availability of at least one staff at the health facility at any point in time, and availability of drugs.

However, a pattern that emerged was the lack of use of feedback in the worst performing facilities in Ondo State, where some health workers had stated that despite receiving the performance feedback, they did not know how to improve on performance in the areas they were performing poorly in. A possible consequence of this is the reduction in performance reflected preliminary results partly due to ineffective use of performance feedback compared to other facilities. This is in line with the study by Ssengooba and colleagues (2012) who found that the discontinuation of the engagement of stakeholders in the discussion of feedback contributed to the failure of a P4P scheme implemented in Uganda, thus suggesting the importance of stakeholder discussion and technical assistance with strategies to improve performance.

Another pattern emerging from the findings is that the health facility managers in top performing facilities in Nassarawa State had implemented unique strategies such as giving practical gifts to patients, structural improvements, and hiring additional staff, some of which the health workers had cited as sources of motivation. This might partly explain why these facilities performed better than others. For example, hiring additional staff helped to meet the increasing demand for health services, reduced waiting times for the clients thereby improving the quality of health services in that facility compared to low performing facilities. This also suggests that health facility managers in the top performing facilities had superior managerial skills superior knowledge/understanding of the scheme, because it appears they were able to recognise, prioritise and meet the needs of the health facility and motivate the health workers.

I am not aware of other studies providing evidence that allows for comparison with my findings. However, it can be said that quality improvements is to a certain degree dependent on the skills or ability of the health facility manager (Ndizeye et al., 2014), and that whatever the strategies used to improve quality of care and performance are, managerial skills and explicit dialogue with the members of staff are important (Elovainio, 2010).

Theme 4: Motivation

The findings of this research suggest that motivation and performance of the health workers have increased as a result of the P4P scheme, by way of bonuses, availability of drugs and equipment, and the acquisition of skills and knowledge. These are consistent with a study exploring health worker motivation in a P4P scheme in Rwanda (Paul,

2009). Other studies on the general motivation of health workers in LMICs also support these findings (Luoma, 2005, Leshabari et al., 2008).

The findings also show that most of the participants thought that the introduction of peer reporting would motivate them to work harder. Though peer reporting was not in practice in the Nigerian P4P scheme at the time of the study, it was of relevance because it is potentially relevant because as a 'non-financial incentive, it is less costly and therefore has the potential of being more cost effective than P4P schemes. In addition, emerging literature suggests that peer reporting used in conjunction with financial incentives could be more effective in driving behaviour change than just financial incentives alone (Jha et al., 2012, Bridgewater et al., 2007, Luoma, 2005, Kolstad, 2013).

The findings suggest that a combination of incentives is responsible for health worker motivation. This provides evidence central to the debate on what motivates health workers in incentive schemes: whether financial (bonuses) or non-financial incentives. The findings are in line with those from the critical review by Henderson and Tulloch (2008) who argue the need for both financial and non-financial incentive to improve quality of care in LMICs. This is because a multifaceted incentive approach is needed in countries with weak health systems where poor motivation of health workers results from a combination of factors such as poor salaries, poor working conditions, inadequate infrastructure, and limited opportunity for career development or training.

Another pattern emerging from the findings was that issues with mobility, lack of infrastructure, and lack of manpower were sources of demotivation for the health workers in both Nassarawa and Ondo States. However, health facilities in Nassarawa State performed on the average better than health facilities in Ondo State. This might be for a number of reasons. One is that some health facility managers in Nassarawa State appeared to take initiative and had swift response to the lack of manpower problem leading to hiring of additional staff, whereas infrastructural problems did not improve in Ondo State (as described earlier). The second reason is that mobility issues in Ondo State hindered patients from utilizing health services and possibly hindered health workers from getting to work, leading to decreased outputs. A third reason is that the challenges facing Ondo state were more costly to control than in Nassarawa State. For example, it seems relatively easy and cheap to hire additional staff compared to buying

a car or carrying out road repairs to improve on the mobility problems in Ondo State. Furthermore, at the time of the study, salaries of the health workers in Ondo State had not been paid for six months. Even though this was not directly related to the P4P scheme, it appeared to have affected the motivation of the health workers participating in scheme in Ondo State. The non-payment of salaries and costly nature of the mobility challenges in Ondo State, therefore, may have put the health facilities there at a disadvantage compared to those in Nassarawa State, leading to reduced performance outputs in Ondo State.

Based on these findings, it appeared that some of the health facilities were not ready to meet the demands of the P4P scheme, which appeared to have led to decreased performance. This is similar to the findings of Locke and Srinivasan (2008) in osteopathic physicians practices in an incentive scheme in USA. Therefore, demonstrating that it is important that all the health facilities are ready to practice, report, and meet reimbursement requirements of the Nigerian P4P scheme.

9.3.3. Strengths and limitations

It is important that the key findings of my research be interpreted in light of the strengths and limitations of this study, I reflect on these now before highlighting the practical implications of my work. This study had four key strengths: novelty, rich detailed data, rigour, and utility, which I now discuss.

This study is the first to explore the influence of contextual and implementation factors in the Nigerian P4P scheme. In addition, previous literature looking at other P4P schemes has rarely considered the influence of contextual and implementation factors in incentive schemes in low-middle income contexts (Van Herck et al., 2010, Eijkenaar, 2012). Hence, my research makes an original and relevant contribution to this literature.

Another key strength of this study lies in the extent, richness, and comprehensiveness of data, which has considerable significance in the applicability or transferability of the findings (Shenton, 2004). I have demonstrated this by providing detailed descriptions and extensive participant quotes from multiple perspectives (including deviant cases). I have also justified methods used and provided essential contextual information in interview data. However, I remain mindful that such quotes are constructs of the

research situation representing partial accounts located in specific interactions (Fontana and Frey, 2000).

Another strength of this study is the steps I have taken to ensure the quality and rigour of the study. Throughout this study, I have provided evidence to demonstrate the rigour and trustworthiness of the study (see chapter 8). This included justification of the research topic and process of data generation and interpretation, comparing multiple data sources, and deviant case analysis. I have also provided detailed transparent description of both the data collection and analysis processes, while acknowledging the multiple ways data could have been collected and analysed. Furthermore, in the next section, I considered how my assumptions and background may have affected the research process and the steps taken to reduce the associated limitations.

Finally, the findings have utility for the consumers (Alvesson, 2009). In section 9.4, I provide recommendations to the World Bank and NPHCDA to inform the implementation of the Nigerian P4P pilot scheme. I also demonstrate how the findings could contribute to planning and implementing incentive programmes in LMICs

Limitations

There were four main limitations in this study. First, my lack of experience in conducting a qualitative piece of work was a potential limiting factor. Experts however, supported this research; my interview skills, data collection and analyses were closely monitored and scrutinised. In addition, my background in and experience in public health (Master's degree), piloting the interviews and my familiarity with the context interviews were strengths that perhaps made up for my lack of experience with conducting interviews.

Second, my familiarity with the context demonstrated by my assumptions and pre-conceived notions about the Nigerian health system such as corruption, distrust in payment system, and poor governance, partly influenced the direction of the research questions/hypotheses. In this study, however, these assumptions, hypotheses, and research questions were supported and verified by extensively by evidence from literature. I also made sure I took steps not to impose my ideas or thoughts on participants in the research. For instance, I asked neutral non-leading questions during

the interviews and I provided and included deviant cases and sought alternative explanations in the data analyses process (see chapter 8).

Third, I was the only person who coded and analysed data, which some researchers have argued might be a limitation due to my familiarity with or closeness to the data' (Flick, 2014). This was ameliorated somewhat, through expert scrutiny of the analysis process (my supervisor and other members of my Thesis Advisory Panel looked over segments of data and coding) (Krefting, 1991). Furthermore, Barbour (2001) argues that regardless of the number of coders or researchers analyzing the data, the most important thing is that a systematic process of data analysis presented transparently and in detail (which I did in chapter 8) .

Finally, I was unable to go the third State (Adamawa) as originally planned due to terrorist attacks. Data from interview of health workers from this State might have provided further insight to answering the research question.

9.4. Recommendations for the implementation of the Nigerian P4P pilot

This research has contributed to the understanding of the influence of contextual and implementation factors in the Nigerian P4P scheme. It has also reinforced existing knowledge of the effects of some of these contextual factors in low and middle-income settings with weak health care systems.

In this section, I consider the implications of the findings of this study for policy and practice, which I present in form of recommendations (which were fed-back to the NPHCDA) to inform the Nigerian P4P scheme, and to a lesser extent, other P4P schemes in low and middle income settings. I outline these recommendations below.

A review of the payment mechanism (how individual health workers earn bonuses)

One of the risks to the effectiveness of the scheme is the degree of the health workers' trust in the payment mechanism. A continuation of problems with transparency, ineffective communication, and delayed access to the incentive may limit the impact of the scheme. A review of the process by which the health workers earn bonuses is required to improve the results of the scheme. The scheme implementers particularly

account and finance officers, P4P consultants[14] in each State, and representatives of the health workers are significant stakeholders in this process and they should be engaged for advice. The following lists the primary points for review:

- To make timely quarterly payments to the each health facility for delivery of services as agreed in the P4P contract.
- To ensure clear communication strategies about changes and difficulties encountered in the scheme to stake holders, particularly to inform and keep the health workers up to date.
- To ensure that the individual assessment tool (basis by which individual health workers earn bonuses) includes a criterion/a set of criteria that clearly captures actual contribution and direct input of the health worker in helping the facility earn money. For example a criterion on outreaches or home visits could be included.
- To provide clear and short guidelines to encourage the use of the individual assessment tool instead of ranks to allocate bonuses to the health workers.
- To move towards 'true pay for performance' (e.g. 50% change in utilisation from baseline) as opposed to pay for reporting.

Develop a plan or guide to foster health worker literacy levels and understanding of the scheme

This study suggests that health worker understanding and knowledge of the P4P scheme is important in improving the impact of the scheme. Furthermore understanding of the scheme by health facility managers appeared to be related to their ability to prioritise the needs of the health facility and health workers needed to improve performance.

It is, therefore, important that health workers' understanding of the scheme be improved. P4P consultants in each state and health facility managers are important stakeholders in this process and should be engaged in planning. In order to improve health workers' understanding of the scheme, the following actions were recommended:

[14] Consultants hired by the NPHCDA to provide technical assistance on the PBF scheme in each State

- To provide training and regular workshops for health workers and equip health facility managers with materials to inform the health workers on how the P4P scheme operates.
- To help health facility managers improve their managerial skills, with a focus on setting priorities, and recognising and meeting the needs of the health facility or how to motivate the health workers (whether it is infrastructure or hiring additional staff).

Effective use of feedback

The findings from the views and experiences of the health workers suggested that feedback on performance was relevant to the effectiveness of the scheme. However, central to the use of feedback is ensuring the health workers utilise the feedback effectively. The recommendation to the NPHCDA on this area was:

- To improve use of performance feedback possibly through discussion of ideas between P4P consultants and the health facility managers on how the health facilities can improve performance.

Start-up resources

The findings suggested that factors such as lack of proper equipment or infrastructure in some health facilities led to decreased performance in these facilities. To improve performance and the effectiveness of the Nigerian P4P scheme, health facilities should be equipped to provide the incentivised health services. While the scheme implementers have expressed their aversion in 'input[15]' funding, it might be important that the health facilities participating in the scheme have the start-up resources or equipment to effectively participate in the scheme i.e. levelling the playing field. Therefore, I recommended that:

- One-off investments could be made in the poorer facilities by either the scheme implementers or the State governments, so as to bring the concerned health facilities to an acceptable standard for a more effective programme.

[15] E.g. buying equipment or improving infrastructure (see chapter 6)

9.5. Recommendations for research in P4P in Nigeria

The literature in the field of P4P in LMICs is underdeveloped. This study was formative and has provided a solid foundation for a future programme of research in Nigeria. I highlight three important specific gaps in research, which require further investigation.

First, in the previous section, I made practical suggestions on how the P4P scheme might be improved to make it more effective. This should also include formative and impact evaluations to ensure that the programme's effectiveness and costs (cost effectiveness study) can continue to be assessed as the scheme evolves until its end in 2018. Thus ensuring scarce resources are being maximised and contributing to the sparse evidence base in this area.

Second, whilst there is considerable evidence supporting the choice of process indicators in the Nigerian P4P scheme, such as deliveries at the health facility, antenatal care, postnatal care, and child growth monitoring, it is important to explore whether the process measures will ultimately translate to better health outcomes (reduction in child and maternal mortality rates), as such research will contribute richly to the selection of incentivised indicators in the future.

Finally, the findings from the study suggested that peer reporting of performance results might improve performance. However, this was not in practice in the scheme at the time this study was conducted. Peer reporting of performance results has been shown to be effective and possibly more cost effective than P4P in a few studies (Kolstad, 2013, Bridgewater et al., 2007). Literature regarding peer reporting is underdeveloped. Therefore, I recommended that peer reporting should be piloted in the Nigerian P4P scheme to explore its potential effect.

9.6. What this chapter adds

In the second part of this thesis, I have used a qualitative approach to explore the influence of contextual and implementation factors on the Nigerian P4P scheme, which has informed the development of recommendations for the Nigerian scheme designers and other similar low and middle-income contexts to inform and improve outcomes of P4P schemes. I also prepared a report for the NPHCDA (see Appendix G1 for executive summary) highlighting recommendations and changes needed for a more effective

scheme, some of which they since effected, such as minimising delays, effective communication, and piloting peer reporting.

In the next chapter, the final chapter of this thesis, I discuss both parts of the thesis, considering not only the influence of contextual and implementation factors, but also the influence of design features on incentive schemes in health care. Particularly, I distil and synthesize the whole thesis in light of its aims and objectives, drawing out the implications of the thesis for policy, practice, and research.

Chapter 10 Discussion and Conclusions

In this chapter, I revisit the background to the thesis, drawing out its aims and objectives. Then, I summarise the research conducted, my findings, and I outline how they relate to existing knowledge. Next I discuss strengths and limitations of the research. Finally, I discuss the relevance of the findings of the thesis for the P4P schemes in health care; providing recommendations for policy, practice, and future research.

10.1. Background and significance of the thesis

The use of incentives schemes to improve quality and efficiency of health care (often referred to as P4P or PBF) has gained popularity over the past few years. However, P4P schemes to some extent have been uncritically implemented, as effectiveness of these schemes has not been convincingly demonstrated (heterogeneous results) despite over a decade of experimenting with P4P.

Researchers suggest that the effectiveness of P4P schemes is related to design features, contexts, and implementation, and that this may explain the heterogeneous results (Epstein, 2012, Van Herck et al., 2010). A few researchers have attempted to narratively review the literature, taking into consideration the design features, conclusions of which were mixed and subjective (Stockwell, 2010, Eijkenaar, 2013, Van Herck et al., 2010). A more sophisticated, quantitative, and systematic approach, which takes into account the design features, contexts, and implementation variables is needed to review this literature in order help make sense of the available evidence.

P4P literature however, lacks a reliable and informed framework to help explore this heterogeneity, quantitatively and systematically, drawing on the range of economic theory on the relationship between behaviour change and incentive that could help guide this. In addition, there is limited knowledge about important aspects of how the impact of P4P may be affected by the context and method of implementation, which requires exploratory qualitative research.

Given that the interest in P4P is unlikely to diminish in the coming years especially in low and middle income countries (LMICs), it is important to try to make better sense of the evidence base and to clarify the features of the scheme which influence effectiveness. In this respect, insight is also necessary in how the implementation and contexts of P4P influences effectiveness of the schemes. Therefore, the aim of this thesis was to improve understanding and provide significant insight and recommendations in these areas.

The objectives of this thesis (chapter one) were:

1. To review the available evidence on P4P and identify the shortcoming of the evidence
2. To develop a reliable framework to categorize P4P schemes in a systematic way to aid evidence synthesis
3. To systematically and quantitatively explore what design features are critical to the effectiveness of P4P
4. To explore the impact of design features, contexts, and implementation factors on effectiveness of the Nigerian P4P scheme

The unique contributions (from each chapter) of this thesis are summarised in the next section.

10.2. Summary of research and findings

Objective 1: to review the available evidence on P4P and identify the shortcoming of the evidence was addressed in chapter two. A systematic search of five literature databases identified 15 relevant reviews analysing evidence of the effectiveness on incentive targeted at health service providers. The quality of the reviews ranged from moderate to good using the AMSTAR criteria. A narrative summary of the evidence of the effects of P4P was produced using the findings from the Identified reviews. Furthermore, 96 P4P primary studies evaluating 65 schemes from different countries were identified from these reviews (in addition to an updated review and other sources). The P4P schemes focused on different aspects of health care, including utilisation of care, smoking cessation, diabetes management, and reducing hospital mortality. However, only 36 of these P4P evaluations were included in the meta-analysis due to

insufficient information to convert the effect sizes to a standardised measure to allow pooling of the estimates, which used a range of measures of effect.

A notable finding was that P4P was more effective in improving process measures such as cancer screening, immunisations, and smoking cessation advice, compared to outcomes such as hospital mortality or smoking quit rates (Hillman et al., 1998, Jha et al., 2012, Twardella and Brenner, 2007). Pooled effect estimate (standardised mean difference) of a subset of studies with process measures was 0.18 (95%CI 0.06, 0.31) compared with the pooled estimate of studies with outcome measures, which was 0.0 (95%CI -0.01, 0.01). This might be because changes in outcomes are harder to achieve because it is less within the control of the clinician. This in turn might also be regarded by the clinician as a 'risky' investment, and therefore, would rather invest their time and resources in a less risky alternative where they are guaranteed desired outcomes. The review also revealed a lack of good evidence to draw valid conclusions on the cost-effectiveness and sustainability of P4P schemes. Furthermore, there was ambiguity regarding the extent of attribution of improvements to P4P due to implementation of other quality improvement strategies alongside P4P and poor evaluation designs or lack of convincing control group in a good number of evaluations studies.

Overall the findings from exploration of literature showed that the evidence on the effectiveness of P4P is mixed (with substantial statistical heterogeneity). Pooled estimates from the meta-analysis also showed that P4P might have a very modest effect on improving the quality of health care (0.15 95%CI 0.03, 0.25). However, this effect is likely to have been over-estimated because findings from further subgroup analyses suggest that RCT evaluations of P4P had a lower pooled effect estimate (0.08 95%CI 0.01, 0.15) compared with quasi-experimental evaluations (0.14 95%CI -0.03, 0.31) and P4P evaluations without adequate controls (0.15 95%CI 0.09, 0.21). However, there was substantial heterogeneity between the pooled studies (I^2=99.9%), which made it difficult to interpret the evidence. This demonstrated the need for a more sophisticated approach to generate evidence (exploration of the heterogeneity).

Objective 2 to develop a reliable framework to categorize P4P schemes in a systematic way in order to explore heterogeneity (to aid evidence synthesis) was addressed in

chapters three and four. I identified and analysed relevant theoretical and empirical literature on the effect of incentives on behaviour, which indicated that several key design features might influence effectiveness. These include who receives the incentives, the type of incentive, the size of incentive, and the perceived risk of not earning the incentive. The analysis also indicated design choices that are likely to produce the desired effects: payment of incentives to groups or organisations as opposed to individuals; payment of large incentives (>5% of usual budget or salary) as opposed to small incentive (<5% of usual budget or salary); levying of fines as opposed to payment of bonuses; and a group of design features that constitute a lower risk or uncertainty (for the recipients) in earning the incentive (minimal time lags between patients, absolute performance measures as opposed to relative measures, and process/structural domains of performance as opposed to outcomes). These design features were then used to create a typology to systematically classify 'types' of schemes with similar design features.

The P4P typology was piloted and then formally tested using 12 health science early-stage researchers (who were trained to use the typology). This showed that the P4P typology was reliable (kappa>0.7 on all four items), and easy to use as a tool to categorise well-reported P4P schemes in healthcare based on their design features. However, in several evaluation studies, the descriptions of P4P designs were poor, incomplete, vague, and non-uniform.

These findings suggest that the P4P typology is ready for use by other researchers as a novel and potentially reliable tool to categorise P4P schemes in health care in an informed way, which would help to explore heterogeneity and make sense of the available evidence of P4P. In addition, the typology could help P4P developers' structure and inform their choice of design features. It could also be used to establish a common language (a reporting template) in which P4P designers, reviewers, and implementers can clearly specify the content of P4P designs in a standardised way and report them, so that reporting of these schemes can be clear, concise and uniform (allowing other people to see what is being done).

Objective 3 to systematically and quantitatively explore what design features are critical to the effectiveness of P4P was addressed in chapter five. I used the typology to

group the previously identified 96 published P4P evaluations into three coherent categories (type A, B, and C) based on their design features: payment of large incentive (>5% of salary or usual budget), payment of incentive to 'groups' (hospitals, clinical groups etc., where individual clinicians may benefit from the incentive), and low risk of earning the incentives (minimal time lags between measurement of performance and payment of incentives, absolute measures, and process domain of performance). Type A schemes (high chance of effectiveness) have all three design features present; type B schemes (medium chance of effectiveness) have two out of three of the design features; and type C (low chance of effectiveness) have one or less than one of the design features present. I then explored the influence of design features and these categories on the effectiveness of the schemes using meta-regression and multilevel logistic regression models on the evaluation studies of P4P identified in chapter two. This study presents the first systematic and quantitative exploration of heterogeneity in results of evaluations P4P schemes in healthcare (using the theoretical typology developed in chapter three).

The findings from both statistical models were similar and they present early steps towards a better understanding of how design features influence the impact of P4P schemes from which several tentative conclusions were drawn. First, P4P schemes (type A and B) with two or more 'adequate' design choices were found to be more effective than 'type C' P4P schemes with less than two of the adequate design features (OR= 3.04 95%CI 1.04,11.76). Another notable finding was that the size of incentive appeared to be the most important design feature influencing the effectiveness of P4P schemes, though very little work has been done regarding optimal size of incentive. Moreover, there is the risk of paying 'too much' for diminishing returns in terms of change in behaviour or consequent health outcomes, rendering P4P inefficient (Evans, 1974). Lastly, it is noteworthy that P4P schemes with poor evaluation design (without adequate controls) appeared more effective than P4P schemes with adequate control groups (OR=24.16 95%CI 6.31, 92.78). Thus reinforcing the need to conduct more rigorous evaluations of upcoming P4P schemes to enlarge the database and strengthen the findings of this thesis.

Robust conclusions on the influence of design features and effectiveness of P4P were not possible for three reasons. First, effectiveness of P4P schemes also likely depends

on contextual and implementation factors. For example, in contexts where corruption or lack of transparency is prevalent, recipients of the incentive are likely to have high levels of uncertainty in earning the incentive, which could translate in minimal behaviour change, despite having 'adequate' design features that reflect lower uncertainty in earning the incentive (minimal time lags between measurement of performance and payment of incentive, absolute measures, and process domains of performance). Second, the efficacy of certain theories (that informed the development of the typology) is context dependent. This aspect was, however not captured in typology. Therefore, the findings were interpreted in light of this. An example is organisational theory that proposes that payment of incentives to groups rather than individuals are more likely to bring desired effects because the organisations are capable of promoting behaviour change in employees through a wide range of strategies e.g. better structures, enacting stricter guidelines and policies etc. (Stewart, 1998). However, this is dependent on how well the organisation is managed. So paying incentives to the organisation as opposed to individuals is only likely to result in greater impact if it is well governed and managed. Third, there is difficulty in implementing certain design features such as fines in some contexts, even though theory suggests that it is likely to bring about a higher impact than payment of bonuses. As a result, fines are a rare design choice in P4P and this impeded the empirical exploration of its influence of P4P effectiveness.

Objective 4 to explore the impact of design features, contexts, and implementation factors on effectiveness of the Nigerian P4P scheme was addressed in chapters six to nine. I conducted a formative evaluation to explore the influence of context and implementation on P4P in Nigeria in order to inform the implementation of P4P on a larger scale in Nigeria. I used a qualitative approach to explore present evidence on the influence of contextual and implementation factors on the Nigerian P4P scheme (a LMIC), which aimed to improve the quality and utilisation of maternal and child health services in rural areas. I reviewed the design features in light of the findings from the quantitative analyses (meta-regression and multilevel regression in chapter five) and an analysis of preliminary results of the Nigerian P4P scheme. I then conducted semi-structured interviews of health workers participating in the scheme to explore their

views and experiences on the influence of contextual and implementation factors on the P4P scheme.

The qualitative study of interview of health workers in the Nigerian P4P scheme (implemented in Ondo and Nassarawa States) suggests that several contextual and implementation factors influenced the scheme. This led to variation in improvements in quality and utilisation of incentivised health services. There were four key themes that captured the views and experiences of health workers: uncertainty of earning the incentive, health worker understanding of the P4P scheme, management and administration of the P4P scheme/role of the health facility manager, and motivation. The first key finding from this study was that uncertainty of earning the incentive (brought on by delay in payment, incomplete payment, doubt in the method of allocation of bonuses to individual health workers, and ineffective communication) decreased motivation of the health workers and/or performance of health facilities. Second, increased health worker understanding of the scheme appeared to be related to increased motivation, which could possibly explain some of the variation in performance between Ondo and Nassarawa States, between top performing and low performing health facilities, and between levels of qualifications. The third interconnected theme was that some health facility managers in top performing facilities or Nassarawa State appeared to have superior managerial skills, which was evident in the strategies they used to motivate the health workers and improve performance in their facilities, such as hiring additional staff and infrastructure improvement, which the participants thought led to improvement in performance reflected in the preliminary results. The fourth major finding was that other factors not inherent to the P4P scheme, such as issues with mobility and lack of infrastructure in Ondo State or lack of manpower in Nassarawa State were sources of demotivation for the health workers. However, some health facility managers in Nassarawa State took initiative and had a swift response to the lack of manpower problem leading to hiring of additional staff, whereas infrastructural problems did not improve in Ondo State. This could possibly explain why health facilities in Nassarawa State appeared to have performed better than those in Ondo State.

On the basis of this fieldwork, I made a series of recommendations to the Nigerian P4P scheme (see appendix G1), which are also relevant (to a lesser extent) to similar LMICs In summary, these were:

- To make timely (quarterly) payments to the each health facility reflecting performance
- To ensure clear communication strategies about changes and difficulties encountered in the scheme between stake holders (to ensure people understand the scheme and encourage effective communication)
- To review the individual assessment tool to include a performance criterion that contributes to earning the incentive (e.g. outreach or home visits)
- To move towards 'true pay for performance' (e.g. 50% change in utilisation from baseline) as opposed to pay per level of activity[16]
- To develop a plan or guide to foster health worker literacy levels and understanding of the scheme
- To consider the use of 'start-up' resources to ensure all the P4P health facilities have the basic infrastructure to provide the incentivised health services.

I now summarise how the aim of the thesis was achieved through the objectives described in previous paragraphs. The aim of this thesis was to better understand important aspects of design, context, and implementation, and to consider their implications on the effectiveness P4P in health care.

First I systematically demonstrated that certain design features (e.g. size of incentive and risk of not earning the incentive) influence the effectiveness of P4P schemes in improving quality of care. Generally, schemes with large size incentives (>5% of salary or budget) and low risk of not earning the incentive (smaller time lag of payment, absolute measures of performance, and process domains of performance) tend to be associated with higher estimates of effect.

[16] Even though there might be risks associated with this such as gaming but measures such as verification and audit trails could be put in place to avoid or address it.

These findings are consistent with theory explored and discussed in previous chapters. For example, theory suggests that large incentives might drive higher performance because of its potential to supplement clinician income and help reach what is known as the target income (Desquins et al., 2009, Evans, 1974) i.e. the larger the size of incentive, the higher the potential of reaching their target income, and the more the clinicians are willing to change behaviour and or improve performance.

Following these findings from theoretical and statistical exploration of the influence of design features on the results of P4P, one would expect that generally speaking, a well-designed P4P scheme has higher chances success compared to poorly designed schemes. However researchers have argued that apart from design features, context and implementation of P4P are likely to influence the effectiveness of P4P (Van Herck et al., 2010, Eijkenaar et al., 2013, Canavan et al., 2008, Toonen et al., 2009). This led me to conduct a formative evaluation on a P4P case study in Nigeria, to better understand how these factors influence P4P schemes, particularly in the Nigerian context.

Early results of the Nigerian P4P scheme suggest that it was successful, in that there were measurable improvements in incentivised services. This was likely as a result of the optimal design features used in the scheme. For example large sized incentive (>10% of salary or budget), paid to groups (health centres where individual clinicians earned bonuses from), with low risk of not earning the incentive (quarterly payments: short time lags, process indicators: within clinicians control, and absolute performance measures). Theoretical and statistical findings discussed in earlier paragraphs (in detail in previous chapters) support and explain why these design features might have resulted in improvements in quality of care. There were however alternative explanations as to why P4P appeared to be effective in Nigeria in particular. These include increased motivation, improvement in transparency, record keeping, and accountability, which are core issues in the Nigerian health system that have led to the failure of past health reforms. For example, the bonuses earned were used in most health centres to buy equipment, drugs and improve infrastructure, and to supplement the salaries of the health workers. All of which most health workers participating in the P4P scheme thought improved their motivation, compared to before the P4P scheme when there was no incentive to come to work either because there were no necessary medications or equipment, in addition to the meagre salary payments. Improved motivation as a result

of the P4P scheme in turn led to reduced health worker absenteeism and improved attitudes towards patients.

Overall the preliminary results of the Nigerian P4P scheme suggest that P4P improved quality of care. However, there was substantial variation in the results between the health centres/practices despite having similar 'optimal design features'. This presented a good case study because the likely reasons for variation were factors other than design features e.g. contextual and implementation factors. This helped to focus the research and explore more efficiently the influence of context and implementation on results of P4P schemes through a formative evaluation.

Findings from the formative evaluation suggest that contextual and implementation factors are indeed important in the effectiveness of P4P schemes, whilst also drawing out the importance of such formative evaluations in other P4P schemes. In the Nigerian scheme in particular, factors such as high uncertainty of earning the incentive, poor health worker understanding of the scheme, and poor management of the scheme (at the health centre level) influenced the impact of the scheme, and limitations in these factors need to be addressed to improve the chances of success of the scheme in Nigeria.
In the next section, I outline the strengths and weaknesses of the thesis before going on to discuss the implications of the findings on policy, practice, and research.

10.3. Strengths and limitations

This thesis makes original and significant contributions to this literature through the development and testing of a potentially reliable typology to aid evidence synthesis of effectiveness of P4P, systematic quantitative exploration of the influence of design features on the effectiveness of P4P schemes (using robust statistical techniques), and using semi-structured interviews to explore the role of context and implementation on P4P (ensuring rigour to improving the validity and reliability of the research).
Furthermore, this thesis has produced policy recommendations for design and implementation of P4P in Nigeria. I highlighted the strengths and limitations associated with each element of research in the previous chapters of this thesis. In this section, I summarise the main limitations and discuss what could have been done differently.

Most of the studies included in the literature review and quantitate analyses (meta-analysis and regression models) were low quality studies (see chapters two and five). The findings of the analyses may have been strengthened by inclusion of only high quality RCT studies. This would have, however, resulted in too few studies to perform the analysis. I tried to compensate for this by either adjusting for evaluation designs or performing sensitivity analyses (as seen in chapter five). Furthermore, in order to bring all the studies together in one analysis, I converted the measures of effect to a standardised measure (standardised mean difference). However, only 36 could be included because of non-uniformity in reporting the effect sizes and inadequate information in most studies to convert the effect sizes to a standardised measure. The study authors could not be contacted to get the additional data needed due to time constraints. The findings and conclusions of the meta-regression (to explore the effect of design choices on effectiveness) would have been stronger (higher power to detect small effect sizes) if all 96 identified P4P evaluation studies were included in the analyses. To increase confidence in the findings, I tried to offset this limitation by adapting the effect size data to binary to accommodate all 96 studies in logistic regression model. The results of both models were similar, which increased confidence in the findings.

Had there been more studies, it would have been worthwhile to unpack further the influence of the different quasi-experimental evaluation designs on the effectiveness of P4P schemes. Literature suggests that some quasi-experimental designs e.g. interrupted time series (ITS) designs are stronger than others e.g. cross sectional designs (Ramsay et al., 2003). For example, an ITS study by Serumaga et al. (2010), which assessed the impact of the UK QOF scheme on management and outcomes of hypertension found that improvements in hypertension management and outcomes were as a result of gradual improvements before the introduction of P4P and was not as a result of P4P. On the other hand, a retrospective cross sectional study by Ryan and Doran (2012), which assessed the impact of the QOF on hypertension management and outcomes concluded that the introduction of P4P improved treatment and management (Ryan and Doran, 2012, Simpson et al., 2011). Whilst there appeared to be an improvement in hypertension management and outcomes in both studies after the introduction of the QOF, the ITS study examined pre-intervention as well as post intervention trends (which were similar), which gave it advantage over the cross-sectional study. This

indicates that further exploration of the influence of different quasi-experimental evaluation designs on the results of evaluations of P4P would be useful in informing appropriate choices for quasi-experimental designs in P4P schemes in cases where RCTs are not feasible.

Regarding the P4P typology (chapter three), there was a trade-off between maximising its comprehensiveness and ensuring its manageability (ease of use). This led to exclusion of some potentially relevant design features from the typology (such as mechanism of payment: absolute or tiered targets and kind of incentive: monetary or non-monetary), and compression of some design variables in the typology (performance measure, timing of payment, and domain of performance measured were compressed to one category namely perceived risk of not earning the incentive 'risk'). Despite the exclusion of these design features from the P4P typology, it is important that evaluations of P4P report these features as they might potentially have an influence on the effectiveness of P4P. A second limitation was that the design features in the typology were assigned equal weights even though their relative importance is likely to vary (as demonstrated in chapter five). These findings were however not strong enough to confidently assign practical weights to the design features of the typology. Also, since the influence of design features on a scheme's effectiveness is likely to vary by the context in which the P4P is being implemented, it is probably better not to assign prior weights.

A third limitation was that the raters used in testing the P4P typology were very similar (consisted mostly of Health Science Masters and PhD students), and they tested the P4P typology only on completely and well-reported P4P evaluations. Ideally, it would have been better to have a broader range of raters using the typology on randomly selected P4P papers to reflect a real life scenario. Therefore, the choice of raters and studies could have overestimated the reliability and ease of use of the P4P typology in categorising P4P schemes in healthcare because the more similar the raters are, the more they are likely to agree. As a result, the typology is being presented as a first step. It will have to be tested further as more rigorously and well-reported evaluations of P4P design emerge.

With respect to the formative evaluation aspect of this thesis (exploring the influence of contexts and implementation on the Nigerian P4P scheme), it would have been interesting to assess the impact of the Nigerian P4P scheme on health outcomes and its value for money. However, this was beyond the scope of this thesis. Furthermore, there were no reliable outcome and cost data for comparison or evaluative purposes at the time the study was conducted. There is however, a planned evaluation (by the scheme implementers) to assess the impact of the scheme on health outcomes at the end of 2018 using an RCT design (NPHCDA, 2012).

10.4. Implications for policy and practice

As this thesis has demonstrated, the effectiveness of P4P requires consideration of several aspects, such as design choices, contexts, and implementation. Overall, expectations regarding the success of P4P schemes in healthcare should be moderate because effectiveness might have been overestimated in the literature. I highlight the implications of the findings of the thesis for designing, implementing and evaluating P4P in health care.

This thesis has shown that better designed P4P schemes are more likely to be effective using the P4P typology. The implication of this is that all P4P scheme designers and implementers can use the P4P typology as a guide to help think about and justify their design choices to enhance effectiveness. Although, undoubtedly, extensive research still needs to be done in this area (see next section), the findings of this thesis represents early steps toward understanding design choices to produce desired effects.

This thesis also demonstrates that evaluations of P4P schemes are poorly reported. The features of the P4P in the evaluation studies need to be uniformly and thoroughly reported to facilitate external scrutiny and to provide possibilities for policymakers in other settings to learn from the results. Evaluation studies of P4P should give adequate description of design features such as who receives the incentives (individuals or groups), type of incentive (fines or bonuses), size of incentive relative to clinicians' earnings (<5%: small, 5-10%: medium, >10% large), performance measure (absolute or relative), timing of payment (short or long time lag), domain of performance measured (structure, processes, outcomes), method of payment (coupled or uncoupled from usual

method of reimbursement), and mechanism of payment (absolute or tiered thresholds) . Once again, the P4P typology could serve as a standard reliable guideline that P4P evaluators can use to describe P4P designs in evaluation studies (see chapter 4 for reporting guideline).

In addition to adequate reporting in evaluation studies, it is important that these studies should also be explicit in reporting other quality improvement strategies that were implemented with P4P and possibly find ways to address or adjust for it, so that the effect of P4P can be clearer in such contexts. This is because in evaluated P4P studies in literature, authors were not clear enough in reporting whether or not there were other quality improvement strategies implemented with P4P and in cases where it was reported, not enough is done to address or adjust for it. This is likely to have introduced biased and confounded the effect of P4P thereby making it difficult to make strong conclusions on the effect of the scheme.

The findings of this thesis suggest the literature of P4P is populated with poorly evaluated P4P schemes, which likely overestimate the effectiveness of the schemes. P4P schemes need to be more rigorously assessed with robust evaluations (such as RCTs and robust quasi-experimental designs) in order to adequately capture the effects of P4P and improve the evidence base.

The variation in results of the preliminary evaluation of the Nigerian the P4P scheme provided useful insight of what was going on within the scheme. P4P literature may benefit from evaluations of large P4P schemes that report effects of P4P on indicators for each participating health service provider, instead of just the average effect of P4P. This would provide some understanding on variation of results within the scheme that goes beyond design features, which could guide exploratory research in this area.

P4P developers need to be sensitive to contextual and implementation factors. This is because whilst P4P in Nigeria was able to address some of the core issues (e.g. underfunding, transparency, and accountability) in the health centres where it was implemented; there were some poor implementation and contextual factors which limited its impact. Findings from my research showed that factors such as infrastructure,

understanding of the schemes, guarantee/certainty in earning the incentive, and level of managerial skills affected the effectiveness of a P4P case study in Nigeria. Throughout this thesis, I have been cautious about the generalizability of these particular findings to other contexts. Some of the findings have, however, been corroborated by other studies both in developed and developing countries: how diminished guarantee of payment of incentives reduces behaviour change of clinicians and effectives of P4P schemes in Australia and Uganda (Stockwell, 2010, Ssengooba et al., 2012); and how clinicians' weak understanding of the scheme reduced the effectiveness of the P4P in USA and Australia (Young et al., 2007, Stockwell, 2010). The implication of this is that similar countries implementing P4P can learn lessons from these findings. Also, it is important that formative evaluations should be carried out for each P4P scheme to take into account the contexts and practicalities of implementation to enhance effectiveness of the P4P in the specific setting. This will also contribute to the evidence base in this area.

Finally, it is important to note that while P4P schemes can produce modest improvements in quality of health care, it is not without the risk of negative consequences (Roland and Campbell, 2014). These include falsification of records and cherry picking (avoiding riskier patients), in addition to the limited evidence on cost effectiveness and sustainability (Gravelle et al., 2008, Jha et al., 2012, Emmert et al., 2012). The implication of this for policy makers especially in developing countries is that P4P should not be treated as a panacea to improve quality of care. Instead, I stand in agreement with Maynard (2012) that P4P should proceed with caution and invest in in robust evaluations that identify both the costs and benefits of change.

10.5. Recommendations for future research

This thesis provided useful directions and first steps in understanding how P4P designs features and evaluation designs influence the effectiveness of P4P schemes in general. This thesis also provides useful insights on the influence of contextual and implementation factors on P4P in Nigeria. However, a number of limitations (presented in section 10.3 and previous chapters) hindered strong conclusions. These form the basis of some of the recommendations for future research. There were three main

implications for research: designs of P4P, evaluations of P4P, and implementation of P4P, which could further contribute to the understanding in these areas.

There is the need for more quantitative and qualitative research specifically designed to explore influence of design features of the effectiveness P4P. This should include who to pay (groups or individuals), size of incentive, domain of performance (what to pay for: process, structures or outcomes), performance measure (absolute or relative measures), mechanism of payment (how to pay: payment on a sliding scale or absolute thresholds), and timing of payments. Particularly, it is important that the appropriate size of incentive (and how best to categorise this into whether small or large) should be explored, as this appeared to be the most important design feature form the findings but however, substantially lacking in empirical and theoretical evidence compared to the other design features (as demonstrated in chapter three and five).

In addition, the meta-regression analyses could be performed with a larger sample size/number of studies (increased power) in the future, where there is more time and resources to contact the authors of such studies for additional data that were originally not or too poorly reported.

As more rigorously evaluated P4P schemes emerge, it might be worthwhile to further explore the relationships between these P4P design features. For example, the findings in chapter 5 suggest that the size of incentive and the risk of not earning the incentive might be related (influencing each other). This will further contribute to the understanding of what combination of design features works in P4P. In addition, future quantitative exploration of heterogeneous results of P4P schemes should to take into consideration other study level characteristics such as risk of bias in order to produce higher a grade of evidence. This way, P4P in health care can be designed, evaluated, and reported with a bit more confidence.

Instrumental to gaining more useful insight for adequate P4P designs is using new/emerging empirical evidence to facilitate evolution of the P4P typology. For example, items of the typology which have been compressed, such as ‘risk of not earning the incentive’ can be unpacked to explore the individualistic and undiluted

effect of the design features (time lag of payment, domain of performance, and performance measure) that were compressed to form that category. Another potential area for future research is further experimentation of the relative weights of the design features included in the typology, which can be possible with the availability of more rigorously evaluated P4P schemes. This way, the typology can be used to predict the outcomes of categorised P4P schemes with a higher level of certainty in the future.

Second, as more evaluations of P4P emerge, it might be valuable to explore in detail effects of evaluation designs on the effectiveness of P4P schemes i.e. to unpack and compare how different quasi-experimental designs influence the evaluation results of P4P. The importance of evaluation design of P4P has been stressed throughout the thesis, and further insight is needed to inform the appropriate choice of quasi-experimental designs in P4P schemes in cases where RCTs are not possible.

Lastly, it is essential that new P4P schemes should first be pilot-tested before large-scale implementation. This will allow exploration of the adequate contextual and implementation variables required for effectiveness. It will also allow for experimenting with various design options (e.g. different incentive sizes) and non-financial incentives such as peer reporting which thus far has shown promising results (Kolstad, 2013, Bridgewater et al., 2007). Such pilots would allow necessary changes to be made and valuable lessons learned to be implemented to ensure maximum effectiveness. In addition, P4P evaluation should assess the long-term impact on health outcomes and costs, which thus far has largely been ignored in literature, as failure to invest in these evaluations might result in wastage of scarce resources especially in LMICs. This is important because in order to maximise the efficiency of resources, the increased costs associated with management and organisation of P4P (alongside the incentives need to be justified by the improvements in health outcomes.

10.6. Conclusion

The originality of my contributions, and the breadth and scope of my approach has made a significant contribution to knowledge in this field. This started with the literature review in which I problematized the fragmented evidence, which led me to develop a reliable theoretically informed framework (P4P typology) to categorise P4P schemes and aid evidence synthesis. This typology was then used in the first (and to

date, the only) quantitative systematic exploration of the influence of design features (and their relative importance) of effectiveness of P4P. My formative evaluation of P4P in Nigeria showed the influence of context and implementation on P4P and useful recommendations for were made for the scaling up of the scheme in Nigeria.

The findings of the thesis suggest that P4P schemes with particular design features are more likely to be effective. These include payment of incentives to groups rather than individuals; payment of large incentive (>5% of salary or budget) rather than small incentives (<5% of salary or budget); short time lag between measurement of performance and payment (<4months) rather than long time lags (<4months); processes/structures incentivised indicators (within clinicians' control) rather than clinical outcomes; and absolute rather than relative performance measure. In addition, effects of P4P schemes evaluated without adequate control groups are likely to over-estimate the effect of P4P. Finally, thesis presents evidence of the influence of contextual and implementation variables on the effectiveness of a P4P case study in Nigeria.

In conclusion, P4P in health care should be implemented with caution and careful consideration of design choices, in addition to reporting P4P schemes using a more structured approach in evaluations. The developed P4P typology could be used to help think about and report the design features of these schemes in a systematic way. In addition, the effectiveness and cost-effectiveness of P4P schemes need to be assessed using rigorous evaluation designs such as RCTs or well-conducted quasi-experimental designs to capture the true effect (and value for money) of P4P. Furthermore, valuable to informing implementation of large-scale P4P studies are preliminary qualitative studies to understand the influence of context and implementation on the effectiveness of P4P.

Appendices

Appendix A

A1. Search strategy output for Cochrane database

Database	**Cochrane**
Host	http://onlinelibrary.wiley.com/cochranelibrary/
Date of search	January 2012-June 2014 last date searched: 26/6/14
Years covered	1990-2014 no date restrictions
Search Strategy	Key word search: Financial incentives, Pay for performance, Performance based financing There are 20 results from 8524 records for your search on 'financial incentive or pay for performance or performance based financing in Title, Abstract, Keywords in Cochrane Reviews' There are 12 results from 30299 records for your search on 'financial incentive or pay for performance or performance based financing in Title, Abstract, Keywords in Other Reviews' There are 3 results from 16096 records for your search on 'financial incentive or pay for performance or performance based financing in Title, Abstract, Keywords in Economic Evaluations'
Language restrictions	None
Number of citations	35
Relevant reviews	8: Huang et al., 2013, Gillam et al., 2012, Reda et al., 2012, Chaix-couturier et al., 2012, Hamilton et al., 2013, Witter et al 2012, Scott et al 2011, Petersen et al 2006,

A2. Search strategy output for PubMed database

Database	**Medline**
Host	http://www.ncbi.nlm.nih.gov/sites/entrez (Pubmed)
Date of search	January 2012-June 2014 last date searched: 26/6/14
Years covered	1990-June 2014 (no date restrictions)
Search Strategy	1. Search **(((((((financial incentive*) OR performance based financing) OR pay for performance) OR paying for performance) OR incentive*) AND Review[ptyp] AND Humans[Mesh] AND English[lang])) AND health**
Language restrictions	None
Number of citations	1453
Relevant reviews	12: Van Herck P et al 2010, de Bruin SR, et al 2011, Witter et al 2012, Scott et al 2011, Petersen et al 2006, Eijkenaar 2012, Christianson et al 2008, Reda et al., 2012, Hamilton et al., 2013, Houle et al., 2012, Gillam et al., 2012, Andrew D Oxman and Atle Fretheim, 2009

A3. Summary of identified reviews

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
Oxman and Fretheim, 2009	The authors undertook a critical appraisal of selected evaluations of incentive (PBF) schemes in the health sector in low and middle-income countries (LMIC)	Key informants were interviewed to identify literature relevant to the use of PBF in the health sector in LMIC, key examples, evaluations, and other key informants. 13 studies were identified but only 4 met their inclusion criteria (which was not explicitly stated in the paper) and were included in the review: two single country cases and two multi-country studies	Quality of studies included in this review was not assessed.	The authors found very limited evidence of PBF having a positive impact and it was impossible to disentangle the effects of financial incentives as one element of PBF. They concluded that when PBF schemes are used, they should be designed carefully, including the level at which they are targeted, the choice of targets and indicators, the type, and magnitude of incentives. In addition, PBF schemes should be monitored for possible unintended effects and evaluated using rigorous study designs	**4/11**
Canavan et al., 2008	The authors explored incentive based approaches adopted in developing countries over the past decade	Search strategy was not described. 5 programs from 5 countries (Democratic Republic of Congo, Rwanda, Burundi, Haiti, Afghanistan), from 8 studies	Quality of included primary studies was not assessed.	The authors found that PBF results showed remarkable improvements in health indicators (utilization, coverage and emergency referral) with associated enhanced quality of health provider performance. They also noted the ambiguity among researchers regarding the extent of	**3/11**

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
				attribution of success, which calls for more rigorous evaluations of these programs.	
Chaix-couturier et al., 2000	The authors' objectives were to identify all the types of financial incentives that have been provided to health care professionals and, when possible, to assess the effects of these incentives on the costs, process or outcomes of health care.	6 databases were searched from January 1993 to May 1999 for English and French publications: MEDLINE, EMBASE, the Health Planning and Administration database, Pascal, International Pharmaceutical Abstracts, and the Cochrane Library. Additional papers were retrieved from the bibliographies of selected articles. It was stated that 89 papers were included in the review, whereas only 36 appeared to directly address the review question	The quality of each study was assessed according to the criteria described by the Cochrane Effective Practice and Organization of Care Group, but the results were not reported in the review.	The authors concluded that financial incentives could be used to reduce the use of health care resources, improve compliance with practice guidelines or achieve a general health target. It may be effective to use combinations of incentives, depending on the target set for a given health care programme. The authors however stated that few studies used the same methodology to assess the impact of the same incentive, thus limiting the external validity of their conclusions.	**6/11**
Christianson et al., 2008	This paper reviews evaluations of recent pay for- performance initiatives instituted by health plans or by provider organizations in cooperation with health plans.	The authors conducted electronic searches of MEDLINE, EMBASE, Cochrane Database of Systematic Reviews, Database of Reviews of Effects, Econlit, the Agency for Healthcare Research and Quality, the Organisation for Economic Co-operation and Development, and the	Quality of included primary studies was not assessed in a standardized way. The authors however stated that most of the studies included in this review were low quality studies (no	The review found that there were improvements in some quality measures, but it was not clear the degree of contribution of pay for performance to these improvements; the incentives typically were implemented in conjunction with other quality improvement efforts, or there was not a convincing comparison group.	**5/11**

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
		World Health Organization. Nine studies were included in this review	adequate control groups).		
de Bruin SR, et al., 2011	This review assessed the effectiveness of P4P schemes used to stimulate delivery of chronic care through disease management with regards to quality and costs.	Only one database was searched (PubMed). In addition to the electronic database search, relevant papers were identified through reference tracking and through a manual literature search on the internet from relevant websites, such as those of health insurers and Ministries of Health. Eight PBF schemes were identified 6 in the USA, 1 in Germany and 1 in Australia. Five of the P4P schemes were part of a larger scheme of interventions to improve quality of care, whereas the other three was implemented as 'standalone' schemes.	Primary studies were not assessed in a standardized way.	Most studies showed positive effects of P4P on healthcare quality. However, there was only one database was searched, and no attempt to identify unpublished literature, important studies that might have influenced the conclusion might have been missed. They authors also found variation in incented entities and the basis for providing incentives. Information about motivation, certainty, size, frequency, and duration of the financial incentives was generally limited.	**6/11**
Eijkenaar, 2012	This review systematically compared pay for performance initiatives in	The author searched Medline through PubMed and searched the Internet via Google and Google	Since this was not an impact evaluation review per se, and	The paper found variations in design and contextual factors between the identified programs. The author concluded that the	**6/11**

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
	the USA to other countries in terms of specific design choices that might contribute to success of PBF programs.	Scholar. The authors also consulted country-specific experts and searched reference list for relevant studies. The author identified 13 programs initiated in 9 countries. Seven programs were regional while six have been implemented nationally.	included studies were used to identify program descriptions, the quality of the studies was not assessed.	designs of these schemes are likely to affect the effectiveness of the schemes. However, the designs of these schemes are lacking in several respects and might be as a result of the limited knowledge about "what works" in P4P. This study has several limitations: some relevant programs were not identified as a result of English language restriction in the search strategy, the study suffers from publication bias as some studies were specifically not included because sufficient information was not found on the programs.	
Gillam et al., 2012	The authors review the growing evidence for the impact of the framework on the quality of primary medical care (QOF) in the United Kingdom.	The authors searched 3 databases: MEDLINE, EMBASE, and PsycINFO. They also searched the reference lists of published reviews and articles. Ninety-four studies were included in the review.	Quality of primary studies were assessed using a modified Downs and Black rating scale for observational studies and a Critical Appraisal Skills Programme rating scale for qualitative studies. The authors however did not report the	The authors found that: Quality of care for incentivized conditions during the first year of the framework improved at a faster rate than the pre-intervention trend and subsequently returned to prior rates of improvement. There were modest cost-effective reductions in mortality and hospital admissions in some domains. Achievement for conditions outside the	**9/11**

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
			quality assessment in this paper.	framework was lower initially and has worsened in relative terms since inception. The person-centeredness of consultations and continuity were negatively affected. Patients' satisfaction with continuity declined, with little change in other domains of patient experience. The conclusions of this study was limited by lack of adequate control groups	
Hamilton et al., 2013	The authors set out to evaluate the effectiveness of providing financial incentives to healthcare professionals for smoking cessation activities.	7 databases were searched till May 2011: MEDLINE, EMBASE, PsycINFO, Cochrane Database of Systematic Reviews, DARE, Cochrane Central Register of Controlled Trials (CENTRAL) and Web of Science. The authors also searched to GreyNet International and Open Grey for grey literature. Reference lists of retrieved articles and relevant reviews were also checked Eighteen studies were included in the review: three RCTs and 15 observational studies.	Primary study quality was assessed using the Downs and Black guidelines for randomised and non-randomised studies of healthcare interventions. Scores ranged from 1 (poor) to 4 (excellent). Included primary studies were considered to be mid-	The Authors found that financial incentives improved some process indicators such as recording smoking status, advice and referrals but not for outcome measures such as smoking quit rates. Studies of QOF program in the UK reported improvements in recording smoking status. One RCT also reported improvements in incentive clinics in the USA. Smoking advice or referral: QOF studies reported an increase in smoking advice.	9/11

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
			range for quality	The QOF studies should however be interpreted with caution because of the lack of adequate control groups Other studies reported mixed findings: two studies reported no differences for financial incentives and some studies reported improvements. Quit rates: Two studies reported no improvements in quit rates as a result of incentives and one study reported mixed effects for outcomes. The authors concluded that financial incentives appeared to improve recording of smoking status and increase provision of cessation advice and referrals to stop smoking services. There was however insufficient evidence to show that financial incentives led to reductions in smoking rates. Limitation: although this review is one of the well-conducted reviews, most data were retrieved from observational studies, which are prone to multiple biases. The authors	

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
				noted that most studies did not account for secular changes during study periods (such as new guidelines for smoking cessation or recent fiscal policy or legislation)	
Houle et al., 2012	This review assessed the effect of Pay-for-Performance remuneration, for individual health care practitioners, on the patient care outcomes.	PubMed, EMBASE, The Cochrane Library, OpenSIGLE, the Canadian Evaluation Society's; Unpublished Literature Bank, and the Grey Literature Collection of the New York Academy of Medicine's Library were searched up to June 2012. Reference lists were also manually searched. Thirty studies were included in the review. Four were RCTs, five were interrupted time series, three were controlled before-and-after studies, one was a non-randomized controlled study, 15 were uncontrolled before-and-after studies, and two were uncontrolled cohort studies.	The primary studies included were assessed, according to the Cochrane risk of bias scale, which included criteria for allocation concealment, similar baseline characteristics, complete outcome reporting, and protection against contamination. The quality of the studies was generally low to moderate; only RCTs had comparable baseline characteristics and only one study had	The authors, taking into consideration the limitations of the uncontrolled studies and the inability to draw reliable conclusions from them; concluded that Pay-for-Performance modestly improved preventive activities, such as immunization rates, but there was little evidence that it was effective for other activities such as mammography referrals and cancer screening.	**10/11**

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
			adequate patient allocation concealment (full results were reported).		
Huang et al., 2013	The authors' objectives were to review and synthesize published evidence of pay-for-performance (P4P) effects on management of diabetes.	Four databases were searched: Ovid MEDLINE, EMbase, PubMed, The Cochrane Library (Issue 3, 2012 12 interrupted time series studies, 7 controlled before-after studies, and 2 cross-sectional studies were included. Additionally, 12 studies were further included for quantitative analysis.	The quality of included primary studies was assessed using Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system. The authors reported that most studies included in the review were low quality studies.	Results of meta-analysis showed that P4P produced generally positive effects in most indicators (e.g. patients with records of total cholesterol or blood pressure). However, these results were inconsistent. The percentage of patients with HbA1c $\leq$ 7% or 53 mmol/mol showed a pooled odds ratio of 0.98 in patients, but a pooled mean difference of 19.71% in the physician groups. The odds ratios of receiving tests/reaching an outcome level were also diverse in patients (odds ratios ranged from 0.98 to 3.32). The authors also found that process indicators had higher rates of improvement than outcome indicators. Limitations: the authors concluded that because of the low quality of included studies, the results of the review should be interpreted with caution.	**8/11**
Petersen et al., 2006,	This review assessed the effects of explicit financial incentives for improving	The search was limited to studies written in English.	The studies were assessed according to a published	The authors found that of the 2 studies that evaluated financial incentives provided at the payment-system level, one found a	**7/11**

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
	performance on health care quality measures.	Seventeen studies were included in the review: 9 randomized controlled trials, 4 controlled trials with before-and-after data and 4 cross-sectional surveys.	methodological quality checklist (by Downs and Black) and graded on a scale of 1 (poor) to 4 (excellent). Six studies were assigned a quality grade of 3, six were assigned a grade of 2, and five were assigned a grade of 1.	positive effect on access to care while the other found a negative effect on access to care for the sickest patients. Of the 9 studies that evaluated the use of financial incentives directed to provider groups, two reported improvements for all quality of care measures, five were classified as partial improvement studies, and two showed no effect of the intervention compared with the control group. Of the 6 studies that evaluated the effects of financial incentives at the physician level, two reported a positive effect of the intervention and three reported some positive effects (partial studies). The authors concluded that incentives at the physician, provider group and payment-system levels have some positive effects, but further research is needed. This review was flawed because only one database was searched and the search was limited to English language papers, which suggests that relevant studies might have been	

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
				missed. Although an attempt was made to obtain unpublished data, publication bias was not assessed. Measures were taken to reduce the risk of bias in study selection.	
Reda et al., 2009	The primary objective of this review was to assess the impact of reducing the costs of providing or using smoking cessation treatment through healthcare financing interventions on abstinence from smoking.	The authors searched the Cochrane Tobacco Addiction Group Specialized Register in April 2012. Eleven studies were included. Of the eleven included studies, six randomly assigned the individual participants to the treatment group and one or two control groups (and three randomly assigned medical practices The two other studies were controlled natural experiments with two and four different benefit groups, respectively.	The quality of primary studies was assessed by The risk of bias of the included studies was assessed using criteria from the Cochrane Collaboration included in the Review Manager software. The Authors reported that most of the included studies had moderate to high risk of bias.	The authors found there was no evidence of an effect on smoking cessation from the results of pooling two trials of financial incentives directed at healthcare providers (RR 1.16, CI 0.98 to 1.37, $I^2 = 0\%$). Limitations: Only one database was searched and potentially important studies could have been missed. In addition, the two primary studies pooled together have relatively different incentive designs (heterogeneity) that were not accounted for.	10/11
Scott et al., 2011	This review assessed the effect of financial incentives on the quality of health care provided by primary care	The authors searched the Cochrane Effective Practice and Organisation of Care (EPOC) Trials Register, Cochrane Central Register of	Quality of included studies was assessed using the Epoc risk of bias guideline. The	Six of the seven studies included in this review showed positive but modest effects on a minority of the measures of quality of care included in the study.	**9/11**

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
	physicians.	Controlled Trials (CENTRAL) and Cochrane Database of Systematic Reviews (CDSR) (The Cochrane Library), MEDLINE, HealthSTAR, EMBASE, CINAHL, PsychLIT, and ECONLIT. Searches of Internet-based economics and health economics working paper collections were also conducted. Finally, studies were identified through the reference lists of retrieved articles, websites of key organisations, and from direct contact with key authors in the field. Articles were included if they were published from 2000 to August 2009. Seven studies were included in this review.	authors reported that there was high risk of bias (low quality) in most of the studies due to poor study designs	The authors concluded that there is insufficient evidence to support or not support the use of financial incentives to improve the quality of primary health care. Implementation should proceed with caution and incentive schemes should be more carefully designed before implementation. In addition to basing incentive design more on theory, there is a large literature discussing experiences with these schemes that can be used to draw out a number of lessons that can be learned and that could be used to influence or modify the design of incentive schemes.	
Van Herck P et al., 2010,	This review summarizes evidence, obtained from studies published between January 1990 and July 2009, concerning P4P effects, as well as evidence on the	The authors looked at papers from 1990- July 2009. They searched the following databases: Cochrane Library, EconLit, Embase, Medline, PsychINFO, and Web of Science. They also screened references,	The vast majority of identified studies was not randomized (only nine were) and roughly 75 studies were either cross-	The authors concluded that P4P programs result in the full spectrum of possible effects for specific targets, from absent or negligible to strongly beneficial and that the effects of P4P interventions varied according to design choices and	**11/11**

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
	impact of design choices and contextual mediators on these effects.	forward citation tracking, and expert consultation to identify studies. Studies that evaluated P4P effects in primary care or acute hospital care medicine were included. They included One hundred twenty-eight evaluation studies	sectional or employed a simple before-and-after design.	characteristics of the context in which it was introduced. This study was however limited because they excluded studies based on quality and this may have produced an overly restrictive analysis.	
Witter et al., 2012	This review assessed the current evidence on the effects of pay for performance on the provision of health care and health outcomes in low and middle-income countries. The studies assessed a mix of both patients' targeted incentives and incentives targeted at health care professionals.	Over 15 databases were searched till June 2011. This includes: the Cochrane Effective Practice and Organisation of Care Group Specialised Register, CENTRAL, MEDLINE, Ovid, EMBASE, EconLit, the Social Sciences Citation Index, ISI Web of Science. They also searched the websites and online resources of numerous international agencies, organisations and universities to find relevant grey literature and contacted experts in the field. Nine studies were included in the	The quality of included studies was assessed using the GRADE Working Group grades of evidence. The authors reported that almost all the studies identified had a high risk of bias. Sources of bias in the primary studies include non-random allocation of interventions, additional	The authors concluded that the evidence base was too weak to draw general conclusions due to validity issues. Only one study out of the nine studies was considered to have low risk of bias, one had a moderate risk of bias and the remaining seven had a high risk of bias. The high and moderate quality study found mixed results: some indicators improved while there was no improvement in others. Two of the studies showed significant improvement for the intervention group, while two showed no significant difference.	11/11

Reviews	Objectives	Search strategy and studies included	Quality of included studies and evaluation design	Results and limitations	Grade of evidence (Amstar score)
		review. There was one randomized trial; six controlled before-after studies and two interrupted time series studies.	funds/structures (other than the PBF schemes) that might have been responsible for the improvements seen, other confounders (e.g. contextual differences between intervention and non-intervention groups), and lack of rigorous evaluations.		

Break down of AMSTAR Checklist and scores

AMSTAR checklist questions	Oxman and Fretheim, 2009	Canavan et al., 2008	Chaix-couturier et al., 2000	Christinson et al., 2008	de Bruin SR, et al., 2011	Eijkenaar, 2012	Gillam et al., 2012	Hamilton et al., 2013	Houle et al., 2012	Huang et al., 2013	Petersen et al., 2006,	Reda et al. 2009	Scott et al., 2011	Van Herck P et al., 2010	Witter et al., 2012
Was an 'a priori' design provided?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Was there duplicate study selection and data extraction?	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Can't answer	Yes	Yes	Yes	Yes
Was a comprehensive literature search performed?	No	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Was the status of	Yes	Can't	No	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

AMSTAR checklist questions	Oxman and Fretheim, 2009	Cana van et al., 2008	Chaix-couturier et al., 2000	Christianson et al., 2008	de Bruin SR, et al., 2011	Eijkenaar, 2012	Gillam et al., 2012	Hamilton et al., 2013	Houle et al., 2012	Huang et al., 2013	Petersen et al., 2006,	Reda et al. 2009	Scott et al., 2011	Van Herck P et al., 2010	Witter et al., 2012
publication (i.e. grey literature) used as an inclusion criterion?		answer													
Was a list of studies (included and excluded) provided?	Yes	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Were the characteristics of the included studies provided?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Was the scientific quality of the included studies assessed and documented?	No	No	Yes	No	No	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes
Was the scientific quality of the included studies used appropriately in formulating conclusions?	Not applicable	Not applicable	No	Not applicable	No	Not applicable	Yes	Yes	Yes	Not applicable	Not applicable	Yes	Yes	Yes	Yes
Were the methods used to combine the findings of studies appropriate?	Not applicable	Not applicable	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Not applicable	Yes	Yes
Was the likelihood of publication bias assessed?	No	No	No	No	No	No	No	No	No	No	No	No	No	Yes	Yes
Was the conflict of interest included?	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Total AMSTAR score	4	3	6	5	6	6	9	9	10	8	7	10	9	11	11

A4. List of identified primary studies

1. An LC, Bluhm JH, Foldes SS, Alesci NL, Klatt CM, Center BA (2008). A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. Archives of Internal Medicine; 168(18):1993-1999.
2. Armour BS, Friedman C, Pitts MM, Wike J, Alley L, Etchason J (2004). The influence of year-end bonuses on colorectal cancer screening. Am J Managed Care; 10(9):617-624
3. Ashworth M, Lea R, Gray H, Rowlands G, Gravelle H, Majeed A (2004). How are primary care organizations using financial incentives to influence prescribing? Journal of Public Health; 26(1):48-51.
4. Basinga P, Gertler P, Binagwaho A, Soucat A, SturdyJ, Vermeersch C. (2011). Paying primary health facilities for performance in Rwanda. World Bank, Washington, DC, Policy research working paper 5190.
5. Beaulieu, N. D., & Horrigan, D. R. (2005). Organizational processes and quality. Putting smart money to work for quality improvement. *HSR: Health Services Research, 40,* 1318-1334.
6. Bardach, N. S., J. J. Wang, et al. (2013). "Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial." Jama 310(10): 1051-1059.
7. Bischoff, K., A. Goel, et al. (2013). "The Housestaff Incentive Program: improving the timeliness and quality of discharge summaries by engaging residents in quality improvement." BMJ Qual Saf 22(9): 768-774.
8. Boland, G. W., E. F. Halpern, et al. (2010). "Radiologist report turnaround time: impact of pay-for-performance measures." AJR Am J Roentgenol 195(3): 707-711.
9. Calikoglu, S., R. Murray, et al. (2012). "Hospital pay-for-performance programs in Maryland produced strong results, including reduced hospital-acquired conditions." Health Aff 31(12): 2649-2658.
10. Calvert M, Shankar A, McManus RJ, Lester H, Freemantle N. (2009). Effect of the quality and outcomes framework on diabetes care in the United Kingdom: retrospective cohort study. BMJ; 338:b1870.
11. Campbell S, Reeves D, Kontopantelis E, et al. (2007). Quality of primary care in England with the introduction of pay for performance. N Engl J Med ;357:181e90.
12. Campbell SM, Reeves D, Kontopantelis E, et al. (2009) Effects of pay for performance on the quality of primary care in England. N Engl J Med;361:368e78.
13. Canavan A, Swai G. (2008). Payment for Performance (P4P) Evaluation: Tanzania Country Report for Cordaid Godfrey Swai, National Consultant Width 1. KIT, Amsterdam.
14. Cattaneo A, Borgnolo G, Simon G. (2001). Breastfeeding by objectives. European Journal of Public Health ; 11(4):397-401.
15. Chang FC, Hu TW, Lin M, et al. (2008). Effects of financing smoking cessation outpatient services in Taiwan. Tob Control ;17:183e9.
16. Chee G, His N, Carlson K, Chankova S, Taylor P. (2007). Evaluation of the first five years of GAVI immunization services support funding. Prepared for the GAVI Alliance Behtesda, MD: Abt Associates Inc,.
17. Chen, J. Y., H. Tian, et al. (2010). "The effect of a PPO pay-for-performance program on patients with diabetes." Am J Manag Care 16(1): e11-19.
18. Chien, A. T., D. Eastman, et al. (2012). "Impact of a pay for performance program to improve diabetes care in the safety net." Prev Med 55 Suppl: S80-85.
19. Chien, A. T., Z. Li, et al. (2010). "Improving timely childhood immunizations through pay for performance in Medicaid-managed care." Health Serv Res 45(6 Pt 2): 1934-1947.
20. Clinical Practice Improvement Centre. (2010). Clinical practice improvement payment: User guide V2.0, pilot scheme—phase two. Brisbane, Australia: Queensland Health
21. Coleman T, Lewis S, Hubbard R, Smith C. (2007) Impact of contractual financial incentives on the ascertainment and management of smoking in primary care. Addiction; 102(5):803-808.
22. CORT, Vadodara. Report on assessment of ASHA/JSY scheme (Rajasthan) (2007). Commissioned by Ministry of Health & Family Welfare, Government of India. Draft.
23. Cupples ME, Byrne MC, Smith SM, et al. (2008). Secondary prevention of cardiovascular disease in different primary healthcare systems with and without pay-for-performance. Heart;94:1594e600.
24. Cutler TW, Palmieri J, Khalsa M, Stebbins M (2007). Evaluation of the relationship between a chronic disease care management program and California pay-for-performance diabetes care cholesterol measures in one medical group. Journal of Managed Care Pharmacy; 13(7):578-588.

25. Doran T, Kontopantelis E, Valderas JM, et al. (2011). Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. *BMJ*.;342:d3590.
26. Eichler R, Auxila P, Antoine U, Desmangles B. (2007). Performance-based incentives for health: six years of results from supply. side programs in Haiti. CGD Working Paper #121.Washington, DC: Center for Global Development
27. Fagan PJ, Schuster AB, Boyd C, Marsteller JA, Griswold M, Murphy SM, et al. (2010). Chronic care improvement in primary care: evaluation of an integrated pay-for-performance and practice-based care coordination program among elderly patients with diabetes. Health Serv Res.;45:1763-82.
28. Fairbrother G, Hanson KL, Friedman S, Butts GC. (1999). The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates. American Journal of Public Health; 89(2):171-175.
29. Fairbrother G, Siegel MJ, Friedman S, Kory PD, Butts GC. (2001). Impact of financial incentives on documented immunization rates in the inner city: results of a randomized controlled trial. Ambul Pediatr. ;1:206-12. [PMID: 11888402]
30. Felt-Lisk S, Gimm G, Peterson S. (2007). Making pay-for-performance work in Medicaid. Health Affairs; 26(4):W516-W527.
31. Friedman, N. L., Kokia, E., & Shemer, J. (2003). Health value added (HVA): Linking strategy, performance, and measurement in healthcare organizations. Israel Medical Association Journal, 5, 3-8.
32. Furth, R. (2006). Zambia Pilot Study of Performance-Based Incentives, USAID http://www.qaproject.org/news/PDFs/ZambiaPerformancePilotStudyInitiatives.pdf (visited June 2008)
33. Gavagan TF, Du H, Saver BG, Adams GJ, Graham DM, McCray R, et al. (2010). Effect of financial incentives on improvement in medical quality indicators for primary care. J Am Board Fam Med.;23:622-31. [PMID: 20823357]
34. Gilmore AS, Zhao YX, Kang N, Ryskina KL, Legorreta AP, Taira DA et al. (2007). Patient outcomes and evidence-based medicine in a preferred provider organization setting: A six-year evaluation of a physician pay-for-performance program. Health Services Research; 42(6):2140-2159.
35. Glickman SW, Ou FS, Delong ER, Roe MT, Lytle BL, Mulgund J et al. (2007). Pay for performance, quality of care, and outcomes in acute myocardial infarction. Jama-Journal of the American Medical Association; 297(21):2373-2380.
36. Grady KE, Lemkau JP, Lee NR, Caddell C. (1999). Enhancing mammography referral in primary care. Preventive Medicine; 26(6):791-800.
37. Greenberg MR, Weinstock M, Fenimore DG, Sierzega GM. (2008). Emergency department tobacco cessation program: staff participation and intervention success among patients. J Am Osteopath Assoc; 108(8):391-396.
38. Gross R, Elhaynay A, Friedman N, Buetow S. (2008). Pay-for-performance programs in P4P programs Israeli sick funds. J Health Organ Manag ; 22(1):23-35.
39. Grossbart, S. R. (2006). What's the return? Assessing the effect of "pay-for-performance" initiatives on the quality of care delivery. Medical Care Research and Review, *63*, 29S-48S.
40. Gulliford MC, Ashworth M, Robotham D, Mohiddin A. (2007). Achievement of metabolic targets for diabetes by English primary care practices under a new system of incentives. Diabetic Medicine 2007; 24(5):505-511.
41. Harries AD, Salaniponi FM, Nunn RR, Raviglione M. (2005). Performance-related allowances within the Malawi National Tuberculosis Control Programme. International Journal of Tuberculosis and Lung Disease; 9(2):138-144.
42. Hillman AL, Ripley K, Goldfarb N, Nuamah I, Weiner J, Lusk E. (1998). Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care. Am J Public Health;88:1699-701. [PMID: 9807540]
43. Hillman AL, Ripley K, Goldfarb N, Weiner J, Nuamah I, Lusk E. (1999). The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care. Pediatrics 1999; 104(4):931-935.
44. Hippisley-Cox, J., Vinogradova, Y., and Coupland, C. Final report for the Information Centre for Health and Social Care: time series analysis for 2001-2006 for selected clinical indicators from the QOF. http://www.qresearch.org/Public_Documents/Time%20Series%Analysis%20for%20selected%20clinical.pdf.
45. Jha, A. K., K. E. Joynt, et al. (2012). "The Long-Term Effect of Premier Pay for Performance on Patient Outcomes." New England Journal of Medicine 366(17): 1606-1615.

46. Kirschner, K., J. Braspenning, et al. (2013). "Assessment of a pay-for-performance program in primary care designed by target users." Fam Pract 30(2): 161-171.
47. Kontopantelis, E., D. Reeves, et al. (2012). "Recorded quality of primary care for patients with diabetes in England before and after the introduction of a financial incentive scheme: a longitudinal observational study." BMJ Quality & Safety.
48. Kruse, G. R., Y. Chang, et al. (2013). "Healthcare system effects of pay-for-performance for smoking status documentation." Am J Manag Care 19(7): 554-561.
49. Kouides RW, Bennett NM, Lewis B, Cappuccio JD, Barker WH, LaForce FM. (1998). Performance-based physician reimbursement and influenza immunization rates in the elderly. American Journal of Preventive Medicine; 14(2):89-95.
50. Kouides RW, Lewis B, Bennett NM, Bell KM, Barker WH, Black ER et al. (1993). A Performance-Based Incentive Program for Influenza Immunization in the Elderly. American Journal of Preventive Medicine; 9(4):250-255.
51. Kuo, R. N. C., Chung, K.-P., & Lai, M.-S. (2011). Effect of the pay-for-performance program for breast cancer care in Taiwan. American Journal of Managed Care, 17(5 Spec No.), e203-e211.Performance Management Program New Zealand 2006
52. Larsen DL, Cannon W, Towner S. (2003). Longitudinal assessment of a diabetes care management system in an integrated health network. J Manag Care Pharm; 9(6):552-558.
53. LeBaron CW, Mercer JT, Massoudi MS, Dini E, Stevenson J, Fischer WM et al. (1999). Changes in clinic vaccination coverage after institution of measurement and feedback in 4 states and 2 cities. Archives of Pediatrics & Adolescent Medicine; 153(8):879-886.
54. Lee, T.-T., Cheng, S.-H., Chen, C.-C., & Lai, M.-S. (2010). A pay-for-performance program for diabetes care in Taiwan: A preliminary assessment. American Journal of Managed Care, 16, 65-69.
55. Levin-Scherz J, DeVita N, Timbie J. Impact of pay-for-performance contracts and network registry on diabetes and asthma HEDIS (R) measures in an integrated delivery network. Medical Care Research and Review 2006; 63(1):14S-28S.
56. Li, J., Hurley, J., DeCicca, P., & Buckley, G. (2010). Physician response to pay-for-performance—Evidence from a natural experiment. Hamilton, Ontario, Canada: McMaster University.
57. Li, Y.-H., Tsai, W.-C., Khan, M., Yang, W.-T., Lee, T.-F., Wu, Y.-C., & Kung, P.-T. (2010).The effects of pay-for-performance on tuberculosis treatment in Taiwan. Health Policy and Planning, 25, 334-341.
58. Lindenauer PK, Remus D, Roman S, Rothberg MB, Benjamin EM, Ma A et al. (2007). Public reporting and pay for performance in hospital quality improvement. New England Journal of Medicine; 356(5):486-496.
59. Lynch M. (1995). Effect of Practice and Patient Population Characteristics on the Uptake of Childhood Immunizations. British Journal of General Practice; 45(393):205-208.
60. MacBride-Stewart SP, Elton R, Walley T. (2008). Do quality incentives change prescribing patterns in primary care? An observational study in Scotland. Family Practice; 25(1):27-32.
61. Magee GM, Hunter SJ, Cardwell CR, Savage G, Kee F, Murphy MC et al. (2010). Identifying additional patients with diabetic nephropathy using the UK primary care initiative. Diabet Med.; 27(12):1372-1378.
62. Mandel KE, Kotagal UR. (2007). Pay for performance alone cannot drive quality. Archives of Pediatrics & Adolescent Medicine; 161(7):650-655.
63. McGovern MP, Boroujerdi MA, Taylor MW, et al. (2008). The effect of the UK incentive-based contract on the management of patients with coronary heart disease in primary care. *Fam Pract*. ;25(1):33-39.
64. McMenamin SB, Schauffler HH, Shortell SM, et al. (2003). Support for smoking cessation interventions in physician organizations: results from a national study. Med Care;41:1396e406.
65. Millett C, Gray J, Saxena S, et al. (2003). Impact of a pay-for-performance incentive on support for smoking cessation and on smoking prevalence among people with diabetes. CMAJ;176:1705e10
66. Morrow RW, Gooding AD, Clark C. Improving physicians' preventive health care behavior through peer review and financial incentives. Arch Fam Med 1995; 4(2):165-169.
67. Norton EC. (1992). Incentive regulation of nursing homes. J Health Econ.;11: 105-28.
68. Oluwatowoju I, Abu E, Wild SH, Byrne CD. (2010). Improvements in glycaemic control and cholesterol concentrations associated with the Quality and Outcomes Framework: a regional 2-year audit of diabetes care in the UK. Diabet Med. ;27(3):354-359.
69. Peabody, J., R. Shimkhada, et al. (2011). "Financial incentives and measurement improved physicians' quality of care in the Philippines." Health Aff 30(4): 773-781.

70. Purdy S, Griffin T, Salisbury C, Sharp D. Emergency admissions for coronary heart disease: a cross-sectional study of general practice, population and hospital factors in England. Public Health. 2011;125(1):46-54.
71. Queensland Health. (2010). Clinical practice improvement payment. Retrieved from http://www.health.qld.gov.au/cpic/service_improve/cpip.asp
72. Rosenthal MB, Frank RG, et al. (2005). Early experience with payfor- performance: From concept to practice. JAMA; 294(14):1788–93.
73. Roski J, Jeddeloh R, An L, et al. (2003). The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines. Prev Med 2003;36:291e9.
74. Ryan AM. (2009). Effects of the Premier hospital quality incentive demonstration on Medicare patient mortality and cost. Health Services Research; 44(3):821-842.
75. Salize HJ, Merkel S, Reinhard I, et al. (2009). Cost-effective primary care-based strategies to improve smoking cessation: more value for money. Arch Intern Med;169:230e5
76. Schauffler HH, Brown C, Milstein A. (1999). Raising the bar: The use of performance guarantees by the Pacific Business Group on Health. Health Affairs; 18(2):134-142.
77. Scott A, Schurer S, Jensen PH, Sivey P (2009). The effects of an incentive program on quality of care in diabetes management. Health Economics , 18(9):1091-1108.
78. Serumaga B, Ross-Degnan D, Avery AJ, Elliott RA, Majumdar SR, Zhang F, et al. (2011). Effect of pay for performance on the management and outcomes of hypertension in the United Kingdom: interrupted time series study. BMJ. ; 342:d108. [PMID: 21266440]
79. Shen Y. (2003). Selection incentives in a performance-based contracting system. Health Serv Res.;38:535-52.
80. Simpson CR, Hannaford PC, Lefevre K, et al. (2006). Effect of the UK incentive-based contract on the management of patients with stroke in primary care. Stroke;37:2354e60.
81. Simpson CR, Hannaford PC, Ritchie LD, Sheikh A, Williams D. (2011). Impact of the pay-for-performance contract and the management of hypertension in Scottish primary care: a 6-year population-based repeated cross-sectional study. Br J Gen Pract.;61:e443-51. [PMID:21722469]
82. Simpson CR, Hippisley-Cox J, Sheikh A. (2010). Trends in the epidemiology of smoking recorded in UK general practice. Br J Gen Pract;60:e121e7.
83. Srirangalingam U, Sahathevan SK, Lasker SS, Chowdhury TA. (2006). Changing pattern of referral to a diabetes clinic following implementation of the new UK GP contract. British Journal of General Practice; 56(529):624-626.
84. Ssengooba, F., B. McPake, et al. (2012). "Why performance-based contracting failed in Uganda – An "open-box" evaluation of a complex health system intervention." Social Science & Medicine 75(2): 377-383.
85. St Jacques PJ, Patel N, Higgins MS. (2004). Improving anesthesiologist performance through profiling and incentives. J Clin Anesth 4;16:523-8. [PMID:15590256
86. Strong M, South G, Carlisle R. (2009). The UK Quality and Outcomes Framework pay-for-performance scheme and spirometry: rewarding quality or just quantity? A cross-sectional study in Rotherham, UK. BMC Health Serv Res; 9:108.
87. Sussman AJ, Fairchild DG, Coblyn J, Brennan TA.(2001). Primary care compensation at an academic medical center: A model for the mixed-payer environment. Academic Medicine; 76(7):693-699.
88. Sutton, M., S. Nikolova, et al. (2012). Reduced mortality with hospital pay for performance in England. N Engl J Med 367(19): 1821-1828.
89. Tahrani AA, McCarthy M, Godson J, et al. (2007). Diabetes care and the new GMS contract: the evidence for a whole county. Br J Gen Pract ;57:483e5.
90. Tsai, W.-C., Kung, P.-T., Khan, M., Campbell, C., Yang, W.-T., Lee, T.-F., & Li, Y.-H. (2010). Effects of pay-for-performance system on tuberculosis default cases control and treatment in Taiwan. Journal of Infection, 61, 235-243.
91. Twardella D, Brenner H. (2007). Effects of practitioner education, practitioner payment and reimbursement of patients' drug costs on smoking cessation in primary care: a cluster randomised trial. Tobacco Control; 16(1):15-21.
92. Vaghela P, Ashworth M, Schofield P, Gulliford MC. (2008). Population intermediate outcomes of diabetes under pay for performance incentives in England from 2004 to 2008. Diabetes Care.
93. Vergeer P, Chansa C. Payment for Performance (P4P) Evaluation: Zambia Country Report for Cordaid. KIT, Amsterdam.
94. Werner, R. M., R. T. Konetzka, et al. (2013). "The effect of pay-for-performance in nursing homes: evidence from state Medicaid programs." Health Serv Res 48(4): 1393-1414.

95. Yao H, Wei X, Liu J, Zhao J, Hu D, Walley JD. (2008). Evaluating the effects of providing financial incentives to tuberculosis patients and health providers in China. International Journal of Tuberculosis and Lung Disease; 12(10):1166-1172.
96. Young G, Meterko M, et al. (2007). Effects of paying physicians based on their relative performance for quality. Journal of General Internal Medicine ;22(6):872–6.

Qualitative studies

1. Aryankhesal, A., T. A. Sheldon, et al. (2013). "Role of pay-for-performance in a hospital performance measurement system: a multiple case study in Iran." Health Policy Plan 28(2): 206-214.
2. Coleman, T. (2010). "Do financial incentives for delivering health promotion counselling work? Analysis of smoking cessation activities stimulated by the quality and outcomes framework." BMC Public Health 10(167): 1471-2458.
3. Gerdes, N., Funke, U. N., Schuwer, U., Kunze, H., Walle, E., Kleinfeld, A., Jäckel, W. H. (2009). Ergebnisorientierte vergutung der rehabilitation nach schlaganfall. Entwicklungsschritte eines modellprojekts 2001-2008 [Pay for performance in rehabilitation after stroke–results of a pilot project 2001-2008]. Die Rehabilitation, 48, 190-201.
4. Gerdes, N., Funke, U., Schuwer, U., Themann, P., Kunze, H., Walle, E., von Ameln, M. (2008). Ergebnis-orientierte vergütung der rehabilitation nach schlaganfall. ergebnisse aus einem modellprojekt [Pay for performance in rehabilitation after stroke-findings from a pilot project]. Retrieved from http://forschung.deutsche-rentenversicherung.de/ForschPortalWeb/ressource?key=05_Gerdes.pdf found
5. Happell, B., C. Palmer, et al. (2010). "Mental Health Nurse Incentive Program: contributing to positive client outcomes." Int J Ment Health Nurs 19(5): 331-339.
6. Hochrhein-Institut. (n.d.). Ergebnisorientierte vergütung der rehabilitation nach schlaganfall (ERGOV) [Pay for performance in rehabilitation after stroke (ERGOV)]. Bad Säckingen, Germany: Hochrhein-Institut für Rehabilitationsforschung. Retrieved from http://www.hri.de/index.php?menuid=0&reporeid=59
7. Plever, S., I. McCarthy, et al. (2012). "Clinical Practice Improvement Payments: incentives for delivery of quality care." Australas Psychiatry 20(5): 407-412.
8. Schlingensiepen, I. (2009, July 23). Gute Reha-Ergebnisse, da gibt einen Bonus [Good outcomes for rehabilitation results in a bonus]. Ärzte Zeitung. Retrieved from http://www.aerztezeitung.de/politik_gesellschaft/krankenkassen/article/559098/gute-reha-ergebnisse-gibts-bonus.html
9. Steel N, Maisey S, Clark A, Fleetcroft R, Howe A. (2007). Quality of clinical primary care and targeted incentive payments: an observational study. Br J Gen Pract; 57(539):449-454.

A5. List of excluded studies

Reason for exclusion: Studies that did not evaluate the effects of P4P on healthcare quality/cost/performance/outcomes e.g. implementation studies or studies exploring the take up of p4p

1. Alshamsan R, Lee JT, Majeed A, Netuveli G, Millett C. Effect of a UK pay-for-performance program on ethnic disparities in diabetes outcomes: interrupted time series analysis. Ann Fam Med. 2012;10:228-34.
2. Anderson, K. K., Sebaldt, R. J., Lohfeld, L., Burgess, K., Donald, F. C., & Kaczorowski, J.(2006). Views of family physicians in southwestern Ontario on preventive care services and performance incentives. Family Practice, 23, 469-471
3. Ashworth M, Armstrong D, de Freitas J, Boullier G, Garforth J, Virji A. The relationship between income and performance indicators in general practice: a cross-sectional study. Health Serv Manage Res 2005; 18(4):258-264.
4. Ashworth M, Armstrong D. The relationship between general practice characteristics and quality of care: a national survey of quality indicators used in the UK Quality and Outcomes Framework, 2004-5. *BMC Fam Pract*. 2006;7:68.

5. Ashworth M, Golding S, Majeed A. Prescribing indicators and their use by primary care groups to influence prescribing. Journal of Clinical Pharmacy and Therapeutics 2002; 27(3):197-204.
6. Ashworth M, Golding S, Shephard L, Majeed A. Prescribing incentive schemes in two NHS regions: cross sectional survey. British Medical Journal 2002; 324(7347):1187-1188.
7. Ashworth M, Lloyd D, Smith RS, Wagner A, Rowlands G. Social deprivation and statin prescribing: a cross-sectional analysis using data from the new UK general practitioner 'Quality and Outcomes Framework'. Journal of Public Health 2007; 29(1):40-47.
8. Ashworth M, Medina J, Morgan M. Effect of social deprivation on blood pressure monitoring and control in England: a survey of data from the quality and outcomes framework. BMJ 2008; 337:a2030.
9. Ashworth M, Schofield P, Seed P, Durbaba S, Kordowicz M, Jones R. Identifying poorly performing general practices in England: a longitudinal study using data from the quality and outcomes framework. *J Health Serv Res Policy*. 2011;16(1):21-27.
10. Ashworth M, Seed P, Armstrong D, Durbaba S, Jones R. The relationship between social deprivation and the quality of primary care: a national survey using indicators from the UK Quality and Outcomes Framework. British Journal of General Practice 2007; 57(539):441-448.
11. Beith A, Eichler R, Weil D. Performance-based incentives for health: A way to improve tuberculosis detection and treatment completion? CGD Working Paper #122. Washington, DC: Center for Global Development, 2007.
12. Bennett NM, Lewis B, Doniger AS, Bell K, Kouides R, LaForce FM et al. A Coordinated, Community-Wide Program in Monroe County, New-York, to Increase Influenza Immunization Rates in the Elderly. Archives of Internal Medicine 1994; 154(15):1741-1745.
13. Bhattacharyya T, Mehta P, Freiberg AA. Hospital characteristics associated with success in a pay-for-performance program in orthopaedic surgery. Journal of Bone and Joint Surgery-American Volume 2008; 90A(6):1240-1243.
14. Bonis PA: Quality incentive payment systems: promise and problems. Journal of Clinical Gastroenterology 2005, 39(4 Suppl 2):S176-182.
15. Bottle A, Gnani S, Saxena S, Aylin P, Mainous AG III, Majeed A. Association between quality of primary care and hospitalization for coronary heart disease in England: national cross-sectional study. *J Gen Intern Med.* 2008;23(2):135-141.
16. Bottle A, Millett C, Xie Y, Saxena S, Wachter RM, Majeed A. Quality of primary care and hospital admissions for diabetes mellitus in England. *J Ambul Care Manage*. 2008;31(3):226-238.
17. Carey IM, DeWilde S, Harris T, Whincup PH, Cook DG. Spurious trends in coronary heart disease incidence: unintended consequences of the new GP contract? British Journal of General Practice 2007; 57(539):486-489.
18. Carey IM, Nightingale CM, Dewilde S, Harris T, Whincup PH, Cook DG. Blood pressure recording bias during a period when the Quality and Outcomes Framework was introduced. *J Hum Hypertens.* 2009;23(11):764-770.
19. Casale AS, Paulus RA, Selna MJ, Doll MC, Bothe AE, McKinley KE et al. "ProvenCare(SM)" a provider-driven pay-for-performance program for acute episodic cardiac surgical core. Annals of Surgery 2007; 246(4):613-623.
20. Casalino L, Gillies RR, Shortell SM, Schmittdiel JA, Bodenheimer T, Robinson JC et al. External incentives, information technology, and organized processes to improve health care quality for patients with chronic diseases. Jama-Journal of the American Medical Association 2003; 289(4):434-441.
21. Chang FC, Hu TW, Lo SY, et al. Quit smoking advice from health professionals in Taiwan: the role of funding policy and smoker socioeconomic status. Tob Control 2010;19:44e9.
22. Chang, H. J. (2004, June 6-8). Quality-based payment—Taiwan's experience. Paper presented at Academy Health Annual Research Meeting, San Diego, CA.
23. Chen, T. T., Chung, K. P., Lin, I. C., & Lai, M. (2011). Unintended consequence of diabetes P4P program in Taiwan: Are patients with more comorbidities or more severe conditions likely to be excluded from the P4P program? Health Services Research, 46, 47-60.
24. Cheng, T. M. (2006, June 25-27). P4P in Taiwan. Paper presented at Academy Health Annual Research Meeting, Seattle, WA.
25. Christensen DB, Neil N, Fassett WE, Smith DH, Holmes G, Stergachis A. Frequency and characteristics of cognitive services provided in response to a financial incentive. J Am Pharm Assoc (Wash). 2000;40:609-17. [PMID: 11029841]

26. Coleman K, Reiter KL, Fulwiler D. The impact of pay-for-performance on diabetes care in a large network of community health centers. Journal of Health Care for the Poor and Underserved 2007; 18(4):966-983.
27. Coleman T, Wynn AT, Barrett S, et al. Intervention study to evaluate pilot health promotion payment aimed at increasing general practitioners' antismoking advice to smokers. BMJ 2001;323:435e6.
28. Coleman T, Wynn AT, Stevenson K, Cheater F. Qualitative study of pilot payment aimed at increasing general practitioners' antismoking advice to smokers. British Medical Journal 2001; 323(7310):432-435.
29. Crawley D, Ng A, Mainous AG, III, Majeed A, Millett C. Impact of pay for performance on quality of chronic disease management by social class group in England. J R Soc Med 2009; 102(3):103-107.
30. Damberg CL, Raube K, Teleki SS, Dela CE. Taking stock of pay-for-performance: a candid assessment from the front lines. Health Aff (Millwood) 2009; 28(2):517-525.
31. de Brantes FS, D'Andrea BG. Physicians respond to pay-for-performance incentives: larger incentives yield greater participation. Am J Manag Care 2009; 15(5):305-310.
32. Doran T, Fullwood C, Gravelle H, et al. Pay-for-performance programs in family practices in the United Kingdom. *N Engl J Med.* 2006;355(4):375-384.
33. Doran T, Fullwood C, Kontopantelis E, Reeves D. Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework. *Lancet.* 2008;372(9640):728-736.
34. Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of patients from pay-for-performance targets by english physicians. New England Journal of Medicine 2008; 359(3):274-284.
35. Doran T. Lessons from early experience with pay for performance. Disease Management & Health Outcomes 2008; 16(2):69-77.
36. Doran, T., & Roland, M. (2010). Lessons from major initiatives to improve primary care in the United Kingdom. Health Affairs, 29, 1023-1029.
37. Downing A, Rudge G, Cheng Y, Tu YK, Keen J, Gilthorpe MS. Do the UK government's new Quality and Outcomes Framework (QOF) scores adequately measure primary care performance? A cross-sectional survey of routine healthcare data. Bmc Health Services Research 2007; 7.
38. Ettner SL, Thompson TJ, Stevens MR, Mangione CM, Kim C, Neil Steers W, et al; TRIAD Study Group. Are physician reimbursement strategies associated with processes of care and patient satisfaction for patients with diabetes in managed care? Health Serv Res. 2006;41:1221-41. [PMID: 16899004]
39. Fairbrother G, Friedman S, Hanson KL, Butts GC. Effect of the vaccines for children program on inner-city neighborhood physicians. Archives of Pediatrics & Adolescent Medicine 1997; 151(12):1229-1235.
40. Fairbrother G, Hanson KL, Butts GC, Friedman S. Comparison of preventive care in Medicaid managed care and medicaid fee for service in institutions and private practices. Ambulatory Pediatrics 2001; 1(6):294-301.
41. Feely J, Moriarty S, O'Connor P. Stimulating reporting of adverse drug reactions by using a fee. BMJ 1990; 300(6716):22-23.
42. Fleetcroft R, Parekh-Bhurke S, Howe A, Cookson R, Swift L, Steel N. The UK pay-for-performance programme in primary care: estimation of population mortality reduction. *Br J Gen Pract.* 2010;60(578):e345-e352.
43. Fleetcroft R, Steel N, Cookson R, Howe A. "Mind the gap!"Evaluation of the performance gap attributable to exception reporting and target thresholds in the new GMS contract: National database analysis. *BMC Health Serv Res*. 2008;8:131.
44. Francis DO, Beckman H, Chamberlain J, Partridge G, Greene RA. Introducing a multifaceted intervention to improve the management of otitis media: How do pediatricians, internists, and family physicians respond? American Journal of Medical Quality 2006; 21(2):134-143.
45. Gemmell I, Campbell S, Hann M, Sibbald B. Assessing workload in general practice in England before and after the introduction of the pay-for-performance contract. J Adv Nurs 2009; 65(3):509-515.
46. Gene-Badia J, Escaramis-Babiano G, Sans-Corrales M, Sampietro-Colom L, Aguado-Menguy F, Cabezas-Pena C et al. Impact of economic incentives on quality of professional life and on end-user satisfaction in primary care. Health Policy 2007; 80(1):2-10. Not an effectiveness evaluation
47. Gosden T, Sibbald B, et al.Paying doctors by salary: a controlled study of general practitioner behaviour in England. *Health Policy* 2003;**64**(3):415–23.

48. Gravelle H, Sutton M, Ma A. Doctor behaviour under a pay for performance contract: Further evidence from the quality and outcomes framework. 34, 1-31. 2008. The University of York, Centre for Health Economics. CHE Research Paper. Ref Type: Report
49. Gray J, Millett C, Saxena S, Netuveli G, Khunti K, Majeed A. Ethnicity and quality of diabetes care in a health system with universal coverage: Population-based cross-sectional survey in primary care. Journal of General Internal Medicine 2007; 22(9):1317-1320.
50. Guthrie, B., McLean, G., & Sutton, M. (2006). Workload and reward in the quality and outcomes framework of the 2004 general practice contract. British Journal of General Practice, 56, 836-841.
51. Halanych JH, Safford MM, Keys WC, Person SD, Shikany JM, Kim YI et al. Burden of comorbid medical conditions and quality of diabetes care. Diabetes Care 2007; 30(12):2999-3004.
52. Helm C, Holladay CL, Tortorella FR. The performance management system: Applying and evaluating a pay-for-performance initiative. Journal of Healthcare Management 2007; 52(1):49-62.
53. Heneghan C, Perera R, Mant D, Glasziou P. Hypertension guideline recommendations in general practice: awareness, agreement, adoption, and adherence. British Journal of General Practice 2007; 57(545):948-952.
54. Herrin J, Nicewander D, Ballard DJ. The effect of health care system administrator pay-for-performance on quality of care. Jt Comm J Qual Patient Saf 2008; 34(11):646-654.
55. Hippisley-Cox J, O'Hanlon S, Coupland C. Association of deprivation, ethnicity, and sex with quality indicators for diabetes: population based survey of 53 000 patients in primary care. British Medical Journal 2004; 329(7477):1267-1269.
56. Hughes E. Payment by results--a model for other diabetes healthcare systems? Prim Care Diabetes 2007; 1(2):111-113.
57. Incentives and rewards best practices primer: Lessons learned from early pilots. 2006. The Leapfrog Group. Ref Type: Report
58. Kantarevic, J., Kralj, B., & Weinkauf, D. (2010). Enhanced fee-for-service model and access to physician services: Evidence from family health groups in Ontario (IZA Discussion Paper No. 4862). Bonn, Germany: Institute for the Study of LaborJ.
59. Karve AM, Ou FS, Lytle BL, Peterson ED. Potential unintended financial consequences of pay-for-performance on the quality of care for minority patients. American Heart Journal 2008; 155(3):571-576.
60. Katz, A., Bogdanovic, B., & Soodeen, R. (2010). Physician integrated network baseline evaluation: Linking electronic medical records and administrative data. Winnipeg, Canada: Manitoba Centre for Health Policy
61. Keating NL, Landrum MB, Landon BE, Ayanian JZ, Borbas C, Robert WF et al. The influence of physicians' practice management strategies and financial arrangements on quality of care among patients with diabetes. Medical Care 2004; 42(9):829-839.
62. Khunti K, Gadsby R, Millett C, Majeed A, Davies M. Quality of diabetes care in the UK: comparison of published quality-of-care reports with results of the Quality and Outcomes Framework for Diabetes. *Diabet Med.* 2007;24(12):1436-1441.
63. Kralewski JE, Rich EC, Feldman R, Dowd BE, Bernhardt T, Johnson C,et al. The effects of medical group practice and physician payment methods on costs of care. Health Serv Res. 2000;35:591-613. [PMID: 10966087]
64. Kwaliteit Gezondheidszorg, UMC St Radboud. Ref Type: Report
65. Lester, H., Schmittdiel, J., Selby, J., Fireman, B., Campbell, S., Lee, J., Whippy, A. & Madvig, P. 2010. The impact of removing financial incentives from clinical quality indicators: longitudinal analysis of four Kaiser Permanente indicators. BMJ, 11.
66. Li R, Simon J, Bodenheimer T, Gillies RR, Casalino L, Schmittdiel J et al. Organizational factors affecting the adoption of diabetes care management process in physician organizations. Diabetes Care 2004; 27(10):2312-2316.
67. Liu X, Mills A. The influence of bonus payments to doctors on hospital revenue: results of a quasi-experimental study. *Applied Health Economics & Health Policy* 2003;**2**:91–8.
68. Looking at lessons on quality from the Medicare pay-for-performance hospital demonstration. Qual Lett Healthc Lead 2005; 17(10):2-13, 1.
69. Ludwig Boltzmann Institut fur Health Technology Assessment. (2009, December). Schweregraddifferenzierung in der neurologischen und trauma-rehabilitation: Internationale erfahrungen zur qualitats-, performancemessung und vergütung (No. 23b) [Classification of

disease severity for neuro- and trauma rehabilitation: International experiences with measurement and remuneration of quality and performance].
70. Maestad O. (2007), Rewarding Safe Motherhood: How can Performance-Based Funding Reduce Maternal and Newborn Mortality in Tanzania?, CMI report series; 2007/17, http://www.cmi.no/publications/file/?2916=rewarding-safe-motherhood
71. Majeed A, Williams J, De Lusignan S, Chan T. Management of heart failure in primary care after implementation of the National Service Framework for Coronary Heart Disease: a cross-sectional study. Public Health 2005; 119(2):105-111.
72. May EL. Take the lead or take your chances: engaging physicians in pay-for-performance. Healthc Exec 2005; 20(2):24-28.
73. McCarlie J, Reid E, Brady AJB. Audit of the new GMS contract Quality and Outcomes Framework: Raising standards in CHD. The British Journal of Cardiology 14, 117-120. 2007. Ref Type: Journal (Full)
74. McDonald, R., White, J., & Marmor, T. R. (2009). Paying for performance in primary medical care: Learning about and learning from "success" and "failure" in England and California. Journal of Health Politics, Policy and Law, 34, 747-776.
75. McElduff P, Lyratzopoulos G, Edwards R, Heller RF, Shekelle P,Roland M. Will changes in primary care improve health outcomes? Modelling the impact of financial incentives introduced to improve quality of care in the UK. *Qual Saf Health Care*. 2004;13(3):191-197.
76. McGovern MP, Williams DJ, Hannaford PC, et al. Introduction of a new incentive and target-based contract for family physicians in the UK: good for older patients with diabetes but less good for women? Diabet Med. 2008;25(9):1083-1089.
77. McLean G, Guthrie B, Sutton M. Differences in the quality of primary medical care for CVD and diabetes across the NHS: evidence from the quality and outcomes framework. Bmc Health Services Research 2007; 7.
78. McLean G, Sutton M, Guthrie B. Deprivation and quality of primary care services: evidence for persistence of the inverse care law from the UK Quality and Outcomes Framework. Journal of Epidemiology and Community Health 2006; 60(11):917-922.
79. McMenamin SB, Schmittdiel J, Halpin HA, Gillies R, Rundall TG, Shortell SA. Health promotion in physician organizations - Results from a national study. American Journal of Preventive Medicine 2004; 26(4):259-264.
80. Mehrotra A, Pearson SD, Coltin KL, Kleinman KP, Singer JA, Rabson B et al. The response of physician groups to P4P incentives. Am J Manag Care 2007; 13(5):249-255.
81. Menachemi N, Struchen-Shellhorn W, Brooks RG, Simpson L. Influence of pay-for-performance programs on information technology use among child health providers: the devil is in the details. Pediatrics 2009; 123 Suppl 2:S92-S96.
82. Millett C, Bottle A, Ng A, et al. Pay for perfomance and the quality of diabetes management in individuals with and without co-morbid medical conditions. J R Soc Med. 2009;102(9):369-377.
83. Millett C, Car J, Eldred D, Khunti K, Mainous AG, Majeed A. Diabetes prevalence, process of care and outcomes in relation to practice size, caseload and deprivation: national cross-sectional study in primary care. Journal of the Royal Society of Medicine 2007; 100(6):275-283.
84. Millett C, Gray J, Bottle A, Majeed A. Ethnic disparities in blood pressure management in patients with hypertension after the introduction of pay for performance. Ann Fam Med 2008; 6(6):490-496.
85. Millett C, Gray J, Wall M, et al. Ethnic disparities in coronary heart disease management and pay for performance in the UK. J Gen Intern Med 2008;24:8e13.
86. Millett C, Netuveli G, Saxena S, Majeed A. Impact of pay for performance on ethnic disparities in intermediate outcomes for diabetes: longitudinal study. Diabetes Care 2008.
87. Mullen KJ, Frank RG, Rosenthal MB. Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. 14886, 1-43. 2009. Cambridge, Massachusetts, National Bureau of Economic Research. NBER Working Paper Series
88. Murray J, Saxena S, Millett C, Curcin V, de Lusignan S, Majeed A. Reductions in risk factors for secondary prevention of coronary heart disease by ethnic group in south-west London: 10-year longitudinal study (1998-2007). Fam Pract. 2010;27(4):430-438.
89. O'Malley AS, Pham HH, Reschovsky JD. Predictors of the growing influence of clinical practice guidelines. Journal of General Internal Medicine 2007; 22(6):742-748.
90. Patel PH, Siemons D, Shields MC. Proven methods to achieve high payment for performance. J Med Pract Manage 2007; 23(1):5-11.

91. Pearson SD, Schneider EC, Kleinman KP, Coltin KL, Singer JA. The impact of pay-for-performance on health care quality in Massachusetts, 2001-2003. Health Affairs 2008; 27(4):1167-1176.
92. Petersen, L. A., Simpson, K., Pietz, K., Urech, T. H., Hysong, S. J., Profit, J., Conrad, D. A., Dudley, R. A. & Woodard, L. D. 2013. Effects of individual physician-level and practice-level financial incentives on hypertension care: a randomized trial. Jama, 310, 1042-50.
93. Pham HH, Landon BE, Reschovsky JD, Wu B, Schrag D. Rapidity and modality of imaging for acute low back pain in elderly patients. Arch Intern Med 2009; 169(10):972-981.
94. Pines JM. Profiles in patient safety: Antibiotic timing in pneumonia and pay-for-performance. Academic Emergency Medicine 2006; 13(7):787-790.
95. Pourat N, Rice T, Tai-Seale M, Bolan G, Nihalani J. Association between physician compensation methods and delivery of guideline-concordant STD care: Is there a link? Am J Managed Care 2005; 11(7):426-432.
96. Ramsay SE, Whincup PH, Lawlor DA, Papacosta O, Lennon LT, Thomas MC et al. Secondary prevention of coronary heart disease in older patients after the national service framework: population based study. British Medical Journal 2006; 332(7534):144-145.
97. Reid GS, Robertson AJ, Bissett C, Smith J, Waugh N, Halkerston R. Cervical Screening in Perth and Kinross Since Introduction of the New Contract. British Medical Journal 1991; 303(6800):447-450.
98. Reiter KL, Nahra TA, Alexander JA, Wheeler JR. Hospital responses to pay-for-performance incentives. Health Serv Manage Res 2006; 19(2):123-134.
99. Reschovsky JD, Hadley J, Landon BE. Effects of compensation methods and physician group structure on physicians' perceived incentives to alter services to patients. Health Services Research 2006; 41(4):1200-1220.
100. Ritchie LD, Bisset AF, Russell D, Leslie V, Thomson I. Primary and Preschool Immunization in Grampian - Progress and the 1990 Contract. British Medical Journal 1992; 304(6830):816-819.
101. Rittenhouse DR, Robinson JC. Improving quality in Medicaid - The use of care management processes for chronic illness and preventive care. Medical Care 2006; 44(1):47-54.
102. Rodriguez HP, Von Glahn T, Rogers WH, Safran DG. Organizational and market influences on physician performance on patient experience measures. Health Services Research 2009; 44(3):880-901.
103. Roland, M. (2006, September). Pay-for-performance: Too much of a good thing? A conversation with Martin Roland. Interview by Robert Galvin. Health Affairs, Web Exclusive, 25(5), w412-419.
104. Rosenthal MB, de Brantes FS, Sinaiko AD, Frankel M, Robbins RD, Young S. Bridges to Excellence - Recognizing High-Quality Care: Analysis of Physician Quality and Resource Use. Am J Managed Care 2008; 14(10):670-677.
105. Rosenthal MB, Fernandopulle R, Song HR, Landon B: Paying for quality: providers' incentives for quality improvement. Health Affairs (Millwood) 2004, 23(2):127-141.
106. Ryan, A. M. and J. Blustein (2011). "The effect of the MassHealth hospital pay-for-performance program on quality." Health Serv Res 46(3): 712-728.
107. Safran DG, Rogers WH, Tarlov AR, Inui T, Taira DA, Montgomery JE et al. Organizational and financial characteristics of health plans - Are they related to primary care performance? Archives of Internal Medicine 2000; 160(1):69-76.
108. Saunders M, Schattner P, Mathews M: Diabetes 'cycles of care' in general practice - do government incentives help? Australian Family Physician 2008, 37(9):781-784.
109. Saxena S, Car J, Eldred D, Soljak M, Majeed A. Practice size, caseload, deprivation and quality of care of patients with coronary heart disease, hypertension and stroke in primary care: national cross-sectional study. Bmc Health Services Research 2007; 7.
110. Schmittdiel J, McMenamin SB, Halpin HA, Gillies RR, Bodenheimer T, Shortell SM et al. The use of patient and physician reminders for preventive services: results from a National Study of Physician Organizations. Preventive Medicine 2004; 39(5):1000-1006.
111. Scott, I. A. (2007). Pay for performance in health care: Strategic issues for Australian experiments. Medical Journal of Australia, 187, 31-35.
112. Shenkman E, Tian LL, Nackashi J, Schatz D. Managed care organization characteristics and outpatient specialty care use among children with chronic illness. Pediatrics 2005; 115(6):1547-1554.
113. Shohet C, Yelloly J, Bingham P, Lyratzopoulos G. The association between the quality of epilepsy management in primary care, general practice population deprivation status and

epilepsy-related emergency hospitalisations. Seizure-European Journal of Epilepsy 2007; 16(4):351-355.
114. Shortell SM, Zazzali JL, Burns LR, Alexander JA, Gillies RR, Budetti PP et al. Implementing evidence-based medicine - The role of market pressures, compensation incentives, and culture in physician organizations. Medical Care 2001; 39(7):I62-I78.
115. Sigfrid LA, Turner C, Crook D, Ray S. Using the UK primary care Quality and Outcomes Framework to audit health care equity: preliminary data on diabetes management. Journal of Public Health 2006; 28(3):221-225.
116. Simon JS, Rundall TG, Shortell SM. Adoption of order entry with decision support for chronic care by physician organizations. Journal of the American Medical Informatics Association 2007; 14(4):432-439.
117. Simpson CR, Hannaford PC, McGovern M, Taylor MW, Green PN, Lefevre K et al. Are different groups of patients with stroke more likely to be excluded from the new UK general medical services contract? A cross-sectional retrospective analysis of a large primary care population. Bmc Family Practice 2007; 8.
118. Smith AL. Merging P4P and disease management: How do you know which one is working? Journal of Managed Care Pharmacy 2007; 13(2):S7-S10.
119. Smith CJ, Gribbin J, Challen KB, Hubbard RB. The impact of the 2004 NICE guideline and 2003 General Medical Services contract on COPD in primary care in the UK. QJM. 2008;101(2):145-153.
120. Soeters R., Habineza C., Peerenboom B.(2006), "Performance-based financing and changing the district health system: experience from Rwanda", Bulletin of the World Health Organization 2006;84:884-889.
121. Soeters, R. Musango, L.Meessen, B. (2005) Comparison of two output based schemes in Butare and Cyangugu provinces with two control provinces in Rwanda
122. Soeters, R. Perrot, J. Lozitto, A. (2006). Purchasing healthcare packages for the poor thorugh performance based contracting; which changes in the district health system does it require?
123. Soeters, R.. Griffiths, F. (2003). Improving government health services through contract management; a case from Cambodia. Health Policy and Planning. V 18; 74-83.
124. Sperl-Hillen JM, O'Connor PJ. Factors driving diabetes care improvement in a large medical group: Ten years of progress. Am J Managed Care 2005; 11(5):S177-S185.
125. Steel N, Bachmann M, Maisey S, Shekelle P, Breeze E, Marmot M et al. Self reported receipt of care consistent with 32 quality indicators: national population survey of adults aged 50 or more in England. British Medical Journal 2008; 337(7667).
126. Stevens VJ, Solberg LI, Quinn VP, et al. Relationship between tobacco control policies and the delivery of smoking cessation services in nonprofit HMOs. J Natl Cancer Inst Monogr 2005;35:75e80.
127. Strong M, Maheswaran R, Radford J. Socioeconomic deprivation, coronary heart disease prevalence and quality of care: a practice-level analysis in Rotherham using data from the new UK general practitioner Quality and Outcomes Framework. Journal of Public Health 2006; 28(1):39-42.
128. Sutton, M., R. Elder, et al. (2010). "Record rewards: the effects of targeted quality incentives on the recording of risk factors by primary care providers." Health Econ 19(1): 1-13.
129. Sutton M, McLean G. Determinants of primary medical care quality measured under the new UK contract: cross sectional study. British Medical Journal 2006; 332(7538):389-390.
130. Tahrani AA, McCarthy M, Godson J, et al. Impact of practice size on delivery of diabetes care before and after the Quality and Outcomes Framework implementation. Br J Gen Pract. 2008;58(553):576-579.
131. Trisolini M, Aggarwal J, Leung M, Pope GC, Kautter J: The Medicare Physician Group Practice Demonstration: lessons learned on improving quality and efficiency in healthcare. The Commonwealth Fund; 2008.
132. Trisolini M, Pope G, Kautter J, Aggarwal J. Medicare physician group practices: innovations in quality and efficiency. 971. 2006. The Commonwealth Fund. Ref Type: Report
133. Tsimtsiou Z, Ashworth M, Jones R. Variations in anxiolytic and hypnotic prescribing by GPs: a cross-sectional analysis using data from the UK Quality and Outcomes Framework. Br J Gen Pract. 2009;59(563):e191-e198.
134. Vamos EP, Pape UJ, Bottle A, Hamilton FL, Curcin V, Ng A, et al. Association of practice size and pay-for-performance incentives with the quality of diabetes management in primary care. CMAJ. 2011;183:E809-16. [PMID: 21810950]

135. Arpana R. Vidyarthi, MD, Adrienne L. Green, MD, Glenn Rosenbluth, and Robert B. Baron (2014), Engaging Residents and Fellows to Improve Institution-Wide Quality: The First Six Years of a Novel Financial Incentive Program, Acad Med, 89**,** 460-8.
136. Vina ER, Rhew DC, Weingarten SR, Weingarten JB, Chang JT. Relationship between organizational factors and performance among pay-for-performance hospitals. J Gen Intern Med 2009; 24(7):833-840.
137. Wang YY, O'Donnell CA, Mackay DF, Watt GCM. Practice size and quality attainment under the new GMS contract: a cross-sectional analysis. British Journal of General Practice 2006; 56(532):830-835.
138. Weber V, Bloom F, Pierdon S, Wood C. Employing the electronic health record to improve diabetes care: A multifaceted intervention in an integrated delivery system. Journal of General Internal Medicine 2008; 23(4):379-382.
139. Whalley D, Bojke C, Gravelle H, Sibbald B. GP job satisfaction in view of contract reform: a national survey. Br J Gen Pract 2006; 56(523):87-92.
140. Whalley D, Gravelle H, Sibbald B. Effect of the new contract on GPs' working lives and perceptions of quality of care: a longitudinal survey. Br J Gen Pract. 2008;58(546):8-14.
141. Wickizer TM, Franklin G, Gluck JV, Fulton-Kehoe D. Improving quality through identifying inappropriate care: THe use of guideline-based utilization review protocols in the Washington state workers' compensation system. Journal of Occupational and Environmental Medicine 2004; 46(3):198-204.
142. Wilkinson E, Randhawa G, Roderick P. Quality and Outcomes Framework (QOF) improves scope for effective management of Type Indo-Asian patients with diabetes who are ten years younger at diagnosis than White European patients. Diabet Med. 2010:27(S1):104.
143. Williams PH, de Lusignan S. Does a higher 'quality points' score mean better care in stroke? An audit of general practice medical records. Inform Prim Care. 2006;14(1):29-40.
144. Wilson, R. (2006, November). Primary care renewal in Ontario—Focus on remuneration. Paper presented at the College of Family Physicians of Canada, Primary Care Forum, Ontario, Canada.
145. Woodson SB. Making the connection between physician performance and pay. Healthc Financ Manage 1999; 53(2):39-42, 44.
146. Wright J, Martin D, Cockings S, Polack C. Overall Quality of Outcomes Framework scores lower in practices in deprived areas. British Journal of General Practice 2006; 56(525):277-279.

No access to full article

1. Amundson, G., Solberg, L. I., Reed, M., Martini, E. M., & Carlson, R. (2003). Paying for quality improvement: Compliance with tobacco cessation guidelines. *Joint Commission Journal on Quality & Safety*, *29*, 59-65.
2. Benavent J, Juan C, Clos J, Sequeira E, Gimferrer N, Vilaseca J. Using pay-for-performance to introduce changes in primary healthcare centres in Spain: first year results. Qual Prim Care 2009; 17(2):123-131.
3. Berenbeim D: The medical group pay-for-performance initiative in California and diabetes care. Managed Care Interface 2003, Suppl C:3-4.
4. Berthiaume, J. T., Chung, R. S., Ryskina, K. L., Walsh, J., & Legoratta, A. (2006). Aligning financial incentives with quality of care in the hospital setting. *Journal for Health Care Quality*, *28*, 36-50
5. Cameron PA, Kennedy MP, Mcneil JJ. The effects of bonus payments on emergency service performance in Victoria. Medical Journal of Australia 1999; 171(5):243-246.
6. Chung RS, Chernicoff HO, Nakao KA, Nickel RC, Legorreta AP. A quality-driven physician compensation model: four-year follow-up study. J Healthc Qual 2003; 25(6):31-37.
7. Fachklinik Herzogenaurach. (2010). Qualitätsbericht rehabilitation [Quality report rehabilitation] 2009. Hopfen am See, Germany: m&i-Klinikgruppe Enzensberg.
8. Hopkins JR. Financial incentives for ambulatory care performance improvement. Jt Comm J Qual Improv 1999; 25(5):223-238.
9. Pedros C, Vallano A, Cereza G, Mendoza-Aran G, Agusti A, Aguilera C et al. An intervention to improve spontaneous adverse drug reaction reporting by hospital physicians: a time series analysis in Spain. Drug Saf 2009; 32(1):77-83.
10. Price Waterhouse Coopers. (2001). Evaluation of primary care reform pilots in Ontario phase 2 interim report. Toronto, Ontario, Canada: Ontario Ministry of Health and Long-Term Care.

11. Rubinstein A, Rubinstein F, Botargues M, Barani M, Kopitowski K. A multimodal strategy based on pay-per-performance to improve quality of care of family practitioners in Argentina. J Ambul Care Manage 2009; 32(2):103-114.
12. Ting HH, Galvin RS, Krumholz HM, Petersen LA, Block PC. Do economic incentives improve quality of health care? Implications for pay-for-performance. ACC Cardiosource Review Journal 16[7], 22-25. 2007. Ref Type: Journal (Full)
13. Walle, E. (2009). ERGOV: Ergebnisorientierte Vergütung in der neurologischen Rehabilitation [Pay for performance in neurological rehabilitation]. Hopfen am See, Germany: m&iKlinikgruppe Enzensberg

Reason for exclusion: Descriptive studies

1. Ahmann AJ: Guidelines and performance measures for diabetes. merican Journal of Managed Care 2007, 13 Suppl 2:S41-46.
2. Aligning incentives to promote quality care: lessons from pay-for-performance initiatives. Qual Lett Healthc Lead 2005; 17(12):2-7, 1.
3. Anonymous: Aligning incentives in bridges to excellence. Managed Care Interface 2004, , Suppl: 5-6, 13.
4. Anonymous: CA health plans collaborate in 'pay for performance' program to boost chronic care management. Disease Management Advisor 2002, 8(4):54-58, 49.
5. Australian National Audit Office. (2010). Practice incentives program (Audit Report No. 5 2010-11). Canberra: Commonwealth of Australia.
6. BlueCross BlueShield Association: The Performance Based Incentive Program (PBIP): A Model for Quality Improvement and Cost.[http://www. bcbs.com/innovations/], Retrieved 28 February 2011 from.
7. Buetow, S. (2008). Pay-for-performance in New Zealand primary health care. Journal of Health Organization and Management, 22, 36-47.
8. Damberg CL, Raube K, Williams T, Shortell SM: Paying for performance: implementing a statewide project in California. Quality Management in Health Care 2005, 14(2):66-79.
9. District Health Boards New Zealand. PHO performance programme. Retrieved from www.dhbnz.org.nz/Site/SIG/pho/Default.aspx
10. Duckett, S., Daniels, S., Kamp, M., Stockwell, A., Walker, G., & Ward, M. (2008). Pay for performance in Australia: Queensland's new clinical practice improvement payment. Journal of Health Services Research & Policy, 13, 174-177.
11. Frohlich, N., Katz, A., De Coster, C., Dik, N., Soodeen, R.-A., Watson, D., & Bogdanovic, B. (2006, August). Profiling primary care physician practice in Manitoba. Winnipeg, Canada: Manitoba Centre for Health Policy.
12. Greb S, Focke A, Hessel F, Wasem J: Financial incentives for disease management programmes and integrated care in German social health insurance. Health Policy 2006, 78(2-3):295-305.
13. Greene RA, Beckman H, Chamberlain J, Partridge G, Miller M, Burden D et al. Increasing adherence to a community-based guideline for acute sinusitis through education, physician profiling, and financial incentives. Am J Managed Care 2004; 10(10):670-678.
14. Griffiths, F. Soeters, R. (2004) Improving workers performance through contract management, a case of Cambodia.
15. Health Care Incentives Improvement Institute: Bridges to Excellence.[http:// www.bridgestoexcellence.org/], Retrieved 28 February 2011
16. Health Net International: (2006) Guidelines for Implementing PBF in Afghanistan.
17. Integrated Healthcare Association: Integrated Healthcare Association (IHA), Pay for Performance (P4P) Program. Overview.[http://www.iha.org/ pdfs_documents/p4p_california/P4PFactSheet_July2010.pdf], Retrieved 28 February 2011 from.
18. Integrated Healthcare Association: Integrated Healthcare Association Pay for Performance (P4P) Program, Program Results.[http://www.iha.org/ program_results.html], Retrieved 28 February 2011 from.
19. Integrated Healthcare Association: Integrated Healthcare Association Pay for Performance (P4P) Program, Awards.[http://www.iha.org/p4p_awards. html], Retrieved 28 February 2011 from.
20. Jaiveer PK, Jaiveer S, Jujjavarapu SB, et al. Improvements in clinical diabetes care in the first year of the new General Medical Services contract in the UK. Brit J Diabetes Vasc Dis. 2006;6(1):45-48.

21. Kautter J, Pope GC, Trisolini M, Grund S. Medicare Physician Group Practice demonstration design: Quality and efficiency pay-for-performance. Health Care Financing Review 2007; 29(1):15-29.
22. Kirschner, K., Braspenning, J., Batenburg, J., van de Rijt, D., Muijrers, P., van Everdingen, C., Grol, R. (2008, March). Value for money: Een model voor honoreren van kwaliteit in de huisartsenpraktijk. project transparantie huisartsenzorg (fase 2) [Value for money: a pay-for-performance program in general practice. Project transparency in general practice (phase 2)]. Nijmegen, Netherlands: Scientific Institute for Quality of Healthcare.
23. Kirschner, K., Braspenning, J., Gootzen, T., van Everdingen, C., Batenburg, J., Verstappen, W., Grol, R. (2009, April). Pay-for-performance in de huisartsenpraktijk [Pay-for-performance in general practice – an experiment in the Southern part of the Netherlands]. Een experiment in Zuid-Nederland. Nijmegen, Netherlands: Scientific Institute for Quality of Healthcare
24. Manitoba Health. (2007). Physician integrated network: Evaluation plan. Winnipeg, Manitoba, Canada: Author.
25. Manitoba Health. (2010). PIN information management guide version 1.5. Winnipeg, Manitoba, Canada: Author.
26. Manitoba Health. (n.d.). Physician integrated network (PIN). Retrieved from http://www.gov.mb.ca/health/phc/pin/index.html
27. Medicare Australia. (2011). Practice incentives program. Retrieved from http://www.medicareaustralia.gov.au/provider/incentives/pip/index.jsp
28. Nalli GA, Scanlon DP, Libby D. Developing a performance-based incentive program for hospitals: a case study from Maine. Health Aff (Millwood) 2007; 26(3):817-824.
29. New Zealand Ministry of Health. (2010). Primary health care. Retrieved from http://www.health.govt.nz/our-work/primary-health-care
30. NHS North West. (2008, November). A North West health system approach to advancing quality. Manchester, England: NHS North West.(not an evalutation)
31. NHS North West. (n.d.). Advancing quality. Retrieved from http://www.advancingqualitynw.nhs.uk
32. Premier, Inc. (2010). NHS Northwest and premier advancing quality program. Composite quality score and outcome methodologies year one. Charlotte, NC: Author
33. Roland, M. (2004). Linking physicians' pay to the quality of care—A major experiment in the United Kingdom. New England Journal of Medicine, 351, 1448-1454.
34. Rosenthal MB, Camillus J. How four purchasers designed and implemented quality-based purchasing activities. 2007. Agency for Healthcare Research and Quality. Ref Type: Report
35. Sutton M, Elder R, Guthrie B, et al. Record rewards: the effects of targeted quality incentives on the recording of risk factors by primary care providers. Health Econ 2010;19:1e13. Not extracted yet
36. The Health and Social Care Information Centre. (2009). General and Personal Medical Services in England: 1998-2008. London, England: Author.
37. Ward, M., Daniels, S. A., Walker, G. J., & Duckett, S. (2007). Connecting funds with quality outcomes in health care: A blueprint for a clinical practice improvement payment. Australian Health Review, 31(Suppl. 1), S54-S58
38. West, D. (2008, November 27). Advancing quality in the North West. Health Service Journal. Retrieved from http://www.hsj.co.uk/advancing-quality-in-the-north-west/1931028.article

Reason for exclusion: Evaluations of P4P targeted at patients

1. Clark RE, Drake RE, McHugo GJ, Ackerson TH. Incentives for community treatment. Mental illness management services. Med Care. 1995;33:729-38. [PMID: 7596211]
2. Boyle RG, Solberg LI, Magnan S, Davidson G, Alesci NL.Does insurance coverage for drug therapy affect smoking cessation?. Health Affairs (Project Hope) 2002;21:162–8.
3. Dey P, Foy R, Woodman M, Fullard B, Gibbs A. Should smoking cessation cost a packet? A pilot randomized controlled trial of the cost-effectiveness of distributing nicotine therapy free of charge. British Journal of General Practice 1999;49:127–8.
4. Curry SJ, Grothaus LC, McAfee T, Pabiniak C. Use and cost effectiveness of smoking-cessation services under four insurance plans in a health maintenance organization. New England Journal of Medicine 1998;339(10):673–9.

5. Halpin HA, McMenamin SB, Rideout J, Boyce-Smith G. The costs and effectiveness of different benefit designs for treating tobacco dependence: results from a randomized trial. Inquiry 2006;43:54–65.
6. Healy J, Sharman E, Lokuge B. Australia: Health system review. Health Systems in
7. Hughes JR, Wadland WC, Fenwick JW, Lewis J, Bickel WK. Effect of cost on the self-administration and efficacy of nicotine gum: a preliminary study. Preventive Medicine 1991;20:486–96.
8. Kaper J, Wagena EJ, van Schayck CP, Severens JL.Encouraging smokers to quit: the cost effectiveness of reimbursing the costs of smoking cessation treatment. Pharmacoeconomics 2006;24:453–64.
9. Kaper J, Wagena EJ, Willemsen MC, van Schayck CP. A randomized controlled trial to assess the effects of reimbursing the costs of smoking cessation therapy on sustained abstinence. Addiction 2006;101:1656–61.
10. Kaper J, Wagena EJ, Willemsen MC, van Schayck CP. Reimbursement for smoking cessation treatment may double the abstinence rate: results of a randomized trial. Addiction 2005;100:1012–20.
11. Schauffler HH, McMenamin S, Olson K, Boyce Smith G, Rideout JA, Kamil J. Variations in treatment benefits influence smoking cessation: results of a randomised controlled trial. Tobacco Control 2001;10:175–80.
12. Bond L, Davie G, Carlin JB, Lester R, Nolan T. Increases in vaccination coverage for children in child care, 1997 to 2000: an evaluation of the impact of government incentives and initiatives. Australian and New Zealand Journal of Public Health 2002; 26(1):58-64.

Reason for exclusion: Poor /unclear reporting of outcomes

1. Balicer, R. D., Shadmi, E., Lieberman, N., Greenberg-Dotan, S., Goldfracht, M., Jana, L., Jacobson, O. (2011). Reducing health disparities: Strategy planning and implementation in Israel's largest health care organization. Health Services Research, 46, 1281-1299
2. Bishop, T. F., A. D. Federman, et al. (2012). "Association between physician quality improvement incentives and ambulatory quality measures." Am J Manag Care 18(4): e126-134.
3. Quy H, Lan N, Lonnroth K, Buu T, Dieu T, Hai T. Publicprivate mix for improved TB control in Ho Chi Minh City,Vietnam: an assessment of its impact on case detection. *International Journal for Tuberculosis and Lung Disease* 2003; 7:464–71.
4. Foels T, Hewner S. Integrating pay for performance with educational strategies to improve diabetes care. Popul Health Manag 2009; 12(3):121-129. Poor reporting: no information whatsoever on the incentive
5. Robinson JC, Casalino LP, Gillies RR, Rittenhouse DR, Shortell SS, Fernandes-Taylor S. Financial incentives, quality improvement programs, and the adoption of clinical information technology. Med Care 2009; 47(4):411-417.
6. Williams TR, Raube K, Damberg CL, Mardon RE. Pay for performance: its influence on the use of IT in physician organizations. J Med Pract Manage 2006; 21(5):301-306. Poor reporting
7. Chen, J. Y., H. Tian, et al. (2011). "Does pay for performance improve cardiovascular care in a "real-world" setting?" Am J Med Qual 26(5): 340-348.
8. Joseph, S. B., M. J. Sow, et al. (2013). "E-prescribing adoption and use increased substantially following the start of a federal incentive program." Health Aff 32(7): 1221-1227.
9. Kiran, T., J. C. Victor, et al. (2012). "The relationship between financial incentives and quality of diabetes care in Ontario, Canada." Diabetes Care 35(5): 1038-1046.
10. Pinnarelli, L., S. Nuti, et al. (2012). "What drives hospital performance? The impact of comparative outcome evaluation of patients admitted for hip fracture in two Italian regions." BMJ Qual Saf 21(2): 127-134.
11. Rodrigues, R., L. Trigg, et al. (2014). "The public gets what the public wants: Experiences of public reporting in long-term care in Europe." Health Policy 116(1): 84-94.
12. Sanada, H., G. Nakagami, et al. (2010). "Evaluating the effect of the new incentive system for high-risk pressure ulcer patients on wound healing and cost-effectiveness: a cohort study." Int J Nurs Stud 47(3): 279-286.

A6. Abstract of systematic review of economic evaluations of P4P

BACKGROUND: Pay-for-performance (P4P) intents to stimulate both more effective and more efficient health care delivery. To date, evidence on whether P4P itself is an efficient method has not been systematically analyzed.

OBJECTIVE: To identify and analyze the existing literature regarding economic evaluation of P4P.

DATA SOURCES: English, German, Spanish, and Turkish language literature were searched in the following databases: Business Source Complete, the Cochrane Library, Econlit, ISI web of knowledge, Medline (via PubMed), and PsycInfo (January 2000-April 2010).

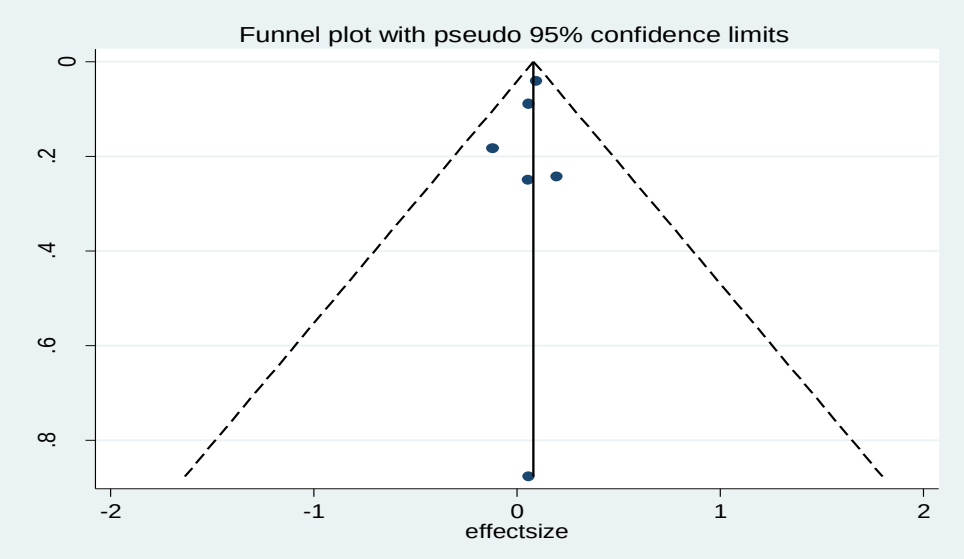
STUDY SELECTION: Articles published in peer-reviewed journals and describing economic evaluations of P4P initiatives. Full economic evaluations, considering costs and consequences of the P4P intervention simultaneously, were the prime focus. Additionally, comparative partial evaluations were included if costs were described and the study allows for an assessment of consequences. Both experimental and observational studies were considered.

RESULTS: In total, nine studies could be identified. Three studies could be regarded as full economic evaluations, and six studies were classified as partial economic evaluations. Based on the full economic evaluations, P4P efficiency could not be demonstrated. Partial economic evaluations showed mixed results, but several flaws limit their significance. Ranges of costs and consequences were typically narrow, and programs differed considerably in design. Methodological quality assessment showed scores between 32% and 65%.

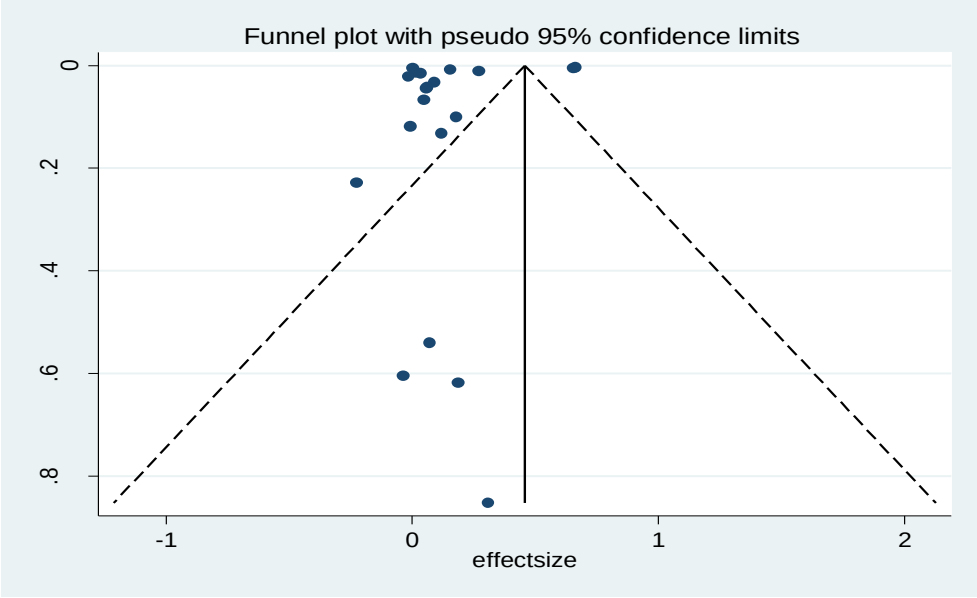
CONCLUSION: The results show that evidence on the efficiency of P4P is scarce and inconclusive. P4P efficiency could not be demonstrated. The small number and variability of included studies limit the strength of our conclusions. More research addressing P4P efficiency is needed.

Additional material such as grades of included studies and detailed literature searches is available at: http://static-content.springer.com/esm/art%3A10.1007%2Fs10198-011-0329-8/MediaObjects/10198_2011_329_MOESM1_ESM.pdf

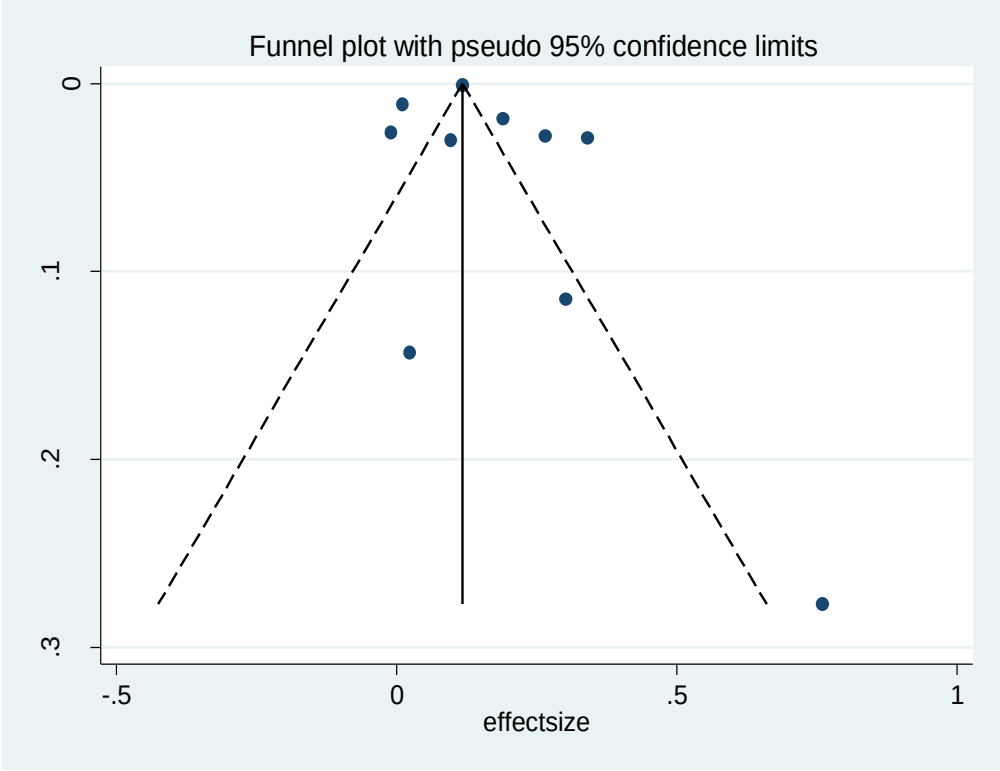
A7. Funnel plot for RCT evaluations of P4P



A8. Funnel plot for quasi-experimental evaluations of P4P



A9. Funnel plot for evaluations of P4P with no control group



Appendix B

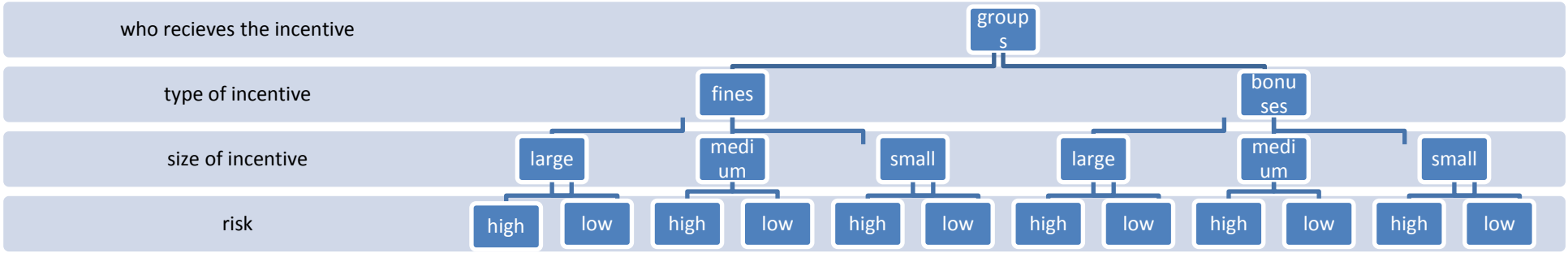
B1. Search strategy output for economic theories to inform the P4P typology

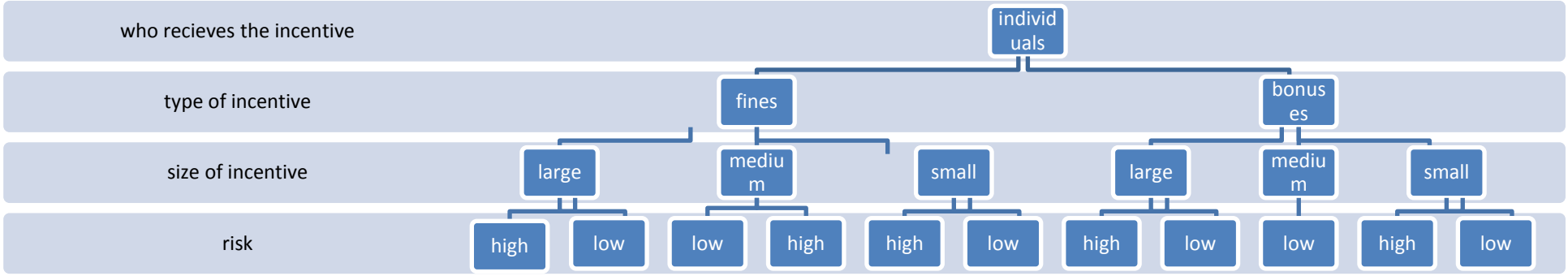
Database	**PubMed, PsycINFO, EconLit,**
Host	http://ovidsp.tx.ovid.com/sp-3.13.1a/ovidweb.cgi
Date of search	January 2012-June 2014 last date searched: 26/6/14
Years covered	1990-2014 no date restrictions
Search Strategy	**You searched:** ((behavioural economics or behavioural theories or incentive theories or economic theories) and incentive).mp. [mp=hw, ab, ti, ct, sh, tn, ot, dm, mf, dv, kw, nm, kf, px, rx, an, ui, tc, id, tm] - *Search terms used:* • behavioural • behavioural economics • behavioural theories • economic • economic theories • economics • incentive • incentive theories • theories
Language restrictions	None
Number of citations	170

B2. Preliminary criteria for identified variable to be potentially included in the typology

Core design features	Variables	Description
Who receives the incentives?	• Individuals • Groups	Individuals: incentive is paid to an individual health care provider e.g. physician Groups: incentive is paid to a group and individual clinicians might not benefit from the incentive directly e.g. hospital trust, clinical team, general physician (GP) practice, NGO, levels of government, faith based organizations
Type of incentive	• Bonus • Fines	Bonus: incentive is in the form of increase in payments, bonus, gifts, peer recognition etc. Fines: negative incentives in the form of reduction in expected payments, penalty, punishment etc.
Type of payment	• Monetary • Non-monetary	Monetary: incentive in form of money Non-monetary: incentives in the form of material things or tangible gifts
Size of incentive	• Large • Medium • Small	Amount or magnitude of monetary or non-monetary reward or fine. Large: >10% Medium: 5-10% Small <5% of salary, budget, or anticipated payment
Payment mechanism	• Absolute • Tiered thresholds	Absolute: incentives are paid as a single payment for an absolute increase in performance for example, an 80 % increase in performance. Tiered thresholds: incentives are paid for a series of target thresholds to meet for example paying increasing incentives for achieving a 65%, an 80%, and a 90% performance threshold.
Method of payment	• Coupled • Decoupled	Coupled: incentives paid are coupled with usual reimbursement e.g. an incentive in form of an increase in salary. Decoupled: incentives are paid separately from the usual reimbursement.
Performance measure/payment scale	• Absolute measure • Relative measure	Absolute: incentive is paid for improvement in performance or behaviour change not dependent on other providers e.g. incentive paid per patient immunized Relative: incentive is paid for attaining a specific rank relative to other providers e.g. incentives paid to clinicians or hospitals in top 2 performing quartiles
Domain of performance measured	• Within clinicians control • Out of clinicians control	Within clinicians control: incentive payments are based on process and structural outcomes e.g. number of children immunized, routine measurement of blood pressure of patients every month Out of clinicians control: payment of incentives to health providers for ultimate health outcomes e.g. reduction in mortality rates from a specific disease
Time lag	• Short • Long	Short time lag: Immediately after measurement of performance: payment of incentives four months or less. Long time lag: Not immediately after measurement of performance: Payment of incentives more than 4months after measurement of performance

B3. Constructing P4P typology





B4. Application of the typology on identified P4P scheme

Program	Perceived risk: high or low	Incentive size: small or large	Who receives the incentive: individuals or groups	Fines or bonuses	Type
Advancing Quality United kingdom 2008	High risk Annually (long time lag) Mostly within Physicians control (2 final outcomes and 26 processes) Relative measure	Small 2-4%	Group	Fines and bonuses	8
Clalit Israel 1998	Low risk Annually (long time lag) Mostly within Physicians control (10 processes and 8 intermediate outcomes) Absolute measure	Dependent on budget savings	Groups	Bonuses	
Clinical Practice Improvement Pay (CPIP) Australia, Queensland (started 2008)	Low risk Semi-annually (long time lag) Within physicians control (12 structures and 7 processes) Absolute measure	Large 8-10%	Group	Bonuses	2
MACCABI Israel 2001	High risk Annually (long time lag) Mostly within Physicians control (12 processes and 5 intermediate outcomes) Relative measure	Size not reported	Group	Bonus	
National Health Insurance P4P (NHI-P4P) Taiwan 2004	High risk Monthly and annually 12 structures, 3 final outcomes, and 2	Large Up to 20%	Individuals and groups	Bonuses	6

Program	Perceived risk: high or low	Incentive size: small or large	Who receives the incentive: individuals or groups	Fines or bonuses	Type
	intermediate outcomes Absolute and relative measures				
Primary care P4P (PC-P4P) Netherlands	High risk Annually (long time lag) Within physicians control (31 processes) Relative measures	Large 8-10%	Individual and groups	Bonuses	6
Primary Care Renewal Models (PCRM) Canada Ontario Started 2007	Low risk Annually Within physicians control (12 processes) Absolute measure	Small 2-4%	Individual and groups	Bonuses	4
Physician Integrated Network (PIN) Canada Manitoba 2004	Low risk Immediately after performance measure (short time lag) Within physicians control (only processes) Absolute	large	Groups	Bonuses	2
Practice Incentive Program (PIP) Australia 1998	Low risk Quarterly, semi-annually and annually , Within physicians control (only structures and processes) Absolute measure	Size not reported relative to income	Group	Bonuses	
Quality and Outcomes Framework (QOF)	Low risk Annually (long time lag) Mostly within physicians control (85% processes) Absolute measure	Large Up to 30-40%	Group	Bonuses	2
Western New York Physician Incentive Program (WNY-PIP) USA	Low risk Annually (long time lag) Mostly process: 6 Process and 3 outcomes Intermediate outcome	Size of varied from $3,000 till $12,000 large	Individuals	bonuses	10

Program	Perceived risk: high or low	Incentive size: small or large	Who receives the incentive: individuals or groups	Fines or bonuses	Type
	Absolute measure				
Kouides et al 1998 Rochester, New York, USA	Low risk Annually (long time lag) Process Absolute measure	Size 'Modest' for just one process? Small	Group	bonuses	4
Ashworth et al 2004 UK 2004	Low risk Annually (long time lag) Process/structure Absolute measure	up to £5000 per GP (large) Up to 5%	Groups but money trickled down to individuals	Bonuses	2
Cattaneo et al 2001 Italy 1998-1999	Low risk Yearly (long time lag) Process Absolute measure	Small 0.5% of annual revenue deducted	Groups	Fines	3
Fair brother et al 1999 New York 12 months	Low risk Annually (long time lag) Process Absolute measure	Bonuses: $1000 (20% improvement from baseline); $2500 (40% improvement); $5000 (80% up-to-date) Large	Individuals	Bonuses	8
Fairbrother et al 2001 USA 16 months	Low risk One off payment after 16 months (long time lag) Process	1000 (30% improvement from baseline); $2500 (45% improvement);	Individual	Bonuses	8

Program	Perceived risk: high or low	Incentive size: small or large	Who receives the incentive: individuals or groups	Fines or bonuses	Type
	Absolute measure	$5000 (80% up-to-date); $7500 (90% up-to-date)			
Grady et al 1997 USA	Low risk Quarterly payments (short time lag) Process Absolute Measure	Token Small? , i.e., $50 for a 50% referral rate. Small up to 1%	Groups	Bonuses	4
Hillman et al 1998	Low risk Every 6 months (long time lag) Process Absolute measure	Large Up to 20% of capitation fees	Individuals and groups	Bonuses	2
Larsen et al 2003	Low risk Time lag not reported Process Absolute measure	Small Size up to 1% of physicians compensation	Individuals	Bonuses	12
Rooski et al 2003 USA	Low risk 3 month time lag in payment Process Absolute measure	Size: up to $10,000 not reported relative to practice budget/income Most likely small.	Groups	Bonuses	4
Harries et al, 2005 Malawi National Tuberculosis Control Programme (four year program/0	Low risk 6month (short time lag) process absolute measure	Size: up to 100% of usual reimbursement	Individual physicians	Bonuses	8

Program	Perceived risk: high or low	Incentive size: small or large	Who receives the incentive: individuals or groups	Fines or bonuses	Type
Chien et al 2012 Hudson Health Plan's P4P program in New York USA	Low risk Both process and outcomes Yearly Absolute	300$ per patient a potential that is: most likely above 5% Second, the bonus amount was set well above typical levels and was substantial compared to office visit fees for a Medicaid population	Groups	Bonuses	2
Hillman et al., 1999 USA	Low risk Process Absolute and relative really Payment frequency: every 6 months	Bonuses based on total compliance score for quality indicators; full and partial bonuses Average bonus, $2,000 (range, $772 to $4682)	Payments to provider groups	Bonuses	4
Christensen et al., 2000 USA	Low risk Timing of payment not reported Process Absolute measure	$4 for cognitive services interventions (_6 min); $6 for _ 6 min; cognitive services	Provider group	Bonuses	
Hillman et al., 1998 USA	Low risk Payment frequency: every 6 months (long time lag) Process Absolute measure	Large Full and partial bonuses (20%; 10% of capitation); range of bonus per site,	Provider group	Bonuses	2

Program	Perceived risk: high or low	Incentive size: small or large	Who receives the incentive: individuals or groups	Fines or bonuses	Type
		$570 to $1260 Large: up to20%			
Thomas F. Gavagan, et al 2010 USA	Low risk Annually (long time lag) Processes Absolute Measure	Small The potential $4000 annual pay-out based on achieving quality targets represented approximately 3% to 4% of a provider's total salary	Individual physicians	Bonuses	12
An LC et al 2008 USA	Low risk Annual (long time lag) Process Absolute measure	5000$ onetime payment at the end of the programme	Groups	Bonuses	4
Glickman et al2007 USA CMS	High risk Yearly (long time lag) Process and outcomes Relative	Small 2%	Groups (hospitals)	Bonuses	7
Levin et al 2006 USA	Low risk Paid monthly (short time lag) Process Relative measure	Up to 20% of budget/salary	Groups	Bonuses	2
Mandel KE, Kotagal UR. 2007	Can't tell: not enough information reported Process	Large 7% fee schedule increase	Practices (groups)	Bonuses	

Program	Perceived risk: high or low	Incentive size: small or large	Who receives the incentive: individuals or groups	Fines or bonuses	Type
Cincinnati USA					
Lindenauer, et al 2007 CMS USA	High risk Annual (long time lag) Processes Relative measure	Up to 2% usual Medicare reimbursement	Group	Bonuses	6
Greenberg et al 2008	Low risk Payment every three months (short time lag) Process	Not enough informtion reported	Individuals	Bonuses	
Yao H et al 2008 China	Not enough information reported Process	$31 694 on incentives to village doctors for providing DOT, $16 011 on incentives for referring TB patients to county TB dispensaries and $15 992 for spreading TB knowledge in villages	Doctors Individuals	Bonuses	
Fagan et al, 2010	Low risk Timing of payment not reported Process and structures Absolute measure	Large Up 20%	Groups	Bonuses	2

Program	Perceived risk: high or low	Incentive size: small or large	Who receives the incentive: individuals or groups	Fines or bonuses	Type
Chien AT et al 2010 USA	Low risk Timing of payment not reported Process Absolute measure	Can't tell a U.S.$200 bonus payment for each fully immunized 2-year-old	Individuals	Bonuses	
Jha et al 2012 CMS	High risk Yearly (long time lag) Process and outcomes Relative measure	2%	Groups hospitals	Bonuses	7
Lynch 1995 UK	Low risk Annually Paid quarterly Absolute (tournament) it would between 70% and 89%; rates below 70% do not qualify for these payments.		Paid to GP practices Groups	Bonuses	
Sussman et al 2000 Boston, Massachusetts USA	Low risk Yearly (long time lag) Process Absolute measure	Large Size: up to 10% of salary	Individuals	Bonuses	10
Norton 1992	Can't tell Timing of payment not reported Outcomes Absolute measure	$126 to $370	Groups	Bonuses	
Shen, 2003 Maine, USA	Low risk Annual payment (long time lag) Process Absolute measure	Not enough reported about size	Groups	Bonuses	
Basinga et al, 2010	Low risk	Large 22-38% of usual budget and	Individuals and groups	Bonuses	2

Program	**Perceived risk: high or low**	**Incentive size: small or large**	**Who receives the incentive: individuals or groups**	**Fines or bonuses**	**Type**
Rwanda	Monthly and quarterly payments (short time lag) Processes Absolute measure	salary			
Canavan A. and Swai G. (2008) Tanzania	Low risk Payment every 6 months (long time lag) Processes Absolute measure	Large 5-10% of hospital budget and clinicians salary	Individuals and groups	Bonuses	2
Sulku, 2011 Turkey	Low risk Monthly payments (short time lag) Process and outcomes Absolute measure	Large Up to 80% of budget and salary	Individuals and groups	Bonuses	2
Vergeer and Chansa, 2008. Zambia	Low risk Absolute measure Quarterly payments (short time lag) Processes	Up to 100% of salary	Individuals and groups	Bonuses	2
Ssengooba, 2012. Uganda	Low risk 6monthly payment (long time lag) Process Absolute measure	Large up to 11% of hospital budget	Groups	Bonuses	2
Cutler USA (California P4P)	High risk Annual payments (long time lag) Processes and intermediate outcomes Relative measure	Large Up to 5% of budget	Groups	Bonuses	6
Gilmore et al Hawaii Medical Services Association	High risk Annual (long time lag) Relative Outcomes	Large Up to 7% of salary	Individuals	Bonuses	14

Program	Perceived risk: high or low	Incentive size: small or large	Who receives the incentive: individuals or groups	Fines or bonuses	Type
Young et al	High risk Annual (long time lag) Processes Relative measure	Large 5% of physician fees was at risk	Individuals	Bonuses	14
Twardella et al	High risk Annual (long time line) Outcome Absolute measure	Small	Individuals	Bonuses	
Li et al Ontario	Low risk Annual (long time lag) Processes Absolute measure	Large: up to 10% of physician revenue	Individuals	Bonuses	8
Kouides 1993	Low risk Annual payment (long time lag) Processes Absolute	Small	Individuals	Bonuses	12
(India) ASHA/JSY	Low risk Payment every three months Processes Absolute measure	$4.94 to $34.58 (small)	health professionals (ASHA's) (individuals)	Bonus	12
Haiti: RBF for NGO	Low risk Quarterly payments Processes Absolute measure	Up to 15% of previous budget of NGO (large)	NGOs: groups/institutions	Bonus	2
GAVI Incentives for national governments	Low risk Time lag not clear Processes Absolute measure	Up to 15% increased immunization funding (large)	National government: institutions/groups	Bonus	2

Appendix C

C1. Summary of existing methods (from studies consulted) for assessing inter-rater reliability for categorization tools in health care

I summarize the methods of investigating inter-rater reliability in the published studies by assessing the kind of tool or instrument been analysed, number of raters included in the study and rationale, how the raters were selected and why, number of subjects (sample size) that the tool is to be applied on, how the subjects were selected, what test statistic was used in computing the reliability, other statistical assumptions, and the software used in analysis.

The first study I assessed is by Lobbestael and colleagues (2011), which examined the inter-rater reliability of the structured clinical interview for the Diagnostic and Statistical Manual of Mental Disorders Axis, I (SCID I). This tool is known as the gold standard of semi-structured assessment instruments for clinical disorders and personality disorders. In their inter-rater reliability analysis, they used two set of raters selected randomly; the first rater consisted of 16 individuals while the second set of raters consisted of 16 individuals with varied levels of professional training. These raters were given the relevant training before using the tool. The raters were split into pairs and each pair assessed audiotaped interviews of 151 randomly selected participants from a large research project. There was no rationale given by the authors for the number of raters or subjects included in the study. To compute the agreement of diagnosis between raters, kappa coefficients were calculated and used to assess agreement of categorical judgements of the SCID 1 and agreement was interpreted according to Fleiss (1981) kappa values lower than 0.40 was interpreted as poor, between 0.41 and 0.75 as fair, and above 0.75 as excellent agreement. The authors did not state the software used in computing kappa or other statistical assumptions made.

The second study by Hartling and colleagues (2012) assessed the Inter-rater reliability of Quality Assessment Instruments (1) the Cochrane Risk of Bias (ROB) tool for randomized controlled trials (RCTs) and the Newcastle-Ottawa Scale (NOS) for cohort studies. The Cochrane ROB tool is used to assess the risk of bias in RCTs and the NOS tool is used to assess the risk of bias in cohort studies. In investigating the inter-rater reliability of the Cochrane ROB tool for RCTs, the authors randomly selected 154 RCTs (subjects) from 616 published trials previously examined for quality reporting. Twelve raters that had experience with Evidence Based Practice Centres (EPC) work were specifically selected. To assess inter-rater reliability, 124 RCTs from this sample between two raters using pairs of raters from 4 EPCs. The authors further assessed the reliability agreement s across the rater pairs using a subset of 30 trials rated by 9 raters split into pairs with one group having 3 raters. For the NOS tool for cohort studies, the authors identified completed meta-analyses of cohort studies through the EPC Programme and Medline. They considered a meta-analysis appropriate if it incorporated at least 10 studies, assessed a dichotomous outcome, and had substantial statistical heterogeneity. The final number of subjects was 131. In order to assess the inter-rater reliability the NOS for cohort studies, two raters each from 4 of the EPCs independently applied the NOS to 131 samples of cohort studies. The authors extracted data from each of the studies (trials and cohort) that might be related to judging the risk of bias which would help the raters use the tool. Inter-rater agreement was calculated for each domain and for overall quality assessment using weighted or unweighted Cohen's kappa statistics, as appropriate and the reliability agreement across the rater pairs was calculated using the Fleiss Kappa and interpretation of the Kappa statistics based was on suggestions from Landis and Koch (1977). No rationale was given for the number of raters used or the number of subjects used.

Another study by Oremus et al 2012 investigated inter-rater reliability between 5 pairs of inexperienced student raters for quality assessments using the Jadad Scale for randomised controlled trials and the Newcastle-Ottawa Scale (NOS) for observational studies. The raters were students taking McMaster Integrative Neuroscience Discovery and Study Programme courses. They received a training session on quality assessment and were randomly assigned into five pairs. Each of the students independently rated the quality of 13 to 20 articles and they number of articles they were given depended on the amount of time each rater could devote to the study. These articles were randomly distributed among the raters and were drawn from a pool of 78 papers examining cognitive impairment following electroconvulsive therapy to treat major depressive disorder. The investigators provided a standardised tabular spread sheet for student raters to use during quality assessment. Raters then independently rated their assigned articles to permit the

authors to examine inter-rater reliability. The authors measured inter-rater reliability for the Jadad and NOS questions using the kappa coefficient calculated for each question and the kappa values were interpreted as follows: >0.80 was very good, 0.61 to 0.80 was good, 0.41 to 0.60 was moderate, 0.21 to 0.40 was fair and <0.21 was poor based on Altman's (1991) suggestions. All statistics were computed using SAS V.9.2 (The SAS Institute) with level of significance <0.05.

The fourth study I reviewed by MacDermid et al (2005) investigated the inter-rater reliability for the AGREE instrument for evaluating Clinical Practice Guidelines (CPG) pertaining to medical care. Three pairs of raters were randomly selected from a pool of 69 physical therapists willing to participate in the research that had varying level of experience with research and educational training but all of them had at least one post-graduate degree. Each rater independently rated a set of 6 (CPGs) from a randomly selected from a pool of 55 CPGs. The pool of CPGs were identified through an inventory that was created by the study authors from a series of systematic searches that included electronic databases, websites, contact of professional associations and guideline developers and the CPGs included in this study were the ones published within the last five years. The reliability between appraisers was determined for each question and each domain of the AGREE instrument using the Kappa coefficient. An unweighted and quadratic weighted kappa was calculated to indicate the agreement within pairs of raters on whether a CPG was appropriate for clinical utilization. Kappa values above 0.75 were considered to represent good, 0.40–0.75 moderate and <0.40 poor reliability based on the suggestions by Fleiss (1986). The authors used the SPSS statistical software for Windows (Version 11.0; SPSS Inc., Chicago, Illinois) for all statistical analyses. P-values of 0.05 or less were considered significant.

The investigators in these studies (explored) did not explicitly give rationales for the number of raters or sample size although, for about three of the studies it appeared that the number of subjects assigned to the raters was a trade-off between the data/resources available and time that the raters could devote to the research. The raters were paired up randomly in all the studies but the total number of raters included in the studies ranged from 2 to about 20. Once again, the investigators gave no rationale for this. Although, it appeared that for some studies, it depended on the number of eligible people that are available, responded or willing to participate in the research and for some, the number of raters used also depended on if they were interested in some things other than just the inter-rater reliability coefficient such as the reliability of the agreement between the raters pairs. One of the studies assessed the reliability of agreement between rater pairs and the investigators computed the 'Fleiss Kappa' to investigate this.

C2. : Ethics approval for study of the inter-rater reliability of the P4P typology

Department of
Health Sciences

c/o Department of Philosophy
Heslington
York YO10 5DD

Telephone (01904) 433253
Fax (01904) 321383
E-mail smh12@york.ac.uk

Dr Stephen Holland

www.york.ac.uk/healthsciences

23 April 2014

Miss Y Ogundeji

University of York

Department of Health Sciences

Heslington

York

YO10 5DD

Dear Yewande

The reliability of a typology tool for categorising descriptions of performance based financing (PBF) schemes

Dear Yewande,

Thank you for resubmitting your study to me for approval by Chair's Action. I am writing to confirm that the study can now go ahead, but would point out two things:

1) Thank you for the telephone call confirming that there will be no transfer of data between yourself, your supervisor and Professor Bland (and that you and Professor Bland will analyse the data together on your password-protected university computer).

2) Regarding the following two parts of the consent form:

I understand that this research is not a formal part of my studies (I will receive compensation for time and effort upon completion on the ratings of 6 academic papers)
I understand that the research has no effect on my studies or assessments(non-completion has no effect on my studies or assessments)

I suggest that you combine the points about the research not being part of the participants' studies, and separate out the point about compensation:

I understand that this research is not a formal part of my studies, that the research has no effect on my studies or assessments, and that withdrawal from the study would have no effect on my studies or assessment
I understand that I will receive compensation for time and effort upon completion on the ratings of 6 academic papers

Good luck with the study and thank you, again, for your resubmission.

Yours sincerely

Stephen Holland

Chair: HSRGC

cc. *Prof Trevor Sheldon*

C3. Letter to potential raters participating in the inter-rater reliability study

Hi,

I am Yewande Ogundeji, a PhD student in the Department of Health Sciences. I am looking for health research volunteers to help test a tool I developed as part of my research on the impact of incentives on health service performance. I hope you can help with this.

The tool is a typology that categorizes incentive programs in health care. I need to test whether the tool is easy to use and reliable; in other words, whether different potential users will come to the same conclusion.

If you choose to participate, I will provide a short training (which might last up to one hour) and give you a comprehensive guideline for use of the tool.

You will then need to read 6 research papers and use the tool to categorize the type of incentive scheme described in the paper.

This is a desk-based exercise, which you can do in the comfort of your own home and in your own time. In addition, you get ample time (three weeks) to complete the task.

As a thank you for your time, you will be given a token of a £10 iTunes/Amazon gift voucher when you complete the task.

I would be grateful if you seriously consider participating in this.

If you are interested in participating and/or would like to know more, please contact me on yo508@york.ac.uk

Many thanks,

Yewande

C4. Participant Information Sheet for the P4P typology inter-rater reliability study

The reliability of a typology tool for categorising descriptions of performance based financing (PBF) schemes

Participant Information Sheet

Title of Study: The reliability of a typology tool for categorising descriptions of performance based financing (PBF) schemes.

I would like to invite you to take part in the above named study but before you decide, please read the following information.

What is the purpose of this study?

The use of incentives in health care has increasingly been adopted in many countries to improve the quality and efficiency of health care across different contexts and different clinical areas (Eldridge and Palmer, 2009).

The evidence base on what works, what does not and why is mixed. Schemes have different designs, contexts, and implementation factors which makes it difficult to synthesise the evidence.

To help categorize these different schemes, I developed and piloted a typology. This typology could help group the different kinds of incentive schemes based on design features, which could make interpretation of evidence easier. The typology was constructed using a range of theory (from behavioural science and economics), concepts, and empirical evidence on how people respond to incentives.

The typology consists of 4 items relevant to design features of PBF resulting16-type typology (the four items are: who receives the incentives, the type of incentives, the size of incentive, and the perceived risk of earning the incentive).

I now wish to empirically test the inter-rater reliability of the instrument to see if different raters produce consistent results across conditions which it is likely to be used (PBF evaluations reported in academic literature) (Atkinson and Nevill, 1998, Streiner and Norman, 1989). .

Who is doing the study?

I, Yewande Ogundeji will be the principal investigator in this study, which is being conducted as a part of my PhD thesis supervised by Trevor Sheldon (Professor of Health Services Research and Policy at York). Data from the study will be analysed by Martin Bland (Professor of Health Statistics, Department of Health Sciences-University of York) and me.

Who is being asked to participate?

The participants being recruited for this study are graduate research students from the University of York-Health Sciences, used as a convenience sample.

This population consists of participants with diverse qualifications, backgrounds, and research experience, all of which are important in assessing inter-rater reliability. This is because the versatility of the participants reflects (as close as possible) a real life scenario of potential users of the tools.

Do I have to take part?

Taking part in this research is entirely voluntary. Whether or not you decide to take part will not have any effect on your studies or assessments.

If you choose to participate, you will receive a token (10 GBP gift voucher) of gratitude upon completion on the ratings of 6 academic papers.

If you decide to participate, you will be asked to sign a consent form to show that your participation is voluntary and has no effect on your studies or assessment i.e. it is not a formal part of your studies.

What will be involved if I take part in this study?

If you decide to participate in this study, after you have signed a consent form, you will be trained on how to use the typology using the training manual developed. The training manual is based on the guideline developed for use of the typology. The estimated time for training session is between 45 minutes to 1 hour. The training session will be instructive and comprehensive with examples and relevant problem solving exercises. It will also be an interactive training session and there will be opportunities for questions and feedback.

After the training, the volunteer raters/participants will be given 6 academic papers which describe PBF schemes from different countries and asked to apply the typology to each scheme to judge what category it falls under. This task is a desk-based exercise that can be done at any suitable location convenient for the participant (you will be given three weeks to finish the task and you will also be asked to report the estimated time taken to apply the typology to each study).

The estimated time to rate one academic paper is 30-45 minutes.

What are the advantages/benefits and disadvantages/risks of taking part?

Some of the advantages of taking part in this study includes: research experience, some knowledge of PBF schemes. In addition, upon completion of the task, £10 gift voucher as a thank you for your time. There are no risks involved in taking part in this research.

Can I withdraw from the study at any time?

If you decide to participate, you can still withdraw from the study at any stage of the research (before training, after training, after you have received the studies to be rated). In addition, you do not have to give a reason for withdrawing.

Will the information obtained in the study be confidential? ***or*** **Will the information I give be kept confidential?**

I, the principal investigator in this study will be in control of and act as custodian for data generated/collected by the study.

I will collect information such as: qualifications, years of research experience, and background. However, no identifiable information such as names of raters will be collected; therefore, all ratings are completely anonymized and no ratings can/will be traced down to any individual rater (participant)

Data will be stored in their original form (categorized studies by the raters, qualifications, years of research experience, and background) for up to 5 years on password protected University computers.

Data generated by this study will be analyzed on a university desktop by both Professor Martin bland and me.

What will happen to the results of the study?

Results of the research will be made available through peer reviewed journals, conferences, and my PhD thesis

Who has reviewed this study?

Research Governance Committee- Department of Health Sciences, University of York

Who do I contact in the event of a complaint?

Trevor Sheldon (supervisor): trevor.sheldon@york.ac.uk

If you agree to take part, would like more information or have any questions or concerns about the study please contact Yewande Kofoworola Ogundeji (PhD Student): yo508@york.ac.uk

C5. Participant consent form for the P4P typology inter-rater reliability study

THE UNIVERSITY of York

The Department of Health Sciences

Title of Study: The reliability of a typology tool for categorizing descriptions of performance based financing (PBF) schemes

	Please confirm agreement to the statements by putting your initials in the boxes below
I have read the invitation email and understand it	
I have had the opportunity to ask questions and discuss this study	
I have received satisfactory answers to all of my questions	
I have received enough information about the study	
I understand what I am required to do in the study	
I understand my participation in the study is voluntary and that I am free to withdraw from the research 1 At any time 2 Without having to give a reason for withdrawing	
I understand that I will receive compensation for time and effort upon completion on the ratings of 6 academic papers	
I understand that this research is not a formal part of my studies, that the research has no effect on my studies or assessments, and that withdrawal from the study would have no effect on my studies or assessment	
I understand that any personal information I provide for the purpose of this research will be kept confidential, stored securely and only accessed by those carrying out the study.	
I agree to take part in this study	
Participant Signature …… Date	
Name of Participant	
Researcher Signature …….. Date	
Name of Researcher	

C6. : Initial guideline for use of the P4P typology

Size **Was the size of the incentive small or large?**	
Criteria for judging **Small**	If the incentive in the PBF program is smaller than 5% of any one of the following: • Salary of individual clinician/health worker/doctor • Anticipated payments (to the health facility/hospital/clinical team) such as: budgets (total budget or budget for the particular intervention in question), fee for service (FFS) and capitation
Criteria for judging **Large**	If the incentive in the PBF program is 5% and above of any one of the following: • Salary of individual clinician/health worker/doctor • Anticipated payments (to the health facility/hospital/clinical team) such as: budgets (total budget or budget for the particular intervention in question), fee for service (FFS) and capitation
Who received the incentive? **Did Individuals or Groups receive the incentive?**	
Criteria for judging **Individuals**	• If the incentives are paid directly to individual health workers/clinicians/doctors • If individual health worker/clinician/doctor's income is supplemented as a result of the incentive (e.g. reflected in the rise of personal income)
Criteria for judging **Groups (including schemes where individuals and groups are paid bonuses)**	If the incentive is paid to a group in which individual clinicians may or may not benefit from the incentive directly Groups include any of the following • Hospital trust • Clinical team • General physician (GP) practice • NGO • Levels of government • Faith based organizations
Type of incentive **Was the incentive in the form of Fines or Bonuses?**	
Criteria for judging **Fines**	If the incentive is negative in the form of reduction in expected payments, penalty, punishment etc.
Criteria for judging **Bonuses**	Bonus: incentive is in the form of increase in payments, bonus, gifts etc.
Performance measure (payment scale) **Absolute or relative measure?**	
Criteria for judging **Absolute measure**	If incentive is paid to the health service provider for any of the following **OR** If penalties/fines are levied on the health service provider for not reaching or achieving any of the following • Improvement in performance typically improvement from some baseline measure. e.g. incentive paid per patient immunized, 0r 70% improvement from baseline • Achieving above a predetermined target • Achieving a proportionate part of predetermined target. • Achieving a predetermined target • Points achieved on the incentive domain(s) of performance • Composite performance score
Criteria for judging **Relative measure**	If incentive payment is based on ranked performance data among participating health service providers in such a way that earning bonuses or being fined is dependent on where the performance health service provider ranks among other providers. • If bonuses are paid for to health service providers in a specific

	performance rank e.g. the providers in top quartile of performance. • If fines are levied on health service providers in certain ranks usually the bottom ranks e.g. the providers the last quartile of performance • If bonuses are paid to health service providers in top ranks and fines are levied on providers at the bottom of the performance ranks
Domain of performance measured **Was the domain of performance measure Mostly within clinicians' control or Mostly out of clinicians' control?**	
Criteria for judging **Mostly within clinicians control**	If incentive payments to health service providers are **mostly/only** based on process and structural outcomes e.g. number of children immunized, routine measurement of blood pressure of patients every month
Criteria for judging **Mostly out of clinicians control**	If incentive payments to health service providers are **mostly/only** for ultimate health outcomes e.g. reduction in mortality rates from a specific disease, blood pressure reduction, patient experience etc.
Time lag **Before or at 4 months OR After 4months?**	
Criteria for judging **Before or at 4 months**	If incentive payment (or penalty) is made Immediately or not more than 4 months after measurement and confirmation of performance
Criteria for judging **After 4months**	If incentive payment (or penalty) is made after 4 months after measurement and confirmation of performance
Risk **High risk or low risk? (based on judgements from the dimensions of Performance measure, Time lag, and Domain of performance measure**	
Criteria for judging **High risk**	If the PBF program has 2 or more of the following features • If incentive payment (or penalty) is made after 4 months after measurement and confirmation of performance • If the domain of performance measure was mostly out of clinicians control • If the perofmance measure (payment scale) is a relative measure
Criteria for judging **Low risk**	If the PBF program has 2 or more of the following features • If incentive payment (or penalty) is made Immediately or not more than 4 months after measurement and confirmation of performance • If the domain of performance measure was mostly within the clinicians control • If the performance measure (payment scale) is an absolute measure

Note: Even though the categories within each dimension are dichotomous, it is possible that the PBF program could have both categories featured in the type of incentive and who receives the incentive category). For example, it is possible that a PBF program pays incentives to the hospital and the doctors/clinicians so both groups and individuals get the incentives. In this situation, it is judged as both individuals and groups.

Criteria for judging 'unclear' in all the dimensions: Insufficient information to permit the judgement of category and/or if the design features are not described in sufficient detail.

C7. P4P studies used in testing the inter-rater reliability of the P4P typology

1. An, L.C., et al., *A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline.* Arch Intern Med, 2008. 168(18): p. 1993-9.
2. Ashworth, M., et al., *How are primary care organizations using financial incentives to influence prescribing?* Journal of Public Health, 2004. 26(1): p. 48-51.
3. Basinga, P., et al., *Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation.* The Lancet, 2011. 377(9775): p. 1421-1428.
4. Beaulieu, N.D. and D.R. Horrigan, *Putting smart money to work for quality improvement.* Health Serv Res, 2005. 40(5 Pt 1): p. 1318-34.
5. Catteneo, A., B. Giulio, and S. Giorgio, *Breasfeeding by objectives.* European Journal of Public Health, 2001. 11: p. 397-401.
6. Fairbrother, G., et al., *The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates.* American Journal of Public Health, 1999. 89(2): p. 171-175.
7. Fairbrother, G., Hanson, K.L., Butts, G.C., Friedman, S., *Comparison of preventive care in medicaid managed care and medicaid fee for service in institutions and private practices* Ambulatory Peadiatrics, 2001. 1: p. 294-301.
8. Harries, A.D., et al., *Performance-related allowances within the Malawi National Tuberculosis Control Programme.* The International Journal of Tuberculosis and Lung Disease, 2005. 9(2): p. 138-144.
9. Jha, A.K., et al., *The Long-Term Effect of Premier Pay for Performance on Patient Outcomes.* New England Journal of Medicine, 2012. 366(17): p. 1606-1615.
10. Kirschner, K., et al., *Assessment of a pay-for-performance program in primary care designed by target users.* Fam Pract, 2013. 30(2): p. 161-71.
11. Kouides, R.W., et al., *Performance-based physician reimbursement and influenza immunization rates in the elderly.* American Journal of Preventive Medicine, 1998. 14(2): p. 89-95.
12. Li, Y.H., et al., *The effects of pay-for-performance on tuberculosis treatment in Taiwan.* Health Policy Plan, 2010. 25(4): p. 334-41.
13. Gavagan, T.F., et al., *Effect of Financial Incentives on Improvement in Medical Quality Indicators for Primary Care.* J Am Board Fam Med, 2010. 23: p. 622– 631.
14. Roski, J., et al., *The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines ☆☆Surveys available upon request from corresponding author.* Prev Med, 2003. 36(3): p. 291-299.
15. Ssengooba, F., B. McPake, and N. Palmer, *Why performance-based contracting failed in Uganda – An "open-box" evaluation of a complex health system intervention.* Social Science & Medicine, 2012. 75(2): p. 377-383.
16. Sutton, M., et al., *Reduced mortality with hospital pay for performance in England.* N Engl J Med, 2012. 367(19): p. 1821-8.
17. Werner, R.M., R.T. Konetzka, and D. Polsky, *The effect of pay-for-performance in nursing homes: evidence from state Medicaid programs.* Health Serv Res, 2013. 48(4): p. 1393-414.

C8. A sample rating template by volunteer users of the P4P typology

Study author	**Who receives the incentive: individuals or groups**	**Type of incentive: fines or bonuses**	**Size of incentive: small or large**	**Time lag: short or long** **Domain of measurement: within clinicians control or out of clinicians control** **Performance measure: absolute or relative measure**	**Perceived risk of not earning the incentive: high or low risk**
An et al (Sample study)	Incentive dollars went into each clinic's general Operating fund. There were no payments to individual administrators, physicians, and staff as part of this project. GROUPs	Incentive dollars went into each clinic's general Operating fund. BONUSES	Clinics that referred 50 smokers would receive a $5000 performance bonus. Clinics would also receive $25 for each referral beyond the initial 50. UNCLEAR Because size was not reported relative to budget or salary. I would day SMALL using my judgment	This project took place from September 1, 2005, through June 31, 2006. Incentive payments were made to clinics in 1 lump sum at the end of the contract period (LONG) The primary outcome measure for this study was the percentage of the clinic's smokers referred to telephone counselling. This was defined as the number of unique individuals referred divided by the estimated number of smokers seen in the clinic. PROCESS Within clinicians' control Blue Cross and Blue Shield of Minnesota modified existing contracts with intervention clinics to provide incentives encouraging quit line referral. ABSOLUTE	Low risk

Jha et al (your exercise study)	Who receives the incentive: individuals or groups	Type of incentive: fines or bonuses	Size of incentive: small or large	Time lag: short or long Domain of measurement: within the clinicians control or out of clinicians control Performance measure: absolute or relative measure	Perceived risk of not earning the incentive: high or low risk
Jha et al	Incentive dollars went to the Medicare programs of all hospitals included in the study. There were no individual, administrators, clinical and non-clinical staff incentives. GROUPS	Bonuses in Medicare payments to hospitals that performed in the top two deciles for the medical conditions to which the incentives were tied. A financial penalty starting in the fourth year of the program was placed on hospitals that under-performed.	1-2% for both bonuses and fines. Hospitals with more Medicare patients with the medical conditions tied to the incentive would receive higher bonuses in comparison to those with less Medicare patients with the conditions hence the size of incentive is dependent on the number of Medicare patients with the medical conditions assigned to the incentive serviced by the hospital. Also, later in the program, additional incentives were offered to hospitals that made substantial improvements in care. SMALL	The study was carried out from the fourth quarter of 2003 through to the fourth quarter of 2009. Bonuses and financial penalty started in the fourth year of the program. LONG The aim of the study was to assess the long term effects of the pay-for-performance program (Premier Hospital Quality Incentive Demonstration) on mortality due to the medical conditions assigned incentives. OUT OF CLINICIANS CONTROL The bonuses were paid to health service providers in the top two deciles and the financial penalties were incurred by health service providers that underperformed. RELATIVE	High risk

		FINES			
Study author	**Who receives the incentive: individuals or groups**	**Type of incentive: fines or bonuses**	**Size of incentive: small or large**	**Time lag: short or long** **Domain of measurement: within the clinicians control or out of clinicians control** **Performance measure: absolute or relative measure**	**Perceived risk of not earning the incentive: high or low risk**
Ashworth et al.	Incentive pounds went to GPs in Primary Care Organizations(PCOs) INDIVIDUALS	Incentive pounds went to GPs. BONUSES	Incentives did not form part of the national pay formula for GPs so the additional money increased their salaries. The average bonus offered in year 2 was 1220 pounds. In my judgment, the incentive is LARGE.	The paper mentioned that the incentives boosted the gross income of the GPs rewarded but did not state if it was awarded on a monthly or yearly basis UNCLEAR The aim of this study was to determine the relationship between financial incentives and prescribing behavior in PCOs. PROCESS Within clinicians'control The incentives paid depended on the attainment of a specified prescribing target. ABSOLUTE	Low
Study author	**Who receives the incentive: individuals or groups**	**Type of incentive: fines or bonuses**	**Size of incentive: small or large**	**Time lag: short or long** **Domain of measurement: within the clinicians control or out of clinicians control** **Performance measure: absolute or relative measure**	**Perceived risk of not earning the incentive: high or low risk**
Basinga et al.	The incentive payments were paid to	Incentive payments were made according	The paper mentioned that services with the highest per-unit payments showed the most improvements	Every 3 months, the facilities submitted quarterly requests for payment of incentives from the committee responsible for issuing payments. SHORT	Low

	the Primary Health Care facilities. GROUPS	to the facility's ability to satisfy the quality criteria. Facilities that do not satisfy all the criteria have their payments reduced accordingly. BONUSES	and alluded to the payments being substantial. LARGE	The P4P scheme was initiated to assess the effect of financial incentives on use and delivery of quality maternal and child health care services. The services included institutional deliveries, antenatal visits, immunization etc. PROCESS Within clinicians' control Incentives were paid to facilities based on their performance in delivery of health care services tied to the P4P scheme. ABSOLUTE	
Study author	**Who receives the incentive: individuals or groups**	**Type of incentive: fines or bonuses**	**Size of incentive: small or large**	**Time lag: short or long** **Domain of measurement: within the clinicians control or out of clinicians control** **Performance measure: absolute or relative measure**	**Perceived risk of not earning the incentive: high or low risk**
Cattaneo et al	Local Health Authorities (LHAs)	0.5% deduction of the	The deducted amounts for the whole region amounts to about 1 billion LIT.	The fine is placed on the annual revenues of their DRG of the LHAs. LONG	Low

	receive penalties for not complying with work plans and achieving targets for improving breastfeeding behavior GROUPS	DRG (Disease Related Groups) annual revenues if they do not achieve work plans and targets. FINES	SMALL	Improvements in breastfeeding rates were attributed to objectives/processes combining policy changes and behaviours of health professionals. Within clinician's control LHAs were penalized if they did not meet targets set for improvement in breastfeeding practices. ABSOLUTE	
Study author	**Who receives the incentive: individuals or groups**	**Type of incentive: fines or bonuses**	**Size of incentive: small or large**	**Time lag: short or long** **Domain of measurement: within the clinicians control or out of clinicians control** **Performance measure: absolute or relative measure**	**Perceived risk of not earning the incentive: low or high**
Harries et al.	Incentive payments are made to programme staff. INDIVIDUALS	Incentives were paid to staff members that achieved the targets of the program. The amount depended on the	All NTP staff depend on the $100 monthly remuneration from the local government. Incentive payments range from $1500 for programme Director to $250, the least, for laboratory supervisors. Therefore, the incentive is a significant boost to their incomes.	The incentive payments were paid after self-assessment forms were submitted every 6 months. LONG The indicators for judging the effect of improved performances were centred on practices of the programme staff. Within clinicians' control The incentives were paid based on improving performance of the programme staff in the control of TB. ABSOLUTE	Low

		number of targets met. BONUSES	LARGE		
Study author	**Who receives the incentive: individuals or groups**	**Type of incentive: fines or bonuses**	**Size of incentive: small or large**	**Time lag: short or long** **Domain of measurement: within the clinicians control or out of clinicians control** **Performance measure: absolute or relative measure**	**Perceived risk of not earning the incentive: low or high**
Kirschner et al.	The incentive payments go to the general practices (GPs) GROUPS	Incentive payments went into each practice's operating fund. BONUSES	The incentives were given in relation to quality scores. The higher the score, the higher the incentive. It is between 5-10% of the practice's income. LARGE	Four months after submitting their quality assessment data, the practices are given their bonuses accordingly. SMALL The indicators measured in this study assess the quality of care offered by the general practices. Individual GPs were assessed using questionnaires on their clinical care practices, practice management and patience experience. Within clinician's control Incentive payments were made to general practices using the quality scale used to assess improvements in delivering services. ABSOLUTE	Low
Study author	**Who receives the incentive: individuals or groups**	**Type of incentive: fines or bonuses**	**Size of incentive: small or large**	**Time lag: short or long** **Domain of measurement: within the clinicians control or out of clinicians control** **Performance measure: absolute or relative measure**	**Perceived risk of not earning the incentive: low or high**
Beaulieu and	The physicians	Incentive payments	The incentive payment ranged from $3000 to	Incentive payments were made on an annual basis to physicians who scored above the predetermined target on the composite performance	Low

Horrigan	receive the incentive payments. INDIVIDUALS	are made in the form of a quality bonus to the physicians. BONUSES	$12,000, depending on the composite score of the physician across the indicator practices. LARGE	index. LONG	
				6 process measures and 3 outcome measures were used for this study hence the process measures outweigh the outcome measures. Within clinicians' control	
				The study aimed to improve patient health by improving performance of the physicians in service delivery. Targets were set to be achieved if incentive was to be awarded. ABSOLUTE	

C9. Questionnaire for information about participants of the P4P typology inter-rater reliability study

1. Have you had any kind of training on incentive programs in health care (if yes, give details)?
2. Have you had any experience with an incentive program in healthcare (if yes, give details)?
3. What is your area of research expertise?
4. How many years of research experience do you have?
5. What is your Highest/present academic qualification?

C10. Raters report of time and ease of use of the typology

Rater	**Average time spent per study**	**Difficulty level of using the typology: easy/moderately difficult/difficult**
1	25	Easy
2	20	Easy
3	20	Easy
4	30	Easy
5	25	Easy
6	15	Easy
7	20	Easy
8	15	Easy
9	20	Easy
10	20	Easy
11	20	Easy
12	15	Easy

Appendix D

D1: Extraction of information from evaluations of P4P schemes

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Advancing Quality United kingdom 2008	High risk Annually (long time lag) Mostly within Physicians control (2 final outcomes and 26 processes) Relative measure	Small 2-4%	Group	Bonuses	Sutton et al, 2012 Pre/post Compared with national average (difference in difference analysis)	Outcomes/clinical/chronic care 30 days in hospital mortality: combined (heart failure, pneumonia, acute myocardial infarction)	General combined results: Risk-adjusted, absolute mortality for the conditions included in the pay-for-performance program decreased significantly, with an absolute reduction of 1.3 percentage points (95% confidence interval [CI], 0.4 to 2.1; P = 0.006) significant impact	C	+**
						Outcome 30 days in hospital mortality for patients admitted for Pneumonia	The largest reduction, for pneumonia, was significant (1.9 percentage points; 95% CI, 0.9 to 3.0; P<0.001) significant impact (positive)		+**
						Outcome 30 days in hospital mortality for patients admitted for myocardial infection	non-significant reductions for acute myocardial infarction (0.6 percentage points; 95% CI, −0.4 to 1.7; P = 0.23)		0**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						30 days in hospital mortality for patients admitted for Heart failure	Non-significant reduction 0.6 percentage points; 95% CI, −0.6 to 1.8; P = 0.30). [positive impact but not significant)		0**
Clalit Israel, 1998	Low risk Annually (long time lag) Mostly within Physicians control (10 processes and 8 intermediate outcomes) Absolute measure	Large Dependent on budget savings	Groups	Bonuses	Gross et al. 2008 pre/post design from 1998 to 2005)	Cost containment (process)	Clinics have managed to reduce 10 percent of budget expenses	A	+
						Mammography rates (process)	Mammography rates had risen from 40 percent to 65 percent		+
						Patient satisfaction (outcome)	Patient satisfaction had risen from about 76 percent to 85 percent of members reporting high satisfaction.		+
						Diabetes control measures (process)	Diabetes control measures have improved from 35 percent to 48 percent		+

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Clinical Practice Improvement Pay (CPIP) Australia, Queensland (started 2008)	Low risk Semi-annually (long time lag) Within physicians control (12 structures and 7 processes) Absolute measure	Large 8-10%	Group	Bonuses	Clinical Practice Improvement Centre (2008, 2010), Queensland Health (2010) Before and after (no control group)	Mental health Sixteen mental health services across Queensland participated and were provided with the opportunity to receive incentive payments during the period between January 2009 and June 2011. Data collection was conducted Using information available on existing Queensland Health databases.	State-wide results showed steady and continual improvement in the indicator over the reporting period.	A	+
MACCABI Israel 2001	High risk Annually (long time lag) Mostly within Physicians control (12 processes and 5 intermediate outcomes) Relative	Most likely large Size not reported	Group	Bonus	Friedman, 2006 Before and after (pre-post) no control group	Mammography rates (process)	Mammography rates had risen from 52 percent in 2002 to 64 percent in 2004	B	+
						Balanced diabetes patients (Intermediate outcome)	An increase in the percentage of balanced diabetes patients (Hba1c , 7) was also noted		+

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
	measure					Vaccination flu rates (process)	Flu vaccination rates had risen from 35 percent to 47 percent		+
National Health Insurance P4P (NHI-P4P) Taiwan 2004	High risk Monthly and annually 12 structures, 3 final outcomes, and 2 intermediate outcomes Absolute and relative measures	Large Up to 20%	Individuals and groups	Bonuses	Chang et al., 2008 Logistic regression/pre/post (no control group) One year	Smoking cessation visits (process)	Odds Ratio (95% CI) Financing policy2004* 2005 0.96 (0.87 to 1.06) This policy increased the annual number of cessation visits per patient.	B	0**
					Tsai et al., 2010: Pre-post design compared with control (non-PBF) for 3 years	Tuberculosis treatment default rate (process)	The treatment default rate after "P4P on TB" was 11.37% compared with the 15.56% before "P4P on TB" implementation. The treatment default rate in P4P hospitals was 10.67% compared to 12.7% in non-P4P hospitals.		+
					Kuo et al., 2011 Pre-post with controls (4 years follow up)	Breast cancer care (BC-P4P) in Taiwan on care quality (process)	BC-P4P enrollees received higher-quality care than nonenrollees (*P* _ .001).		+**
						Breast cancer care (BC-P4P) in Taiwan on patient survival	BC-P4P enrollees had better 5-year overall survival (odds ratio, 0.167; *P* _ .001)		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						(outcome)			
						Breast cancer care (BC-P4P) in Taiwan on recurrence (outcome)	Less recurrence (odds ratio, 0.370; *P* _ .002)		+**
					Li et a.l, 2010 Pre-post compared with controls: 4 years	Tuberculosis cure rate (intermediate outcome)	Cure rate: Number cured (cure rate) p4p:18 377 (68.1) non p4p: 2778 (42.4) <0.01 (%) p4p:N 26 977 (80.4) non p4p 6559 (19.6) P4P hospital 0.2911 1.338 (1.159–1.544) <0.0001 cure rate odds ratio 95% CI		+**
					Lee at al., 2010 One year: Pre-post design with control groups	Diabetes care (diabetes specific tests and exams) (process)	Patients in the P4P program (received significantly more diabetes-specific exams and tests after enrolment (3.8 vs 6.4, P <.001) than patients not enrolled in the program (3.5 vs 3.6, P <.001).		+**
						Physician visits for diabetes (process)	Patients in the intervention group had an average of 2 more physician visits for diabetes than those in the comparison group (P <.001).		+**
						Diabetes related hospitalizations (intermediate outcome)	Conversely, the intervention group had fewer diabetes-related hospitalizations (−0.027, P = .003).		+**
Primary care P4P	High risk Annually	Large 8-10%	Individual and groups	Bonuses	*Kirschner et al 2013	Mean score diabetes (9 process indicators)	10.4* (*=significant, p less than 0.05)	B	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
(PC-P4P) Netherlands	(long time lag) Within physicians control (31 processes) Relative measures				Pre-post design evaluation after one year` with control group	Blood pressure controlled	5.9*		+**
						Total cholesterol controlled	8.8*		+**
						HbA1c controlled (≤7.0%) (Intermediate outcome)	7.7*		+**
						Asthma management (4 process indicators)	11.5*		+**
						Asthma outcome	4.4		0**
						Mean score COPD (5 process indicators)	8.1*		+**
						COPD outcome	2.5		0**
						Influenza vaccination (process)	-1.2 (negative impact although not significant)		0**
						Cervical cancer screening (process)	0.6 (no significant impact)		0**
						CRVM process	14.7**		+**
						CRVM outcomes	8.4**		+**
Primary Care Renewal	Low risk Annually Within	Small 2-4%	Individual and groups	Bonuses	Li et al., 2010 Difference in difference	Pap smear	0.003*** pless than 0.005	B	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Models (PCRM) Canada Ontario Started 2007	physicians control (12 processes) Absolute measure				estimates Cross sectional design /time series(with control group)data collected from 1998-2008	Influenza vaccination	0.009		0**
						Mammograms	0.073***		+**
						Childhood immunizations	-0.008		0**
						Colorectal screening	0.092***		+**
Physician Integrated Network (PIN) Canada Manitoba 2004	Low risk Immediately after performance measure (short time lag) Within physicians control (only processes) Absolute	Maximum payment unknown but likely large	Groups	Bonuses	PIN evaluation report, 2012. Pre post design (no control group)	Colon cancer screening	38.7%	A	+
						Dyslipidaemia screening	35.4%		+
						Cervical cancer screening	11.1%		+
						Breast cancer screening	12.3%		+
						Nephropathy screening	29.6%		+
						Lipid profile	22%		+

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						Obesity screening	14.8%		+
						HGBA1C screening	12.5%		+
						Blood pressure test	5%		+
						Renal dysfunction test	11.5%		+
Practice Incentive Program (PIP) Australia 1998	Low risk Quarterly, semi-annually and annually, Within physicians control (only structures and processes) Absolute measure	Size not reported relative to income but likely small	Group	Bonuses	PIP Audit report No 5 2010-2011 Before and after (with control group)	Diabetes	20%points	B	+*
						Prescribing	No significant effect		0**
						Information technology	No significant effect		0**
Quality and Outcomes	Low risk Annually (long time lag)	Large Up to 30-40%	Group	Bonuses	Calvert et al., 2009	Diabetes management Change in HbA1c levels >10%	The introduction of the quality and outcomes framework did not lead to improvement in the management of	A	0**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Framework (QOF)	Mostly within physicians control (85% processes) Absolute measure				Retrospective cohort design (no control group)	Reduction Intermediate outcome	patients with type 1 diabetes, nor to a reduction in the number of patients with type 2 diabetes who had HbA1c levels greater than 10%.		
						HbA1c levels of ≤7.5% Intermediate outcome	Odds ratio 1.05 (95% confidence interval 1.01 to 1.09; P=0.02).		+**
					Campbell et al., 2007 Adequate control	Coronary heart disease Mean Difference (95% CI) P Value Intermediate outcome	0.53 (−0.01 to 1.08) 0.054		0**
						Asthma Intermediate outcome	0.03 (−0.45 to 0.51) 0.904		0**
						Type 2 diabetes management Intermediate outcome	0.08 (−0.32 to 0.49) 0.682		0**
					Taggart et al., 2012 2000-2008 Before and after: no control group	Smoking cessation advice process	Rapid increases in recording smoking status and advice occurred around the QOF's introduction in April 2004. Subsequently, compliance to targets has been sustained, although rates of increase have slowed.		+
					Millet et al., 2009 Before and after with no control group	Achievement of diabetes treatment targets for blood pressure (< 140/80 mm Hg), HbA1c (# 7.0%) and cholesterol	Patients with co-morbidity remained significantly more likely to meet treatment targets for cholesterol and HbA1c than those without after the introduction of pay for performance		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						(# 5 mmol/L). Intermediate outcome			
					MacBride-Stewart, et al,. 2008 Before and after ITS Adequate control	Changes in prescription pattern Process	QOF significant reduction in prescribing pattern compared to a non-significant increase in prescribing pattern for the Non QOF control group.		+**
					Strong et al., 2009 Before and after with no control group	Accurate spirometry in the management of COPD process	There was no association between quality, as measured by adherence to BTS spirometry standards, and either QOF COPD9 achievement (Spearman's rho = -0.11), or QOF COPD10 achievement (rho = 0.01).		0**
					Vaghela et al,. 2008 Before and after: no control group	A1C <or=7.5%,	The estimated annual increase in percent of diabetes subjects achieving targets was 3.03% (95% CI 2.95–3.10; P 0.001) for the A1C target		+**
						Blood pressure <or=145/85 mmHg Process	The estimated annual increase in percent of diabetes subjects achieving targets was 3.26% (3.18–3.34; P 0.001) for the blood pressure target		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						Cholesterol <or=5 mmol/l was determined. Process	The estimated annual increase in percent of diabetes subjects achieving targets was 3.99 % (3.92– 4.07; P 0.001) for the cholesterol target.		+**
					Tahrani et al., 2007 Before and after with no control group PCTs	Process indicators	95% CI April 2004- March 2006 all p values less than < 0.001		
						BMI Record	-19.2 to -14.5		+**
						Smoking record	-54.7 to -47.3		+**
						HBA 1c Record	-22.5 to -15.0		+**
						Retinal screening record	-42.9 to -32.5		+**
						Peripheral pulses record	-63.6 to -52.7		+**
						Neuropathy testing record	-64.2 to -53.2		+**
						BP record	-10.8 to -8.2		+**
						Micro albumin testing record	-74.8 to -65.9		+**
						Creatinine record	-15.0 to -11.2		+**
						Cholesterol record	-17.3 to -13.6		+**
						Outcome indicators	95% CI April 2004- March 2006 all p values less than < 0.001		
						Smoking cessation advice	-15.2 to -9.2		+**
						HbA1c< 7.4	-24.1 to -16.2		+**
						HbA1c< 10	-22.6 to -16.4		+**
						BP< 145/85mmHg	-20.3 to -15.9		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						TC<5	-25.9 to -22.0		+**
						Influenza vaccine	-24.6 to -18.1		+**
					Serumaga et al., 2011 Design Interrupted time series.	Blood pressure monitoring (no change) process	After accounting for secular trends, no changes in blood pressure monitoring (level change 0.85, 95% confidence interval −3.04 to 4.74, P=0.669 and trend Change −0.01, −0.24 to 0.21, P=0.615), control (−1.19, −2.06 to 1.09, P=0.109 and −0.01, −0.06 to 0.03, P=0.569)		0**
						Treatment intensity (no change) process	Treatment intensity (0.67, −1.27 to 2.81, P=0.412 and 0.02, −0.23 to 0.19, P=0.706) Good quality of care for hypertension was stable or improving before pay for performance was introduced. Pay for performance had no discernible effects on processes of care or on hypertension related clinical outcomes.		0**
					Cupples et al., 2008 2004-2006 Cross-sectional Study Control group	Blood pressure,	More RoI than NI participants had systolic blood pressure >140 mm Hg (37% vs 28%, P = 0.01)		+**
						Cholesterol	More RoI than NI participants had cholesterol >5 mmol/L (24% vs 17%, P		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
							= 0.02)		
						Medications	Fewer participants in the RoI (55% vs 70%) were prescribed β-blockers. ACE inhibitor prescribing was similar for both groups (41%; 48%); high proportions were prescribed statins (84%; 85%) and aspirin (83%; 77%)		+**
						Smoking status 1	-62.1 (-67.0 to -56.3)		0**
						Smoking status 2	-22.7 (-26.4 to -19.0)		-**
						Smoking status 3	3.5 (-1.8 to 8.6)		0**
						Smoking status 4	-3.1(-8.4 to 1.8)		0**
					Coleman, 2007 1990-2005 Retrospective longitudinal survey	Smoking status recording	Compared with the first quarter of 2003, recording of smoking status increased up to the first quarter of 2004 in (rate ratio = 1.88; 95% CI, 1.87–1.89)		+**
						Brief advice to smokers	Compared with the first quarter of 2003, and in brief advice to smokers increased up to (RR = 3.03; 95% CI, 2.98–3.09),		+**
					Campbell, et al., 2009 1998-2007 Before and after	Coronary heart disease	Mean change in rate of improvement - 0.250, 95% CI, -0.401 to 0.100, pvalue=0.001		0*
						Asthma	Mean change in rate of improvement -		0*

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
					study Interrupted time series		0.468, 95% CI, -0.748 to -0.187, pvalue=0.001		
						Diabetes	Mean change in rate of improvement -0.220, 95% CI, -0.313 to -0.127, pvalue=0.001		0*
						Continuity of care	Mean change in rate of improvement 0.091, 95% CI, 0.025 to 0.157, pvalue=0.001		+**
					Hippisiley-cox, et al., 2007 2001-2006 Interrupted time series However, absolute mean changes were reported	Coronary heart disease	This is equivalent to a relative increase of 50% (95% CI 37%-63%) over the five year study period as shown in the graph below		+**
						Stroke patients with cholesterol < 5 mmol	356% relative increase (95% CI 182-637%) in the percentage of stroke patients with cholesterol < 5 mmol/l in the preceding 15 months		+**
						Stroke patients with a blood pressure reading < 150/90 mm hg	There was a 68% relative increase (95% CI 55-83%) in the percentage of patients with a blood pressure reading < 150/90 mm hg in the preceding 15 months		+**
						Diabetes recorded prevalence	Using the new 2006/7 definitions, there was a 117% (95% CI 115-120) relative increase in the recorded prevalence of diabetes (Diabetes1).		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						percentage of diabetes patients with cholesterol < 5 mmol/	there was a 132% relative increase (95% CI 95-176%) in the percentage of diabetes patients with cholesterol < 5 mmol/l in the preceding 15 months.		+**
						Diabetics with a blood pressure reading < 145/85 mm hg	There was a 56% relative increase (95% CI 47-66%) in the percentage of patients with a blood pressure reading < 145/85 mm hg in the preceding 15 months.		+**
						Diabetic High blood pressure recorded	There was a 35% (95% CI -41 - 209) relative increase in the recorded prevalence of hypertension (BP1).		0**
						Diabetic High blood pressure controlled	There was a 65% (95% CI 51-79%) relative increase in the percentage of patients with controlled blood pressure levels		+**
						Chronic kidney disease chronic kidney disease and blood pressure recorded	there was a 20% relative increase (95% CI 3-32%) in the percentage of patients with chronic kidney disease and blood pressure recorded in preceding 15 months.		+**
						Chronic Kidney disease percentage of patients with a blood pressure reading < 140/85	There was an 89% relative increase (95% CI 59-124%) in the percentage of patients with a blood pressure reading < 140/85 mm hg in the preceding 15 months.		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
					Magee, 2010 Interrupted time series	Nephropathy prevalence	Nephropathy prevalence was 15.1% and 11.5%, respectively.		+
						The median ACR testing rate	The median ACR testing rate was 82% compared with a historic figure of 41% in 2001/2002		+
					Milliet, et al.,2007 2003-2005 Longitudinal cross-sectional survey	Record of smoking status	Significantly more patients with diabetes had their smoking status ever recorded in 2005 than in 2003 (98.8% vs 90.0%, P <.001).		+**
						Smoking cessation advise	The proportion of patients with documented smoking cessation advice also increased significantly over this period, from 48.0% to 83.5% (P <.001).		+**
						Prevalence of smoking/quit rates	The prevalence of smoking decreased significantly from 20.0% to 16.2% P <.001)		+**
					McGovern, 2008 200-2005: serial cross sectional study		Recording and prescribing increased by mean 17.1% after the introduction of the GMS contract		+
					Oluwatowoju, et al., 2010	Diabetes HbA1c <7.5%);	In 2006, 39.7% of adults had glycemic control within the QOF threshold		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
					2006-2008 Retrospective retrieval of computer-held biochemical measurements		(HbA1c <7.5%); by 2008, this proportion had risen to 52.1% (P <.001).		
						Diabetes HbA1c >10.0%	In 2006, 11.8% of subjects had poor glycemic control (HbA1c >10.0%); by 2008, this proportion had decreased to 10.1% (P <.001).		+**
						Diabetes (both HbA1c <7.5% and total cholesterol ≤5.0 mmol/L)	The proportion of subjects achieving HbA1c and cholesterol targets (both HbA1c <7.5% and total cholesterol ≤5.0 mmol/L) was 30.2% in 2006; in 2008 this proportion had increased to 43.7% (P <.001)		+**
					Srirangalingam et al., (2006) Before and after cross sectional study	Diabetes	Increase in referrals for poor glycaemic control, and the glycaemic threshold for referral with poor glycaemic control has reduced (9.7% vs 10.6%, P= .006, mean difference = 0.9%, 95% CI, 0.4-1.3%).		+**
					Simpson et al., 2010 Before and after	Smoking status reporting	The proportion of people with smoking status recorded increased by 32.9% (from 46.6% in2001/2 to 79.5% in 2006/7, OR 4.45, 95% CI 4.43 to 4.46)		+**
						Smoking cessation advise	There was a large increase in provision of smoking cessation advice (43.6% in		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
							2001/2, 84%in 2006/7, OR 6.75, 95% CI 6.66 to 6.85)		
						Smoking cessation referral	The proportion of patients referred to stop smoking clinics increased (from 0.95% to 6.56%, OR 7.32, 95% CI 6.92 to 7.73)		+**
						Quit rates	The proportion of people recorded as being a smoker reduced from 28.4% in 2001/2 to 22.4% in 2006/7 (OR 0.73, 95% 0.72 to 0.73)		+**
					Simpson et al., 2011 No control group	Hypertension	Increasing treatment for hypertension (absolute difference [AD] 9.2%; 95% confidence interval [CI] = 9.0 to 9.5) occurred throughout the study period.		+**
					Gulliford, et al., 2007	Diabetes	HbA1c≤7.4% Among 26 practices in South London, the median practice-specific proportion of patients achieving HbA1c≤7.4% each year increased: 2000,22%; 2001, 32%; 2002, 37%; 2003, 38% and in 2005 from QOF, 57%.		+
					Kontopantelis et al., 2012 Interrupted time series analysis	Diabetes	Recorded quality of care improved for all subgroups in the pre-incentive period. In the first year of the incentives, composite quality improved		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
					Adequate control		over-and-above this pre-incentive trend by 14.2% (13.7–14.6%).		
							By the third year the improvement above trend was smaller, but still statistically significant, at 7.3% (6.7–8.0%).		+**
Western New York Physician Incentive Program (WNY-PIP) USA	Low risk Annually (long time lag) Mostly process: 6 Process and 3 outcomes Intermediate outcome Absolute measure	Size of varied from $3,000 till $12,000 large	Individuals	Bonus	Beaulieu ND and Horrigan DR (2005) 8months pre-post with a control group Even though they stated that there was a control group, the results presented are absolute so I will treat as no control group	Diabetes control: HbA1c test (process)	HbA1c test (1) no significant difference Significance: p<0.0001 (for all)	B	0**
						Lipid test (process)	Lipid test: significant increase		+**
						HbA1c < 9.5 (intermediate outcome)	HbA1c < 9.5: significant increase		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						LDL <130 (Intermediate outcome)	LDL <130: significant increase		+**
						Diabetes control: HbA1c test (process)	HbA1c test (1) no significant difference Significance: p<0.0001 (for all)		0**
Kouides et al., 1998 Rochester, New York, USA	Low risk Annually (long time lag) Process Absolute measure	Size 'Modest' for just one process ?	Group	Bonus	PBF vs. non PBF Before PBF vs. After PBF Control group	Influenza immunization rates	Absolute increase in immunization rates (from 1990 [baseline] to 1991) was 6.8%; *P* _ 0.03 Change in immunization rates (1991-1990) intervention:10.3% , control: 3.5% p=0.3	A	0**
Ashworth et al., 2004 UK 2004	Low risk Annually (long time lag) Process/structure Absolute measure	up to £5000 per GP (large) Up to 5%	Groups but money trickled down to individuals	Bonus	Before and after incentive (no control group)	Change in use of prescription budget (overspent/underspent) of primary care organization (PCO)	PCO prescribing budgets were, on average, overspent by 4.5 per cent in the first year and marginally under spent by 0.6 per cent in the second year. Many PCOs had successfully turned a first year prescribing overspend into a	A	+

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
							second year under spend. PCOs that successfully reversed their overspend (49 out of 84; 58 per cent)		
Cattaneo et al., 2001 Italy 1998-1999	Low risk Yearly (long time lag) Process Absolute measure	Small 0.5% of annual revenue deducted	Groups	Fines	Before and after study (no control)	Change in breast feeding rates (intermediate outcome)	Significant increase in breast feeding rates	B	+**
Fairbrother et al., 1999 New York 12 months	Low risk Annually (long time lag) Process Absolute measure	$1000 Large	Individuals	Bonus plus feedback	Before and after study with control group July 1995-July 1996	Childhood immunization coverage rates (process)	Bonus group improved significantly in documented up-to-date immunization status, with an overall change of 25.3% (P _ 0.01),	B	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Fairbrother et al., 2001 USA 16 months	Low risk One off payment after 16 months (long time lag) Process Absolute measure	1000 usd	Individual	Bonus	Comparison of Preventive Care in Medicaid Managed Care and Medicaid Fee for Service in Institutions and Private Practices Control group	Change in documentation of up-to-date immunization status.	The bonus group improved significantly in documented up-to-date immunization status, with an overall change of 5.9% (P _ 0.05) compared with the control group. N=57 physicians (24 bonus; 12 FFS; 21 control)	B	+**
Grady et al., 1997 USA	Low risk Quarterly payments (short time lag) Process Absolute	Token Small?, i.e., $50 for a 50% referral rate.	Groups	Bonus with education	Mammography referral rates (process)	Mammography referral rates (process)	No significant difference between the two groups	B	0**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
	Measure	Small up to 1%							
Hillman et al., 1998	Low risk Every 6 months (long time lag) Process Absolute measure	Large Up to 20% of capitation fees	Individuals and groups	Bonus and feedback 18 months: no effect	RCT 2 years	Cancer screening: breast, cervical and colorectal Mean compliance score	No significant difference between the intervention and control groups for pap test	A	0**
							No significant difference between the intervention and control groups for colorectal screening		0**
							No significant difference between the intervention and control groups for mammography		0**
							No significant difference between the intervention and control groups for breast exam		0**
Larsen et al., 2003	Low risk Time lag not reported Process Absolute measure	Small Size up to 1% of physicians compensation	Individuals	Bonus	Four years pre-post: no control group	Diabetes care: LDL < 130	Significant difference p<0.001 from 1998-2002 39.9% To 69.8% pvalue less than 0.001	C C	+**
						Average HbA1c	Reduction of 8.1-7.3		+**
						HbA1c>9.5	Reduction of 34.6-21.4		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						HbA1c < 7 (Intermediate outcome)	33.5%%To 52.8%		+**
						Annual HbA1c	78.5-90.5%		+**
						Bi annual LDL	Increase of 65.9-91.7		+**
						Annual eye exam	From 52-62%		+**
LeBaron et al., 1999 USA	Not enough information reported on the costs and nature of incentives			Bonuses	Before and after (no control group)	Childhood immunization coverage rates	Mean change +3 percentage points From 1994-1996 75 (74-76)- 78 (77-79) (95% CI))	Can't tell	+**
Ritchie et al., 1991 Scotland: UK	Low risk Quarterly payments (short time lag) Process Absolute measure	Not enough information reported on size	Groups Clinical practices	Bonuses	Before and after study Study period: one year no control group	Percentage immunized by practice/ immunization rates	Percentage of children aged 5 years given preschool boosters in Grampian region, 1987-91 rose from 78- 93% (p<0-0001). All 95 general practices in Grampian region (313 general practitioners). Those aged 5 years on the first day of the relevant quarter, with an average population of 6600	B	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Rooski et al., 2003 USA	Low risk 3 month time lag in payment Process Absolute measure	Size: up to $10,000 not reported relative to practice budget/income Most likely large.	Groups	Bonuses	RCT 12 Months (unbalanced)	Adherence to smoking cessation clinical practice guidelines and patients' smoking cessation behaviours.	Percentage of patients, tobacco use status identified in the last visit (Process) 14.1 vs 6.2(incentive vs control)	A	+*
							Percentage of smokers who received advice to quit in the last visit (Process)24.2 vs 18.3 (incentives vs control)		+*
							Percentage of smokers who were offered assistance to quit in the last visit (Outcome) 14.3 vs 8.8 (incentives vs control)		+*

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
							Quitting rates did not differ statistically significantly between the experimental conditions.		0*
Harries et al., 2005 Malawi National Tuberculosis Control Programme (four year program/0	Low risk 6month (short time lag) process absolute measure	Size: up to 100% of usual reimbursement	Individual physicians	Bonuses	before and after study with control groups	Tuberculosis control and other outcome measure.	Percentage of patients documented as smear-positive in the laboratory register that are subsequently registered for treatment in the TB register. Target set at or above 90%	B	0
							Percentage of patients aged 15 years and above registered in the TB register as smear-negative PTB patients who have had Sputum smears examined (data from laboratory register). Target set at or above 85%.		+

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
							Percentage of new smear-positive PTB patients who default from treatment/transfer out or who complete treatment with no smears examined. Target set at or below 10%.		+
							Percentage of relapse smear-positive PTB patients for whom sputum specimens arrived at the mycobacterial central reference laboratory, Lilongwe, for culture and drug sensitivity testing. Target set at or above 60%.		0
Chien et al., 2012 Hudson Health Plan's P4P program	High risk Both process and outcomes Yearly Absolute	300$ per patient	Groups	Bonus	Four years (2003–2007) Design: case-comparison difference-in-difference study	Lipid testing (process)	+4%points	B	0*

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
in New York					using plan-level administrative data; (2) a patient-level claims data analysis; and (3) a cross-sectional survey (control group)	HbA1c <9	+8%points		0*
						Hba1c testing (process)	+2%points		0*

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Hillman et al., 1999 USA	Low risk Process Absolute and relative really Payment frequency: every 6 months	Bonuses based on total compliance score for quality indicators; full and partial bonuses Averag	Payments to provider groups	Bonus and feedback	RCT 18MONTHS RCT (3 arms); 1993 to 1995; 49 PC sites (19 FB_I; 15 FBO; 15 controls)	Rate of paediatric immunization: randomly assigned primary care sites serving children in a Medicaid HMO to one of three groups: a feedback group (where physicians received written feedback about compliance scores), a feedback and incentive group (where physicians received feedback and	However, no significant differences were observed between either intervention group and the control group, for compliance scores	A	0**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
		e bonus, $2,000 (range, $772 to $4682)				a financial bonus when compliance criteria were met), and a control group. They evaluated compliance with paediatric preventive care guidelines through semi annual chart audits during the years	However, no significant differences were observed between either intervention group and the control group, for immunization rates		0**
Christensen et al., 2000 USA	Low risk Timing of payment not reported Process Absolute measure	$4 for cognitive services	Provider group	Bonuses	RCT (2 arms); February 1994 to September 1995 200 pharmacies (110 interventions; 90 control)	Dosage with CS	Student *t*-test Mean rate, 1.59 interventions per 100 Medicaid prescriptions (study pharmacies) vs. 0.67 (controls); *P* _ 0.001 Pharmacists practicing in 110 study (financial incentive) and 90 control community pharmacies. Study pharmacists documented an average of 1.59 CS interventions per	A	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
							100 prescriptions over a 20-month period, significantly more than controls, who documented an average of 0.67 interventions (P < .05) per 100 prescriptions.		
Hillman et al., 1998 USA	Low risk Payment frequency: every 6 months (long time lag) Process Absolute measure	$1260 Large: up to20%	Provider group	Bonuses	RCT (2 arms); 1993 to 1995; 52 PC sites (26 intervention; 26 control)	Compliance with cancer screening for women age >50 y; aggregate compliance scores and improvement in scores over time	Repeated-measures ANOVA Absolute increase in total mean compliance scores for intervention group from baseline was 26.3%; control group was 26.4%. No significant differences between the groups Aggregate compliance scores and improvement in scores over time.	A	0**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Gavagan, et al., 2010 USA	Low risk Annually (long time lag) Processes Absolute Measure	Small approximately 3% to 4% of a provider's total salary	Individual physicians	Bonuses	A retrospective review of administrative data (2003-2007) was done to evaluate a natural quasi-experiment With a control group	Rates of Papanicolaou screening	Overall, there was no clinically significant effect of incentives on performance	C	0** (non-significant difference)
						Rates of mammography	Overall, there was no clinically significant effect of incentives on performance		0** (non-significant difference).
						Rates of child immunizations	Overall, there was no clinically significant effect of incentives on performance		0**
An et al., 2008 USA	Low risk Annual (long time lag) Process Absolute measure	Small 5000$ onetime payment at the end of	Groups	Bonuses	RCT Clinical randomized trial? Compared with what: non PBF, standalone	Smoking cessation referral rates	Intervention clinics referred a mean of 11.4% (95% CI, 8.0%-14.9%) of their smokers compared with 4.2% (95% CI, 1.5%-6.9%) of smokers visiting usual care clinics (t47=3.45; P=.001) significant difference	B	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
		the programme			scheme Intervention clinics				
Glickman et al.,2007 USA CMS Premier program	High risk Yearly (long time lag) Process and outcomes Relative	Small 2%	Groups (hospitals)	Bonuses	Patients were treated between July 1, 2003, and June 30, 2006, at 54 hospitals in the CMS program and 446 control hospitals 3 years pre-post with control group	Aspirin prescription rate	Pvalue of comparison of intervention group to control group 0.12	C	0**
						Smoking cessation counselling rates	0.05		+**
						In hospital mortality	0.21		0**
						Aspirin at discharge	0.04		+**
						Beta blockers at arrival	0.91		0**
						Beta blockers at discharge	0.98		0**
						ACE inhibitor at discharge	0.51		0**
						CMS composite score	0.16		0**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Levin et al., 2006 USA	Low risk Paid monthly (short time lag) Process Relative measure	Up to 20% of budget/ salary	Groups	Bonuses	Two year program Pre-post design with control group	HbAIC screening	PCHI's performance in HbAIC screening in the index health plan improved over 2 years by 7 percentage points, compared with a statewide improvement of 4.9 percentage points ($p < .05$).	A	+**
						Eye exams	For diabetic eye exams, PCHI's performance improved 18.7 percentage points, compared to a slight decline in statewide performance ($p < .05$).		+**
						LDL screening	For diabetic LDL screening, PCHI improved by 13.2 percentage points, almost twice that of the state average ($p < .05$),		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						Nephropathy screening	Nephropathy screening rate improved by 15.2 percentage points, over twice the state-wide improvement ($p < .05$).		+**
						Paediatric asthma controller use	(PCHI improvement 1.7 percentage points, state improvement 3.9 percentage points, $p > .05$), 3.8* mean change (process drug).		0*

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Mandel et al., 2007 Cincinnati USA	Can't tell: not enough information reported Process	Large 7% fee schedule increase	Practices (groups)	Bonuses	Between October 1, 2003,andNovember30,2006 No control group but interrupted time series desing. Good quality, so will count as control	Asthma improvement in children Influenza vaccination rates	all-payer asthma population receiving "perfect care" increased from 4% to 88%, with 18 of 44 practices (41%) achieving a perfect care percentage of 95% or greater influenza vaccine increased from 22% at baseline (2003- 2004 season [September 1 through March 31]) to 41% for the 2004-2005 season, to 62% for the 2005-2006 season, with 7 of 44 practices (16%) achieving an influenza vaccination percentage of 80% or greater for the 2005- 2006 season.		+ +

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Lindenauer et al., 2007 CMS USA	High risk Annual (long time lag) Processes Relative measure	Up to 2% usual Medicare reimbursement	Group	Bonuses	2 years Natural experiment: pre-post with control. multivariable modeling to estimate the improvement attributable to financial incentives p4p implementedd with public reporting	Aspirin on arrival	Percentage change 3.3**	C	+**
						Aspirin on discharge	0.9		0**
						ACE inhibitor	9.9**		+**
						Beta blocker on arrival	2.8**		+**
						Beta blocker on discharge	2.8**		+**
						LV assessment	5.1**		+**
						Ace inhibitor for LVSD	2.0		0**
						Antibiotic timing for pneumonia patients	4.3**		+**
						Vaccination for pneumonia patients	10.9**		+**
						Oxygen assessment	0.6		0**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						Appropriate care for MI	7.5**		+**
						Appropriate care for heart failure	6.0**		+**
						Appropriate care for pneumonia	7.1**		+**
						Composite process scores all 10 measures	4.3**		+**
Greenberg et al., 2008	Low risk Payment every three months (short time lag) Process	Not enough informtion reported	Individuals	Bonuses	Before and after design with no control group	Smoking cessation referral rates	Staff referrals increased with program incentives (P=.008), with a total of 150 interventions occurring in the 3-month span.	CANT tell	+**
Yao H et al., 2008 China	Not enough information reported Process	$31 694 for spreading TB knowledge in villages	Doctors Individuals	Bonuses	Implemented with a demand side intervention Pre-post design with control group	TB case detection and treatment	The project achieved its case detection target: the total number of new smear-positive TB cases identified in the intervention counties during the whole project period (November 2004–October 2005) was 7736, which was 136% of the project target established	Can't tell	0*

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
					One year period evaluation		in the proposal, according to the baseline data of the intervention group. However, no improvement on TB case finding and case holding was found in the intervention group compared with the control group (Table 2). At baseline, the intervention group had a significantly higher case notification rate (P < 0.01).		
Fagan et al., 2010	Low risk Timing of payment not reported Process and structures Absolute measure	Large Up 20%	Groups	Bonuses	2004-2007 Quasi experimental 9before after and control group)	Influenza vaccine	Odds ratio 1.79 (1.37-2.35)	A	+**
						Haemoglobin testing	0.44 (0.33-0.65)		-**
						Eye exam	0.98(0.61-1.58)		0**
						Ldl test	0.62(0.44-0.86)		_**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						Nephropathy test	0.96(0.62-1.46)		0**
						Management of hypertension with diabetes	1.11(0.58-2.13)		0**
Chien et al., 2010 USA	Low risk Timing of payment not reported Process Absolute measure	Large	Individuals	Bonuses	Study Design. Case-comparison and interrupted times series 2003–2007	Childhood Vaccination rates	Hudson Health Plan members or by private practices were also significantly more likely to be immunized (Table 2, high number of Hudson enrollees OR 5 1.65–1.73, po.001	B	+** on the long run
Jha et al., 2012 CMS	High risk Yearly (long time lag) Process and outcomes Relative	2%	Groups hospitals	Bonuses	Pre-post with control group.	Premier vs non premier Mortality rates for different conditions 30-day mortality	The rates of decline in mortality per quarter at the two types of hospitals were also similar (0.04% and 0.04%, respectively; difference, −0.01 percentage points; 95% CI, −0.02 to 0.01),	C	0*

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
	measure						and mortality remained similar after 6 years under the pay-for-performance system (11.82% for Premier hospitals and 11.74% for non-Premier hospitals; difference, 0.08 percentage points; 95% CI, −0.30 to 0.46). 0.36 for interaction)		0*
							We found that the effects of pay for performance on mortality did not differ significantly among conditions for which outcomes were explicitly linked to incentives: acute myocardial infarction		0*
							CABG		0*
							Congestive heart failure		0*
							Pneumonia		0*
Lynch et al.,1995	Annually Paid quarterly Absolute		Paid to GP practices	Bonuses	1990 general practitioners contract	Uptake of childhood immunizations	While this has led to an increase in the number of general practitioners providing the services	A	+

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
	(tournament) it would between 70% and 89%; rates below 70% do not qualify for these payments. Low risk		Groups						
Sussman et al., 2000 Boston, Massachusetts USA	Low risk Yearly (long time lag) Process Absolute measure	Large Size: up to 10% of salary	Bonuses	Groups	Before and after study (no control group)	Percentage of the wRVU productivity-	After the first year of operation of this plan, there was an overall 20% increase in PCP productivity.	A	+
Norton et al.,1992	High risk Can't tell Timing of payment not reported: yearly Outcomes Absolute measure	Large $126 to $370	Groups	Bonuses	RCT (2 arms); November 1980 to April 1983; 36 SNFs (18study facilities; 18 control facilities) Up to 4 years	Improvement in health status	Patients in experimental homes were more likely to be discharged to home or to an ICF and had less likelihood of hospital admission or death (P _ 0.001)	CANT TELL	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Shen et al., 2003 Maine, USA	Low risk Annual payment (long time lag) Process Absolute measure	Not enough reported about size	Groups	Bonuses	CBA; FY 1991 to 1995	Substance abuse treatment	The percentage of OSA outpatient clients classified as most severe users dropped by 7 percent (po50.001) after the innovation of performance based contracting compared to the increase of 2 percent for Medicaid clients	A	+**
Werner et al., 2012 CMS USA	High risk Yearly (long time lag) Process and outcomes Relative measure Yearly HIGH RISK	Small 2%	Groups	BONUSES	Pre-post design with control group 5 years	In house mortality rates	The performance of the hospitals in the project initially improved more than the performance of the control group: More than half of the pay-for performance hospitals achieved high performance scores, compared to fewer than a third of the control hospitals. However, after five years, the two groups' scores were virtually identical. Improvements were largest among hospitals that were eligible for larger bonuses, were well financed, or operated in less competitive markets	C	0*
Basinga et al., 2011 Rwanda	Low risk Monthly and quarterly payments	Large 22-38% of usual budget and	Individuals and groups	Bonuses	Pre-post with control groups	Any prenatal care	0·002 p= 0•875	A	0**
						Four or more prenatal care visits	0·008 p= 0•875		0**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
	(short time lag) Processes Absolute measure	salary				Institutional delivery	0·081 p= 0•017		+**
						Tetanus vaccine during prenatal visit	0·051 p= 0•057		0**
						Standardised total quality score	0·157 p= 0•020		+**
						Younger than 23 months preventive visit, previous 4 weeks	0·119 p= 0•004		+**
						24–59 months preventive visit, previous 4 weeks	0·111 p< 0•0001		+**
						12–23 months fully immunised	−0•055 p= 0•390		0**
Canavan A. and Swai G. (2008) Tanzania	Low risk Payment every 6 months (long time lag) Processes	Large 5-10% of hospital budget and	Individuals and groups	Bonuses	Pre-post with control groups 3 years	In patient department	IPD RR: 0.82 (0.76-0.86) P<0.00001	A	0**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
	Absolute measure	clinicians salary				Change in utilization	Utilization in health facilities RR: 0.94 (0.83 to 1.08) p>0.40)		No significant impact
Sulku, 2011 Turkey	Low risk Monthly payments (short time lag) Process and outcomes Absolute measure	Large Up to 80% of budget and salary	Individuals and groups	Bonuses	Pre-post with control group 5 years	Mortality rates	Hospital mortality rates (increased non significantly: 0.01-0.012 p>0.05)	A	0 no significant impact
						Mean outpatient visits	Mean outpatient visits increase by 78% significantly p<0.01		+**
Vergeer and Chansa,	Low risk Absolute	Up to 100% of	Individuals and groups	Bonuses	Pre-post with control group.	ANC	No significant change in ANC, 4. No significant difference in intervention and control hospitals in relation to	A	0 **

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
2008. Zambia	measure Quarterly payments (short time lag)/Processes	salary					IPD/OPD. Variety of patterns across facilities		
Ssengooba et al., 2012. Uganda	Low risk 6monthly payment (long time lag) Process Absolute measure	Large up to 11% of hospital budget	Groups	Bonuses	Pre-post with control group	Maternal and child health process measures	After 2½ years and three survey rounds, the study found no discernable impact of bonuses on the provision of health services by the PNFP providers (group C). Twenty-two out of 23 facilities receiving performance bonuses did reach at least one performance target, and 12 reached all three, but service levels at group B institutions similarly improved. If anything, facilities in the bonus group performed slightly worse than the facilities receiving only the untied base grant and about as well as the facilities in the control group.	A	0** no significant impact. If anything, bonus group performed slightly worse although not significant

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Cutler et al., 2007 USA (California P4P)	High risk Annual payments (long time lag) Processes and intermediate outcomes Relative measure	Large Up to 5% of budget	Groups	Bonuses	Retrospective study: before and after (with control group)	Diabetes testing	The LDL-C testing rate for patients in the CDCM program was 91.5% versus 67.8% for the routine care group). The LDL-C goal attainment rate for the CDCM program was 78.2%, significantly higher than the 55.7% rate for the routine care group ($P < 0.001$	B	+** +**
Rosenthal et al., 2005 USA California p4p						Cervical screening	Compared with physician groups in the Pacific Northwest, the California network demonstrated greater quality improvement after the pay-for-performance intervention only in cervical cancer screening (a 3.6% difference in improvement [P=.02]).	B	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						Mammography	Difference in difference result not significant		0**
						Haemoglobin	Difference in difference result not significant		0**
Gilmore et al., 2007 Hawaii Medical Services Association	High risk Annual (long time lag) Relative Outcomes	Large Up to 7% of salary	Individuals	Bonuses	Before and after with control group	Patient satisfaction on recommended care	We found a consistent, positive association between having seen only program-participating providers and receiving recommended care for all 6 years with odds ratios ranging from 1.06 to 1.27 (95 percent confidence interval: 1.03–1.08, 1.09–1.40)	C	+**
Young et al., 2007	High risk Annual (long time lag) Processes Relative measure	Large 5% of physician fees was at risk	Individuals	Fines	Before and after with control group/similar to an interrupted time series design	Diabetes measures	Based on the absence of a significant interaction term for each measure in this context, the post-intervention trends were not different from the pre-intervention trends, indicating that the overall pattern of performance did not change after program	C	0**
Twardella and Brenner, 2007	High risk Annual (long time line) Outcome	Unclear	Individuals	Bonuses	RCT	Smoking cessation	Self-reported smoking abstinence obtained at 12 months follow-up and validated by serum cotinine. In intention-to-treat analysis, smoking	C	0**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
	Absolute measure						abstinence at 12 months follow-up as 3% (2/74), 3% (5/ 144), 12% (17/140) and 15% (32/219) in the usual care, and interventions		
Scott et al., 2009 PIP					Before and after with control group	Diabetes test HbA1c test	Model (1) of Table II shows a statistically significant effect of 20% (1% level) for *Treatment group 1*. This marginal effect suggests that the average GP working in an average practice of the sample that joined the PIP program is more than 20 percentage points more likely to order an HbA1c test than a comparable GP in a practice that has not joined		+**
Schauffler et al., 1999 California USA	Low risk Annual (long time lag) Processes Absolute measure	Small up to 2% of premiums at risk	Groups	Fines	Before and after (no control group)	CHILDHOOD IMMUNIZATIONS	The majority of the HMOs exceeded their negotiated targets for most of the quality-of care measures However, they fell considerably short on childhood immunizations, and nearly half missed their targets on mammograms and Pap smears as well. Eight plans missed their targets for	B	+
						CESAREAN SECTIONS.			-
						MAMMOGRAPHIES.			+

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						PAP SMEARS	childhood immunizations, falling short by 3–12 percent. The five plans that met their targets exceeded them on average by 9.3 percent, with individual plans exceeding it by 2–19 percent. Only four plans missed their targets for cesarean section rates, and they were only about 0.7 percent off target.		+
						PRENATAL CARE			-
Kouides et al., 1993	Low risk Annual payment (long time lag) Processes Absolute	Unclear	Individuals	Bonuses	RCT	Immunization rates	For practices in the incentive group, the mean immunization rate was 68.6% (SD 16.6%) compared with 62.7% (SD 18.07 o) in the control group practices (P = .22). The median practice-specific improvement in immunization rate was +10.3% in the incentive group compared with +3.5% in the control group (P = .03).	B	+**
St Jacques et al., 2004	low risk Monthly payment Processes Relative	Large Up to 500 dollars per month	Individuals	Bonuses	Before and after No control group N= 31 anaesthesiologists,	percentage of first cases of the day in the room at or before the scheduled in-room time	shows that the percentage of first cases of the day meeting the goal of being in the OR at or before their scheduled start time was significantly higher during the sixth month of the study (19 ± 15% vs. 61 ± 19%, p < 0.01),	B	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						percentage of cases with an anesthesia prep time less than a target	and that the percentage of cases meeting the goal of an anesthesia preparation time of less than 15 minutes increased over the study period (57 ± 18 vs. 73 ± 14, p < 0.01).		+**
						percentage of cases delayed due to waiting for an anesthesiology patient evaluation	delays from waiting for an anesthesia attending were not significantly changed, whereas delays from lengthy anesthesia preparation or emergence time were decreased (14 ± 9 vs. 3 ± 3, p < 0.01) during the study period.		+**
Salize et al., 2009	High risk Payment after a year Outcomes (quit rate) Absolute	financial incentive of (€130)	Individuals	Bonuses	Cluster-randomised smoking cessation trial. Main outcome was cost-effectiveness but abstinence rates also compared with mixed logistic regression	Smoking cessation	The TI intervention was not effective compared with TAU. The point prevalence of abstinence at 12 months was 3.5% vs 2.7%, OR 1.29, 95% CI 0.25 to 6.84, p=0.75	C	0*
McMenamin et al., 2003	Low risk Process Absolute	Not reported	Groups	Bonuses	Cross-sectional survey Control group	Numbers of HMOs providing smoking cessation advice and	OR 3.63 (95% CI 1.70 to 7.76, p<0.001), providing NRT starter kit OR 2.75 (95% CI 1.33 to 5.65,	A	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
						other interventions such as self help materials and NRT	p=0.006), providing written materials: on pharmacotherapy OR 2.13 (95% CI 1.04 to 4.33, p=0.034), counselling OR 3.11 (95% CI 1.50 to 6.44, p=0.002), self-help OR 2.33 (95% CI 0.93 to 5.84)		
Chee et al, 2007 GAVI Incentives for national governments	Low risk Time lag not clear Processes Absolute measure	Up to 15% increased immunization funding (large)	National government: institutions /groups	Bonus	the evaluators utilized a regression model for 52 countries that received ISS funds from 1995 to 2005 and in-depth qualitative studies in six countries (3 matched pairs of countries with similar circumstances and starting baseline coverage and different results).		A relationship was found between ISS funding and in- creased immunization coverage.	A	+

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Eichler et al., 2007 Haiti: RBF for NGO	Low risk Quarterly payments Processes Absolute measure	Up to 15% of previous budget of NGO (large)	NGOs: groups/institutions	Bonus	Before and after with no control group	Immunization coverage for children	6.2%	A	+
						Percentage of pregnant women receiving at least 3 prenatal care visits	2.2%		+
						Percentage of deliveries assisted by a trained attendant	3%		+
						Percentage of women receiving a postnatal care visit	7.8%		+
CORT 2007	Low risk Payment every three months Processes Absolute measure	$4.94 to $34.58 (large as per Indian standards)	health professionals (ASHA's) (individuals)	Bonuses	The program was evaluated using a mix of quantitative (survey) and qualitative (interviews) methods Before and after with no control group	Institutional deliveries	The proportion of institutional deliveries increased from 32.5% to 65.1% and the number of institutional deliveries in the public sector in Rajasthan state increased by 36% the year after the JSY was established compared to a slight decrease (−0.25%) the previous year (	A	+

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Armour et al., 2004	Low risk End of the year payments Processes Absolute measure	Size unknown	Individuals	Bonuses	Before and after: no control group.	Cancer screening	Results: From 2000 to 2001, CRC screening use increased from 23.4% to 26.4% (P < .01). Results from the multivariate logistic regression analysis revealed that the probability that a patient received a CRC screening was approximately 3 percentage points higher in the bonus year, 2001 (P < .01).	Unknown	+**
Chen et al., 2010	Low risk Annually Processes Absolute	Large Up to 7.5% of salary	Individuals	Bonuses	Longitudinal study with control groups	Diabetes care	Patients with diabetes who saw P4Pparticipating physicians were more likely to receive quality care than those who did not (odds ratio, 1.16; 95% confidence interval, 1.11-1.22; P <.001).	B	+**
							Patients with diabetes who received quality care were less likely to be hospitalized than those who did not (incident rate ratio, 0.80; 95% confidence interval, 0.80-0.85; P <.001).		+**
							During 1 year, there was no difference in hospitalization rates between patients with diabetes who saw P4P-participating physicians versus those who did not.		0*

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
							However, patients with diabetes who saw P4P-participating physicians in 3 consecutive years were less likely to be hospitalized than those who did not (incident rate ratio, 0.75; 95% confidence interval, 0.61-0.93; P <.01).		+**
Greene et al., 2004	High Yearly Process and outcomes Relative	Large Up to 20% of capitation fees	Individuals	Withholds Fines	Before and after with control group Stated that they had used a historical control but reported results for before and after studies N= approximately 900 credentialed primary care physicians as of December 1999, October 2000,	Proper hospital care	A statistical process control chart showed a shift toward recommended treatment patterns after our intervention. The rate of exceptions per episode of acute sinusitis decreased 20%, from 326 exceptions per 1000 episodes between January 1, 1999, and October 31, 2000, to 261 between November 1, 2000, and December 31, 2001. P < .005.	C	+**
							Decreased use of less effective or inappropriate antibiotics accounted for most of the change (199 to 136 exceptions per 1000 episodes [32% change]). Azithromycin use decreased 30%, from 97 to 68 prescriptions per 1000 episodes. P < .005.		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
					and December 2001.		Firstline antibiotic (amoxicillin and doxycycline) use increased 14%, from 451 to 514 prescriptions per 1000 episodes.		+**
							Inappropriate radiology use decreased 20%, from 15 to 12 per 1000 episodes. These changes were significant at P < .005.		+**
Bardach et al., 2014	Low risk Unclear timing of payment Processes Absolute measure	large	Groups	bonuses	Rct Participating practices (n=42 for each group) had similar baseline characteristics, with a mean (median) of 4592 (2500) patients at the incentive group practices and	Aspirin therapy, with IVD or DM	Odds ratio 1.28 (1.10 to 1.50) Pvalue= .001	A	+**
						Blood pressure controlNo IVD or DM	1.23 (1.05 to 1.44) Pvalue=.01		+**
						Blood pressure control IVD	0.71 (0.40 to 1.24) Pvalue=0.23		0*
						Blood pressure control DM	1.52 (1.12 to 2.07) Pvalue=.007		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
					3042 (2000) at the control group practices.	Blood pressure control IVD or DM	1.37 (1.07 to 1.75) Pvalue=.01		+**
						Cholesterol control	0.86 (0.67 to 1.09) Pvalue=.22		0*
						Smoking cessation intervention	1.30 (1.04 to 1.63) Pvalue= .02		+**
Bischoff et al, 2012	Low risk Payment after a year Processes Absolute	Unclear	Groups	Bonuses	Before and after No control group N=123 residents	Completion of discharge summary	With implementation of the bundle, the average time from patient discharge to completion of the discharge summary fell from 3.5 to 0.61 days (p<0.001).	A	+**
						Percentage of summaries completed on day of discharge	The percentage of summaries completed on the day of discharge rose from 38% to 83% (p<0.001)		+**
						The percentage of summaries that included all recommended elements	The percentage of summaries that included all recommended elements increased from 5% to 88% (p<0.001).		+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
Boland et al., 2010	Low risk Payment at 6 months intervals Processes Absolute measure e	Up to $5000 annually Large	Individuals	Bonuses	Before and after no control group N=81 radiologist	Radiologist report turnaround time	The mean C–F times for all radiologists significantly decreased from the baseline (42.7 hours) to the immediate period (31.6 hours) to the post period (16.3 hours) ($p < 0.0001$).	B	+**
							Similarly the mean C–P time also declined for all three periods from 20.0 hours at baseline to 19.0 hours at the immediate period to 11.9 hours during the post period ($p < 0.0001$).		+**
Kruse et al., 2013	Low risk Payment after 2 years Processes Absolute	Large Approximately 5%	Groups	Bonuses	Before and after with control group	Smoking status documentation	Documentation increased from 48% of 207,471 patients before P4P to 71% of 227,574 patients after P4P. Improvement occurred both among P4P-eligible patients, 56% to 83% (AOR, 3.6; 95% CI, 2.9 to 4.5) and the comparable subset of non-P4P-eligible patients, 56% to 80% (AOR, 3.0; 95% CI, 2.3 to 3.9). The difference in improvement between groups was significant (AOR, 1.3; 95% CI, 1.1 to 1.4, p=0.009).	A	+**
Peabody et al.,	Low risk Payment date	Large approxi	Groups and individuals	Bonuses	Controlled trial N = 10 for both	Composite scores of about 4 process measures	at thirty-six months after the intervention, bonus sites were 9.7	A	+**

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

Program	Perceived risk	Incentive size	Who receives the incentive	Fines or bonuses	Author/Evaluation design	Objectives /clinical area	Results Effect size	Types	Effect
2011	no known Process Absolute	mately 5% of clinicians salary			populations		percentage points higher than baseline ($p < 0:001$).		

Types (C: low, B: med, A: high); Effect (+=Positive impact but not statistically significant, + statistically significant positive effect , 0**=no statistically significant effect, -** statistically significant negative effect)**

D2: Formulas and calculations used to convert effect estimates of P4P to standardized mean difference

Formulas

Conversion from percentage or number of events to odds ratio

Where sample size (N) and percentages or number of events were reported, I was able to estimate odds ratio (OR) and associated standard errors (SE), using the formulas below:

$OR = (N_{ei} / N_i - N_{ei}) / (N_{ec} / N_c - N_{ec})$
Where:

- N_{ei} = number of events in intervention group
- N_i = total sample size in intervention group
- N_{ec}= number of events in control group
- N_c= total sample size in control group

$SE\ (logOR) = \sqrt{\{ (1/ N_{ei}) + (1/ Ni - Nei) + (1/N_{ec}) + (1/ N_c - N_{ec})\}}$

Conversion from odds ratio (OR) or mean difference (MD) to standardized mean difference (d)

$d = logoddsratio * \sqrt{3}/ \pi$

$Variance_d = Variance\ of\ log\ odds * 3/ \pi^2$

d= mean difference/SD

$SE_d = \sqrt{Variance_d}$

$SE_d = SE*\sqrt{3}/ \pi$

Combining effect sizes for multiple outcomes within a study

Summary effect for two outcomes in a study

$$\overline{Y} = \frac{1}{2}(Y_1 + Y_2).$$

Variance

$$V_{\overline{Y}} = \frac{1}{2}V(1 + r).$$

Or

Summary effect for more the two outcomes in a study

$$\overline{Y} = \frac{1}{m}\left(\sum_{j}^{m} Y_j\right),$$

Variance

$$V_{\bar{Y}} = \frac{1}{m} V(1 + (m-1)r).$$

Where v= mean of all variance, r= mean of all correlations.

Variance inflation factor (VIF)= Variance * VIF

$$VIF = 1 + (m-1)r,$$

Where m is the number of outcomes and r is the correlation

Other important formulas used in the conversion

If a 95% confidence interval is available for an absolute measure of intervention effect (e.g. SMD, risk difference, rate difference), then the standard error can be calculated as

SE = (upper limit CI – lower limit CI) / 3.92.

Variance =SE^2

SE = √Variance

Standard deviation (SD) = √N * (upper CI limit-lower CI limit)/3.92 (FOR 95% CI)

SD= √N * SE

D3. List of included studies in the meta-regression analyses

1. An LC, Bluhm JH, Foldes SS, Alesci NL, Klatt CM, Center BA et al. A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. Archives of Internal Medicine 2008; 168(18):1993-1999.
2. Bardach, N. S., J. J. Wang, et al. (2013). "Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial." Jama 310(10): 1051-1059.
3. Basinga P, Gertler P, Binagwaho A, Soucat A, SturdyJ, Vermeersch C. Paying primary health facilities for performance in Rwanda. World Bank, Washington, DC, Policy research working paper 5190.
4. Beaulieu, N. D., & Horrigan, D. R. (2005). Organizational processes and quality. Putting smart money to work for quality improvement. *HSR: Health Services Research, 40,* 1318-1334.
5. Bischoff, K., A. Goel, et al. (2013). "The Housestaff Incentive Program: improving the timeliness and quality of discharge summaries by engaging residents in quality improvement." BMJ Qual Saf 22(9): 768-774.
6. Boland, G. W., E. F. Halpern, et al. (2010). "Radiologist report turnaround time: impact of pay-for-performance measures." AJR Am J Roentgenol 195(3): 707-711.
7. Calvert M, Shankar A, McManus RJ, Lester H, Freemantle N. Effect of the quality and outcomes framework on diabetes care in the United Kingdom: retrospective cohort study. BMJ 2009; 338:b1870.
8. Campbell S, Reeves D, Kontopantelis E, et al. Quality of primary care in England with the introduction of pay for performance. N Engl J Med 2007;357:181e90.
9. Campbell SM, Reeves D, Kontopantelis E, et al. Effects of pay for performance on the quality of primary care in England. N Engl J Med 2009;361:368e78.
10. Chang FC, Hu TW, Lin M, et al. Effects of financing smoking cessation outpatient services in Taiwan. Tob Control 2008;17:183e9.
11. Chen, J. Y., H. Tian, et al. (2010). "The effect of a PPO pay-for-performance program on patients with diabetes." Am J Manag Care 16(1): e11-19.
12. Chien, A. T., Z. Li, et al. (2010). "Improving timely childhood immunizations through pay for performance in Medicaid-managed care." Health Serv Res 45(6 Pt 2): 1934-1947.
13. Coleman T, Lewis S, Hubbard R, Smith C. Impact of contractual financial incentives on the ascertainment and management of smoking in primary care. Addiction 2007; 102(5):803-808.
14. Cupples ME, Byrne MC, Smith SM, et al. Secondary prevention of cardiovascular disease in different primary healthcare systems with and without pay-forperformance. Heart 2008;94:1594e600.
15. Cutler TW, Palmieri J, Khalsa M, Stebbins M: Evaluation of the relationship between a chronic disease care management program and California pay-for-performance diabetes care cholesterol measures in one medical group. Journal of Managed Care Pharmacy 2007, 13(7):578-588.
16. Fagan PJ, Schuster AB, Boyd C, Marsteller JA, Griswold M, Murphy SM, et al. Chronic care improvement in primary care: evaluation of an integrated pay-for-performance and practice-based care coordination program among elderly patients with diabetes. Health Serv Res. 2010;45:1763-82. [PMID:20849553]
17. Gilmore AS, Zhao YX, Kang N, Ryskina KL, Legorreta AP, Taira DA et al. Patient outcomes and evidence-based medicine in a preferred provider organization setting: A six-year evaluation of a physician pay-for-performance program. Health Services Research 2007; 42(6):2140-2159.
18. Glickman SW, Ou FS, Delong ER, Roe MT, Lytle BL, Mulgund J et al. Pay for performance, quality of care, and outcomes in acute myocardial infarction. Jama-Journal of the American Medical Association 2007; 297(21):2373-2380.
19. Greene RA, Beckman H, Chamberlain J, Partridge G, Miller M, Burden D et al. Increasing adherence to a community-based guideline for acute sinusitis through education, physician profiling, and financial incentives. Am J Managed Care 2004; 10(10):670-678.
20. Hippisley-Cox, J., Vinogradova, Y., and Coupland, C. Final report for the Information Centre for Health and Social Care: time series analysis for 2001-2006 for selected clinical indicators from the QOF. http://www.qresearch.org/Public_Documents/Time%20Series%Analysis%20for%20selected%20clinical.pdf.
21. Jha, A. K., K. E. Joynt, et al. (2012). "The Long-Term Effect of Premier Pay for Performance on Patient Outcomes." New England Journal of Medicine 366(17): 1606-1615.

22. Kontopantelis, E., D. Reeves, et al. (2012). "Recorded quality of primary care for patients with diabetes in England before and after the introduction of a financial incentive scheme: a longitudinal observational study." BMJ Quality & Safety.
23. Kouides RW, Bennett NM, Lewis B, Cappuccio JD, Barker WH, LaForce FM. Performance-based physician reimbursement and influenza immunization rates in the elderly. American Journal of Preventive Medicine 1998; 14(2):89-95.
24. Kruse, G. R., Y. Chang, et al. (2013). "Healthcare system effects of pay-for-performance for smoking status documentation." Am J Manag Care 19(7): 554-561.
25. Kuo, R. N. C., Chung, K.-P., & Lai, M.-S. (2011). Effect of the pay-for-performance program for breast cancer care in Taiwan. American Journal of Managed Care, 17(5 Spec No.), e203-e211.
26. Lee, T.-T., Cheng, S.-H., Chen, C.-C., & Lai, M.-S. (2010). A pay-for-performance program for diabetes care in Taiwan: A preliminary assessment. American Journal of Managed Care, 16, 65-69.
27. Li, J., Hurley, J., DeCicca, P., & Buckley, G. (2010). Physician response to pay-for-performance—Evidence from a natural experiment. Hamilton, Ontario, Canada: McMaster University.
28. Li, Y.-H., Tsai, W.-C., Khan, M., Yang, W.-T., Lee, T.-F., Wu, Y.-C., & Kung, P.-T. (2010).The effects of pay-for-performance on tuberculosis treatment in Taiwan. Health Policy and Planning, 25, 334-341.
29. Lindenauer PK, Remus D, Roman S, Rothberg MB, Benjamin EM, Ma A et al. Public reporting and pay for performance in hospital quality improvement. New England Journal of Medicine 2007; 356(5):486-496.
30. McMenamin SB, Schauffler HH, Shortell SM, et al. Support for smoking cessation interventions in physician organizations: results from a national study. Med Care 2003;41:1396e406.
31. Salize HJ, Merkel S, Reinhard I, et al. Cost-effective primary care-based strategies to improve smoking cessation: more value for money. Arch Intern Med 2009;169:230e5; discussion 35e6.
32. Simpson CR, Hannaford PC, Ritchie LD, Sheikh A, Williams D. Impact of the pay-for-performance contract and the management of hypertension in Scottish primary care: a 6-year population-based repeated cross-sectional study. Br J Gen Pract. 2011;61:e443-51. [PMID:21722469]
33. Simpson CR, Hippisley-Cox J, Sheikh A. Trends in the epidemiology of smoking recorded in UK general practice. Br J Gen Pract 2010;60:e121e7.
34. Srirangalingam U, Sahathevan SK, Lasker SS, Chowdhury TA. Changing pattern of referral to a diabetes clinic following implementation of the new UK GP contract. British Journal of General Practice 2006; 56(529):624-626.
35. St Jacques PJ, Patel N, Higgins MS. Improving anesthesiologist performance through profiling and incentives. J Clin Anesth. 2004;16:523-8. [PMID:15590256
36. Sutton, M., S. Nikolova, et al. (2012). "Reduced mortality with hospital pay for performance in England." N Engl J Med 367(19): 1821-1828.
37. Twardella D, Brenner H. Effects of practitioner education, practitioner payment and reimbursement of patients' drug costs on smoking cessation in primary care: a cluster randomised trial. Tobacco Control 2007; 16(1):15-21.
38. Vaghela P, Ashworth M, Schofield P, Gulliford MC. Population intermediate outcomes of diabetes under pay for performance incentives in England from 2004 to 2008. Diabetes Care 2008.

D4: Extraction of raw numbers used in the meta-analyses and meta-regression

Program	Study	Effect type	Outcome	Intervention Data	Control Data	Reported effect size	LCI	UCI	d Standardized mean difference)	V_d Standardized variance)	SE_d Standardized standard error
Kouides et al 1998	Kouides et al 1998	% Change	Immunization rates in the elderly	The mean immunization rate was 68.6% (SD 16.6%) N=53	62.7% (SD 18.0%) in the control group practices (P = .22). N=82				0.197		0.243
An et al., 2008	An et al., 2008	%Change	Smoking cessation referral rates	11.4% (95% CI, 8.0%-14.9%) N=25	4.2% (95% CI, 1.5%-6.9%) N=24				0.059		0.089
Premier program	Glikman et al., 2007	Odds ratio	CMS composite measure	0.91 (95% CI 0.84-0.99) N=54	0.97 (95% CI 0.94-0.99) N=446				-0.015		0.022
California P4P	Cutler et al., 2007	% Change	Diabetes care ldl test	72.8% N=165	55.7% N=1694				0.180		0.100
	Rosenthal et al., 2005	Mean difference	Cervical screening		N=300	3.6			0.115	0.003	0.058
			Mammography			1.7			0.065	0.003 r estimated at 0.5 V_d= 0.001	0.058 SE_d =0.032

Program	Study	Effect type	Outcome	Intervention Data	Control Data	Reported effect size	LCI	UCI	d Standardized mean difference)	V_d Standardized variance)	SE_d Standardized standard error
St Jacques, et al, 2004	St Jacques, et al, 2004	% change	percentage of first cases of the day in the room at or before the scheduled in-room time	61 ± 19%, (SD) ±6.5% (CI) N-1439	19 ± 15% (SD) ±4.5% (CI) N= 1261				0.454	0.002	0.049
			percentage of cases with an anesthesia prep time less than a target	73 ± 14% (SD) ±5.1% (CI) N-1439	57 ± 18% (SD) ±5.3% (CI) N= 1261				0.171	0.002	0.045
			percentage of cases delayed due to waiting for an anesthesiology patient evaluation	3 ± 3% (SD) ±1% (CI) N-1439	14 ± 9%% (SD) ±2.9% (CI) N= 1261				0.399 Dtotal= 0.341	0.009 r= 0.75 $V_{d\ total}$ =0.0008	0.096 SE_d= 0.029

Program	Study	Effect type	Outcome	Intervention Data	Control Data	Reported effect size	LCI	UCI	d Standardized mean difference)	V_d Standardized variance)	SE_d Standardized standard error
Bischoff et al, 2012	Bischoff et al, 2012	(%) Before and after data	Percentage of summaries completed on day of discharge	38% N=563	83% N=2560				0.497	0.003	0.056
			Inclusion of all recommended elements on summary	5% N=80	88% N=80				1.03 Dtotal=0.76	0.101 VD= 0.077	0.318 0.227
National Health Insurance P4P (NHI-P4P) Taiwan	Lee et al., 2010	Mean difference	Essential diabetes exams and tests	All patients in the P4P program (n = 12,499).	Comparison group (n = 26,172)	2.450			0.655		0.005
Rwanda PBF program	Basinga et al., 2011	Mean difference	Any prenatal care	N=80	N=86	0·002	−0•021	0•025	0.013	0.006	0.079
			Four or more prenatal care visits			0·008	−0•063	0•079	0.017	0.005	0.077
			Institutional delivery			0·081	0·015	0·146	0.035	0.005	0.077

Program	Study	Effect type	Outcome	Intervention Data	Control Data	Reported effect size	LCI	UCI	d Standardized mean difference)	V_d Standardized variance)	SE_d Standardized standard error
			Tetanus vaccine during prenatal visit			0·051	−0·002	0·103	0.148	0.006	0.078
			Standardized total quality score			0·157	0·026	0·289	0.188	0.006	0.078
			Younger than 23 months preventive visit, previous 4 weeks			0·119	0·041	0·198	0.243	0.006	0.078
			24–59 months preventive visit, previous 4 weeks			0·111	0·059	0·162	0.178	0.006	.079
			12–23 months fully immunized			−0·055	−0·184	0·074	-0.065 d=0.095	0.006 r=0.5 0.002	0.078 0.041
QOF	Campbell et al.,	Mean difference	Coronary heart disease			-0.250 n=42	-0.401	0.100	-0.302	0.024	.155

Program	**Study**	**Effect type**	**Outcome**	**Intervention Data**	**Control Data**	**Reported effect size**	**LCI**	**UCI**	**d Standardized mean difference)**	**V_d Standardized variance)**	**SE_d Standardized standard error**
	2009										
			Asthma			-0.468 n=42	-0.748	0.187	. -0.302	0.024	.154
			Diabetes			-0.220 n=42	-0.313	-0.127	-0.717	0.023	0.153
			Continuity of care			0.091 n=42	0.025	0.157	0.413 d=-0.227	0.023 r=0.5 0.053	0.153 0.229
AQ	Sutton et al, 2012	Percentage points	30 day Mortality for CABG and other heart related diseases	N 134435 Percentage change -1.8%	N 722139 Percentage change -0.9%	1.3	0.4	2.1	0.166		0.013
Premier	Jha et al., 2012	Percentage points	30 day Mortality for CABG and other heart related diseases	11.82% N= 137287	11.74% Control=1094 034	0.08	−0.30	0.46	0.002		0.005
Premier	Lindenauer et al.,	Percentage points	Composite measure of	N= 116613	N=192381	4.3	3.0	5.7	0.155		0.008

Program	Study	Effect type	Outcome	Intervention Data	Control Data	Reported effect size	LCI	UCI	d Standardized mean difference)	V_d Standardized variance)	SE_d Standardized standard error
	2007		process indicators								
QOF	Kontopantelis et al. 2012	Percentage points	Composite quality score on diabetes in the first year	67.3%	Ntotal= 23,780 60%	7.3	6.7	8.0	0.270		0.011
QOF Before and after QOF design	Simpson et al., 2011	OR	Blood pressure below target <150/90	1.11 (1.04 to 1.19)	0.74 (0.67 to 0.82)	N=315			0.097		0.030
QOF	Srirangalingam et al., 2006	Percentage points	Number with HbA1c >7.4% (%)	No (%) 32, 296 9.7% 0.031 0.003	No (%) 34, 285 10.6% 0.029 0.004	0.9	- 0.4,	1.3	1.104 0.259 d=0.024		0.143
QOF	Cupples et al., 2014	Percentage points	Smoking status documentation	No (%)76 (16.9) N=449 76/449-76=0.204 0.013, 0.003	N (%) 40 (13.4) N=299 40/229-40=0.212 0.025 0.005	3.5	-1.8	8.6	OR=0.962. Se= 0.214 D= -0.009		0.118
QOF	Vaghela et al,	Percentage points	Diabetes outcome	N =2087478 N reaching	N =1764063 N reaching				0.086	0.001^2	0.001

Program	Study	Effect type	Outcome	Intervention Data	Control Data	Reported effect size	LCI	UCI	d Standardized mean difference)	V_d Standardized variance)	SE_d Standardized standard error
	2008		target A1C <or=7.5%,	target=1186695	target=845522						
			Blood pressure <or=145/85 mmHg	N =2087478 N reaching target=1518780	N =1764063 N reaching target=1064995				0.134	0.001^2	0.001
			Cholesterol <or=5 mmol/l was determined	N =2087478 N reaching target=1545301	N =1764063 N reaching target=1092954	3.99	3.92	4.07	0.134 d=0.118	0.001^2 0.000002	0.001 0.001
National Health Insurance P4P (NHI-P4P) Taiwan	Chang et al, 2008	OR	Smoking cessation	N= 3446	N=1823	0.96	0.87	1.06	SMD= -0.010 SE of log odds=0.048		SE (d)= 0.026
National Health Insurance P4P (NHI-P4P) Taiwan	Kuo et al., 2011	OR (control vs intervention) And mean in intervention and	Quality of care of breast cancer (enroless vs non enrollees)	0.70 N= 4,528 patients in total	0.63	0.062	0.050	0.074	0.664		0.003

Program	**Study**	**Effect type**	**Outcome**	**Intervention Data**	**Control Data**	**Reported effect size**	**LCI**	**UCI**	**d Standardized mean difference)**	**V_d Standardized variance)**	**SE_d Standardized standard error**
		control									
National Health Insurance P4P (NHI-P4P) Taiwan	Li et al., 2010	OR	TB cure rate in the first 12 months	N= 25754	N= 33536	1.338	1.159	1.544	**0.070** SE=0.098		0.054
Hawaii medical group	Gilmore et al., 2007	OR	Recommended care (a composite score from 11 indicators)	N was not reported		1.27	1.09	1.40	**0.057** **SE = 0.079**		0.044
Hudson health plan	Chien et al., 2010	OR	Childhood vaccinations	N=155	N=16	1.65			**0.120** SE= 0.24		0.132
McMenamin et al, 2003	McMenamin et al, 2003	OR	Smoking cessation advise			3.63 N=1104	1.7	7.76	0.309 SE= 1.546		0.852
Salize et al 2009	Salize et al 2009	OR	Smoking abstinence	N=20 We might need patient sample	N=21	1.28	0.25	6.48	0.059 SE= 1.589		0.876

Program	**Study**	**Effect type**	**Outcome**	**Intervention Data**	**Control Data**	**Reported effect size**	**LCI**	**UCI**	**d Standardized mean difference)**	**V_d Standardized variance)**	**SE_d Standardized standard error**
				here							
Twardella and Brenner, 2007	Twardella and Brenner, 2007	OR	Smoking cessation		Participants: 577 patients in 82 practices	1.26 N=557	0.65	2.43	0.055 SE= 0.454		0.250
Kruse et al., 2013	Kruse et al., 2013	OR	Smoking status Documentation	N =227574	N 207,471	1.3	1.1	1.4	0.062 SE= 0.077		0.042
Chen et al., 2010	Chen et al., 2010	OR	Quality of diabetes care	19,193	32,365	1.16	1.11	1.22	0.035 SE= 0.028		0.015
QOF	Coleman et al., 2007	OR	Brief advise to smokers	No N		3.03	2.89	3.09	0.265 SE= 0.051		0.028
QOF	Calvert et al., 2009	OR	HbA1c levels of ≤7.5%		N=147	1.05	1.01	1.09	0.011 SE= 0.020		0.011
QOF	Simpson	OR	Smoking	Total N= 525		4.45	4.43	4.46	0.357	0.004^2	0.004

Program	Study	Effect type	Outcome	Intervention Data	Control Data	Reported effect size	LCI	UCI	d Standardized mean difference)	V_d Standardized variance)	SE_d Standardized standard error
	et al., 2010		status reporting	R=0.75							
			Smoking cessation advise			6.75	6.66	6.85	0.457	0.026^2	0.026
			Smoking cessation referral			7.32	6.92	7.73	0.467	0.013	0.114
			Quit rates			0.73	0.72	0.73	-.075 d=0.3015	0.001^2 r= 0.75 0.013	0.001 0.115
Bardach et al, 2014 Fagan et al., 2010 .	Bardach et al, 2014	OR	Aspirin therapy, with IVD or DM	N=42 R=0.75	N=42	1.28	1.10	1.50	0.059	0.003	0.056
			Blood pressure control No IVD or DM			1.23	1.05	1.44	0.050	0.003	0.055
			Blood pressure control IVD			0.71	0.4	1.24	-0.82	0.014	0.118
			Blood			1.52	1.12	2.07	0.100	0.018	0.134

Program	Study	Effect type	Outcome	Intervention Data	Control Data	Reported effect size	LCI	UCI	d Standardized mean difference)	V_d Standardized variance)	SE_d Standardized standard error
			pressure control DM								
			Blood pressure control IVD or DM			1.37	1.07	1.75	0.075	0.009	0.096
			Cholesterol control			0.86	0.67	1.09	-0.36	0.003	0.059
			Smoking cessation intervention			1.30	1.04	1.63	0.063 mean d=-0.119	0.007 0.033	0.083 0.183
	Fagan et al., 2010	OR	Influenza vaccine	N= 1587 Around diabetes= 0.75	N=19356	1.79	1.37	2.35	0.139	0.019	0.138
			Hemoglobin testing			0.44	0.33	0.65	-0.196	0.031	0.176
			Eye exam			0.98	0.61	1.58	-0.005	0.285	0.534
			Ldl test			0.62	0.44	0.86	-0.114	0.053	0.231
			Nephropathy test			0.96	0.62	1.46	-.010 dtotal= -0.0372	0.184 r=0.75 0.366	0.429 0.605
Gavagan et al, 2010	Gavagan et al, 2010	0R	Pap smears						0.162	0.043	0.208
			Mammograms						0.093	0.096	0.309

Program	Study	Effect type	Outcome	Intervention Data	Control Data	Reported effect size	LCI	UCI	d Standardized mean difference)	V_d Standardized variance)	SE_d Standardized standard error
			Pediatric immunization						0.426 r=0.5 dtotal= 0.187	0.721 0.382	0.849 0.618
Larsen et al , 2003	Larsen et al , 2003	% change	Diabetes care	N=9436 52.85	N= 5785 33.5%				0.190		0.019
Tsai et al., 2010	Tsai et al., 2010	% change	Tb treatment	N= 16434 89.96% no default in treatment	N= 638 87.30% no default in treatment				0.047		0.067

D5. Estimates of effect sizes, standard errors, and study characteristics used in Meta-regression analyses

Study name	Outcome	Effect size (standardized mean difference: d)	SE_d	Who revives the incentive	Size of incentive	Risk	TYPE	Control group
1. Kouides et al 1998	Immunization rates in the elderly	0.197	0.243	Groups	Large	Low	A	Yes
2. An et al., 2008	Smoking cessation referral rates	0.059	0.089	Groups	Small	Low	B	Yes
3. Glikman et al., 2007	CMS composite measure on mortality	-0.015	0.022	Groups	Small	High	C	Yes
4. Cutler et al., 2007	Diabetes care ldl test	0.180	0.100	Groups	Large	High	B	Yes
5. Rosenthal et al., 2005	Summary of cancer screening	0.09	0.032	Groups	Large	High	B	Yes
6. St Jacques, et al, 2004	Summary measure for anesthesia care	0.341	0.029	Individuals	Large	Low	B	No
7. Bischoff et al, 2012	Summary of discharge processes	0.76	0.277	Groups	Large	Low	C	No
8. Lee et al., 2010	Essential diabetes exams and tests	0.655	0.005	Groups	Large	High	B	Yes
9. Basinga et al., 2011	Summary measure for maternal health indicators	0.095	0.041	Groups	Large	Low	A	Yes
10. Campbell et al., 2009	Summary measure for diabetes process indicators	-0.227	0.229	Groups	Large	Low	A	Yes
11. Sutton et al, 2012	30 day Mortality for CABG and other heart related diseases	0.016	0.013	Groups	Small	High	C	Yes
12. Jha et al., 2012	30 day Mortality for CABG and other heart related diseases	0.002	0.005	Groups	Small	High	C	Yes

Study name	Outcome	Effect size (standardized mean difference: d)	SE_d	Who revives the incentive	Size of incentive	Risk	TYPE	Control group
13. Lindenauer et al., 2007	Composite measure of process indicators	0.155	0.008	Groups	Small	High	C	Yes
14. Kontopantelis et al. 2012	Composite quality score on diabetes in the first year	0.270	0.011	Groups	Large	Low	A	Yes
15. Simpson et al., 2011	Summary measure for improvement in hypertension	0.097	0.030	Groups	Large	Low	A	No
16. Srirangalingam et al., 2006	Number with HbA1c >7.4% (%)	0.024	0.143	Groups	Large	Low	A	No
17. Cupples et al., 2014	Smoking status documentation	-0.009	0.118	Groups	Large	Low	A	Yes
18. Vaghela et al, 2008	Summary measure for improvements in diabetes intermediate outcomes	0.118	0.001	Groups	Large	Low	A	No
19. Chang et al, 2008	Smoking cessation	-0.010	0.026	Groups	Large	High	B	No
20. Kuo et al., 2011	Quality of care of breast cancer	0.664	0.003	Groups	Large	High	B	Yes
21. Li et al.,2010	TB cure rate in the first 12 months	0.070	0.54	Groups	Large	High	B	Yes
22. Gilmore et al., 2007	Recommended care (a composite score from 11 indicators)	0.057	0.044	Individuals	Large	High	C	Yes
23. Chien et al., 2010	Childhood vaccinations	0.120	0.132	Individuals	Large	Low	B	Yes
24. McMenamin et al, 2003	Smoking cessation advise	0.309	0.852	Groups	Large	Low	A	Yes
25. Salize et al.,	Smoking abstinence	0.059	0.876	Individuals	Small	High	C	Yes
26. Twardella and Brenner, 2007	Smoking cessation	0.055	0.250	Individuals	Small	High	C	Yes

Study name	Outcome	Effect size (standardized mean difference: d)	SE_d	Who revives the incentive	Size of incentive	Risk	TYPE	Control group
27. Kruse et al., 2013	Smoking status Documentation	0.062	0.042	Groups	Large	Low	A	Yes
28. Chen et al., 2010	Quality of diabetes care	0.035	0.015	Individuals	Large	Low	B	Yes
29. Coleman et al., 2007	Brief advise to smokers	0.265	0.028	Groups	Large	Low	A	No
30. Calvert et al., 2009	HbA1c levels of ≤7.5%	0.011	0.011	Groups	Large	Low	A	No
31. Simpson et al., 2010	Summary effect of smoking related indicators	0.3015	0.115	Groups	Large	Low	A	No
32. Bardach et al, 2014	Summary measure for heart disease process	-0.119	0.183	Groups	Large	Low	A	Yes
33. Fagan et al., 2010	Summary measure of essential tests and exams for diabetes	-0.0372	0.605	Groups	Large	Low	A	Yes
34. Larsen et al., 2003	Diabetes care	0.190	0.019	Individuals	Small	Low	C	No
35. Gavagan et al., 2010	Cancer screening	0.187	0.618	Individuals	Small	Low	C	Yes
36. Tsai et al., 2010	Tb treatment	0.042	0.067	Groups	Large	High	B	

D6. Statistical output for the multifactorial multilevel logistic regression analysis (Model B variant 2)

	Univariate	Multivariate
Type 2	3.41[0.98-11.89] p=0.054	1.72[0.63-4.69] p=0.291
Type 3	3.99[0.58-27.67] p=0.161	2.36[0.47-11.80] p=0.296
Evaluation		17.91[4.76-67.47] p<0.0001

Appendix E

EI. Detailed description of incentivized health services in the Nigerian P4P scheme

No	Name MPA Service	Description	Primary Data Collection Tools1	Secondary Data Collection Tools2
1	New outpatient consultation	Any new curative care visit during the past month	Curative Care Register	Original prescription for drugs dispensed kept at the pharmacy which includes cost of drugs. Drugs register and stock cards conform.
2	New outpatient consultation of an indigent patient	During the past month, indigents who have been consulted as an outpatients. Indigents are locally identified. Maximum of 20% of all new curative consultations during the previous month.	Indigent outpatient register	Proceedings indigent committee Community Client Satisfaction Survey: post-identification questionnaire application
3	Minor Surgery	Any new minor surgical intervention during the past month. Minor Surgery defined as (i) Suture; (ii) incision and drainage; (iii) minor excisions.	Minor Surgery Register	Original prescription for drugs and medical consumables dispensed kept at the pharmacy which includes cost of drugs/consumables. Drugs register and stock cards conform
4	Referred patient arrived at the General Hospital	Counter-referral slip available at the Health facility. Fully filled in by the MD. The number of valid counter-referral slips is counted.	Original of counter-referral slip available at the Health facility.	Copy of the counter-referral slip available at the General Hospital. Referred patient registered in the outpatient's department register.
5	Completely Vaccinated Child	Child less than 12 months old which has received all vaccines according to the national protocol (BCG; DTP3; Measles)	Vaccination Register	Under-five card with vaccination records, held by the mother.

6	Growth monitoring visit Child	Any new quarterly growth monitoring visit of a child less than five years old during the past month. These growth monitoring visits ought to be monthly according to the protocol, however, here, a quarterly visit is remunerated.	Under-five clinic/Nutrition Register	Under-five card with growth curve plotted, held by the mother
7	2 - 5 Tetanus Vaccination of Pregnant Woman	Each second to fifth TT vaccination of a pregnant woman during the past month	ANC register Individual Card kept at the HF	ANC card held by the mother Vaccination register
8	Postnatal consultation	A post natal consultation held within 48 hours after giving birth, during the past month.	Delivery register	Partogram or inpatient form
9	First ANC consultation before four months pregnancy	A first ANC consultation occurs before 4 month's pregnancy, during the past month.	ANC register Individual Card kept at the HF	ANC card held by the mother
10	ANC standard visit (2-4)	Any 2-4th standard visit according to the focused antenatal care visit schedule and approach. Second visit between 24-28 weeks; third visit at 32 weeks and the fourth visit at 36 weeks. During the past month.	ANC register Individual Card kept at the HF	ANC card held by the mother Medical prescriptions for Ferrosulphate, Vermox and SP kept at the pharmacy. Drugs register and stock cards conform
11	Second dose of SP provided to a pregnant woman	The second dose of SP (IPTp), according to the protocol, during the past month.	ANC register Individual Card kept at the HF	ANC card held by the mother; medical prescription for SP kept at the pharmacy. Drugs register and stock card conform.
12	Normal delivery	A delivery attended by a trained attendant at the health facility during the past month.	Delivery Register	Partogram; eventual drugs and medical consumables dispensed through the prescriptions kept at the pharmacy; drugs register and stock cards conform.

13	FP: total of new and existing users of modern FP methods	Any new or existing user of injectable contraceptive or oral contraceptive pills, during the past month. An injection represents three month's protection and a FP visit for OAC should provide three months' worth of pills.	FP register Individual Card kept at the HF	Eventual drugs and medical consumables dispensed through the prescriptions kept at the pharmacy; drugs register and stock cards conform.
14	FP: implants and IUDs	Any new user of implant or IUD, during the past month.	FP register Individual Card kept at the HF	Eventual drugs and medical consumables dispensed through the prescriptions kept at the pharmacy; drugs register and stock cards conform.
15	VCT/PMTCT test	Any new VCT or PMTCT test carried out during the past month.	VCT register	Laboratory register; stock records
16	PMTCT: HIV+ mothers and children born to are treated according to protocol	Any new HIV+ mother and newborn child treated according to the PMTCT protocol, during the past month.	ARV register; delivery room register	PMTCT register; laboratory register; stock records.
17	STD treated	Any new STD treated according to syndromic treatment protocol, during the past month	Curative Care Register	Drugs and medical consumables dispensed through the prescriptions kept at the pharmacy; drugs register and stock cards conform.
18	New AAFB+ PTB patient	A new AAFB sputum positive Pulmonary Tuberculosis patient diagnosed, at the facility, during the past month.	Tuberculosis register	Laboratory register. Slides kept for counter-verification/quality assurance.
19	PTB patient completed treatment and cured	A former AAFB+ PTB patient completed DOTS, and cured after treatment proven by negative sputum examinations, during the past month.	Tuberculosis register	Laboratory register. Slides kept for counter-verification/quality assurance. Drugs register.
20	ITN Distributed	ITN distributed, during the past month.	ITN register	Stock control card conform. Proof of acquisition, purchase of ITNs available.

21	New family using a latrine	During the past month, the individual effort of a community health worker who has delivered a package of behavioral change communication including on hygiene to a family. The objectively verifiable measure of his BCC is the use of a newly constructed latrine in the catchment area (Ward), next to the household of this family, during the past month. Construction according to the norms. Maximum one latrine per household.

E2. Quality checklist of the Nigerian P4P scheme explained

1	**General Management [max 11 points]**	**YES**	**NO**
1.1	**Presence of map of health facility catchment area**	**1**	**0**
1.1.1	**Health map of the health area available and on the notice board of HF showing villages, main roads, natural barriers, special points and distance**		
1.2	**HMIS reports - business plan - minutes of meetings and patient cards well stored**	**2**	**0**
1.2.1	**In cupboard and in box files and accessible by duty manager**		
1.3	**Staff duty roster available and well displayed up to date for current month and visible for staff and patients**	**1**	**0**
1.4	**Technical meetings with staff conducted monthly and minutes available**	**3**	**0**
1.4.1	**Each monthly minutes contain: (i) date of the meeting; (ii) signed list of participants; (iii) follow-up of decisions taken during the previous meeting; (iv) there is a list of developed recommendations or decisions taken; (v) each month the monthly financial balance is discussed; (vi) minutes of the meeting are signed by the chair. Each report according to norms = 1 p**		
1.5	**Standard Sheets for referral available**	**1**	**0**

1.5.1	At least 10 sheets		
1.6	**Availability of radio or mobile phone for communication between health facility and general hospital**	**1**	**0**
1.6.1	Radio or mobile phone functional with batteries and/or call credit and contact details on the phone		
1.7	**HMIS reports are filled, updated and transmitted to the LGA on schedule**	**1**	**0**
1.7.1	After verification of the SPHCDA of the monthly MPA invoice and signed receipt of acknowledgement available		
1.8	**HMIS data analysis report for the quarter being assessed concerning priority problems**	**1**	**0**
1.8.1	Three priority health problems are followed each quarter and data have been updated up to the month prior to the supervisor's visit		
	Total Points (11)	../11	xxxx

2	**Business Plan [max 9 points]**	**YES**	**NO**
2.1	**Quarterly business plan for the current period made and accessible**	**2**	**0**
2.1.1	Valid and renegotiated		
2.2	**Business plan prepared with key stakeholders**	**2**	**0**
2.2.1	Facility RBF Committee Members involved		

2.2.2	Representative (s) of subcontracted private clinics or health posts involved (if applicable)		
2.3	**Business plan contains convincing geographic coverage plan**	**1**	**0**
2.3.1	Strategies for sub-contracts (e.g. villages at more than one hour by foot)		
2.3.2	Mobile strategies (EPI, FP; PNC, LITN distribution, latrines)		
2.4	**Business plan analyses presence of untrained informal practitioners in catchment area**	**1**	**0**
2.4.1	HF treats this subject in the BP, and suggests a strategy for discouraging		
2.5	**Business plan analyses presence of trained practitioners operating without any permission**	**1**	**0**
2.5.1	BP may suggest to include them or to discourage if quality conditions are not met		
2.6	**Business plan shows a plan to assure financial accessibility for the population**	**2**	**0**
2.6.1	Business plan shows negotiated rates between HF, committee and community		
2.6.2	Business plan shows planning for care for the indigents		
	Total Points (9)	../9	xxxx

3	**Finance [max 10 points]**	**YES**	**NO**
3.1	**Financial and accounting documents available and well kept**	**2**	**0**
3.1.1	Monthly report of treasury available and correctly filled		

3.1.2	Theoretical balance of cash-book corresponds to liquidity in cash		
3.2	**Document available to show that quarterly calculation of incomes, running costs, investments and variable performance subsidies are done**	**3**	**0**
3.2.1	This document guarantees running costs: = salaries, purchase of drugs and equipments, subcontracts, petty cash fro small expenditures, social marketing, maintenance and rehabilitation		
3.2.2	This document calculates the performance bonus according to the formula: performance bonuses = income of the quarter - running costs		
3.3	**Contract salaries and benefits + performance bonuses do not exceed 50% of total HF income through PBF**	**2**	**0**
3.4	**Existence of fixed basic salaries and monthly performance bonus system is know by staff**	**3**	**0**
3.4.1	Established criteria for the performance bonus calculation through (i) basic performance index + (ii) seniority + (iii) responsibility + (iv) overtime hours worked - hours lost + (v) quarterly performance evaluation		
	Total Points (10)	../10	xxxx

N_R	Revenue Categories	Revenues	N_E	Expense Categories	Expenses
1	Cost recovery (user-charges)		9	Salaries	
2	Cost recovery (pre-payment)		10	Performance bonuses	
3	Salaries from Gov. & other sources		11	Drugs and medical consum.	

4	PBF Subsidies from fund holders		12	Subsidies for sub-contracts	
5	Contributions from other sources		13	Cleaning and office costs	
6	Other		14	Transport costs	
7	Cash in hand		15	Social marketing	
8	Bank balance at the end of the quarter		16	Infrastructure rehabilitation	
	Total Revenue		17	Equipment and furniture	
			18	Other	
			19	Put into reserve	
				Total Expenses	
				Total Revenue - Total Expenses	

4	**Care for the Indigents [max 7 points]**	**YES**	**NO**
4.1	**Planning for Care for the Indigents expenditures**	**1**	**0**
4.1.1	20% of curative consultations of the previous month: documented quantity in monthly management meetings		
4.2	**Indigent committee meets monthly**	**6**	**0**
4.2.1	The Indigent committee meets monthly to review the Care for the Indigent Category use. Each monthly minutes contain: (i) date of the meeting; (ii) signed list of participants; (iii) follow-up of decisions taken during the previous meeting; (iv) there is a list of developed recommendations or decisions taken; (v) each month the monthly financial balance is discussed; (vi) minutes of the meeting are signed by the chairman. Each report according to norms = 2 p		
	Total Points (7)	../7	xxxx

5	**Hygiene and Sterilization [max 25 points]**	**YES**	**NO**

5.1	**Fence health facility available and well-maintained**	**1**	**0**
5.1.1	Fence exists, can be closed at night and there are no holes		
5.2	**Availability of a garbage bin in the courtyard**	**1**	**0**
5.2.1	Bin with lid accessible to clients which is not full		
5.3	**Presence of sufficient latrines/toilets which are well-maintained**		
5.3.1	At least two latrines/toilets	**1**	**0**
5.3.2	Floor without fissures with single hole and lid	**0.5**	**0**
5.3.3	Recently cleaned without visible fecal matter	**0.5**	**0**
5.3.4	Door lockable from the inside, super structure with roofing, without flies and no smell	**0.5**	**0**
5.3.5	Smells of disinfectant	**0.5**	**0**
5.4	**Presence of sufficient showers which are well-maintained**		
5.4.1	At least one bathing facility	**1**	**0**
5.4.2	Bathing facility with running water, or container with at the least 20 L of water	**0.5**	**0**
5.4.3	Evacuation of the waste water in a sanitation pit	**0.5**	**0**
5.5	**Waste pit for Heath Care Waste is available and according to the norms**		
5.5.1	Waste disposal pit minimum 2 meters deep, lined with clay, concrete or brick or plastic, it is fenced and has a bright flag.	**6**	**0**
5.5.2	The waste pit is a minimum of 15 meters from the health facility, minimum of 50 meters from a household, and 100 meters from a water source		
5.5.3	Health Care Waste is not visible (covered by at the least 10 cm of soil or lime)		
5.5.4	The health facility maintains a register indicating the date of the creation of the pit(s), and the location (s)		

5.6	**Courtyard clean**	**1**	**0**
5.6.1	No waste or medical waste in the courtyard		
5.7	**Sterilization according to norms using a pressure sterilizer**	**3**	**0**
5.7.1	Sterilizer functional		
5.7.2	Sterilization protocol available and utilized		
5.9	**Hygienic conditions assured during wound dressing and injections**	**2**	**0**
5.9.1	Yellow and Red Bins for medical waste with lid and foot pedal, lined		
5.9.2	Security box for needles well positioned, and used		
5.9.3	Needle cutter available and used		
5.9.4	Container/bowl with lid containing disinfectant used for putting used instruments		
5.10	**Disposal of Health Care Waste according to National Norms**	**6**	
5.10.1	Waste disposal of non-contaminated waste in Black Bin with lid and foot pedal, lined		
5.10.2	Waste disposal of contaminated HCW in Yellow Bins with lid and foot pedal, lined		
5.10.3	Waste disposal of organically HCW in Red Bins with lid and foot pedal, lined		
5.10.4	Protective gear for personnel managing HCW available; boots, plastic shorts, thick plastic/rubber gloves		
	Total Points (25)	../25	xxxx

6	**Curative Consultations [max 34 points]**	**YES**	**NO**
6.1	**Good conditions in waiting area**	**1**	**0**
6.1.1	Sufficient benches and or chairs protected against sun and rain and waiting area is not inside room		

6.2	**Unit fees of drugs displayed to the public**	**1**	**0**
6.2.1	Easily visible in the waiting area, updated, with (i) unit price per item; (ii) price for a standard treatment of the drug		
6.2.2	Drugs are all generics		
6.3	**Existence of waiting card system with numbers**	**1**	**0**
6.4	**Consultation room in good condition**	**3**	**0**
6.4.1	Walls with durable materials well painted, floor paved with cement without fissures, undamaged ceiling		
6.4.2	Consultation room and waiting space separated assuring confidentiality		
6.4.3	Windows with curtains		
6.4.4	Functional door with lock		
6.5	**Consultation room (where emergencies are received) has 24/7 light**	**1**	**0**
6.5.1	Electricity or solar light **or** functioning high pressure kerosene light present		
6.6	**Consultations are done by skilled staff**	**2**	**0**
6.6.1	Identification of consulting staff in register		
6.7	**Consulting staff is well-dressed**	**1**	**0**
6.7.1	Clean blouse and footwear		
6.8	**Correct numbering of registers**	**1**	**0**
6.8.1	Correct numbering and closed at the end of the month		
6.9	**Service availability 7/7**	**1**	**0**
6.9.1	Supervisor verifies entries in register for the last three Sundays		

6.10	**Malaria protocol put on wall and accessible for staff**	**1**	**0**
6.10.1	National protocol for diagnosis and treatment of simple and severe malaria		
6.11	**Simple malaria correctly treated**	**1**	**0**
6.11.1	Register see last five cases of simple malaria and review treatment acc protocol		
6.12	**WHO flow diagram for ARI put on wall and accessible for staff**	**1**	**0**
6.13	**ARI protocol applied**	**1**	**0**
6.13.1	See last five cases of ARI and review treatment acc protocol		
6.14	**WHO protocol for Diarrhea put on wall and accessible for staff**	**1**	**0**
6.15	**Diarrhea protocol applied**	**1**	**0**
6.15.1	See last five cases of Diarrhea and review treatment acc protocol		
6.16	**Proportion of consultancies treated with antibiotics <30%**	**4**	**0**
6.16.1	See last 100 cases in register, check diagnosis and calculate the rate (< 30 cases)		
6.17	**MSF treatment guidelines available in consultancy room**	**1**	**0**
6.18	**Knowledge of tuberculosis danger signs and criteria for referral**	**1**	**0**
6.18.1	Select any available qualified medical staff, and ask the question on TB dangers signs		
6.18.2	Answer must contain at least 4 of the following signs: (i) weight loss; (ii) loss of appetite; (iii) fever; (iv) cough of more than 15 days duration; (v) night sweating		
6.19	**Stethoscope and BP machine available and functional**	**1**	**0**
6.19.1	Let nurse check BP and review measure		
6.20	**Thermometer available and functional**	**1**	**0**

6.21	**Otoscope available and functional**	**1**	**0**
6.22	**Examination bed available with mattress**	**1**	**0**
6.22.1	Non-torn, plastic cover, specific for the OPD consultations only		
6.23	**Weighing scale available and functional**	**1**	**0**
6.23.1	Inspect in comparison with known weight of supervisor: after weighing, the balance should return to zero		
6.24	**Integrated Management of Childhood Illnesses strategy is applied**	**2**	**0**
6.24.1	Protocol is available in the consultation room		
6.24.2	The last five IMCI cases are traced in the register and comply with the IMCI strategy		
6.25	**Determination of nutritional status**	**2**	**0**
6.25.1	Determination of nutritional status of all children under 5 who come for consultation		
6.25.2	Determination of nutritional status of all women with a sick child under 6 months of age		
6.25.3	Screening record of nutritional status available, up to date and properly filled out		
	Total Points (34)	../34	xxxx

7	**Family Planning [max 22 points]**	**YES**	**NO**
7.1	**At least one qualified staff trained in Family Planning**	**2**	**0**
7.2	**Confidentiality in consultancy room assured**	**2**	**0**
7.2.1	Room with closed doors, curtains at windows or non transparent glass		
7.3	**Family planning methods available and visible in demonstration box for potential users**	**2**	**0**
7.3.1	Condoms; OAC; Injectable; Implant; IUD; beads are available in the demonstration box		

7.3.2	Penis model available on the desk; box with condoms available with at the least 50 condoms		
7.4	**Staff correctly calculates number of clients expected monthly for oral and injectable contraceptives**	**1**	**0**
7.4.1	For example for 10.000 population (target is entire ward catchment pop) = 10.000 * 22.5% * 25%/12 * 4 * 90% (assuming 25% unmet need; 22.5% target population; 90% of oral/inject AC at HC level		
7.5	**Business plan contains strategy to achieve FP targets**	**3**	**0**
7.5.1	Collaboration with public sector, private sector and social marketing, mobile strategies, advocacy among local leaders etc		
7.5.2	Involvement of HF staff in strategies		
7.6	**Stock of oral and injectable contraceptives in adequate**	**2**	**0**
7.6.1	for example for 10.0000 pop 72 doses of oral (3 month cycles) and injectable methods combined		
7.7	**IUD available and staff trained to use it**	**3**	**0**
7.7.1	at least five IUDs and at the least one staff trained to use it		
7.8	**Implant method available and staff trained to use it**	**3**	**0**
7.8.1	at least five implants available and staff trained to use it		
7.9	**Strategies available for transfer of persons to hospital seeking permanent FP methods**	**2**	**0**
7.9.1	Referral system worked out - strategy to reduce prices; mobile strategy for surgery?		
7.10	**FP individual cards available and filled according to the format**	**2**	**0**
7.10.1	Check at least five cards for BP, hepatomegaly, varices, weight		

	Total Points (22)	../22	xxxx
8	**Laboratory [max 10 points]**	**YES**	**NO**
8.1	**Laboratory technician or technologist available**	**1**	**0**
8.2	**Laboratory is open every day of the week**	**1**	**0**
8.2.1	Supervisor verifies the last 4 Sundays in laboratory register		
8.3	**List of laboratory examinations visible for the public with fees**	**1**	**0**
8.4	**Results recorded correctly in laboratory register and match with results in inpatient sheets or OPD examination cards**	**1**	**0**
8.4.1	Supervisor verifies last five results		
8.5	**Availability of parasites demonstrations**	**1**	**0**
8.5.1	On plastic paper, in a color book, or put on wall		
8.5.2	Blood smear: Vivax, Ovale, Falciparum and Malariae		
8.5.3	Stools: Ascaris, entamboeabae, ankylostoma and schistosome		
8.6	**Microscope available and functional**	**1**	**0**
8.6.1	functional objectives; immersion oil available, mirror or electricity		
8.6.2	blades, cover glass, GIEMSA available		
8.7	**Malaria rapid tests available**	**1**	**0**
8.7.1	At the least 20 tests available in the laboratory; non-expired		
8.8	**Centrifuge available and functional**	**1**	**0**
8.9	**Waste evacuation correctly carried out**	**1**	**0**

8.9.1	Organic waste in a bin with lid with disinfectant		
8.9.2	Security box for sharp objects available and destroyed according to waste disposal guidelines		
8.10	**Personnel adequately washes dirty pipettes in containers with disinfectant**	**1**	**0**
	Total Points (10)	../10	xxxx

9	**In-patient Wards [max 10 points]**	**YES**	**NO**
9.1	**Guard duty roster clearly visible for staff and followed up**	**1**	**0**
9.1.1	Supervisor verifies guard duty's report - names and signatures		
9.2	**Furniture available and in good state**	**2**	**0**
9.2.1	Each bed has a (i) plastic covered mattress, (ii) mosquito net, (iii) clean sheets, (iv) night table		
9.3	**Patient comfort and hygiene**		
9.3.1	The wards are clean: no debris on the floor; and wards smell of disinfectant	**0.5**	**0**
9.3.2	Space between the beds is at the least one meter	**0.25**	**0**
9.3.3	Each ward has access to drinking water	**0.25**	**0**
9.4	**Light available in each ward**	**2**	**0**
9.4.1	Electricity; solar light or rechargeable battery lamp		
9.5	**Confidentiality**	**1**	**0**
9.5.1	Women in separate ward from men; the inside of the wards are not visible from the outside		
9.6	**In patient register available and is well maintained**	**2**	**0**
9.6.1	check identity and hospital bed days		
9.7	**Recording forms for hospitalizations available and well filled and well stored**	**1**	**0**
9.7.1	At least 10 blanks; supervisor verifies 5 filled forms		

9.7.2	Weight, temperature, and eventual laboratory exams recorded		
9.7.3	Treatment monitoring checked		
	Total Points (10)	../10	xxxx

10	Essential Drugs Management [max 20 points]	YES	NO
10.1	**Staff maintains stock cards for ED showing security stock levels = monthly average consumption / 2**	**4**	**0**
10.1.1	Supply in register corresponds with physical supply: random sample of three ED		
10.2	**Health facility purchases drugs, equipment and consumables from the Pharmaceutical Council of Nigeria certified distributor, approved by SMOH/SPHCDA**	**3**	**0**
10.2.1	Latest Pharmaceutical Council of Nigeria certified distribution center list for the State available		
10.2.2	Last procurement list is shown which shows the certified distributor which sold the drugs		
10.2.3	All drugs and medical consumables are (i) NAFDAC certified and (ii) Generic		
10.3	**Main pharmacy store delivers drugs to health facility departments according to requisition**	**10**	**0**
10.3.1	Supervisor verifies whether quantity requisitioned equals quantity served		
10.3.2	Drugs to clients are uniquely dispensed through prescriptions. Prescriptions are stored and accessible		
10.3.3	Drugs and medical consumables prescribed, are all in generic form		
10.4	**Drugs stored correctly**	**2**	**0**
10.4.1	Clean place, well ventilated with all drugs on cupboards, labeled shelves		
10.4.2	Drugs and medical consumables stored on alphabetical order, first in - first out basis		

10.5	**Absence of out of date drugs or drugs with unreadable labels**	**1**	**0**
10.5.1	Supervisor verifies randomly three drugs and 2 consumables		
10.5.2	Out of date drugs well separated from stock		
10.5.3	Destruction protocol for out of date drugs available and applied		
	Total Points (20)	../20	xxxx

11	**Tracer Drugs (min. stock = Monthly Av. Consumption / 2) [max 30 points]**	Available YES > MAC / 2	Available NO < MAC / 2
11.1	**Paracetamol 500 mg tab**	**1**	**0**
11.2	**Ibuprofen 200 mg caps**	**1**	**0**
11.3	**Promethazine 25 mg tab**	**1**	**0**
11.4	**Oxytocin 10IU/ml vial**	**1**	**0**
11.5	**Mebendazole 100 mg tab**	**1**	**0**
11.6	**Ferrous Sulfate 325 mg tab**	**1**	**0**
11.7	**Penicillin V 250 mg tab**	**1**	**0**
11.8	**Amoxicillin 500 mg tab**	**1**	**0**
11.9	**Amoxicillin 200 mg/5ml suspension**	**1**	**0**
11.10	**Co-trimoxazol 480 mg tab**	**1**	**0**
11.11	**Co-trimoxazol 40mg/200mg - 5ml susp**	**1**	**0**
11.12	**Doxycycline 100 mg caps**	**1**	**0**

11.13	**Erythromycin 250 mg tab**	**1**	**0**
11.14	**Co-artemeter 20/120 mg tab**	**1**	**0**
11.15	**Sulfadoxine/pyrimethamine 500 mg tab**	**1**	**0**
11.16	**ORS sachet**	**1**	**0**
11.17	**Condom**	**1**	**0**
11.18	**Metronidazol 250 mg tab**	**1**	**0**
11.19	**Sterile gloves**	**1**	**0**
11.20	**Venflon 18G**	**1**	**0**
11.20.1	Min stock = 10; MAC applies only when higher than 10		
11.21	**Venflon 22G**	**1**	**0**
11.21.1	Min stock = 10; MAC applies only when higher than 10		
11.22	**IV giving set**	**1**	**0**
11.22.1	Min stock = 10; MAC applies only when higher than 10		
11.23	**Ringers lactate 1L**	**1**	**0**
11.23.1	Min stock = 5L; MAC applies only when higher than 5L		
11.24	**Dextrose 5% 1L**	**1**	**0**
11.24.1	Min stock = 5L; MAC applies only when higher than 5L		
11.25	**IV colloids 500 ml**	**1**	**0**
11.25.1	Min stock = 5 bags; MAC applies only when higher than 5 bags		

11.26	**Syringe 5ml**	**1**	**0**
11.27	**Syringe 10ml**	**1**	**0**
11.28	**Needle 18G**	**1**	**0**
11.29	**Needle 22G**	**1**	**0**
11.30	**ITN**	**1**	**0**
	Total Points (30)	…./30	xxxx

12	**Maternity [max 21 points]**	**YES**	**NO**
12.1	**Sufficient water with soap in delivery room**	**1**	**0**
12.1.1	A functioning water source or at the least 20L		
12.2	**Light in delivery room 24 hours**	**1**	**0**
12.2.1	Electricity, solar light or rechargeable battery lamp or kerosene lamp filled with kerosene		
12.3	**Waste from Maternity correctly handled**	**1**	**0**
12.3.1	Bin with lid and safe needle disposal container, specific for the maternity room use only		
12.4	**Delivery room is well-maintained**		
12.4.1	Walls with durable materials and painted	**1**	**0**
12.4.2	Curtain between delivery bed and door	**1**	**0**
12.4.3	Delivery room smells of disinfectant	**1**	**0**
12.4.4	Floor level cement, without fissures and ceiling not damaged	**1**	**0**
12.4.5	Windows with curtains and functional door	**1**	**0**
12.5	**Availability and use of the partogramme**	**1**	**0**

12.5.1	At the least 10 forms available for use		
12.5.2	Verify three randomly selected partograms whether filled according to the norms		
12.6	**Deliveries performed by skilled personnel**	**2**	**0**
12.6.1	Identification of the obstetrician from names in the register		
12.7	**Availability of scales for weight/length, an obstetrical stethoscope and an aspirator**		
12.7.1	Tape to measure length	**1**	**0**
12.7.2	Scale to measure weight (check functionality)	**1**	**0**
12.7.3	Aspirator plunged into a non-irritating disinfectant or functional manual/electric aspirator	**1**	**0**
12.8	**Availability of at the least 10 pairs of sterile gloves**	**1**	**0**
12.9	**Availability of at the least 2 sterilized obstetrical boxes**	**2**	**0**
12.9.1	Content at the least 1 pair of scissors, 2 pliers and one needle holder		
12.10	**Availability of at the least one episiotomy box**	**1**	**0**
12.10.1	One sterilized box with needle holder, needles, 1 anatomical plier and 1 surgical plier		
12.10.2	Catgut and nylon sutures; antiseptic, local anesthetics, sterile swaps		
12.11	**Delivery table in good state**	**1**	**0**
12.11.1	Table in two parts with removable non-torn plasticized mattress and two functional leg supports		
12.12	**Available equipment for care of the newborn**	**1**	**0**
12.12.1	Sterile tying string or clip for umbilical cord		
12.12.2	1% tetracycline eye ointment		
12.13	**Adequate in-patient rooms**	**1**	**0**
12.13.1	Mattress covered in impermeable plastic		

12.13.2	Sheets, blankets and mosquito nets on each occupied bed		
	Total Points (21)	../21	xxxx

13	**EPI and Pre-School Consultation [max 18 points]**	**YES**	**NO**
13.1	**Personnel calculates correctly target for fully vaccinated children**	**1**	**0**
13.1.1	Target = population * 4% / 12		
13.1.2	The target population concerns the ward population		
13.2	**EPI fridge**	**3**	**0**
13.2.1	Presence of a fridge - temp form available, filled twice a day including the day of the visit		
13.2.2	Temperature remains between 2 and 8C in register sheet		
13.2.4	Supervisor verifies functionality of thermometer		
13.2.5	Temperature between 2 and 8C also according to the thermometer		
13.3	**Chemical Temperature Indicator**	**1**	**0**
13.3.1	Presence of a chemical temperature indicator which shows temperature acc to the norms		
13.4	**Appropriate storage of vaccines**	**1**	**0**
13.4.1	Freezing compartment: Measles		
13.4.2	Non-freezing compartment: BCG, DTP + HepB, TT, thinners		
13.4.3	Absence of vaccines which are expired		
13.4.4	Readable labels on all vaccines		
13.5	**Appropriate stock of vaccines**	**1**	**0**
13.5.1	BCG, DPT, Polio, Yellow Fever, HBV, Measles, Tetanus		
13.5.2	Presence of stock control cards for all vaccines; concordance paper and physical stock verified		

13.6	**Cold Chain maintenance**	**1**	**0**
13.6.1	If kerosene fridge: stock of at the least 14L Kerosene; if solar fridge: battery not damaged		
13.7	**Cold packs are well frozen**	**1**	**0**
13.7.1	At the least 5		
13.8	**Syringes available**	**1**	**0**
13.8.1	Auto-blocking at least 30; for dilution - at least 3		
13.9	**Waste collection availability of safe disposal box**	**1**	**0**
13.10	**Stock of U5 growth cards available**	**1**	**0**
13.10.1	At the least 10		
13.11	**Child immunization register well maintained**	**1**	**0**
13.11.1	System is capable of identifying drop outs and Fully Vaccinated Children		
13.12	**Conditions in waiting area for immunization services**	**1**	**0**
13.12.1	Sufficient benches and or chairs, protected against sun and rain		
13.13	**Patients receive numbered waiting buttons according to their arrival**	**1**	**0**
13.14	**Baby weighing scale available and in working condition**	**1**	**0**
13.14.1	Balance calibrated to zero + pants available, clean and in good condition		
13.15	**Group IEC/BCC**	**1**	**0**
13.15.1	Group meeting held before vaccinations		
13.15.2	Existence of updated IEC report with (a) topic, (b) number of participants, © leader of activity, (d) date and (e) signature		

13.16	**Existence of a system to recover drop-outs**	**1**	**0**
13.16.1	Schedule, record of appointments, classified invidual charts		
	Total Points (18)	../18	xxxx
14	**Antenatal Care [max 12 points]**	**YES**	**NO**
14.1	**Business plan contains convincing strategies to effectively conduct ANC for all pregnant women in catchment area**	**1**	**0**
14.1.1	Fixed strategy; and advanced strategy for distant villages: catchment area covers entire ward		
14.2	**Weighing scale present, functional and calibrated to zero**	**1**	**0**
14.3	**ANC form for HF available and well filled in: last five forms verified**	**3**	**0**
14.3.1	All: Examinations: weight - BP, Size, Parity, Date of last menstruation		
14.3.2	All: Laboratory: albuminuria, glucose		
14.3.3	All: Obstetrical examination done: Fetal heart rate, Uterine height, presentation, Fetal movement recorded		
14.4	**ANC form for HF shows the administration of Ferrous Sulphate/Folic Acid and Mebendazole and SP (for the last five forms above)**	**2**	**0**
14.5	**ANC cards for mother available: at least 10 in stock**	**1**	**0**
14.6	**ANC register available and well filled in**	**2**	**0**
14.6.1	Complete identity, state of vaccinations, date visit, whether high risk pregnancy or not/danger signs		
14.6.2	All columns well filled including the identification of problems if any, and actions taken		
14.7	**ANC conducted by qualified personnel**	**1**	**0**
14.7.1	Nurse; midwife CHO or CHEW, verified on ANC cards		

14.8	**Group IEC/BCC**	**1**	**0**
14.8.1	Group meeting held before FP consultation		
14.8.2	Existence of updated IEC report with (a) topic, (b) number of participants, (c) leader of activity and (d) date and (e) signature		
	Total Points (12)	../12	xxxx

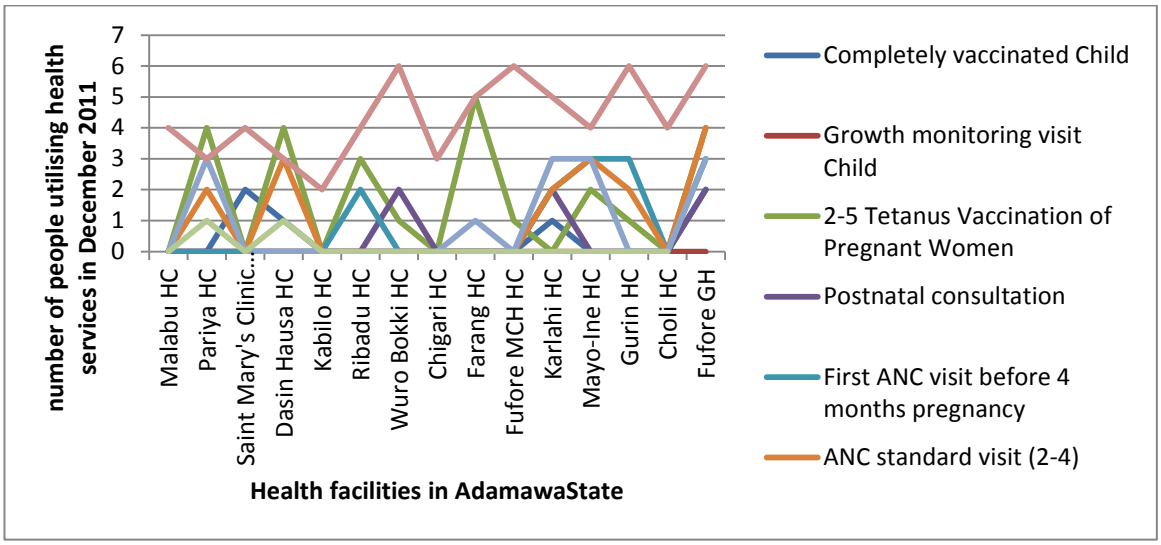
15	**HIV/TB [max 10 points]**	**YES**	**NO**
15.1	**Well-equipped HIV counseling room ensuring privacy:**	**1**	**0**
15.1.1	Plastered and painted wall of solid material		
15.1.2	Smooth cement floor		
15.1.3	Ceiling in good condition		
15.1.4	Windows with glass and curtains		
15.1.5	Doors that close		
15.2	**Availability of IEC/BCC material related to HIV**	**1**	**0**
15.2.1	Penis model on the table		
15.2.2	A box of condoms on the table which has at the least 50 condoms		
15.3	**Existence of a VCT/PMTCT councelling register and lab register acc norms**	**1**	**0**
15.4	**Staff trained in councelling**	**1**	**0**
15.4.1	At the least one staff trained as a councilor		
15.4.2	All councelling done by a trained councilor		
15.5	**Referral system and follow up for HIV clients**	**1**	**0**

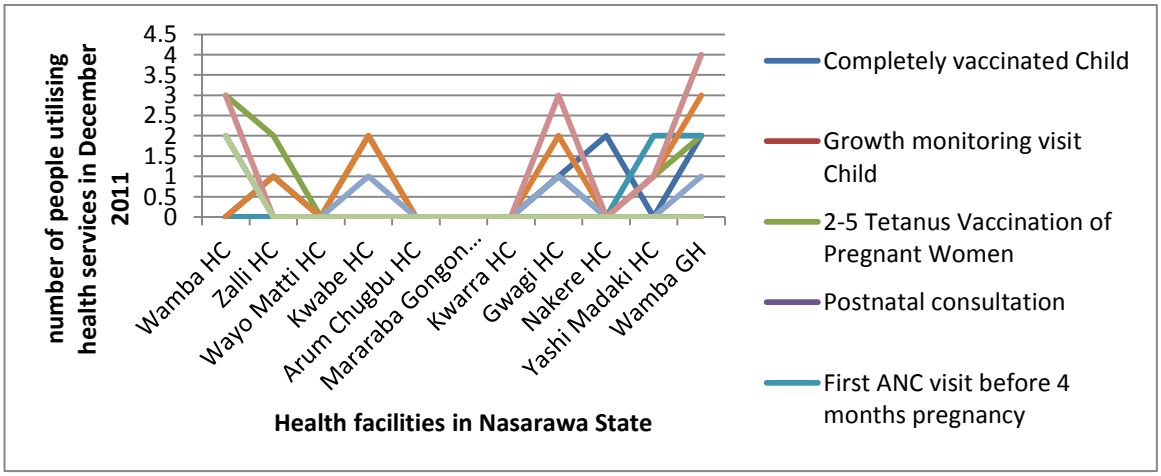
15.5.1	Individual client cards available; planning for CD4 cell counts		
15.6	**Referral system and follow up for TB patients**	**2**	**0**
15.6.1	Each AAFB PTB patient has a person attached to him/her who supervises DOTS: proof of in register; mobile phone number of such a supervisor is registered		
15.6.2	[Define further composite criteria]		
15.7	**Laboratory equipment for testing for PTB**	**1**	**0**
15.7.1	[Define reagents for AAFB testing; stock control cards for reagents; slides etc]		
15.7.2	[Define measures for quality assurance testing of slides]		
15.8	**Availability of anti-tuberculosis drugs**	**1**	**0**
15.8.1	Rifampicine-isoniazide-pyrazinamide : cp120+50+300mg		
15.8.2	Streptomycin 1 gr		
15.8.3	Etambutol tabs 400 mg		
	Total Points (10)	../10	xxxx

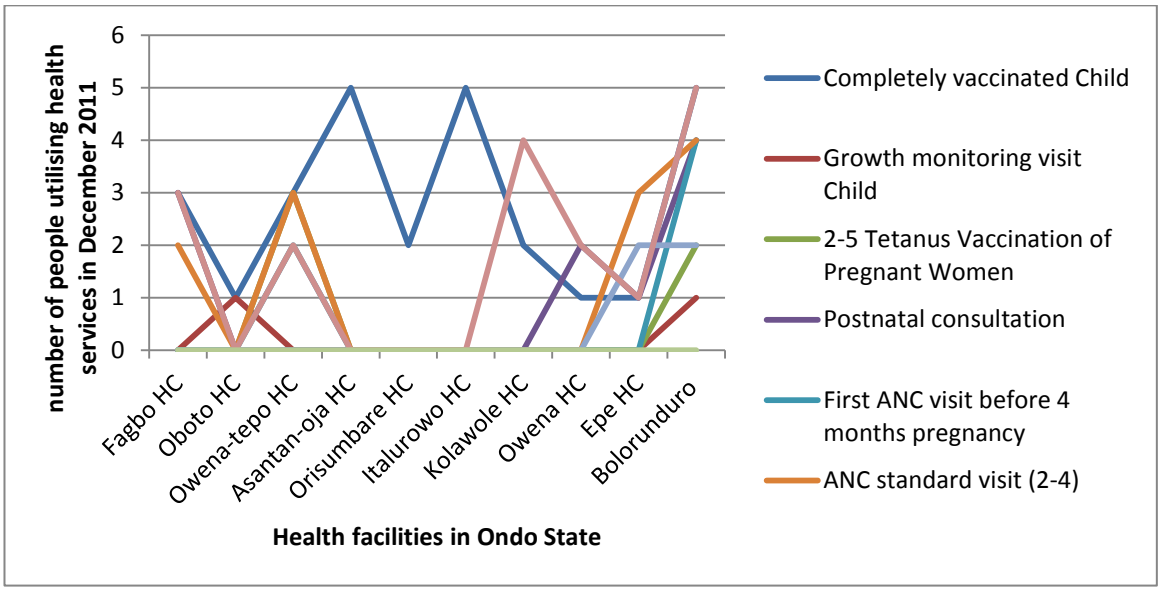
E3. A sample performance verification and tariff for each incentivized health service in the Nigerian P4P scheme

INDICATOR	CLAIMED QTT.	VALIDATED QTT.	TARIF	AMOUNT (NAIRA)
New outpatient consultation by a Doctor	547	403	150	60,450
Counter-referral slip arrived at the Health facility	9	9	350	3,150
Minor Surgery	12	10	900	9,000
Major Surgery (ex CS)	4	4	3,000	12,000
Normal delivery	36	36	1,850	66,600
Assisted delivery	0	0	2,000	0
CS	0	0	3,250	0
Impatient Day	308	308	400	123,200
Postnatal consultation	31	31	350	10,850
First ANC consultation before 4 months pregnancy	13	12	350	4,200
ANC standard visit (2-4)	49	49	350	17,150
FP: total of new users of modern FP methods	12	12	750	9,000
FP: implants and IUDs	0	0	1,300	0
FP: vasectomy and bilateral tuba ligation	0	0	2,000	0
VCT/PMTCT/PIT test	354	354	150	53,100
PMTCT: HIV+ pregnant mothers and children born to are treated according to protocol	1	1	3,000	3,000
STD treated	0	0	1,000	0
New Client put under ARV treatment	20	20	2,000	40,000
New AFB+PTB patient	3	3	3,500	10,500
PTB patient completed treatment and cured	2	2	7,000	14,000

E4. Baseline trends for incentivized maternal and child health services in health facilities in Adamawa, Nassarawa, and Ondo States







Appendix F

FI. Consideration of other methods of data collection in the qualitative study of the formative evaluation of the Nigerian P4P scheme

Observation: There are two main methods of observation: participant and non-participant observation. Participant observation where the researcher is an active member (participating and interacting with other members of the group which is being observed and non-participant observation is a method where the researcher does not interact with the group that is being observed (Barbour, 2013). While this method allows for a richly detailed insight of the topic or theme investigated (Kawulich, 2005), it is difficult to observe 'why' participants do what they do (reasons for their attitudes or behaviours) (Gobo, 2011, Creswell, 1998), which makes it quite impractical for this doctoral thesis, as I am interested in 'why' performance varies across different health facilities in three Nigerian States, which will require large amounts of time and resources.

Document analysis : Document analysis is a method generating data in which existing documents such as: public records, personal records, and physical evidence (found within the study setting) are interpreted by the researcher to give meaning around the investigated topic (Atkinson and Coffey, 2011). In contexts where documents relevant to the research question exist, they usually are not regarded as a completely accurate representation of data for the research topic, but are seen as valuable sources of data in situations where direct observation or questioning of the key participants is not possible (Bowling, 2014a). This method data collection was not employed in this study because there were no publicly available documents relevant to the research question, as this is a new area of inquiry in the Nigerian context and it will be possible to engage the key participants directly, which is regarded as a superior method of collecting data.

Focus groups : Focus group is a method that exploits communication between research participants, collecting data from several people simultaneously (Barbour, 2013). In focus groups, the participants are encouraged to talk to one another: asking questions, exchanging anecdotes and commenting on each other's experiences and points of view (Bowling, 2014b). Both focus group discussions and in-depth interviews are particularly useful in exploring participants' knowledge and experiences, and can be used to examine not only what people think but how they think and why they think that way (Carter and Henderson, 2005). Focus groups are more appropriate where interaction between participants is likely to provide more insight on the topic being investigated and/or where the purpose of the research is study group norms or group meanings (Barbour, 2013), which is not completely relevant to this study, as I am not only interested in group meanings or norms but also in individual meanings and experiences. In addition, focus group discussions fall short in the area of maintaining confidentiality and anonymity within the group, making participants likely to be hesitant in expressing their thoughts (Carter and Henderson, 2005, Bowling, 2014b, Barbour, 2013). Hence, it was not a suitable or effective method for exploring some of the sensitive areas in this study such as uncertainty in payment and role of management, both of which are likely to involve corruption and transparency issues. Furthermore, focus group discussions might produce biased outputs and unlikely to achieve the 'in-depthness' obtained from interviews as a result of dominating participants within the group, which is often a problem with inexperienced focus groups moderators (Finch and Lewis, 2004, Bowling, 2014b).

In depth interviews (unstructured and semi-structured): In-depth interviews are encounters between the researcher and participant, expressed through words and relay the participant's thoughts, feelings and motivations (Barbour, 2013). The two main types used in qualitative research are unstructured and semi-structured interviews (Carter and Henderson, 2005).

Unstructured interviews : Unstructured interviews are entirely participant led; the participants tell their own stories in their own words (to give a deep meaning of the participants' world) with little or no direction from the researcher, and in most cases, the researcher returns to one participant over a period of time to build up a vivid picture (Bowling, 2014b). The interviewer approaches the interview with the aim of discussing a limited number of topics, sometimes as few as one or two which are covered in great detail (Miller and Glassner, 2011). This method of interviewing is mostly used in history projects or project involving people's lifetimes or biographies (Carter and Henderson, 2005). This method was not used in this study for two main reasons. First, it is very time consuming, which limits the sample size and areas of enquiry of the research (Bowling, 2014b). Second, the data produced, though richly detailed is often non-comparable to other participants (Carter and Henderson, 2005, Bowling, 2014b), which is needed to achieve some of the objectives in this study.

F2. Evidence of Ethics Approval for the qualitative study of the formative evaluation of the Nigerian P4P scheme

THE UNIVERSITY of York

DEPARTMENT OF HEALTH SCIENCES

c/o Department of Philosophy
Heslington
York YO10 5DD

Telephone (01904) 433253
Fax (01904) 321383
E-mail smh12@york.ac.uk

Dr Stephen Holland

www.york.ac.uk/healthsciences

21 May 2013

Miss Y K Ogundeji
University of York
Department of Health Sciences
YO10 5DD

Dear Yewande

Attitudes and beliefs of Nigerian health workers participating in the Nigerian Performance Based Financing (PBF) pre-pilot program

Thank you for submitting the above research study to the Health Sciences Research Governance Committee and for attending the committee's meeting on Monday, 20 May 2013 to respond to our queries.

The committee have approved the study but asked me to reiterate the following points:

1. The researchers name and (work) contact email should be added to the Patient Information Sheet.

2. Some of the bullet points on the consent form could be combined to make it shorter and clearer. On the other hand, the option of being interviewed but not audio-recorded (answers being written by the researcher by hand) should be listed on the consent form, with a 'tick box'.

3. You should reconsider whether you want to use the data gathered from participants who withdraw during the study. If so, this will need to be made explicit in the Patient Information Sheet and Consent Form.

The committee is happy for you to take up this feedback with your supervisor and does not need to see the study again, but if you make any substantial amendments to the study then you might need further approval. If you have any questions regarding the committee's decision then please contact me.

Yours sincerely

S. Holland

Stephen Holland
Chair: HSRGC

cc. Professor Trevor Sheldon

F3. Original Information sheet for potential participants in the qualitative study of the formative evaluation of the Nigerian P4P scheme

THE UNIVERSITY of York

Information Sheet

Date: 24/04/2013

Study title

Attitudes and beliefs of Nigerian health workers participating in the Nigerian Performance Based Financing (PBF) pre-pilot program

Introduction

My name is Yewande Ogundeji, a PhD student at The University of York in the United Kingdom. I am collaborating with the federal ministry of health on the implementation of Performance Based Financing (PBF) pilot programmes in Nigeria. I am interested in ways to improve the quality of healthcare in Nigeria. I am conducting research on how pay for performance programmes (the use of incentives) in healthcare can be better designed and implemented for maximum impact on health care.

You are being invited to take part in my research study. Before you decide if you want to take part in this research or not, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and ask me any question(s) if there is anything that is not clear or if you would like more information about the research. Take time to decide whether or not you wish to take part.

What is the purpose of the study?

Performance Based Financing has recently been proposed as a way to encourage the changes needed to improve the quality of care in Nigeria. The proposed PBF program is currently in its pre-pilot stage. Evidence shows certain factors affect the success of the program (such as: the size of incentive, degree of uncertainty in achieving targets and receiving incentives, and the way the scheme is implemented).Therefore, the purpose of this study is to explore the views (and thoughts) of the health workers and health facility mangers participating in the new PBF scheme in the Nigerian context regarding issues (such as payments, administration, management, and experience with the scheme) in order to understand how these views affect the impact of the scheme on quality of care and health facility performance. The results of this research will be used to inform new design and implementation of the PBF pilot schemes in Nigeria.

Why have I been approached to participate?

You have been approached because the PBF program is being implemented in the health facilities where you work.

What will happen to me if I take part?

This research involves one on one interview sessions that will around one hour and would be audio-recorded with your permission. If you want to take part in the research and are not willing to be audio-recorded, this can be arranged and I will take down written notes only. I will be asking you questions about your about your roles at work, management, changes in the health facility, and I will also ask about your thoughts and opinions about new incentive program. You do not have to answer any questions that you do not feel comfortable with. If you decide to take part, you are still free to withdraw at any time and without giving a reason. A decision to withdraw at any time, or a decision not to take part, will not affect your job.

Will my taking part in this study be kept confidential?

Your privacy is very important to me and participation in this research does not put you at risk in any way. If you are willing to be a part of the research and be interviewed, I will not collect your name during the interview sessions (I will give you identification numbers instead). This way, your rights are protected and whoever reads the findings of this research will not be able to link anything in the research to you.

Do I have to take part?

Taking part in this research is entirely voluntary. It is up to you to decide whether or not to take part. If you do decide to take part you will be asked to sign a consent form. As stated earlier, if you decide to take part you are still free to withdraw at any time and without giving a reason.

What will happen to the results of the research study?

I will provide a brief summary/feedback of the research on a Local Government Area level (not specific to any individual or health facility) at its completion and it will be made available through the health facility manager.

The main findings of this research would be submitted as part of my PhD thesis; results may also be published in academic journals and World Bank reports. Some extracts of the interviews may appear in a published paper but you would not be identified and it will not be traced back to you. The audio recordings would be deleted after I accurately transcribe the audio recordings but the transcripts will then be stored in secure folders on a password-protected computer for up to 5 years after the end of the research. Only my supervisor and I will have access to the raw transcripts.

If you have any questions about this research, please ask me now.

If you have any concerns about confidentiality and anonymity, please do not hesitate to contact my supervisor Professor Trevor Sheldon at trevor.sheldon@york.ac.uk.

F4. Participant Consent form for participants in the qualitative study of the formative evaluation of the Nigerian P4P scheme

THE UNIVERSITY of York

Participant Consent Form

Title of research: Attitudes and beliefs of Nigerian health workers participating in the Nigerian Performance Based Financing (PBF) pre-pilot program

Health facility (LGA):

Name of researcher: Yewande Ogundeji

Participant ID:

Thank you for reading the information about this research project. If you would like to take part, please read, tick the appropriate box, and sign this form.

	YES	NO
• I have read the information sheet dated 24/04/2013 for the above study, I have a copy to keep, and I understand what the research is about.	☐	☐
• I understand that my participation is entirely voluntary and I can withdraw from this study at any time	☐	☐
• I understand that refusal to participate does NOT affect my job in any way	☐	☐
• I understand that if I participate I am still free to refuse to answer any questions	☐	☐
• I understand that the interview sessions will be audio recorded with my permission	☐	☐
• I understand that the audio recordings will be fully transcribed and kept as secure computer files	☐	☐
• I understand that all data will be kept strictly confidential and anonymous	☐	☐
• I understand Excerpts from the research may appear in publications but under no circumstances will my name be included in the report.	☐	☐
• I am willing to take part in this research	☐	☐

Name …… Signature ……

Date ……

Name of researcher ……..…….…. Signature ……

Date ……

F5. Amended information sheet for participants qualitative study of the formative evaluation of the Nigerian P4P scheme

THE UNIVERSITY of York

Information Sheet (amended sheet for the health workers)

Date: 24/04/2013

Study title

Attitudes and beliefs of Nigerian health workers participating in the Nigerian Performance Based Financing (PBF) pre-pilot program

My Name is Yewande Ogundeji a PhD student from the university of York working in conjunction with the NPHCDA on the PBF program.

The PBF program is still very new and it is in its implementation of the stage. The implementers of the program are still open and willing to make adjustments in some of the design features and the way the scheme is implemented, in order to make the program more successful. So the purpose of my research is to explore some areas of the design on the program, to hear your thoughts and views, which will help inform design of the program in the scale up program.

Taking part in this interview is entirely voluntary and if you decide you do not want to be a part of it, it will not affect your job in any way. However, if you do decide to take part; you will be asked to sign a consent form, the interview sessions will be one on one, will last around one hour, and would be audio-recorded with your permission. If you do not wish to be audio recorded, only written notes would be taken. In addition, you do not have to answer any questions that you do not feel comfortable with and you are still free to withdraw at any time without giving a reason. A decision to withdraw at any time, or a decision not to take part, will not affect your job.

No one apart from me will have access to the recordings and whatever you say here will be in confidence and will not be traced back to you. The audio recordings would be deleted after I write down the necessary information, which would be stored securely.

I will also write a report but it will be a general report that will include information from other health workers in the other health facilities and this report would be made available through your health facility manager. In addition, the main findings of this research would be submitted as part of my PhD thesis; results may also be published in academic journals and World Bank reports. Some extracts of the interviews may appear in a published paper but you would not be identified and it will not be traced back to you.

If you have any questions about this research, please ask me now.

Thank you

Evidence of ethical approval for amendment on the information sheet

Stephen Holland <stephen.holland@york.ac.uk> 28/08/2013

to me, trevor.sheldon, cath.jackson, Sandi

Dear Yewande,

Many thanks for informing us of this amendment to your study. I am writing to approve by Chair's Action the change to the information sheet. If you require a more formal letter to this effect, let me know. Otherwise, good luck with the study.

Best wishes,

Stephen Holland
Chair HSRGC

F6. Final/refined Interview Questions for the qualitative study of the formative evaluation of the Nigerian P4P scheme

In the interview, I asked about experience with the scheme so far and changes in the health facility since the start of the scheme. Then we moved on to talk about views about the payment systems, delay in payments, and the assessment of performance. Listed below is a set of semi structured questions that guided me through the areas I am interested in covering for the interview. These questions are numbered but might not necessarily follow that sequence as the participant may talk about something I had the intention of asking later on in the interview.

1. Tell me about what you do (your role) in the health facility?
2. So what do the health workers think about the program?
3. What has changed since the start of the program?
   [Prompts]
   a) Effects of the program on the health facility/patient outcomes
   b) Administrative burden
   c) Motivation
   d) Supervision
   e) Has it had any effect on you/change professionally
   f) Any negative effects
4. Can you please tell me more about how this scheme works in this health facility?
   [Prompts]
   a) How did you get to know about it?
   b) Aim of the program
   c) Training received
   d) What activities attract payment of bonuses?
   e) What do you have to do to earn bonuses
   f) Size of potential incentives?

5. We will move on to talk about how the health facility has changed so far; Performance data collected so far shows that performance has improved for some activities. Can you tell me about some of the approaches that have been used in this health facility to improve performance?

Follow up questions/can also serve as prompt depending on the situation with the participants

- For example, you performed so well (judging from your scores before and after PBF) in (insert activity), could you please tell me some of the approaches you used in improving performance.
- So what happened in this particular activity (insert activity), the health facility didn't perform so well in it. Can you tell me what happened? (Prompts: mobility, competition from other facilities, distance, cost, and staff strength)

[Prompts]

a) Approaches e.g. Supervision and monitoring, training, health promotion, investing in infrastructure, drugs etc.
b) Challenges/ease
c) Have you had to change any attitude or behaviour?
d) How has the community perceived or received it?

6. Tell me about the use of feedback in this health facility

{Prompt] How has this knowledge of how you have performed compared to other health facilities impacted you in anyway (feedback)?

7. How you would feel about sharing information about your ranks or performance results with other health facilities.

Prompt: do you think it will it have any effect on performance at all?

8. I would like to hear about the incentive; how it is used in this facility

[For the health facility manager]: How do you decide how to utilize the incentives earned?

9. I've heard that payments are sometimes delayed or have been delayed in the past; can you please share your experiences with the delay in payments of incentives.

Prompts:

a) How does this make you feel?
b) Did this affect the performance of the facility at all?

10. Are explanations for the delays in payments communicated to you through the manager? [for the managers:….to you through the scheme implementers]

Follow up question: … and how does this affect the health care workers?

11. What about explanations for any other changes in the program?

Follow up question: … and how does this make you feel?

12. Have you been satisfied with the payments so far?

Follow up: Why?

Now, let us move on to talk about how individual performance is measured for payment of bonuses.

13. What are your thoughts about the way the bonuses are shared to the health workers

In cases where is sense hesitation: What do the other health workers in the health facility feel about this method of deciding how much to pay the individual workers?

Prompts:

a) Does it have any effect on performance or health worker motivation
b) Are there any other way(s) of measuring individual performance that you would prefer?
c) If yes, can you explain it to me?

14. What are your thoughts toward assessment by the contribution you make to performance (increase in activities) or to the quality of care in order for the health facility to earn incentives:
    Prompt: How will that affect your performance?

Is there any other thing you'd like us to discuss regarding this scheme?

Thank you very much, for taking part of this research.

P.S. all questions ended with "is there any other thing you would like to add"

F7. A list of the health facilities in each State (and their characteristics) positioned according to performance

Adamawa

1. Gurin (Health Center): Responsable: Mrs. Mable M. Zubairu, Phone number: +234-7082662007, Population: 41,945 (2012), Status: Public, Staff size: 9
2. Farang (Health Center): Responsable: Mrs. Ruth Simon, Phone number: +2348098162148, Population: 25,798 (2012), Status: Public, Staff size: 8
3. Furore MCH (Health Center): Responsable: Mrs. Aishatu Ahmadu Tukur, Phone number: +234-8032920598/8052, Population: 13,631 (2012), Status: Public, Staff size: 9
4. Wuro Bokki (Health Center): Responsable: Mrs. Ungopwa Dauda, Phone number: +234-8039092743/8051, Population: 12,598 (2012), Status: Public, Staff size: 6
5. Mayo-Ine (Health Center): Responsable: Mrs. Aishatu Kadiri, Phone number: +234-8082534289, Population: 10,149 (2012), Status: Public, Staff size: 5
6. Chigari (Health Center): Responsable: Mrs. Uwami Ayuba, Phone number: +234-8058782153/8050, Population: 20,160 (2012), Status: Public, Staff size: 6
7. Furore General Hospital (General Hospital): **Responsable:** Dr. Peter Tihze Kanu, **Phone number:** +2348054178797, **Population:** 240,160 (2012), **Status:** Public, **Staff size:** 20
8. Malubu: Responsable: Mrs. Damaris Bilison, Phone number: +234-8074496353, Population: 12,598 (2012), Status: Public, Staff size: 8
9. Pariya (Health Center): Responsable: Mrs. Aishatu Hayatu A., Phone number: +234-8130001718, Population: 40,060 (2012), Status: Public, Staff size: 13
10. Dasin Hausa (Health Center): Responsable: Mrs. Peace S. Audu, Phone number: +234-8037890837, Population: 10,475 (2012), Status: Public, Staff size: 5
11. Kabilo (Health Center): Responsable: Mr. Tadawus Tula, Phone number: +234-8067554243, Population: 10,011 (2012), Status: Public, Staff size: 3
12. Ribadu (Health Center): Responsable: Mrs. Pwanagoshi Emmanuel, Phone number: +234-8161592783/8134, Population: 10,475 (2012), Status: Public, Staff size: 6
13. Choli (Health Center): **Responsable:** Mr. HarunaI. Domlek, **Phone number:** +234-8022789389, **Population:** 10,379 (2012), **Status:** Public, **Staff size:** 2
14. MCH Yadim/st mary? (Health Center), Status: Public, Staff size: 2
15. Karlahi (Health Center): Responsable: Mrs. Suzana E. Yagah, Phone number: +234-8063919391/7052, Population: 10,283 (2012), Status: Public, Staff size: 7

Ondo

1. Bolorunduro (General Hospital): **Responsable:** Dr Agosile, **Phone number:** 08052159504, **Population:** 90,454 (2011), **Status:** Public, **Staff size:** 11

2. Fagbo (Health Center): Responsable: Mrs. Cocker, Phone number: 07030710798, Population: 8,353 (2011), Status: Public, Staff size: 5

3. Orisumbare (Health Center): Responsable: Mrs Akinboye, Phone number: 08032136636, Population: 9,394 (2011), Status: Public, Staff size: 3

4. Kolawole (Health Center): Responsable: Mrs Adegbayemu, Phone number: 08033697742, Population: 5,538 (2011), Status: Public, Staff size: 3

5. Asatan (Health Center): Responsable: Mr Adeoyin, Phone number: 08101121200, Population: 8,261 (2011), Status: Public, Staff size: 3

6. Epe (Health Center): **Responsable:** Mrs. Awosika, **Phone number:** 08039559530, **Population:** 13,528 (2011), **Status:** Public, **Staff size:** 3

7. Owena Bridge (Health Center): Responsable: Mrs. Adesiyan, Phone number: 08032469936, Population: 9,486 (2011), Status: Public

8. Oboto (Health Center): Responsable: Mr. Falekulo, Phone number: 07033524686, Population: 8,592 (2011), Status: Public, Staff size: 4

9. (8)Owena Tepo (Health Center): Responsable: Mrs Oyewole, Phone number: 08072662377, Population: 5,862 (2011), Status: Public, Staff size: 2 size: 3

10. Italuworo (Health Center): Responsable: Mrs Olafusi, Phone number: 08060867543, Population: 12,308 (2011), Status: Public, Staff size: 3

Nassarawa

1. Wamba (General Hospital): **Responsable:** Dr Usman A.T., **Phone number:** +2348069569208, **Population:** 85,328 (2011), **Status:** Public, **Staff size:** 32

2. Zalli (Health Center): Responsable: Jacob Momana Audu, Phone number: +2348024608614, Population: 18,163 (2011), Status: Public, Staff size: 7

3. Nakere (Health Center): Nakere Ward, Responsable: Laiatu Mathiew, Phone number: +2348065467694, Population: 7,520 (2011), Status: Public, Staff size: 14

4. Wamba (Health Center): P.o .Box 59 Wamba, Responsable: Juliana Umaru, Phone number: +2348035045456, Population: 6,803 (2011), Status: Public, Staff size: 20

5. (4 as well)Wayo Matti (Health Center): Responsable: Ludya A. Tanze, Phone number: +2348036232353, Population: 6,989 (2011), Status: Public, Staff size: 8

6. Kwabe (Health Center): Responsable: Monha D. Haruna, Email: mangachawa@yahoo.com, Phone number: +23480332500161, Population: 6,677 (2011), Status: Public, Staff size: 5

7. Arum Chugbu (Health Center): Responsable: Rebecca Thomas, Phone number: +2347080173450, Population: 4,012 (2011), Status: Public, Staff size: 6

8. Mararaba Gongon (Health Center): Gitta Ward, Responsable: Rose Joshua, Phone number: +2348020860860, Population: 12,224 (2011), Status: Public, Staff size: 13

9. Gwagi (Health Center): Responsable: Shehu Usman Zakari, Phone number: +2348133216819, Population: 7,177 (2011), Status: Public, Staff size: 9

10. Kwarra (Health Center): Responsable: Danjuma Nubu, Phone number: +2348029601166, Population: 8,549 (2011), Status: Public, Staff size: 11

11. Yashi Madaki (Health Center): Responsable: Polycarp Danjuma, Phone number: +2348029186098, Population: 7,214 (2011), Status: Public, Staff size: 8

F8. Appointment forms for potential participants in the qualitative study of the formative evaluation of the Nigerian P4P scheme

THE UNIVERSITY of York

Appointment form

- Please if you wish to take part in this interview, pick an appropriate date and time.
- Please note that I might need to reschedule your appointment due to clashes.
- Your interview date would be confirmed with a phone call.
- If you do not wish to participate in this research, please leave this appointment form blank but still put it in the sealed envelope and hand it to your manager.
- If you are willing to participate, please bring along your information sheet along to the interview.

Name:

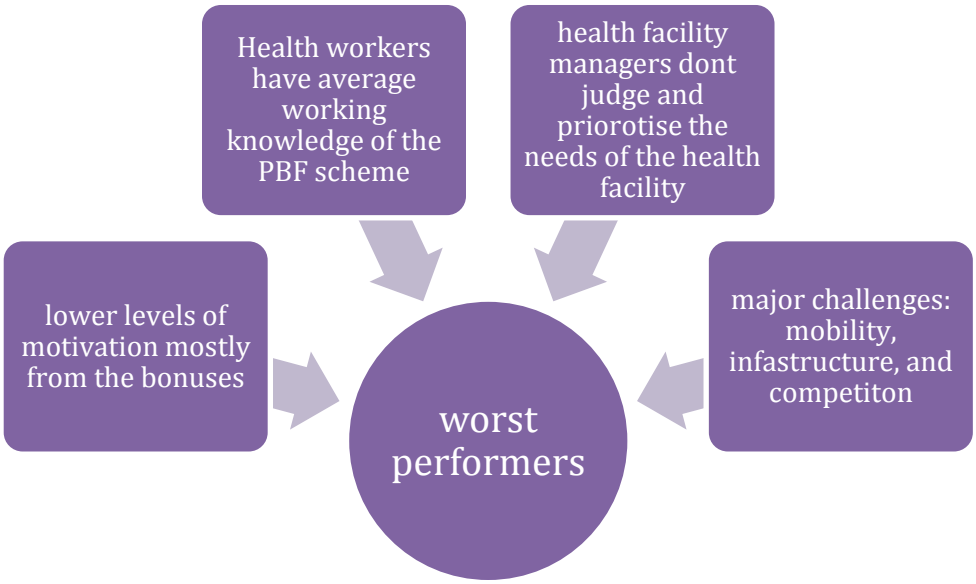
Telephone number:

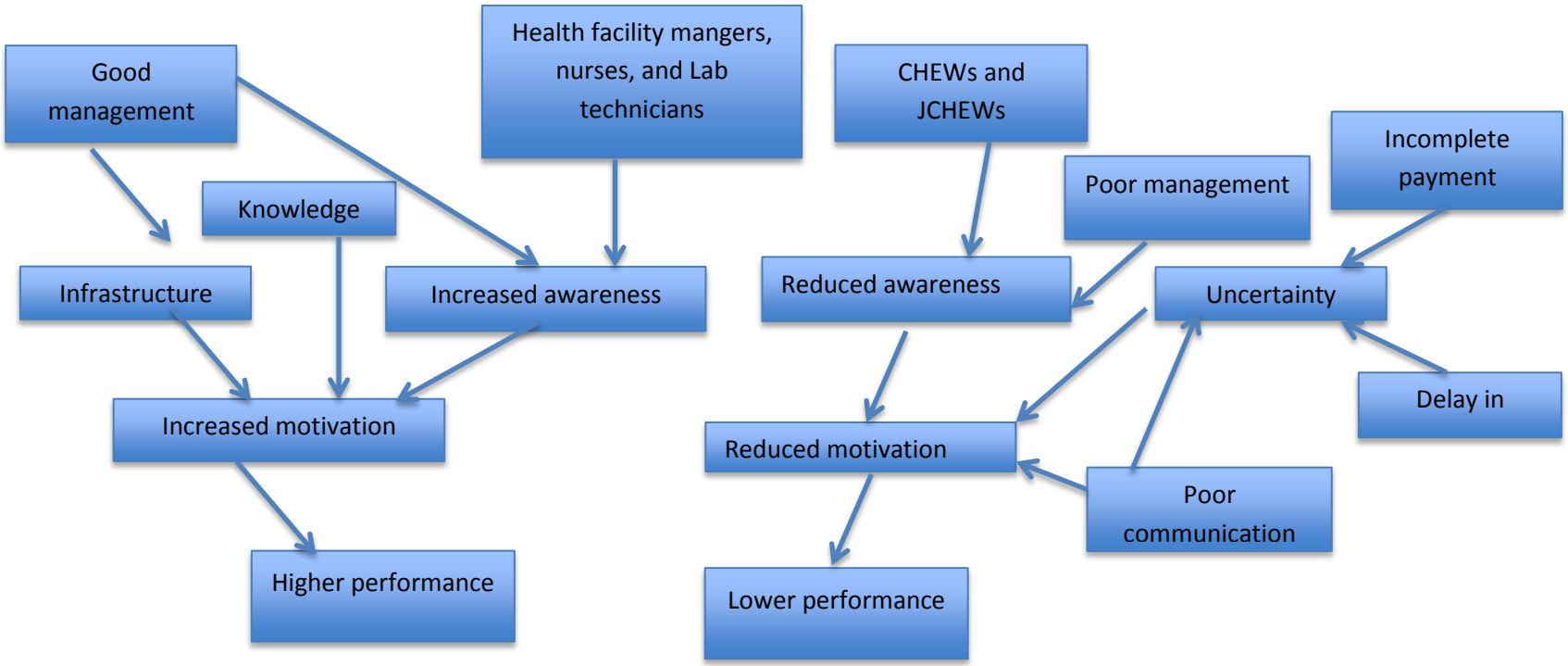
(Please circle what days and times you would be available for the interview)

Available days: Monday Tuesday Wednesday Thursday Friday Saturday

Available times: 10-11am 11:30-12:30am 1-2pm 2:30-3:30pm

F9. Factors linked with worst performing facilities (sample chart)





F10. Relationship between themes, categories and concepts (sample chart)

APPENDIX G

G1. Executive summary of report submitted to the NPHCDA on the formative evaluation of the Nigerian P4P scheme

Introduction

Pay for Performance (P4P) has been proposed as a way to incentivise the changes in behaviour needed to improve the quality of care and health indicators in three selected States (Ondo, Nassarawa, and Adamawa) in Nigeria.
Evidence on effectiveness of P4P schemes is mixed (with heterogeneous results), and preliminary results of the Nigerian P4P scheme show similar variation in results among the three sites in the selected States.
A detailed consideration of the literature and theory from behavioural economics expounding on how people respond to incentives, suggest certain key aspects of P4P schemes, such as: design, context, and implementation are likely to affect the impact of the scheme and possibly explain the heterogeneous results.
The aim of this study was to carry out a formative evaluation to explore differences in implementation and contextual factors between the States and health facilities.

Methods

Face to face in-depth interviews were then conducted on health workers participating in the scheme in Ondo and Nassarawa State to explore their views, thoughts and attitudes on factors such as uncertainty in payment, role of management, infrastructure, and understanding of the scheme.
The study sample of health workers was drawn from top, average, and worst performing facilities in each State, reflecting diversity in performance to facilitate comparisons.
Data obtained from interviews were transcribed verbatim and analysed using the framework approach.

Results

Thirty-six health workers comprising of 13 health facility mangers, 7 nurses, 6 laboratory technicians, 6 Community Health Extension Workers (CHEWs) and 4 Junior Community Health Extension Workers (JCHEWs) were interviewed.

The findings from this study suggest that delay in payment, doubt in the method of allocation of bonuses to individual health workers, and ineffective communication led to uncertainty and distrust in the P4P payment system, which reduced motivation of health workers and or performance of health facilities. Another key finding was that health worker understanding of the scheme appeared to be related to motivation and performance and participants in top performing facilities had a better understanding of the scheme. A third interconnected theme was that some health facility mangers in top performing facilities or Nassarawa State appeared to have superior managerial skills, which was evident in the unique strategies they used to motivate the health workers and improve performance in their facilities such as hiring additional staff and infrastructural improvement, which led to improvement in performance reflected in the preliminary results. The fourth major finding was that other factors not inherent to the PBF scheme such as issues with mobility and lack of infrastructure in Ondo State or lack of manpower in Nassarawa State were sources of demotivation for the health workers. However, some health facility managers in Nassarawa State took initiative and had a swift response to the lack of manpower problem leading to hiring of additional staff whereas infrastructural problems did not improve in Ondo State, possibly explaining why Nassarawa State appeared to have performed better than facilities in Ondo State.

Conclusion and Recommendations

As the Nigerian PBF program is set to last till 2018, the continual evolution of the program to maximize effectiveness and cost effectiveness is necessary. Based on the findings from this study, the following recommendations were made for consideration in the scaling up of the P4P scheme.

- To make timely quarterly payments to the each health facility for delivery of services as agreed in the PBF contract.
- To ensure clear communication strategies about changes and difficulties encountered in the scheme between stakeholders, particularly to inform and keep the health workers up to date.

- To ensure that the assessment tool (basis by which individual health workers earn bonuses) includes a criterion/a set of criteria that clearly captures actual contribution and direct input of the health worker in helping the facility earn money. For example a criterion on outreaches or home visits can be added on, which this study has shown is instrumental in increasing utilisation of health services.
- To ensure clearer and shorter guidelines are also needed to encourage the use of the tool instead of ranks to allocate bonuses to the health workers.
- To carry out training and regular workshops for health workers and equipping health facility managers with materials to properly inform the health workers on how the P4P scheme operates.
- To help health facility managers improve their managerial skills, with a focus on their autonomy, setting priorities, and recognising and meeting the needs of the health facility or how to motivate the health workers, whether it is infrastructure or hiring additional staff.
- To introduce ideas and suggestions on how the health facilities can improve performance on areas they are lagging behind should be introduced to the feedback forms.
- To move towards 'true pay for performance' (e.g. change in utilisation from baseline) as opposed to pay for reporting.

List of Keywords

Pay for Performance (P4P)

Incentive

P4P evidence

P4P effectiveness

Heterogeneity

Evaluation design

P4P design

P4P context

P4P implementation

P4P Typology

Inter-rater reliability

Meta-regression

Multilevel logistic regression

Nigerian-P4P

Formative evaluation

Framework analysis

Clinician attitude

Clinician experience

List of Abbreviations

ANC: Antenatal Care

CHEW: Community Health Extension Worker

FMOH: Federal Ministry of Health

GP: General Physician

JCHEW: Junior Community Extension Worker

LGA: Local Government Area

LMICs: Low and middle-income countries

MDGs: Millennium development goals

MSS: Midwives service scheme

NHIS: National health insurance scheme

NICS: National immunisation coverage scheme

NPHCDA: National Primary Health Care Development Agency

P4P: Pay for performance

PBF: Performance based financing

PHC: Primary health care

PMTCTHIV: Prevention of mother to child transmission of HIV

PNC: Postnatal Care

RCT: Randomised controlled trials

SBA: Skilled Birth Attendant

SMD: Standardised mean difference

SMOH: State Ministry of health

SMOH: State Ministry of Health

UK: United Kingdom

USA: United States of America

VCT: Voluntary Counselling and Testing

WHO: World Health Organisation

References

ABANIDA, E. A. Nigeria Immunization Programme-Increasing Access. Nigeria At the 2nd Dialogue & Retreat of the Alliance of Southern Civil Society in Global Health 2012 Abuja, Nigeria. https://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CDAQFjAA&url=http%3A%2F%2Fwww.chestrad-ngo.org%2Findex.php%3Foption%3Dcom_docman%26task%3Ddoc_download%26gid%3D26%26Itemid%3D174&ei=1f15UvaIO6iN7Abqm4GoCw&usg=AFQjCNF0ejf28e5StwHlFDPa_DjOpLnZNA&sig2=EOlUSeaMRCzobvaQcpackg&bvm=bv.55980276,d.ZGU.

ABDULRAHEEM, I. S., OLAPIPO, A. R. & AMODU, M. O. 2012. Primary health care services in Nigeria: Critical issues and strategies for enhancing the use by the rural communities. *Journal of Public Health and Epidemiology,* 4**,** 5-13.

ABIMBOLA, S., OKOLI, U., OLUBAJO, O., ABDULLAHI, M. J. & PATE, M. A. 2012. The Midwives Service Scheme in Nigeria. *PLOS MEDICINE,* 9.

ADLER, P. A. 2012. How many qualitative interviews is enough? *In:* BAKER, S. E. & EDWARDS, R. (eds.) *Expert voices and early career reflections on sampling and cases in qualitative research.* http://eprints.ncrm.ac.uk/2273/4/how_many_interviews.pdf National Centre for Research Methods

AKINWALE, A. 2010. The menace of inadequate infastructure in Nigeria. *African Journal of Science, Technology, Innovation, and Development* 2**,** 207-208.

AKINYANDENU, O. 2013. Counterfeit drugs in Nigeria: A threat to public health. *African Journal of Pharmacy and Pharmacology,* 7**,** 2571-6.

AKINYEMI, O. & ATILOLA, O. 2013. Nigerian resident doctors on strike: insights from and policy implications of job satisfaction among resident doctors in a Nigerian teaching hospital. *The International Journal of Health Planning and Management,* 28**,** e46-e61.

AKWATAGHIBE, N., SAMARANAYAKE, D., LEMIERE, C. & DIELEMAN, M. 2013. Assessing health workers' revenues and coping strategies in Nigeria -- a mixed-methods study. *BMC Health Services Research,* 13**,** 387.

ALTMAN, D. G. 1991. *Practical statistics for medical research,* London, Chapman and Hall.

ALVESSON, M. 2009. *Reflexive methodology : new vistas for qualitative research,* Thousand Oaks, CA, SAGE Publications.

AN, L. C., BLUHM, J. H., FOLDES, S. S., ALESCI, N. L., KLATT, C. M., CENTER, B. A., NERSESIAN, W. S., LARSON, M. E., AHLUWALIA, J. S. & MANLEY, M. W. 2008. A randomized trial of a pay-for-performance program targeting clinician referral to a state tobacco quitline. *Arch Intern Med,* 168**,** 1993-9.

APPLEBY, J., HARRISON, T., HAWKINS, L. & DIXON, A. 2012. Payment by Results: How can payment systems help to deliver better care? The King's Fund.

ARROW, K. J. 1965. The theory of risk aversion. *In:* YRJO JAHNSSONIN SAATIO, H. (ed.) *Aspects of the Theory of Risk Bearing.* Chicago: Markham Publ. Co.

ARYANKHESAL, A., SHELDON, T. A. & MANNION, R. 2013. Role of pay-for-performance in a hospital performance measurement system: a multiple case study in Iran. *Health Policy Plan,* 28**,** 206-14.

ASUZU, M. C. 2005. The necessity for a health systems reform in Nigeria. *Journal of Community Medicine & Primary Health Care,* 16**,** 1-3.

ATKINSON, G. & NEVILL, A. 1998. Statistical Methods For Assessing Measurement Error (Reliability) in Variables Relevant to Sports Medicine. *Sports Medicine,* 26**,** 217-238.

ATKINSON, P. & COFFEY, A. 2011. Analysing Document Realities. *In:* SILVERMAN, D. (ed.) *Qualitative research : issues of theory, method and practice.* London: SAGE.

BAILEY, K. D. 1994. *Typologies and taxonomies an introduction to classification techniques,* Thousand Oaks, Calif.

BAKER, S. E. & EDWARDS, R. 2012. How many qualitative interviews is enough? *Expert voices and early career reflections on sampling and cases in qualitative research.* http://eprints.ncrm.ac.uk/2273/4/how_many_interviews.pdf National Centre for Research Methods

BARBOUR, R. S. 2001. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *BMJ,* 322**,** 1115-7.

BARBOUR, R. S. 2013. *Introducing qualitative research : a student's guide,* London, Sage.

BARCLAY, J. H. & HARLAND, L. K. 1995. Peer performance appraisals: the impact of rater competence, rater location, and rating correctability on fairness perceptions. (includes appendix). *Group & Organization Management,* 20**,** 39.

BASINGA, P., GERTLER, P. J., BINAGWAHO, A., SOUCAT, A. L. B., STURDY, J. & VERMEERSCH, C. M. J. 2011. Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation. *The Lancet,* 377**,** 1421-1428.

BERKMAN, N. D., LOHR, K. N., MORGAN, L. C., KUO, T.-M. & MORTON, S. C. 2013. Inter- rater Reliability Testing of the AHRQ EPC Approach to Grading Strength of Evidence in Systematic Reviews. *Journal of Clinical Epidemiology*.

BHUTTA, Z. A., DARMSTADT, G. L. & RANSOM, E. I. 2003. Using evidence to save new-born lives. *Policy brief.* Washington, DC: Population Reference Bureau.

BLAND, J. M. 2008. *Cohen's Kappa* [Online]. University of York Department of Health Sciences http://www-users.york.ac.uk/~mb55/msc/clinimet/week4/kappash2.pdf. [Accessed February 13 2014].

BLAND, M. 2000. *An introduction to medical statistics,* Oxford, Oxford University Press.

BOADEN, R., HARVEY, G., MOXHAM, C. & PROUDLOVE, N. 2008. Quality Improvement: Theory and Practice in Healthcare. http://www.institute.nhs.uk/service_transformation/quality_improvement/quality_improvement%3A_theory_and_practice_in_healthcare.html#sthash.hMmiNdPN.dpuf: NHS Institute for Innovation and Improvement

BOOTH, C. M. & TANNOCK, I. F. 2014. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *Br J Cancer,* 110**,** 551-5.

BORENSTEIN, M., HEDGES, L. V., JULIAN, P. T. H. & ROTHSTEIN, H. R. 2009. *Introduction to meta-analysis,* Chichester, West Sussex, U.K., John Wiley & Sons.

BOWLING, A. 2014a. *Research methods in health [electronic resource] : investigating health and health services,* Maidenhead, Open University Press.

BOWLING, A. 2014b. Research methods in health: investigating health and health services 3rd ed. ed. Maidenhead: Open University Press.
BOWLING, A. & EBRAHIM, S. 2005. *Handbook of health research methods: Investigation, Measurement and Analysis.*, Berkshire, England. , Open University Press, Mcgraw-Hill.
BREDENKAMP, C., SOETERS, R., NDIZEYE, C., MEESSEN, B., FRITSCHE, G. B. & HETEREN, G. V. 2014. Health Facility Financial Management and the Indice Tool. *In:* FRITSCHE, G. B. (ed.) *Performance-Based Financing Toolkit.*
BRIDGEWATER, B., GRAYSON, A. D., BROOKS, N., GROTTE, G., FABRI, B. M., AU, J., HOOPER, T., JONES, M. & KEOGH, B. 2007. Has the publication of cardiac surgery outcome data been associated with changes in practice in northwest England: an analysis of 25,730 patients undergoing CABG surgery under 30 surgeons over eight years. *Heart,* 93**,** 744-8.
BRITTEN, N. 1995. Qualitative interviews in medical research. *BMJ,* 311**,** 251-3.
BURNARD, P., GILL, P., STEWART, K., TREASURE, E. & CHADWICK, B. 2008. Analysing and presenting qualitative data. *Br Dent J,* 204**,** 429-432.
CAMERON, J., BANKO, K. M. & PIERCE, W. D. 2001. Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues. *Behavior Analyst,* 24**,** 1-44.
CAMPBELL, S., KONTOPANTELIS, E., HANNON, K., BURKE, M., BARBER, A. & LESTER, H. 2011. Framework and indicator testing protocol for developing and piloting quality indicators for the UK quality and outcomes framework. *BMC Family Practice,* 12**,** 85.
CAMPBELL, S. M., REEVES, D., KONTOPANTELIS, E., SIBBALD, B. & ROLAND, M. 2009. Effects of Pay for Performance on the Quality of Primary Care in England. *New England Journal of Medicine,* 361**,** 368-378.
CANAVAN, A. & SWAI, G. 2008. Payment for Performance (P4P) Evaluation 2008 Tanzania Country Report for Cordaid. Mauritskade, Amsterdam: KIT Development Policy and Practice.
CANAVAN, A., TOONEN, J. & ELOVAINIO, R. 2008. Performance Based Financing: An international review of the literature Mauritskade, Amsterdam: KIT Development Policy & Practice
CARTER, S. & HENDERSON, L. 2005. Approaches to qualitative data collection in social science. *In:* BOWLING, A. & EBRAHIM, S. (eds.) *Handbook of research methods: Investigation, measurement, and analysis.* Berkshire, England.: Open University Press.
CASH, A. H., HAMRE, B. K., PIANTA, R. C. & MYERS, S. S. 2012. Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *EARLY CHILDHOOD RESEARCH QUARTERLY,* 27**,** 529-542.
CHAIX-COUTURIER, C., DURAND-ZALESKI, I., JOLLY, D. & DURIEUX, P. 2000. Effects of financial incentives on medical practice: results from a systematic review of the literature and methodological issues. *Int J Qual Health Care,* 12**,** 133-42.
CHARTER, R. A. & FELDT, L. S. 2002. The Importance of Reliability as it Related to True Score Confidence Intervals. *Measurement & Evaluation in Counseling & Development* 35**,** 104-112.
CHEE, G., HIS, N., CARLSON, K., CHANKOVA, S. & TAYLOR, P. 2007. Evaluation of the first five years of GAVI immunization services support funding. Behtesda, MD: GAVI Alliance.

CHEN, T. T., CHUNG, K. P., LIN, I. C. & LAI, M. S. 2011. The unintended consequence of diabetes mellitus pay-for-performance (P4P) program in Taiwan: are patients with more comorbidities or more severe conditions likely to be excluded from the P4P program? *Health Serv Res,* 46**,** 47-60.

CHUNG, S., PALANIAPPAN, L., WONG, E., RUBIN, H. & LUFT, H. 2010. Does the frequency of pay-for-performance payment matter?--Experience from a randomized trial. *Health Serv Res,* 45**,** 553-64.

CLARK, R. E. & ESTES, F. 2002. *Turning research into results: A guide to selecting the right performance solutions,* Atlanta GA, CEP Press.

COHEN, J. 1960. A coefficient of agreement for nominal scales *Educational and Psychological Measurement,* 20**,** 37-46. .

COHEN, J. 1988. *Statistical power analysis for the behavioral sciences,* Hillsdale, N.J., L. Erlbaum Associates.

COMETTO, G. 2008. Discussion paper on performance-based financing. *Countdown Conference.* Cape Town: Save the Children Policy Department Health & HIV Team.

COOPER, C. L., DYCK, B. & FROHLICH, N. 1992. Improving the Effectiveness of Gainsharing: The Role of Fairness and Participation. *Administrative Science Quarterly,* 37**,** 471-490.

COVINGTON, M. & MÜELLER, K. 2001. Intrinsic Versus Extrinsic Motivation: An Approach/Avoidance Reformulation. *Educational Psychology Review,* 13**,** 157-176.

CRESWELL, J. W. 1998. *Qualitative inquiry and research design : choosing among five traditions,* Thousand Oaks, Calif., Sage Publications.

CRESWELL, J. W. 2009. *Research design : qualitative, quantitative, and mixed methods approaches,* Thousand Oaks, CA, Sage Publications.

CRESWELL, J. W. & PLANO CLARK, V. L. 2011. *Designing and conducting mixed methods research,* Los Angeles, SAGE.

CRIFO, P. & DIAYE, M. 2007. *Incentives in Agency relationships: To be monetary or non-monetary* [Online]. http://epee.univ-evry.fr/EPEE/colloques/CrifoDiaye-EPEE.pdf. [Accessed 8 August 2012].

DE BRUIN, S. R., BAAN, C. A. & STRUIJS, J. N. 2011. Pay-for-performance in disease management: A systematic review of the literature. *BMC Health Services Research,* 11.

DECI, E. L., KOESTNER, R. & RYAN, R. M. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychol Bull,* 125**,** 627-68.

DEflOOR, T. & SCHOONHOVEN, L. 2004. Inter-rater reliability of the EPUAP pressure ulcer classification system using photographs. *Journal of clinical nursing,* 13**,** 952–959.

DESQUINS, B., HOLLY, A. & HUGUENIN, J. 2009. Physicians' working practices: target income, altruistic objectives or a maximization problem? Lausanne, Switzerland: Institute of Health Economics and Management (IEMS), University of Lausanne

DHILLON, R., BONDS, M., FRADEN, M., NDAHIRO, D. & RUXIN, J. 2012. The impact of reducing financial barriers on utilisation of a primary health care facility in Rwanda. *Global Public Health,* 7**,** 71-86.

DIXON-WOODS, M., SHAW, R. L., AGARWAL, S. & SMITH, J. A. 2004. The problem of appraising qualitative research. *Quality and Safety in Health Care,* 13**,** 223-225.

DORAN, T. 2008. Lessons from Early Experience with Pay for Performance. *Disease Management and Health Outcomes,* 16**,** 69-77.
DORAN, T., KONTOPANTELIS, E., VALDERAS, J. M., CAMPBELL, S., ROLAND, M., SALISBURY, C. & REEVES, D. 2011. *Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework.*
DULEY, L., GULMEZOGLU, A. M. & HENDERSON-SMART, D. J. 2003. Magnesium sulphate and other anticonvulsants for women with pre-eclampsia. *Cochrane Database Syst Rev,* 2.
EBEIGBE, P. N. 2013. Reducing maternal mortality in Nigeria: the need for urgent changes in financing for maternal health in the Nigerian health system. *Niger Postgrad Med J,* 20**,** 148-53.
ECCLES, M., STEEN, N., GRIMSHAW, J. & CAMPBELL, M. 2000. Experimental and quasi-experimental designs for evaluating guideline implementation strategies.
EICHLER, R. 2006. Can "Pay for Performance" Increase Utilization by the Poor and Improve the Quality of Health Services? Discussion paper for the first meeting of the Working Group on Performance-Based Incentives Centre for Global Development.
EICHLER, R. & LEVINE, R. 2009. Performance incentives for global health: potentials and pitfalls. Washington D.C.: Centre for global development.
EIJKENAAR, F. 2012. Pay for performance in health care: an international overview of initiatives. *Med Care Res Rev,* 69**,** 251-76.
EIJKENAAR, F. 2013. Key issues in the design of pay for performance programs. *EUROPEAN JOURNAL OF HEALTH ECONOMICS,* 14**,** 117-131.
EIJKENAAR, F., EMMERT, M., SCHEPPACH, M. & SCHOFFSKI, O. 2013. Effects of pay for performance in health care: a systematic review of systematic reviews. *Health Policy,* 110**,** 115-30.
EISELE, T. P., LARSEN, D. A., WALKER, N., CIBULSKIS, R. E., YUKICH, J. O., ZIKUSOOKA, C. M. & STEKETEE, R. W. 2012. Estimates of child deaths prevented from malaria prevention scale-up in Africa 2001-2010. *Malar J,* 11**,** 1475-2875.
EISENHARDT, K. M. 1989. Agency Theory: An Assessment and Review. *The Academy of Management Review,* 14**,** 57-74.
ELDRIDGE, C. & PALMER, N. 2009. Performance-based payment: some reflections on the discourse, evidence and unanswered questions. *Health Policy Plan,* 24**,** 160-6.
ELMAN, C. 2005. Explanatory Typologies in Qualitative Studies of International Politics. *International Organization,* 59**,** 293-326.
ELOVAINIO, R. 2010. Performance incentives for health in high-income countries: key issues and lessons learned. *Health systems financing.* Geneva, Switzerland: World Health Organisation.
EMMANUEL, N. K., GLADYS, E. N. & COSMAS, U. U. 2013. Consumer knowledge and availability of maternal and child health services: a challenge for achieving MDG 4 and 5 in Southeast Nigeria.(Research article). *BMC Health Services Research,* 13**,** 53.
EMMERT, M., EIJKENAAR, F., KEMTER, H., ESSLINGER, A. S. & SCHOFFSKI, O. 2012. Economic evaluation of pay-for-performance in health care: a systematic review. *Eur J Health Econ,* 13**,** 755-67.

ENGELS, E. A., SCHMID, C. H., TERRIN, N., OLKIN, I. & LAU, J. 2000. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med,* 19**,** 1707-28.
ENO, V. B. 2010. Governance constraints and health care delivery in Nigeria: The case of primary health care services in Akwa Ibom State. *Public Administration and Management,* 15**,** 342-364.
EPSTEIN, A. M. 2012. Will Pay for Performance Improve Quality of Care? The Answer Is in the Details. *New England Journal of Medicine,* 367**,** 1852-1853.
EVANS, R. G. 1974. Supplier-Induced Demand: Some Empirical evidence and Implications. *In:* PERLMAN, M. (ed.) *The Economics of Health and Medical Care* New York: Wiley.
FAPOHUNDA, B. & OROBATON, N. 2014. Factors influencing the selection of delivery with no one present in northern Nigeria: Implications for policy and programs. *International Journal of Women's Health,* 6**,** 171-183.
FEDER, M. 2008. *Encyclopedia of survey research methods [electronic resource],* London, SAGE.
FEDERAL GOVERNMENT OF NIGERIA 2011. National health bill 2009. http://www.internationalhealthpartnership.net/fileadmin/uploads/ihp/Documents/Country_Pages/Nigeria/Nigeria%20National%20Strategic%20Health%20Development%20Plan%20Framework%202009-2015.pdf.
FEDERAL REPUBLIC OF NIGERIA 2010. Ministry of Health Budget. *In:* FEDERATION, B. O. O. T. (ed.). http://www.budgetoffice.gov.ng/2010%20budget%20messeage/HEALTH.pdf.
FELT-LISK, S., GIMM, G. & PETERSON, S. 2007. Making pay-for-performance work in Medicaid. *Health Aff,* 26**,** 26.
FERRITER, M. & HUBAND, N. 2005. Does the non-randomized controlled study have a place in the systematic review? A pilot study. *Crim Behav Ment Health,* 15**,** 111-20.
FINCH, H. & LEWIS, J. 2004. Focus Groups *In:* RITCHIE, J. & LEWIS, J. (eds.) *Qualitative Research Practice: A guide for social science students and researchers* London: SAGE.
FIRESTONE, W. A. 1993. Alternative Arguments for Generalizing from Data as Applied to Qualitative Research. *Educational Researcher,* 22**,** 16-23.
FLEISS, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin.,* 76**,** 378-382.
FLEISS, J. L. 1973. *Statistical methods for rates and proportions,* New York, Wiley.
FLICK, U. 2014. *The SAGE handbook of qualitative data analysis,* Los Angeles, SAGE.
FLOM, P. 2014. *The Advantages & Disadvantages of a Multiple Regression Model* [Online]. http://www.ehow.com/info_12070171_advantages-disadvantages-multiple-regression-model.html. [Accessed 16 September 2014].
FOLLAND, S., GOODMAN, A. C. & STANO, M. 1993. *Economics of Health and Health Care,* New York, Macmillan.
FONTANA, A. & FREY, J. 2000. The interview: from structured questions to negotiated text. *In:* DENZIN, N. K. & LINCOLN, Y. S. (eds.) *The handbook of qualitative research.* 2nd ed. ed. London: Sage Publications.
FURNHAM, A. & ARGYLE, M. 1998. *The Psychology of Money,* London, Routledge.
GALE, N., HEATH, G., CAMERON, E., RASHID, S. & REDWOOD, S. 2013. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology,* 13**,** 117.

GARUBA, H., KOHLER, J. & HUISMAN, A. 2009. Transparency in Nigeria's public pharmaceutical sector: perceptions from policy makers. *Globalization and Health,* 5**,** 14.
GELMAN, A. 2006. Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. *American Statistical Association and the American Society for Quality,* 48**,** 432-435.
GILLAM, S. J., SIRIWARDENA, A. N. & STEEL, N. 2012. Pay-for-performance in the United Kingdom: impact of the quality and outcomes framework: a systematic review. *Ann Fam Med,* 10**,** 461-8.
GLASZIOU, P. P., BUCHAN, H., DEL MAR, C., DOUST, J., HARRIS, M., KNIGHT, R., SCOTT, A., SCOTT, I. A. & STOCKWELL, A. 2012. When financial incentives do more good than harm: a checklist. *BMJ,* 13.
GLICKMAN, S. W., OU, F. S., DELONG, E. R., ROE, M. T., LYTLE, B. L., MULGUND, J., RUMSFELD, J. S., GIBLER, W. B., OHMAN, E. M., SCHULMAN, K. A. & PETERSON, E. D. 2007. Pay for performance, quality of care, and outcomes in acute myocardial infarction. *Jama,* 297**,** 2373-80.
GNEEZY, U. & RUSTICHINI, A. 2000. Pay Enough or Don't Pay at All. *The Quarterly Journal of Economics,* 115**,** 791-810.
GOBO, G. 2011. Ethnography *In:* SILVERMAN, D. (ed.) *Qualitative research : issues of theory, method and practice.* 3rd ed. London: SAGE.
GORMAN, C. A. & RENTSCH, J. R. 2009. Evaluating Frame-of- Reference Rater Training Effectiveness Using Performance Schema Accuracy. *JOURNAL OF APPLIED PSYCHOLOGY,* 94**,** 1336-1344.
GRAHAM, M., A., M. & J., M. 2012. Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings. Centre for Educatior Compensation Reform
GRAVELLE, H., SUTTON, M. & MA, A. 2008. Doctor behaviour under a pay for performance contract: further evidence from the quality and outcomes framework. . *Working papers.* York, UK: Centre for Health Economics, University of York.
GREENE, W. H. & HILBE, J. M. 2008. Count response regression models. *In:* RAO, C. R., MILLER, J. & RAO, D. C. (eds.) *Handbook of Statistics: Epidemiology and Medical Statistics.* Amsterdam Elsevier.
GROL, R. 2013. *Improving patient care : the implementation of change in clinical practice,* Edinburgh, Elsevier / Butterworth Heinemann.
GROSS, R., ELHAYNAY, A., FRIEDMAN, N. & BUETOW, S. 2008. Pay-for-performance programs in P4P programs Israeli sick funds. *J Health Organ Manag,* 22**,** 23-35.
GUBA, E. 1981. Criteria for assessing the trustworthiness of naturalistic inquiries. *ECTJ,* 29**,** 75-91.
GUEST, G., BUNCE, A. & JOHNSON, L. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field Methods,* 18**,** 59-82.
GUSTAFSSON-WRIGHT, E. & VAN DER GAAG, J. 2008. An Analysis of Nigeria's Health Sector by State: Recommendations for the Expansion of the Hygeia Community Health Plan. http://www.aiid.org/uploads/File/publications/14_Analysis%20of%20Nigeria%5C's%20Health%20Sector.pdf?PHPSESSID=843e721de6b3e0880444023a29242f1b Amsterdam Institute for International Development (AIID)

GWET, K. 2002. Statistical Methods For Inter-Rater Reliability Assessment. *Kappa Statistic is not Satisfactory for Assessing the Extent of agreement Between Raters* MD: STATAXIS Consulting.
GWET, K. L. 2010. *Inter-Rater Reliability Discussion: Sample Size Determination* [Online]. http://agreestat.com/blog_irr/sample_size_determination.html [Accessed 10 December 2012].
HADDON, B. 2013. Some Lessons on Health Sector Reform, from DFID Health Programmes in Nigeria. http://www.prrinn-mnch.org/documents/lessonslearntonhealthsectorreform.pdf: DFID.
HAHN, J. 2006. Pay-for-Performance in Health Care. Washington, DC: Congressional Research Services.
HAIR, J. F., BLACK, W. C., BABIN, B. J., ANDERSON, R. J. & TATHAM, R. L. 2006. *Multivariate data analysis,* Upper Saddle River, N.J. London, Prentice Hall PTR.
HALLGREN, K. A. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol,* 8**,** 23-34.
HAMILTON, F. L., GREAVES, F., MAJEED, A. & MILLETT, C. 2013. Effectiveness of providing financial incentives to healthcare professionals for smoking cessation activities: systematic review. *Tob Control,* 22**,** 3-8.
HARBORD, R. M. & HIGGINS, J. P. T. 2008. Meta-regression in Stata. *Stata Journal,* 8**,** 493-519.
HARGREAVES, S. 2002. Time to right the wrongs: improving basic health care in Nigeria. *Lancet,* 359**,** 2030-5.
HARRELL, F. E. 2010. *Regression modeling strategies : with applications to linear models, logistic regression and survival analysis,* New York, Springer-Verlag New York Inc.
HARTLING, L., HAMM, M., MILNE, A., VANDERMEER, B., SANTAGUIDA, P. L., ANSARI, M., TSERTSVADZE, A., HEMPEL, S., SHEKELLE, P. & DRYDEN, D. M. 2012. *Validity and inter-rater reliability testing of quality assessment instruments,* Rockville MD, Agency for Healthcare Research and Quality.
HEALTH SYSTEMS 2012. Public Budgeting and Expenditure Management in Three Nigerian States: Challenges for Health Governance. file://userfs/yo508/w2k/Nigeria_Governance_Fin.pdf: USAID.
HEATH, C., LARRICK, R. P. & WU, G. 1999. Goals as reference points. *Cogn Psychol,* 38**,** 79-109.
HECHT, R., BATSON, A. & BRENZEL, L. 2004. Making healthcare accountable: why performance-based funding of health services in developing countries is getting more attention. *Finance and Development* 41**,** 16-19.
HENDERSON, L. N. & TULLOCH, J. 2008. Incentives for retaining and motivating health workers in Pacific and Asian countries. *Hum Resour Health,* 6**,** 1478-4491.
HIGGINS, J. P. & GREEN, S. 2011. *Cochrane handbook for systematic reviews of interventionsVersion 5.1.0* [Online]. www.cochrane-handbook.org. : The Cochrane Collaboration. [Accessed 20 June 2013].
HILLMAN, A. L., RIPLEY, K., GOLDFARB, N., NUAMAH, I., WEINER, J. & LUSK, E. 1998. Physician Financial Incentives and Feedback: Failure to Increase Cancer Screening in Medicaid Managed Care. *American Journal of Public Health,* 88**,** 1699-1701.

HOGAN, D., LIU, L. & MATHERS, C. 2014. CHERG-WHO methods and data sources for child causes of death 2000-2013. http://www.who.int/healthinfo/global_burden_disease/ChildCOD_method_2000_2013.pdf?ua=1: World Health Organisation.
HOSMER, D. W. & LEMESHOW, S. 2000. *Applied logistic regression,* New York Chichester, Wiley.
HOX, J. J. 2010. *Multilevel analysis : techniques and applications,* New York, Routledge.
HOYT, W. T. & KERNS, M. D. 1999. Magnitude and moderators of bias in observer ratings: A meta- analysis. *PSYCHOLOGICAL METHODS,* 4**,** 403-424.
HSEE, C. K. & ZHANG, J. 2010. General evaluability theory. *Perspectives on Psychological Science,* 5**,** 343-355
HUANG, J., YIN, S., LIN, Y., JIANG, Q., HE, Y. & DU, L. 2013. Impact of pay-for-performance on management of diabetes: a systematic review. *Journal of evidence-based medicine* 6**,** 173-84.
HULL, C. L. 1932. The Goal-Gradient Hypothesis and Maze Learning. . *Psychological Review,* 39**,** 25-43.
IGBOKWE-IBETO, C. J. 2012. Issues and challenges in local government project monitoring and evaluation in Nigeria: The way forward. *European Scientific Journal,* 8**,** 180-195.
ILESANMI, O. S., ADEBIYI, A. O. & FATIREGUN, A. A. 2014. National health insurance scheme: how protected are households in Oyo State, Nigeria from catastrophic health expenditure? *Int J Health Policy Manag,* 2**,** 175-80.
INSTITUTE OF MEDICINE 2001. Crossing the Quality Chasm: A New Health System for the 21St Century. http://www.ncbi.nlm.nih.gov/books/NBK222274/: National Academies Press, US.
IRELAND, M., PAUL, E. & DUJARDIN, B. 2011. Can performance-based financing be used to reform health systems in developing countries? *Bull World Health Organ,* 89**,** 695–698.
JEFFERY, S. 2010. *The Benefits of Tangible Non-Monetary Incentives* [Online]. http://www.customdesignmkt.com/docs/The%20Benefits%20of%20Tangible%20Non%20Monetary%20Incentives.pdf The Incentive Research Foundation Resource Center. [Accessed August 12 2012].
JHA, A. K., JOYNT, K. E., ORAV, E. J. & EPSTEIN, A. M. 2012. The Long-Term Effect of Premier Pay for Performance on Patient Outcomes. *New England Journal of Medicine,* 366**,** 1606-1615.
JONES, G., STEKETEE, R. W., BLACK, R. E., BHUTTA, Z. A. & MORRIS, S. S. 2003. How many child deaths can we prevent this year? *Lancet,* 362**,** 65-71.
KACZOROWSKI, J., GOLDBERG, O. & MAI, V. 2011. Pay-for-performance incentives for preventive care: views of family physicians before and after participation in a reminder and recall project (P-PROMPT). *Can Fam Physician,* 57**,** 690-6.
KASTNER, M., LOTTRIDGE, D., MARQUEZ, C., NEWTON, D. & STRAUS, S. E. 2010. Usability evaluation of a clinical decision support tool for osteoporosis disease management.(Research article)(Clinical report). *Implementation Science,* 5**,** 96.
KAWULICH, B. B. 2005. Participant observation as a data collection method. *Forum Qualitative Sozialforschung,* 6.

KHAN, K. S., WOJDYLA, D., SAY, L., GULMEZOGLU, A. M. & VAN LOOK, P. F. 2006. WHO analysis of causes of maternal death: a systematic review. *Lancet,* 367**,** 1066-74.
KHEMANI, S. 2004. Local Government Accountability for Service Delivery in Nigeria Washington, DC The World Bank
KIDWELL, R. E. & BENNETH, N. 1993. Employee Propensity to Withhold Effort: A Conceptual Model to Intersect Three Avenues of Research. *The Academy of Management Review,* 18**,** 429-456.
KINOTI, S. 2011. Effects of Performance-Based Financing on Maternal Care in Developing Countries: Access, Utilization, Coverage and Health Impact. Rapid Review of the Evidence. . *TRAction Project.* Washington, DC: USAID
KLEINBAUM, D. G. & KLEIN, M. 2010. *Logistic Regression: A Self-Learning Text,* New York, NY.
KLUGE, S. 2000. *Empirically Grounded Construction of Types and Typologies in Qualitative Social Research.*
KOHN, A. 1987. *Studies Find Reward Often No Motivator: Creativity and intrinsic interest diminish if task is done for gain* [Online]. http://naggum.no/motivation.html [Accessed 30 January 2013].
KOLSTAD, J. T. 2013. Information and Quality When Motivation is Intrinsic: Evidence from Surgeon Report Cards. *AMERICAN ECONOMIC REVIEW,* 103**,** 2875-2910.
KONTOPANTELIS, E., SPRINGATE, D., REEVES, D., ASHCROFT, D. M., VALDERAS, J. M. & DORAN, T. 2014. *Withdrawing performance indicators: retrospective analysis of general practice performance under UK Quality and Outcomes Framework.*
KREFTING, L. 1991. Rigor in qualitative research: the assessment of trustworthiness. *Am J Occup Ther,* 45**,** 214-22.
LANDIS, J. R. & KOCH, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics,* 33**,** 159-174.
LAU, J., IOANNIDIS, J. P. A., TERRIN, N., SCHMID, C. H. & OLKIN, I. 2006. The case of the misleading funnel plot. *BMJ : British Medical Journal,* 333**,** 597-600.
LAWANSON, A. O., OLANIYAN, O. & SOYIBO, A. 2012. National Health Accounts estimation: lessons from the Nigerian experience. *Afr J Med Med Sci,* 41**,** 357-64.
LESHABARI, M. T., MUHONDWA, E. P., MWANGU, M. A. & MBEMBATI, N. A. 2008. Motivation of health care workers in Tanzania: a case study of Muhimbili National Hospital. *East Afr J Public Health,* 5**,** 32-7.
LINCON, Y. & GUBA, G. 1985. *Naturalistic Inquiry,* Beverly Hills, CA, Sage.
LINDENAUER, P. K., REMUS, D., ROMAN, S., ROTHBERG, M. B., BENJAMIN, E. M., MA, A. & BRATZLER, D. W. 2007. Public reporting and pay for performance in hospital quality improvement. *N Engl J Med,* 356**,** 486-96.
LLANOS, K. & ROTHSTEIN, J. 2007. Physician pay-for-performance in Medicaid: A guide for states. Hamilton, NJ: Center for Health Care Strategies, Inc.
LOBBESTAEL, J., LEURGANS, M. & ARNTZ, A. 2011. Inter-rater reliability of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID I) and Axis II Disorders (SCID II). *Clin Psychol Psychother,* 18**,** 75-9.
LOCKE, R. G. & SRINIVASAN, M. 2008. Attitudes toward pay-for-performance initiatives among primary care osteopathic physicians in small group practices. *J Am Osteopath Assoc,* 108**,** 21-4.

LOEWENSTEIN, G. & PRELEC, D. 1992. Anomalies in Intertemporal Choice: Evidence and an Interpretation. *The Quarterly Journal of Economics,* 107**,** 573-597.
LUKE, D. A. 2004. *Multilevel modeling,* Thousand Oaks, Calif. London, SAGE.
LUOMA, M. 2005. The effects of the public posting of performance data on healthcare workers in Kyrgyzstan, PRIME II Project. http://www.prime2.org/prime2/pdf/PRIME%20II%20Final%20Report.pdf.
MACDERMID, J. C., BROOKS, D., SOLWAY, S., SWITZER-MCINTYRE, S., BROSSEAU, L. & GRAHAM, I. D. 2005. Reliability and validity of the AGREE instrument used by physical therapists in assessment of clinical practice guidelines. *BMC Health Serv Res,* 5**,** 18.
MANTZOUKAS, S. 2005. The inclusion of bias in reflective and reflexive research: A necessary prerequisite for securing validity. *Journal of Research in Nursing,* 10**,** 279-295.
MASON, M. 2010. Sample Size and Saturation in PhD Studies Using Qualitative Interviews. *Qualitative Social Research,* 11.
MAUGER, B., MARBELLA, A., PINES, E., CHOPRA, R., BLACK, E. R. & ARONSON, N. 2014. Implementing quality improvement strategies to reduce healthcare-associated infections: A systematic review. *American Journal of Infection Control,* 42**,** S274-S283.
MAXWELL, J. A. 2010. Using Numbers in Qualitative Research. *Qualitative Inquiry,* 16**,** 475-482.
MAYS, N. & POPE, C. 2000. Qualitative Research in Health Care: Assessing Quality in Qualitative Research. *BMJ: British Medical Journal,* 320**,** 50-52.
MCCABE, C., CLAXTON, K. & CULYER, A. J. 2008. The NICE cost-effectiveness threshold: what it is and what that means. *Pharmacoeconomics,* 26**,** 733-44.
MCCOY, D., BENNETT, S., WITTER, S., POND, B., BAKER, B., GOW, J., CHAND, S., ENSOR, T. & MCPAKE, B. 2008. Salaries and incomes of health workers in sub-Saharan Africa. *The Lancet,* 371**,** 675-681.
MCDONALD, R., BOADEN, R., ROLAND, M., KRISTENSEN, S. R., MEACOCK, R., LAU, Y.-S., MASON, T., TURNER, A. J. & SUTTON, M. 2014. Evaluation of The Advancing Quality Pay for Perfomance Programme in the NHS north west http://nets.nihr.ac.uk/__data/assets/pdf_file/0020/131726/FLS-08-1809-250.pdf: National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre.
MCDONALD, R. & ROLAND, M. 2009. Pay for performance in primary care in England and California: comparison of unintended consequences. *Ann Fam Med,* 7**,** 121-7.
MCGINN, T., WYER, P. C., NEWMAN, T. B., KEITZ, S., LEIPZIG, R. & GUYATT, G. 2005. Tips for teachers of evidence-based medicine: 3. Understanding and calculating kappa (vol 171, pg 1369, 2004). *CANADIAN MEDICAL ASSOCIATION JOURNAL,* 173**,** 18-18.
MCHUGH, M. L. 2012. Interrater reliability: the kappa statistic. *BIOCHEMIA MEDICA,* 22**,** 276-282.
MCKINNEY, J. C. 1966. *Constructive typology and social theory,* New York, Appleton-Century-Crofts.
MCNAMARA, P. 2006. Foreword: payment matters? The next chapter. *Med Care Res Rev,* 63**,** 5S-10S.
MEACOCK, R., KRISTENSEN, S. R. & SUTTON, M. 2014. THE COST-EFFECTIVENESS OF USING FINANCIAL INCENTIVES TO IMPROVE

PROVIDER QUALITY: A FRAMEWORK AND APPLICATION. *Health Economics,* 23**,** 1-13.
MEESSEN, B., MUSANGO, L., KASHALA, J. P. & LEMLIN, J. 2006. Reviewing institutions of rural health centres: the Performance Initiative in Butare, Rwanda. *Trop Med Int Health,* 11**,** 1303-17.
MEHROTRA, A., SORBERO, M. E. & DAMBERG, C. L. 2010. Using the lessons of behavioral economics to design more effective pay-for-performance programs. *Am J Manag Care,* 16**,** 497-503.
MEYRICK, J. 2006. What is Good Qualitative Research?: A First Step towards a Comprehensive Approach to Judging Rigour/Quality. *Journal of Health Psychology,* 11**,** 799-808.
MILES, M. B. & HUBERMAN, A. M. 1994. *Qualitative Data Analysis,* Thousand Oaks, CA, Sage.
MILLER, E. C. & SALKIND, N. J. 2002. *HOW RESEARCHERS CREATE THEIR OWN SCALES: AN ACTIVITY OF LAST RESORT.,* Thousand Oaks, CA, SAGE Publications, Inc.
MILLER, J. & GLASSNER, B. 2011. Qualitative research : issues of theory, method and practice. *In:* SILVERMAN, D. (ed.) 3rd ed. London: SAGE.
MOERBEEK, M. 2004. The consequence of ignoring a level of nesting in multilevel analysis. *MULTIVARIATE BEHAVIORAL RESEARCH,* 39**,** 129-149.
MOHAMMED, U. 2013. Corruption in Nigeria: a challenge to sustainable development in the Fourth Republic. *European Scientific Journal,* 9**,** 118.
MOLD, J. W., HAMM, R. M. & MCCARTHY, L. H. 2010. The law of diminishing returns in clinical medicine: how much risk reduction is enough? *J Am Board Fam Med,* 23**,** 371-5.
MORSE, J. M. 1994. Designing funded qualitative research. *In:* NORMAN, K., DENZIN & LINCOLN, Y. S. (eds.) *Handbook of qualitative research.* 2nd ed. Thousand Oaks, CA: Sage.
MORTON, S. C., ADAMS, J. L., SUTTORP, M. J. & SHEKELLE, P. G. 2004. *Meta-regression Approaches: What, Why, When, and How?* [Online]. http://www.ncbi.nlm.nih.gov/books/NBK43897/: Rockville (MD): Agency for Healthcare Research and Quality (US). [Accessed 7 July 2014].
MURPHY, E., DINGWALL, R., GREATBATCH, D., PARKER, S. & WATSON, P. 1998. Qualitative research methods in health technology assessment : a review of the literature. *Health Technol Assess,* 2**,** 1-274.
MYFORD, C. M. & WOLFE, E. W. 2009. Monitoring Rater Performance Over Time: A Framework for Detecting Differential Accuracy and Differential Scale Category Use. *JOURNAL OF EDUCATIONAL MEASUREMENT,* 46**,** 371-389.
NAHRA, T. A., REITER, K. L., HIRTH, R. A., SHERMER, J. E. & WHEELER, J. R. 2006. Cost-effectiveness of hospital pay-for-performance incentives. *Med Care Res Rev,* 63**,** 49S-72S.
NAIR, D. 2011. Project Information Document Appraisal stage: Nigeria State Health Programme Investment Credit. http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2011/11/11/000001843_20111114084158/Rendered/PDF/NSHPIC0Appraisal0PID0November010002011.pdf: The World Bank.
NAIR, D. 2012. Project Appraisal Document on a Proposed Credit. http://www-wds.worldbank.org/external/default/WDSContentServer/WDSP/IB/2012/03/27/000386194_20120327001611/Rendered/PDF/672230PAD0Box300900IDA0R20120008101.pdf: The World Bank.

NATIONAL BUREAU OF STATISTICS 2011. Nigeria Multiple Indicator Cluster Survey 2011 Summary Report.
NATIONAL HEALTH INSURANCE SCHEME 2012. NHIS Verification Exercise Report Prepared by the Health Reform Foundation of Nigeria. http://www.herfon.org/downloads/nhis_reports/nigeria_health_insurance_scheme_verification_exercise_2012.pdf.
NATIONAL POPULATION COMMISSION 2009. Nigeria Demographic and Health Survey (DHS). http://www.measuredhs.com/pubs/pdf/SR173/SR173.pdf.
NDIZEYE, C., SOETERS, R., FRITSCHE, G. B., HETEREN, G. V., BREDENKAMP, C. & MEESSEN, B. 2014. Health Facility Autonomy and Governance. *In:* FRITSCHE, G. B. (ed.) *Performance-Based Financing Toolkit.*
NIGERIA COUNT DOWN TO 2015 2012. Maternal, newborn, and child survival. http://www.countdown2015mnch.org/country-profiles/nigeria
NILAKANT, V. & RAO, H. 1994. Agency Theory and Uncertainty in Organizations: An Evaluation. *Organization Studies,* 15**,** 649-672.
NPHCDA 2012. Performance Based Financing User Manual. Abuja, Nigeria: https://nphcda.thenewtechs.com/cside/contents/docs/NSHIP-_PBF_manual_2012_version.pdf.
NYANDEKWE, M., NZAYIRAMBAHO, M. & BAPTISTE KAKOMA, J. 2014. Universal health coverage in Rwanda: dream or reality. *Pan Afr Med J,* 17.
O'NEILL, O. 2004. Accountability, trust and informed consent in medical practice and research. *Clin Med. ,* 4**,** 269-76.
OKAFOR, U. V. 2009. Challenges in critical care services in Sub-Saharan Africa: perspectives from Nigeria. *Indian J Crit Care Med.,* 13**,** 25-27.
OKOLI, U., ABDULLAHI, M. J., PATE, M. A., ABUBAKAR, I. S., ANIEBUE, N. & WEST, C. 2012. Prenatal care and basic emergency obstetric care services provided at primary healthcare facilities in rural Nigeria. *Int J Gynaecol Obstet,* 117**,** 61-5.
OREMUS, M., OREMUS, C., HALL, G. B. & MCKINNON, M. C. 2012. Inter-rater and test-retest reliability of quality assessments by novice student raters using the Jadad and Newcastle-Ottawa Scales. *BMJ Open,* 2**,** 2012-001368.
ORTLIPP, M. 2008. Keeping and using reflective journals in the qualitative research process.(Report). *The Qualitative Report,* 13**,** 695.
OSBORNE, J. W. (ed.) 2008. *Best practices in quantitative methods,* Los Angeles, Calif.: Sage.
ØVRETVEIT, J. 1998. *Evaluating health interventions : an introduction to evaluation of health treatments, services, policies, and organizational interventions,* Buckingham [England], Open University Press.
OXMAN, A. D. & FRETHEIM, A. 2009a. Can paying for results help to achieve the Millennium Development Goals? A critical review of selected evaluations of results-based financing. *J Evid Based Med,* 2**,** 184-95.
OXMAN, A. D. & FRETHEIM, A. 2009b. Can paying for results help to achieve the Millennium Development Goals? Overview of the effectiveness of results-based financing. *J Evid Based Med,* 2**,** 70-83.
OYEKALE, A. S. & ELUWA, C. G. 2009. Utilization of Health-Care and Health Insurance among Rural Households in Irewole Local Government, Osun State, Nigeria. *International Journal of Tropical Medicine,* 4**,** 70-75.
PALYS, T. & FRASER, S. 2008. Purposive Sampling. . *In:* GIVEN, L. M. (ed.) *The SAGE Encyclopedia of Qualitative Research Methods.* Thousand Oaks, CA: SAGE Publications, Inc.

PARTNERSHIPS FOR TRANSFORMING HEALTH SYSTEMS 2009. Strengthening Voice and Accountability in the Health Sector http://www.healthpartners-int.co.uk/our_expertise/documents/Voiceandaccountability.pdf: DFID.
PAUL, F. 2009. Health Worker Motivation and the Role of Performance Based Finance Systems in Africa: A Qualitative Study on Health Worker Motivation and the Rwandan Performance Based Finance Initiative in District Hospitals. *LSE International Development Working Papers.* London, United Kingdom: LSE Department of International Development (ID).
PIERCE, R. G., BOZIC, K. J. & BRADFORD, D. S. 2007. Pay for performance in orthopaedic surgery. *Clin Orthop Relat Res,* 457**,** 87-95.
POPE, G. C. 2011. Overview of Pay for Performance Models and Issues. *In:* CROMWELL, J., TRISOLINI, M. G., POPE, G. C., MITCHELL, J. B. & GREENWALD, L. M. (eds.) *Pay for Performance in Health Care:Methods and Approaches.* North Carolina: Research Traiangle Press.
PORTER, S. 2007. Validity, trustworthiness and rigour: Reasserting realism in qualitative research. *Journal of Advanced Nursing,* 60**,** 79-86.
POTVIN, M. C. 2007. Psychometric Properties: Validity. *Evidence Based Journal Club,* 1.
POWELL, A. A., WHITE, K. M., PARTIN, M. R., HALEK, K., CHRISTIANSON, J. B., NEIL, B., HYSONG, S. J., ZARLING, E. J. & BLOOMFIELD, H. E. 2012. Unintended Consequences of Implementing a National Performance Measurement System into Local Practice. *Journal of General Internal Medicine,* 27**,** 405-412.
PRICE, C. 1993. *Time, Discounting, and Value,* Oxford, Blackwell.
RAMSAY, C. R., MATOWE, L., GRILLI, R., GRIMSHAW, J. M. & THOMAS, R. E. 2003. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. *Int J Technol Assess Health Care,* 19**,** 613-23.
REDA, A. A., KAPER, J., FIKRELTER, H., SEVERENS, J. L. & VAN SCHAYCK, C. P. 2009. Healthcare financing systems for increasing the use of tobacco dependence treatment. *Cochrane Database Syst Rev,* 15.
RIMAN, H. B. & AKPAN, E. S. 2012. Healthcare Financing and Health outcomes in Nigeria: A State Level Study using Multivariate Analysis. *International Journal of Humanities and Social Science,* 2**,** 296-309.
RITCHIE, J. & LEWIS, J. 2003. *Qualitative research practice : a guide for social science students and researchers,* London, Sage.
RITCHIE, J., LEWIS, J. & ELAM, G. 2003. Designing and selecting samples. *In:* RITCHIE, J. & LEWIS, J. (eds.) *Qaulitative research practice.* Londaon: Sage.
RITCHIE, J. & SPENCER, L. 1994. Qualitative data analysis for applied policy research. *In:* BRYMAN & BURGESS (eds.) *Analysing Qualitative Data.* London: Routledge.
RIZZO, J. A. & BLUMENTHAL, D. 1994. Physician income targets: new evidence on an old controversy. *Inquiry,* 31**,** 394-404.
RIZZO, J. A. & ZECKHAUSER, R. J. 2003. Reference Incomes, Loss Aversion, and Physician Behaviour. *Review of Economics and Statistics* 85**,** 909-922.
ROBYN, P. J., BÄRNIGHAUSEN, T., SOUARES, A., TRAORÉ, A., BICABA, B., SIÉ, A. & SAUERBORN, R. 2014. Provider payment methods and health worker motivation in community-based health insurance: A mixed-methods study. *Social Science & Medicine,* 108**,** 223-236.

ROLAND, M. & CAMPBELL, S. 2014. Successes and Failures of Pay for Performance in the United Kingdom. *New England Journal of Medicine,* 370**,** 1944-1949.
ROLFE, G. 2006. Validity, trustworthiness and rigour: quality and the idea of qualitative research. *J Adv Nurs,* 53**,** 304-10.
ROODENBEKE, E. D., LUCAS, S., ROUZAUT, A. & BANA, F. 2011. Outreach services as a strategy to increase access to health workers in remote and rural areas. *Increasing access to health workers in rural and remote areas.* WHO Library Cataloguing-in-Publication World Health Organisation.
RUDNICKA, A. R. & OWEN, C. G. 2012. An introduction to systematic reviews and meta-analyses in health care.
RUSA, L. & FRITSCHE, G. 2010. *Rwanda: Performance-Based Financing in Health. Sourcebook* [Online]. http://www.mfdr.org/sourcebook/2ndEdition/4-3RwandaPBF.pdf [Accessed 10 March 2012].
RYAN, A. M. & DORAN, T. 2012. The effect of improving processes of care on patient outcomes: evidence from the United Kingdom's quality and outcomes framework. *Med Care,* 50**,** 191-9.
RYAN, R. M. & DECI, E. L. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemp Educ Psychol,* 25**,** 54-67.
SAINI, P., LOKE, Y. K., GAMBLE, C., ALTMAN, D. G., WILLIAMSON, P. R. & KIRKHAM, J. J. 2014. *Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews*.
SANDELOWSKI, M. 2001. Real qualitative researchers do not count: the use of numbers in qualitative research. *Res Nurs Health,* 24**,** 230-40.
SCOTT-EMUAKPOR, A. 2010. *The evolution of health care systems in Nigeria: Which way forward in the twenty-first century*.
SCOTT, A., SCHURER, S., JENSEN, P. H. & SIVEY, P. 2009. The effects of an incentive program on quality of care in diabetes management. *Health Econ,* 18**,** 1091-108.
SEALE, C. & SILVERMAN, D. 1997. Ensuring rigour in qualitative research. *The European Journal of Public Health,* 7**,** 379-384.
SHADISH, W. R., COOK, T. D. & CAMPBELL, D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference,* Boston, Houghton Mifflin.
SHEA, B., GRIMSHAW, J., WELLS, G., BOERS, M., ANDERSSON, N., HAMEL, C., PORTER, A., TUGWELL, P., MOHER, D. & BOUTER, L. 2007. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology,* 7**,** 10.
SHENTON, A. K. 2004. Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information,* 22**,** 63–75
SHEPPERD, J. A. 1993. Productivity loss in performance groups: a motivation analysis. *Psychol. Bull.,* 113**,** 67-81.
SHETTERLY, D. R. 2000. The Influence of Contract Design on Contractor Performance: The Case of Residential Refuse Collection. *Public Performance & Management Review,* 24**,** 53-68.
SIM, J. & WRIGHT, C. C. 2005. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy,* 85**,** 257-268.
SIMPSON, C. R., HANNAFORD, P. C., RITCHIE, L. D., SHEIKH, A. & WILLIAMS, D. 2011. Impact of the pay-for-performance contract and the management of hypertension in Scottish primary care: a 6-year population-based repeated cross-sectional study. *Br J Gen Pract,* 61.

SMAILL, F. & HOFMEYR, G. J. 2002. Antibiotic prophylaxis for cesarean section. *Cochrane Database Syst Rev,* 3.
SMITH, J. & FIRTH, J. 2011. Qualitative data analysis: the framework approach. *Nurse Res,* 18**,** 52-62.
SNIJDERS, T. A. B. & BOSKER, R. J. 2012. *Multilevel analysis : an introduction to basic and advanced multilevel modeling,* Los Angeles, London, SAGE.
SOCIETY FOR MONITORING AND EVALUATION NIGERIA. Establishing a VOPE in a plural developing economy. EvalPartners Forum, 2012 Chiang Mai, Thailand. http://www.ioce.net/download/national/Nigeria_SMEAN_Presentation.pdf.
SOETERS, R., HABINEZA, C. & PEERENBOOM, P. B. 2006. Performance-based financing and changing the district health system: experience from Rwanda. *BULLETIN OF THE WORLD HEALTH ORGANIZATION,* 84**,** 884-889.
SPITZNAGEL, E. L. 2008. Logsitic regression. *In:* RAO, C. R., MILLER, J. & RAO, D. C. (eds.) *Handbook of Statistics: Epidemiology and Medical Statistics.* Amsterdam: Elsevier.
SSENGOOBA, F., MCPAKE, B. & PALMER, N. 2012. Why performance-based contracting failed in Uganda – An "open-box" evaluation of a complex health system intervention. *Social Science & Medicine,* 75**,** 377-383.
STATA LIBRARY. 2014. *Analyzing Correlated (Clustered) Data* [Online]. http://www.ats.ucla.edu/stat/stata/library/cpsu.htm: UCLA: Statistical Consulting Group. [Accessed 23 March 2014].
STERNE, J. A. C., SUTTON, A. J., IOANNIDIS, J. P. A., TERRIN, N., JONES, D. R., LAU, J., CARPENTER, J., RÜCKER, G., HARBORD, R. M., SCHMID, C. H., TETZLAFF, J., DEEKS, J. J., PETERS, J., MACASKILL, P., SCHWARZER, G., DUVAL, S., ALTMAN, D. G., MOHER, D. & HIGGINS, J. P. T. 2011. *Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials*.
STEWARDSON, A. J., ALLEGRANZI, B., PERNEGER, T. V., ATTAR, H. & PITTET, D. 2013. Testing the WHO Hand Hygiene Self-Assessment Framework for usability and reliability. *Journal of Hospital Infection,* 83**,** 30-35.
STEWART, R. 1998. *Management of health care,* Aldershot, Hants, England, Ashgate/Dartmouth.
STIGLER, S. M. 1997. Regression towards the mean, historically considered. *Statistical Methods in Medical Research,* 6**,** 103-114.
STOCKWELL, A. 2010. *Evaluation of Financial Incentives as a Quality Improvement Strategy in the Public Hospital Context: Clinicians Attituides, Design Variables, and Economic Costs.* Doctor of Health Science, Queensland University of Technology.
STRAUSS, A. L. 1987. *Qualitative analysis for social scientists,* Cambridge [Cambridgeshire], Cambridge University Press.
STREINER, D. L. & NORMAN, G. R. 1989. *Health Measurement Scales: A practical guide to their development and use,* Oxford, Oxford University Press.
SUTTON, M., NIKOLOVA, S., BOADEN, R., LESTER, H., MCDONALD, R. & ROLAND, M. 2012. Reduced mortality with hospital pay for performance in England. *N Engl J Med,* 367**,** 1821-8.
SWINGLER, G. H. 2001. Observer variation in chest radiography of acute lower respiratory infections in children: A systematic review.
TAHRANI, A. A., MCCARTHY, M., GODSON, J., TAYLOR, S., SLATER, H., CAPPS, N., MOULIK, P. & MACLEOD, A. F. 2008. Impact of practice size on

delivery of diabetes care before and after the Quality and Outcomes Framework implementation. *Br J Gen Pract,* 58**,** 576-9.
THALER, R. 1985. Mental Accounting and Consumer Choice. *Marketing Science,* 4**,** 199-214.
THALER, R. H. 1999. Mental accounting matters. *Journal of Behavioral Decision Making,* 12**,** 183-206.
THE COMPLETE LAWS OF NIGERIA 1997. National Programme on Immunization Act. http://www.placng.org/lawsofnigeria/node/497.
THE WORLD BANK. 2013. *Nigeria: country at a glance* [Online]. http://data.worldbank.org/country/nigeria. [Accessed 16 October 2013].
THOMPSON, S. G. & SHARP, S. J. 1999. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med,* 18**,** 2693-708.
THURMOND, V. A. 2001. The point of triangulation. *J Nurs Scholarsh,* 33**,** 253-8.
TILLING, K., STERNE, J., BROOKES, S. & PETERS, T. 2005. Features and designs of randomized controlled trials and non-randomized experimental designs *In:* BOWLING, A. & EBRAHIM, S. (eds.) *Handbook of health research methods: Investigation, Measurement and Analysis.* Berkshire, England: Open University Press, Mcgraw-Hill.
TIRYAKIAN, E. A. 1968. *Typologies,* New York, Macmillan.
TOONEN, J., CANAVAN, A., VERGEER, P. & ELOVAINIO, E. 2009. Learning lessons on implementing Performance Based Financing, from a multi country evaluation. Mauritskade, Amsterdam

KIT (Royal Tropical Institute) in collaboration with Cordaid and WHO.
TOWN, R., WHOLEY, D. R., KRALEWSKI, J. & DOWD, B. 2004. Assessing the influence of incentives on physicians and medical groups. *Med Care Res Rev,* 61**,** 80S-118S.
TRANSPARENCY INTERNATIONAL. 2008. *Corruption perceptions index* [Online]. http://www.transparency.org/research/cpi/cpi_2008/0/. [Accessed January 8 2013].
TRICCO, A. C., ANTONY, J., IVERS, N. M., ASHOOR, H. M., KHAN, P. A., BLONDAL, E., GHASSEMI, M., MACDONALD, H., CHEN, M. H., EZER, L. K. & STRAUS, S. E. 2014. Effectiveness of quality improvement strategies for coordination of care to reduce use of health care services: a systematic review and meta-analysis.(Research). *CMAJ: Canadian Medical Association Journal,* 186**,** E568.
TRISOLINI, M. G. 2011. Theoretical Perspectives on Pay for Performance *In:* CROMWELL, J., TRISOLINI, M. G., POPE, G. C., MITCHELL, J. B. & GREENWALD, L. M. (eds.) *Pay for Performance in Health Care: Methods and Approaches.* North Carolina: Research Triangle Press.
TROCHIM, M. K. 2006a. *Measurement Validity Types* [Online]. http://www.socialresearchmethods.net/kb/measval.php. [Accessed 29 July 2014].
TROCHIM, M. K. 2006b. *Social research method* [Online]. http://www.socialresearchmethods.net/kb/contents.php [Accessed 12 January 2013].
TVERSKY, A. & KAHNEMAN, D. 2004. Loss Aversion in Riskless Choice: A Reference –Dependent Model *In:* TVERSKY, A. (ed.) *Preference, Belief, and Similarity: Selected Writings.* . http://cseweb.ucsd.edu/~gary/PAPER-

SUGGESTIONS/Preference,%20Belief,%20and%20Similarity%20Selected%20Writings%20(Bradford%20Books).pdf
TWARDELLA, D. & BRENNER, H. 2007. Effects of practitioner education, practitioner payment and reimbursement of patients' drug costs on smoking cessation in primary care: a cluster randomised trial. *Tob Control,* 16**,** 15-21.
UBERSAX, J. 2010. *Kappa coefficients, a critical appraisal* [Online]. http://www.john-uebersax.com/stat/kappa.htm. [Accessed 10 December 2013].
UMUKORO, N. 2012. Governance and Public Health Care in Nigeria. *Journal of Health Management,* 14**,** 381-395.
UNEKE, C. J., EZEOHA, A. E., NDUKWE, C. D., OYIBO, P. G. & ONWE, F. 2010. Development of health policy and systems research in Nigeria: lessons for developing countries' evidence-based health policy making process and practice. *Healthc Policy,* 6**,** e109-26.
UNEKE, C. J., EZEOHA, A. E., NDUKWE, C. D., OYIBO, P. G. & ONWE, F. 2013. Promotion of health sector reforms for health systems strengthening in Nigeria: perceptions of policy makers versus the general public on the Nigeria health systems performance. *Soc Work Public Health,* 28**,** 541-53.
VAN DEN NOORTGATE, W., OPDENAKKER, M. C. & ONGHENA, P. 2005. The effects of ignoring a level in multilevel analysis. *SCHOOL EFFECTIVENESS AND SCHOOL IMPROVEMENT,* 16**,** 281-303.
VAN HERCK, P., DE SMEDT, D., ANNEMANS, L., REMMEN, R., ROSENTHAL, M. & SERMEUS, W. 2010. Systematic review: Effects, design choices, and context of pay-for-performance in health care. *BMC Health Services Research,* 10**,** 1-13.
VERGEER, P. & CHANSA, C. 2008. Payment for Performance (P4P) Evaluation 2008 Zambia Country Report for Cordaid Amsterdam: KIT Development Policy & Practice
VIERA, A. J. & GARRETT, J. M. 2005. Understanding interobserver agreement: the kappa statistic. *Fam Med,* 37**,** 360-3.
VITTINGHOFF, E., GLIDDEN, D. V., SHIBOSKI, S. C. & MCCULLOCH, C. E. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models,* Springer New York.
VROOM, V. H. 1964. *Work and Motivation,* New York, McGraw Hill.
WAGSTAFF, A., CLAESON, M., HECHT, R. M., GOTTRET, P. & FANG, Q. 2006. Millennium Development Goals for Health: What Will It Take to Accelerate Progress? *In:* JAMISON, D. T., BREMAN, J. B., MEASHAM, A. R., ALLEYNE, G., CLAESON, M., EVANS, D. B., JHA, P., MILLS, A. & MUSGROVE, P. (eds.) *Disease Control Priorities in Developing Countries.* 2nd ed. Washington, DC: The World Bank.
WALTER, S. D., ELIASZIW, M. & DONNER, A. 1998. Sample size and optimal designs for reliability studies. *Stat Med,* 17**,** 101-10.
WELCOME, M. O. 2011. The Nigerian health care system: Need for integrating adequate medical intelligence and surveillance systems.
WERNER, R. M., KOLSTAD, J. T., STUART, E. A. & POLSKY, D. 2011. The effect of pay-for-performance in hospitals: lessons for quality improvement. *Health Aff,* 30**,** 690-8.
WERNER, R. M., KONETZKA, R. T. & POLSKY, D. 2013. The effect of pay-for-performance in nursing homes: evidence from state Medicaid programs. *Health Serv Res,* 48**,** 1393-414.

WHO. 2012. *Nigeria Health Profile* [Online]. http://www.who.int/gho/countries/nga.pdf. [Accessed 10 July 2012].
WHO. 2013. *The Nigerian Health system* [Online]. http://www.who.int/pmnch/countries/nigeria-plan-chapter-3.pdf [Accessed 4 January 2013].
WILLIS-SHATTUCK, M., BIDWELL, P., THOMAS, S., WYNESS, L., BLAAUW, D. & DITLOPO, P. 2008. Motivation and retention of health workers in developing countries: a systematic review. *BMC Health Services Research,* 8**,** 247.
WITTER, S., FRETHEIM, A., KESSY, F. L. & LINDAHL, A. K. 2012. Paying for performance to improve the delivery of health interventions in low- and middle-income countries. *Cochrane Database Syst Rev,* 15.
WOEHR, D. J. & HUFFCUTT, A. I. 1994. Rater training for performance appraisal: a quantitative review. (includes appendices). *Journal of Occupational and Organizational Psychology,* 67**,** 189.
WYNIA, M. K. 2009. The Risks of Rewards in Health Care: How Pay-for-performance Could Threaten, or Bolster, Medical Professionalism. *Journal of General Internal Medicine,* 24**,** 884-887.
XU, K., SAKSENA, P., JOWETT, M., INDIKADAHENA, C., KUTZIN, J. & EVANS, D. B. 2010. Exploring the thresholds of health expenditure for protection against financial risk. *Health Sysytems financing.* World Health Organisation
YAKOOB, M. Y. & BHUTTA, Z. A. 2011. Effect of routine iron supplementation with or without folic acid on anemia during pregnancy. *BMC Public Health.*
YOUNG, G. J., METERKO, M., WHITE, B., BOKHOUR, B. G., SAUTTER, K. M., BERLOWITZ, D. & BURGESS, J. F., JR. 2007. Physician attitudes toward pay-for-quality programs: perspectives from the front line. *Med Care Res Rev,* 64**,** 331-43.
YOUNG, G. J., WHITE, B., BURGESS, J. F., JR., BERLOWITZ, D., METERKO, M., GULDIN, M. R. & BOKHOUR, B. G. 2005. Conceptual issues in the design and implementation of pay-for-quality programs. *Am J Med Qual,* 20**,** 144-50.