

Applying Process-Oriented Data Science to Dentistry

Frank Gerard Fox

Submitted in accordance with the requirements of the degree of
Doctor of Philosophy

The University of Leeds

School of Dentistry

School of Computing

January 2019

Intellectual Property and Publication Statements.

The candidate confirms that the work submitted is his own except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in Section 10.19.3 of the thesis has appeared in publications as follows:

“A Data Quality Framework for Process Mining of Electronic Health Record Data”, **F.Fox**, V.R.Aggarwal, H.Whelton, O.Johnson, IEEE International Conference on Healthcare Informatics Proceedings, P.12-21, New York, 2018.

FF & OJ conceived and planned the theoretical framework with support from VA & HW. FF applied the framework to the use case data and identified the key data quality issues. FF wrote initial drafts & final draft of manuscript with input from all authors. All authors discussed results and commented on the manuscript generally.

“The ClearPath Method for Care Pathway Process Mining and Simulation”. Owen Johnson, Angelina Prima Kurniati, **Frank Fox**, Eric Rojas and Thamer Ba Dhafari, International Workshop on Process-Oriented Data Science for Healthcare, Sydney, 2018. FF contributed the section on data quality and discussed results and commented on the manuscript generally.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgment.

The right of Frank G. Fox to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

Acknowledgements

My heartfelt thanks to my three supervisors, Dr Vishal Aggarwal, Mr Owen Johnson and Professor Helen Whelton, first, for their faith in me and second, for their guidance and encouragement over the last three years. I would like to say a big thank you to my last principal supervisor, Vishal, for his support over the last year when getting the job finished was a real challenge, to Helen for her inspirational dedication to improving population oral health and to Owen for sharing his vast experience in applying data analytics in healthcare settings. They made my PhD experience challenging, inspiring, and fun.

Thanks to everyone in the University of Leeds School of Dentistry ADVOCATE office, especially Amy and Heather, for your help and all the fun we have had in the last few years. I would also like to thank my friends in both Ireland and the UK for all your support in helping keep things running smoothly during my studies. I would like to thank my friend and business partner Brendan Burke for supporting me throughout this work. Thanks to the staff of the Oral Health Services Research Centre in University College Cork, Mairead Harding, Patrice James and Maria Tobin for their help and support in this venture. I owe a debt of gratitude to the late Dr Gerard Meehan, Galway and Dr Niall O'Neil, Cork for the initial risks they took and their inspiration and support over many years.

I owe a huge thank you to my family. I couldn't wish for better: my parents, Frank and Maura, for their love, support and belief in me always; my sons Eamonn, Joey, and Tom for being loving, responsible, and trustworthy lads. I could not have undertaken this work without knowing I could rely on them every day I was away. And to Majella, for her constant love, support and good humour in the face of frequent absences, and for always encouraging me take the opportunities presented, whatever the cost.

Abstract

Background: Healthcare services now often follow evidence-based principles, so technologies such as process and data mining will help inform their drive towards optimal service delivery. Process mining (PM) can help the monitoring and reporting of this service delivery, measure compliance with guidelines, and assess effectiveness. In this research, PM extracts information about clinical activity recorded in dental electronic health records (EHRs) converts this into process-models providing stakeholders with unique insights to the dental treatment process. This thesis addresses a gap in prior research by demonstrating how process analytics can enhance our understanding of these processes and the effects of changes in strategy and policy over time. It also emphasises the importance of a rigorous and documented methodological approach often missing from the published literature. **Aim:** Apply the emerging technology of PM to an oral health dataset, illustrating the value of the data in the dental repository, and demonstrating how it can be presented in a useful and actionable manner to address public health questions. A subsidiary aim is to present the methodology used in this research in a way that provides useful guidance to future applications of dental PM. **Objectives:** Review dental and healthcare PM literature establishing state-of-the-art. Evaluate existing PM methods and their applicability to this research's dataset. Extend existing PM methods achieving the aims of this research. Apply PM methods to the research dataset addressing public health questions. Document and present this research's methodology. Apply data-mining, PM, and data-visualisation to provide insights into the variable pathways leading to different outcomes. Identify the data needed for PM of a dental EHR. Identify challenges to PM of dental EHR data. **Methods:** Extend existing PM methods to facilitate PM research in public health by detailing how data extracts from a dental EHR can be effectively managed, prepared, and used for PM. Use existing dental EHR and PM standards to generate a data reference model for effective PM. Develop a data-quality management framework. **Results:** Comparing the outputs of PM to established care-pathways showed that the dataset facilitated generation of high-level pathways but was less suitable for detailed guidelines. Used PM to identify the care pathway preceding a dental extraction under general anaesthetic and provided unique insights into this and the effects of policy decisions around school dental screenings. **Conclusions:** Research showed that PM and data-mining techniques can be applied to dental EHR data leading to fresh insights about dental treatment processes. This emerging technology along with established data mining techniques, should provide valuable insights to policy makers such as principal and chief dental officers to inform care pathways and policy decisions.

Table of Contents

ABSTRACT	4
1 INTRODUCTION.....	13
1.1 Background	15
1.2 Research Technology Terms.....	16
1.3 Dental Domain Terms	21
1.4 Linking Process Mining to Care Pathways and Clinical Guidelines.....	31
1.5 Structure of the Thesis	33
2 LITERATURE REVIEW	34
2.1 Previous Literature Reviews & Related work.....	34
2.2 Process Mining Tools, Discovery Algorithms and Techniques	36
2.3 Data Mining in Dentistry	38
2.4 Review Method	39
2.5 Review Results	41
2.6 This Research’s Process Mining Vocabulary	49
3 RESEARCH AIMS, OBJECTIVES, AND RESEARCH QUESTIONS	64
3.1 Aims	64
3.2 Objectives	64
3.3 Research Questions.....	65
4 THE DATA AND THE RESEARCH DATA ENVIRONMENT	67
4.1 Research Data Description.....	67
4.2 Data Pipeline Environment.....	87
4.3 System Environment & Architecture.....	88
5 CHALLENGES WHEN APPLYING PM TO ROUTINE DENTISTRY DATA.....	90
5.1 Introduction	90
5.2 Data Access.....	91
5.3 Data Quality Management in this Research.....	93
5.4 Process Model Quality.....	99
5.5 Data Transforms.....	98
6 METHODOLOGY.....	104
6.1 Introduction	104
6.2 How Process Mining in Dentistry Fits in the Research Landscape.	104

6.3	Process Mining Project Methods	107
6.4	Extending the Existing Methods for Dentistry Research.....	120
6.5	Policy and Strategy Questions Methodological Approach (RQ4)	130
6.6	Conclusion.....	131
7	VALIDATION OF THE METHODOLOGY: EXPERIMENTS AND RESULTS	132
7.1	Introduction.....	132
7.2	Assessing Compliance with Care Pathways and Clinical Guidelines	132
7.3	Establishing the Treatment Pathway for a Specific Outcome.....	144
7.4	Assessing the Impact of ‘frequency of screening’ Policies	152
7.5	Assessing the Impact of ‘age at first screening’ Policies	169
7.6	Rejected Validating Question.....	182
7.7	What Data is Needed in an EHR for Effective PM? (RQ5)	183
8	DISCUSSION	194
8.1	Introduction and Overview	194
8.2	Reflections on the Approach	194
8.3	Managing the Data Environment, the Data Quality, and the Data Analysis	195
8.4	Principal Outputs of this Research	200
8.5	Limitations of the Study	204
8.6	What unique insights does PM bring to analysing healthcare processes?	205
8.7	Meanings and Implications for Clinicians and Policymakers	206
9	CONCLUSIONS	208
9.1	Review of Research Questions	208
9.2	Future Research Opportunities	210
10	APPENDICES	222
10.1	Data Management Plan.....	222
10.2	Data Mappings For standardisation and SNOMED	224
10.3	BridgesPM1 Data attributes	229
10.4	Ethical Approval & Data-Owner Permission	234
10.5	Other Governing Documents.....	235
10.6	Anonymisation Standard Planning Record.....	236
10.7	Sample Bridges EHR Application screen	236
10.8	Medical Questionnaire Questions (Alphabetically)	237
10.9	How DMFT is calculated in this research.	237
10.10	Posters and Oral Presentations	238
10.11	Code Reuse Guide	242
10.12	RQ1 SQL cohort selection code (as sample).....	243

10.13	Frequency of Screening Details	244
10.14	Age at first screening details	258
10.15	Screening Base Data	263
10.16	Age at First Screening Process Mining Output	265
10.17	Data Quality Issues	280
10.18	Data Transforms	286
10.19	Data Quality Framework	288
10.20	Dental Literature Review Details	303
10.21	Application of the ADF	310

List of Figures

Figure 1-1: The Data Science Process (O'Neil & Schutt, 2014, p. 41)	17
Figure 1-2: The Data Scientist's Role (O'Neil & Schutt, 2014, p. 44)	17
Figure 1-3: Data and Process Science Skills (van der Aalst, 2016, p. 18)	18
Figure 1-4: Data Mining in the Knowledge Discovery from Databases chain (Dragon1.com, 2018)	19
Figure 1-5: Machine Learning Types (Mathworks, 2018)	19
Figure 1-6: Process Mining Types and Environment (Mans, et al., 2015, p. 22)	21
Figure 1-7: Main kinds of organisational healthcare processes (Mans, et al., 2015, p. 13)	23
Figure 1-8: Factors involved in caries development (Selwitz, et al., 2007)	26
Figure 2-1: Literature search results and removal criteria	40
Figure 2-2: Process Mining in Dentistry adapted from Mans et al. (2012)	42
Figure 2-3. Evaluating the impact of IT using Process Mining adapted from Mans et al. (2013)	42
Figure 2-4: Is your upgrade worth it? adapted from van Genuchten et al. (2014)	43
Figure 2-5: Example Process Model	53
Figure 2-6: Process Mining Data-level Vocabulary Model	54
Figure 4-1: Organisation of Public Dental Services for children in Ireland. (Irish Oral Health Services Guideline Initiative, 2012, p. 12)	67
Figure 4-2: Sample Bridges Odontogram	68
Figure 4-3: Entity Relationship Overview of the BridgesPM1 database	72
Figure 4-4: Clients' year of birth histogram for data collected 1998-2014	75
Figure 4-5: Clients' Nationalities	76
Figure 4-6: Procedure counts 1998-2014	77
Figure 4-7: Distribution of DMFT values for patients examined in 2007	78
Figure 4-8: Ages when treatment received for data collected 2000-2015	78
Figure 4-9: Ages at first School Screening Histograms	79
Figure 4-10: Geo-map of HSE South high DMFT values (>3).	80
Figure 4-11: Population Density (Central Statistics Office (Ireland), 2012, p. 12)	80
Figure 4-12: Recorded Medical Conditions and DMFT Distribution	82
Figure 4-13: Overall DMFT Values (2000-2014)	83
Figure 4-14: DMFT Values, by tooth number for two sample years 2005 & 2014	84
Figure 4-15: DMFT Values for the 6s, for the years 2000-2015	84
Figure 4-16: DMFT Heat map, All Areas, All DMFT values, starting 2007	86
Figure 4-17: Data Pipeline Environment	87
Figure 4-18: Process Mining Environment (Adapted from Mans, et al. (2015, p. 22)) ..	88
Figure 4-19: System Architecture representation adapted from Santos, et al. (2013, p. 275)	89
Figure 6-1: The Research Onion (University of Derby, 2018)	105

Figure 6-2: Deduction (top down) & Induction (bottom up) approaches to research...	106
Figure 6-3: Phases of the methodology (Bozkaya, et al., 2009, p. 23)	108
Figure 6-4: Proposed Method for BPA in healthcare (Rebuge & Ferreira, 2012, p. 107)	110
Figure 6-5: The Sequence Clustering Analysis subprocess (Rebuge & Ferreira, 2012, p. 108)	110
Figure 6-6: L* life-cycle methodology (IEEE, 2011).....	112
Figure 6-7: PM ² Method Steps.....	113
Figure 6-8: Method for the analysis of medical treatment processes (Rovani, et al., 2015)	115
Figure 6-9: Question-Driven Methodology for Analyzing Emergency Room Process Using Process Mining (Rojas, et al., 2017).....	116
Figure 6-10: Policy and Strategy Questions Methodology with Example.....	131
Figure 7-1: Proposed Dental Care Pathway (NHS England, 2009, p. 45).....	133
Figure 7-2: Event Log Characteristics (Care Pathway Compliance)	134
Figure 7-3: Routine and Urgent Care Pathway generated from a single week of data.	136
Figure 7-4L Urgent Care Pathway generated from a single week of data.	137
Figure 7-5: Routine Care Pathway generated from a single week of data.....	137
Figure 7-6: Oral Health Assessment Program Proposal (adapted from Irish Oral Health Services Guideline Initiative (2012, pp. 6, 7))	141
Figure 7-7: Fissure Sealant Cycle (Irish Oral Health Services Guideline Initiative, 2010, p. 6)	143
Figure 7-8: Number of GA Extractions by age (2004-2014).....	146
Figure 7-9: Number of Prescriptions by age (2004-2014).....	146
Figure 7-10: Number of Prescriptions by age - followed by GA (2004-2014).....	146
Figure 7-11: Process mining frequency analysis of General Anaesthetic Extractions. Temporal sequence for teeth extracted under general anaesthetic between 2004 and 2014 and all preceding events.	147
Figure 7-12: Process Mining performance analysis of General Anaesthetic Extraction. Temporal sequence for teeth extracted under general anaesthetic between 2004 and 2014 and all preceding events.	148
Figure 7-13: Detail of paths taken between Amalgam Filling and GA Extraction.....	149
Figure 7-14: Frequency of Screening Outcomes Calculation	158
Figure 7-15: Default Frequency Model for 2 screenings. Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.	160
Figure 7-16: Default Performance Model for 2 screenings. Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.	160
Figure 7-17: Frequency model enhanced with 'rank' for 2 Screenings (Kerry). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.....	163
Figure 7-18: Performance model enhanced with 'rank' for 2 Screenings (Kerry). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.....	163
Figure 7-19: Frequency model enhanced with 'rank' for 3 Screenings (West Cork). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.....	164
Figure 7-20: Performance model enhanced with 'rank' for 3 Screenings (West Cork). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.....	165
Figure 7-21: Age at First Screening, Outcomes calculation	173

Figure 7-22: DMFT at 12/13 by age-at-first-screening for patients with a baseline DMFT=0	174
Figure 7-23: DMFT at 12/13 by age-at-first-screening for patients with a baseline DMFT>0	175
Figure 7-24: Distribution of times between screenings,	180
Figure 7-25: Distribution of times between screenings,	180
Figure 7-26: DMFT at age 12/13 & time between.....	180
Figure 7-27: HRM comparison to BridgesPM1	185
Figure 7-28: Data Model extension to cater for Care Pathways	189
Figure 7-29: Data Model extension to cater for Diagnosis-Treatment pairs	190
Figure 7-30: Data Model extension to cater for CPITN	191
Figure 7-31: Data Model extension with Procedure Mappings	192
Figure 7-32: Proposed Dental Data Reference Model	192
Figure 8-1: Proposed System Architecture adapted from Santos et al. (2013, p. 275).	202
Figure 10-1: Event Log Characteristics (Frequency of school screening).....	244
Figure 10-2: Frequency model enhanced with 'rank & DMFT' for 2 Screenings (Kerry). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.....	246
Figure 10-3: Performance model enhanced with 'rank & DMFT' for 2 Screenings (Kerry). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.....	247
Figure 10-4: Frequency model enhanced with 'rank & DMFT' for 3 Screenings (West Cork). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.....	248
Figure 10-5: Performance model enhanced with 'rank & DMFT' for 3 Screenings (West Cork). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.....	249
Figure 10-6: Policy & Screening Profile for Area (Kerry)	251
Figure 10-7: Policy & Screening Profile for Area (North Cork)	252
Figure 10-8: Policy & Screening Profile for Area (North Lee)	253
Figure 10-9: Policy & Screening Profile for Area (South Lee)	254
Figure 10-10: Policy & Screening Profile for Area (West Cork)	255
Figure 10-11: Event Log Characteristics (Age at 1 st school screening).....	258
Figure 10-12: Results for 2 & 3 Screenings, Initial DMFT=0.....	260
Figure 10-13: Results for 2 & 3 Screenings, Initial DMFT > 0.....	261
Figure 10-14: Default output from Disco for 6-year olds. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008	265
Figure 10-15: Performance output from Disco for 6-year olds. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008	266
Figure 10-16: Disco output showing 100% detail, excerpt below right.	267
Figure 10-17: Process model detail for age at first screening = 6. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008.	268
Figure 10-18: Process model detail for age at first screening = 7. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008.	269
Figure 10-19: Process model detail for age at first screening = 8. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008	270
Figure 10-20: Process model detail for age at first screening = 9. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008.	271

Figure 10-21: First Screening Age 6 – Frequency. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008	272
Figure 10-22: First Screening Age 6 – Performance. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008.	273
Figure 10-23: First Screening Age 7 – Frequency. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008.	274
Figure 10-24: First Screening at age 7 – Performance. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008.	275
Figure 10-25: First Screening at age 8 - Frequency. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008	276
Figure 10-26: First Screening at age 8 – Performance. Temporal sequence for patients receiving first screening between January 1 st , 2004 and December 31 st , 2008.	277
Figure 10-27: First Screening at age 9 – Frequency. Temporal sequence for patients receiving first screening	278
Figure 10-28: First Screening at Age 9 – Performance. Temporal sequence for patients receiving first screening	279
Figure 10-29: CP-DQF Entity Relationship Diagram.....	290
Figure 10-30: CP-DQF (Step 1).....	291
Figure 10-31: CP-DQF (Step 2).....	292
Figure 10-32: CP-DQF (Step 3).....	293
Figure 10-33: Example of Research Data with Metadata added	300
Figure 10-34: Data Flow between multiple environments.....	310

List of Tables

Table 2-1: Summary of dental process mining literature analysis	45
Table 2-2: Summary of Literature's PM 'types' and 'perspectives'	57
Table 2-3: Process Mining Types and Perspectives.....	58
Table 4-1: BridgesPM1 Data Classes	71
Table 4-2: Recorded Medical Conditions and DMFT Distribution	81
Table 5-1: Event Logging Guidelines, adapted from van der Aalst (2016, p. 152))	98
Table 6-1: Process Mining Methods Summary.....	119
Table 6-2: Extended Methodology Steps.....	128
Table 6-3: Policy and Strategy Questions Methodology	130
Table 7-1: Classes targeted by area, 2005 (from (UCC/HRB, 2005/6)	153
Table 7-2: Level of policy attainment* (%) based on the number of patients seen in first targeted year having DMFT=0.....	155
Table 7-3: Level of policy attainment* (%) based on the number of patients seen in first targeted year having all DMFT values.....	155
Table 7-4: Initial DMFT and Final DMFT for Frequency of Screenings	158
Table 7-5: DMFT Distribution at 2 nd or 3 rd Screening.....	167
Table 7-6: Summary of the Age at first screening Process Model Characteristics (2004-2008), 3 Screenings, Baseline DMFT=0.....	178
Table 7-7: Average number of months between 1 st & 2 nd Screening, and between 2 nd & 3 rd Screening related to DMFT outcome at 3 rd screening, broken down by age at 1 st screening	179
Table 7-8: HRM Classes mapped to BridgesPM1 Classes	186
Table 7-9: HRM Process Steps Classes mapped to BridgesPM1 Classes	186
Table 7-10: HRM Medication Classes mapped to BridgesPM1 Classes.....	187
Table 10-1: Base frequency of screening data with patients having initial DMFT=0..	256
Table 10-2: Base frequency of screening having initial DMFT >0 values	257
Table 10-3: Starting DMFT=0, Number of school screenings =2	261
Table 10-4: Starting DMFT=0, Number of school screenings =3	262

Table 10-5: Starting DMFT > 0, Number of school screenings =2	262
Table 10-6: Starting DMFT >0, Number of school screenings =3	262
Table 10-7: Base data with patients having initial DMFT=0.....	263
Table 10-8: Base data with patients having initial DMFT > 0.....	264
Table 10-9: 27 Data Quality Issues (adapted from (Mans, et al., 2015).....	295
Table 10-10: Process Characteristics leading to DQ issues adapted from Bose, et al. (2013).....	296
Table 10-11: Sample of Data Quality Registry entries	301
Table 10-12: Data Features from ADF	311
Table 10-13: Data Extracted and Anonymisation Steps	312

Glossary

ADF	Anonymisation Decision Framework
ANN	Artificial Neural Network
AUC	Area Under the Curve
BPR	Business Process Reengineering
BPI	Business Process Improvement
BPMN	Business Process Modelling Notation
BPM	Business Process Management
BSODR	British Society for Oral and Dental Research
CART	Classification and Regression Tree
CD	Compact Disk
CG(s)	Clinical Guideline(s)
CAD/CAM	Computer Aided Design/Computer Aided Manufacturing
CP-DQF	Care Pathway Data Quality Framework
CPI	Continuous Process Improvement
CPITN	Community Periodontal Index of Treatment Needs
CRM	Customer Relationship Management
CSV	Comma Separated Value
Disco	Data Mining Software Product
DEHR	Dental Electronic Health Record
DM	Data Mining
DMFT	Decayed Missing Filled Teeth (permanent), D is caries to dentinal threshold - D ₃
dmft	Decayed Missing Filled Teeth (deciduous), d is caries to dentinal threshold - d ₃
DQ	Data Quality
ECC	Early Childhood Caries
EHR	Electronic Health Record
EL	Event Log
ER	Entity Relationship
ERP	Enterprise Resource Planning
ETL	Extract Transform Load
EU	European Union
FDI	World Dental Federation
FS	Fissure Sealant
GA	General Anaesthetic
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
GUID	Globally Unique Identifier
HRM	Healthcare Reference Model
HSE	Health Service Executive
ICDAS	International Caries Detection and Assessment System
IE	Initial Examination

IRC	Integrated Research Centre
IT	Information Technology
KPI	Key Performance Indicator
LIDA	Leeds Institute for Data Analytics
NHS	National Health Service
OHI	Oral Health Instruction
OHRSC	Oral Health Research Services Centre
PCRC	Primary Care Research Committee
PDM	Process Diagnostics Method
PHN	Public Health Nurse
PODS	Process-Oriented Data Science
PM	Process Mining
ProM	Data Mining Software Product
QoL	Quality of Life
RCT	Randomised Control Trial
ROC	Receiver Operating Characteristic curves
RQ(s)	Research Question(s)
SNODENT	Standard Nomenclature for Dentistry
SNOMED	Standard Nomenclature for Medicine
SQL	Structured Query Language
TFA	Topical Fluoride Application
TQM	Total Quality Management
UCC	University College Cork
UoL	University of Leeds
VSM	Value Stream Mapping
WHO	World Health Organisation

1 Introduction

Process-oriented data science is an emerging discipline dedicated to extracting high-level process knowledge from low-level event data commonly available in organisations' information systems. It combines traditional process analysis and data-centric analysis and its key tool, process mining (PM), delivers unique insights into the way healthcare is delivered by facilitating the discovery of treatment pathways and the creation of their associated process models. These insights assist in discovering the true care pathways experienced by patients, and subsequently monitoring and enhancing these - a task central to the continuous improvement of care delivery. PM also facilitates both checking the conformance of these models with established models and also the models' enhancement using additional information from event data such as performance and resource details. This research demonstrates how process-oriented data science techniques can extract information about clinical activity from Dental Electronic Health Records (EHRs) and generate visualisations and process models, providing policy makers with unique, actionable insights into the dental treatment process.

Although these technologies have been applied to healthcare generally, dentistry has been largely ignored. Specifically, the application of these technologies to large datasets such as those available from dental public health EHRs or insurance databases has not been explored. This thesis showcases how PM can be used to illustrate the value of the data in these repositories and how it can be presented in a useful and actionable manner to address public health questions. As an example, identifying the cause of a population's oral health problems and planning effective interventions is a key function of a dental public health service, however, evaluating the pathways of delivery of such interventions in primary dental care, where the majority of dentistry is delivered, has proven difficult and time consuming. This is where PM shows its worth. As part of a structured methodology documented in this research, PM not only facilitates discovery of the treatment processes experienced by patients but also contextualises this within a strict data provenance protocol and a comprehensive data description and profile.

Importantly, PM is just one step in this research. It does not stand alone. As it is an emerging technology using EHR data, it is anchored in existing, established technologies and PM research methods. The work in this thesis documents the key steps in a robust end-to-end methodology for the application of PM to a dataset extracted from an EHR. For convenience, this methodology is known as PM4D (Process Mining for Dentistry) and is supported by a rigorous data quality assessment. PM4D has distinct steps, each consisting of actions, inputs and outputs, and documentation and artefacts: planning, data

modelling, ethics and permissions, research environment definition and preparation, data extraction, data pre-processing, data quality assessment, data description and profiling, incorporation of EHR considerations, data transforms, PM and analysis, evaluation, and process improvement and support. The methodology is used to address the following questions:

Research Question 1: *Can PM discover care pathways, from a dental EHR?*

Research Question 2: *Can PM help assess compliance of real-world processes with recommended care pathways and clinical guidelines?*

Research Question 3: *Can PM discover dental care pathways associated with a specific outcome – e.g. extraction under general anaesthetic?*

Research Question 4: *Is PM and PM4D capable of assessing the impact of policy changes on service delivery and oral health outcomes, from the dental EHR.*

To answer these questions, modern data and process mining technologies are being applied to a dataset extracted from an Irish public health dental EHR known as Bridges. This data extract, known as BridgesPM1, contains dental clinical and administrative data on over 200,000 children who accessed Ireland's dental public health system. It is hoped that this work will inform Irish dental public health policy and be generalisable to the U.K.'s National Health System and other international public health datasets to inform care pathways and policy decisions. The dataset is described in detail in Chapter 4.

The research shows that PM can provide valuable insights and information to stakeholders on the delivery of dental services. Assessing the effects of strategy and policy changes on oral health status and outcomes can be assisted using PM and data mining techniques. These techniques could inform the drive towards optimal service delivery strategies such as remuneration methods, dental contracts, avoidance of unnecessary treatments, and compliance with guidelines and evidence-based principles. The findings will feed back to the Irish public health service and will be generalizable to international public health providers.

Resulting from the application of the methodology and the validating experiments, this research provides a number of valuable developments and potentially publishable advances in the domain:

- Initiated development of a consistent vocabulary for PM

- Documented an enhanced methodology for PM of dental EHR forming the basis for a method capable of managing the specific requirements of dental research. This would benefit from validation with further datasets.
- Visualisation and profiling of a public health dental EHR.
- Proposed data reference model for dental PM.
- Addressing data quality of a public health dental EHR.
- Developed and implemented a Data Quality Framework.
- Architecture and environment specification used for PM of dental EHR data.
- Demonstration of the flexibility of using EHR data in research. The applied techniques and methods demonstrate flexibility and agility and form the basis for a data product capable of providing ongoing, robust, and actionable insights to domain stakeholders.
- Application of the PM4D to data from an Irish public health EHR, validating PM4D and showing how it can be generalised to U.K.'s NHS and other international datasets.

This research does not carry out a detailed comparison of PM products, nor a detailed comparison of PM algorithms. It also does not do a detailed assessment of process model quality using formal metrics.

1.1 Background

One of the first uses of computers in medicine in 1959 can be attributed to a dentist, Robert S. Ledley (November, 2011). From his career as an army dentist through his work at the dental materials section of the National Bureau of Standards he advocated for the application of operations research techniques and computing to medicine. He argued that the vast amounts of medical diagnostic and treatment data could only be exploited using operations research information management techniques and he endeavoured to have these ideas accepted and implemented in the U.S. medical community (Ledley & Lusted, 1959). Early work involved a notched-card system to assist in the diagnosis of disease and this was adapted by Homer R Warner as a Bayesian scheme, again using the notched card system, to assist in the diagnosis of congenital heart disease in the LDS hospital in Salt Lake City (November, 2011). Despite the successful trials of these techniques, they were met with scepticism in the larger medical community. Nonetheless, they laid the groundwork for the extensive uses of information technology in medicine and hospital operation. Their ideas around using existing data to analyse symptoms and develop diagnoses are a precursor to modern day evidence-based medicine. Their beliefs around

the use of information generated as a by-product of operational activities underlies the principles of data mining (November, 2011) and the emergent technology of interest here, process mining.

In this thesis, 60 years later, similar ideas to theirs, but using modern data and process mining technologies, are being applied to a large dataset extracted from an Irish dental public health EHR.

1.2 Research Technology Terms

1.2.1 Data Science

Data science is a multidisciplinary field using scientific processes, algorithms, methods, and systems to extract knowledge, patterns and actionable insights from data. It incorporates skills from many fields including statistics, information science, computer science and mathematics. In 2007, Jim Gray termed it the ‘fourth paradigm’ of science after empirical, theoretical, and computational (Hey, et al., 2009) and anticipated computational analysis of large data being a primary scientific method. He identified three basic activities of data science: capture, curation, and analysis, and suggested that everything about science would change due to the impact of information technology.

In 2013, the IEEE Task Force on Data Science and Advanced Analytics was founded and in 2013, the "European Conference on Data Analysis (ECDA)" was first organised. The first international conference, the IEEE International Conference on Data Science and Advanced Analytics was launched in 2014.

A definition by Dhar (2013) describes data science as the study of the generalizable extraction of knowledge from data with the requirement that this knowledge is actionable for decision making and prediction, not just explaining the past. The scale of the available data often renders traditional database models and computational methods inadequate. There is a need to provide actionable, robust patterns with predictive power, and patterns that are likely to occur in the future.

In their book “Doing Data Science”, O’Neil & Schutt (2014) propose a data science process model as in Figure 1-1 below where data represents the traces of real-world processes gathered by data collection or sampling methods. They make the important point that building models and working with data is not value neutral. Researchers choose which problems to address, they make assumptions, chose metrics and design the algorithms. In their view, the data-scientist turns the world into data, and this is a subjective, not objective, process. They emphasise the value of exploratory data analysis

for building intuition for the research data, where the creation of histograms, scatterplots, written descriptions, graphs, and summary statistics constitute a vital step before using the data to prove or disprove anything to others.

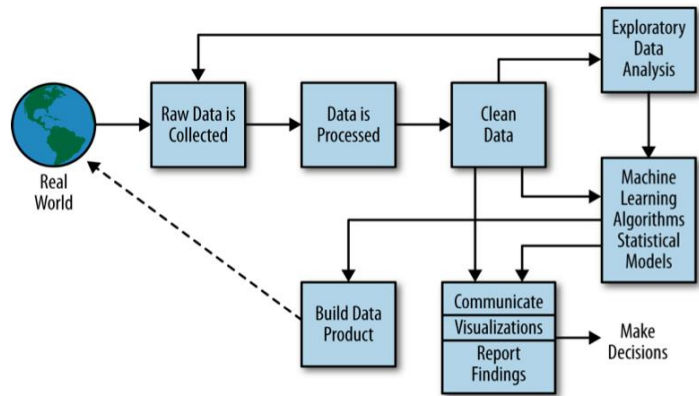


Figure 1-1: The Data Science Process (O'Neil & Schutt, 2014, p. 41)

O'Neil & Schutt (2014) state that describing and understanding these data-generating processes is often part of the solution to the problems being addressed and point out that in the case of data products, a feedback loop is being created where our behaviour changes the product and the product changes our behaviour and as such brings with it ethical responsibilities. They identify the data-scientist as being involved in all the stages of the data science process as in Figure 1-2 below. i.e. “...a data-savvy, quantitatively minded, coding-literate problem-solver... trained in anything from social science to biology, who works with large amounts of data, and must grapple with computational problems posed by the structure, size, messiness and the complexity and nature of the data, while simultaneously solving a real-world problem.”

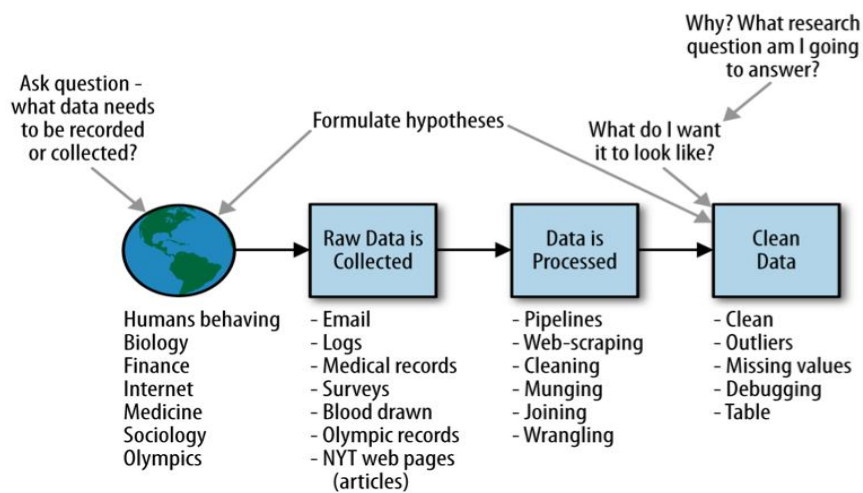


Figure 1-2: The Data Scientist's Role (O'Neil & Schutt, 2014, p. 44)

Dhar (2013) outlines a range of skills required of a data-scientist: statistics, machine-learning, computer science, and coding. This range of skills is required due primarily to the volume and variety of the data being analysed today. Wil van der Aalst (2016) proposed an outline of the skills employed by data scientists shown in Figure 1-3 as well as its link to process science through PM.

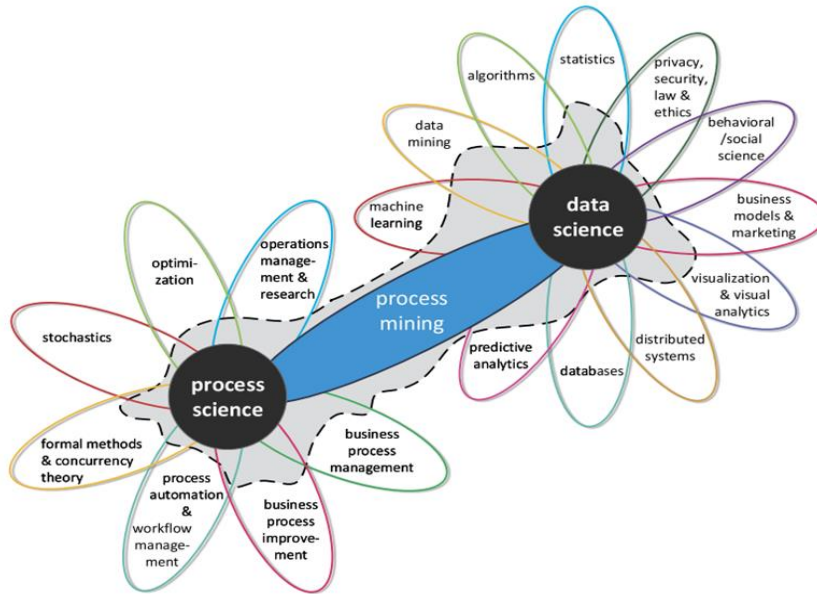


Figure 1-3: Data and Process Science Skills (van der Aalst, 2016, p. 18)

Wil van der Aalst (2016) defined data science as ‘...an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.’

This thesis incorporates the steps outlined in the data science process in Figure 1-1: The real-world data originates from a dental EHR. A comprehensive pre-processing phase to prepare the data for exploratory analysis and profiling was carried out, and machine learning in the form of PM was executed. Many of the skills identified in Figure 1-3 are employed in the course of the research: knowledge of databases and algorithms to extract, transform and load data in preparation for analysis, data mining, PM, and visualisation for exploration and analysis, some statistics for evaluation of results, domain knowledge for formulation of the research questions (RQs), and for discussion of the outcomes and results. The resulting applied techniques and methods demonstrate flexibility capable of providing ongoing, robust and actionable insights to domain stake holders, satisfying a key data science attribute of having a clear focus on its organisations’ goals.

1.2.2 Data Mining

Data mining is the process of seeking and extracting patterns from previously incomprehensible large datasets and this author views data mining as a key subset of the

data science process, primarily involving the steps: data pre-processing, data cleaning, exploratory data analysis and machine learning. It is also commonly referred to as a step in the Knowledge Discovery from Databases (KDD) process using data analysis and discovery algorithms to yield patterns (or models) based on the data (Santos, et al., 2013). Data mining can be broadly categorised into ‘descriptive’ and ‘predictive’, the former includes association, classification and clustering activities and the latter, correlation and regression.

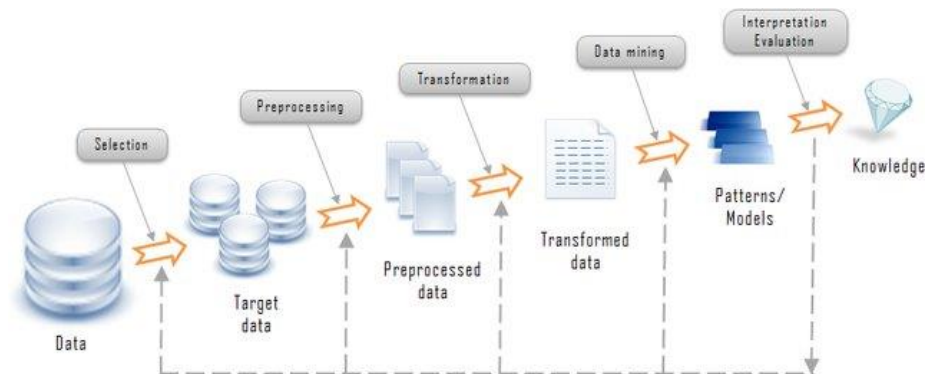


Figure 1-4: Data Mining in the Knowledge Discovery from Databases chain (Dragon1.com, 2018)

This author views PM as a data mining technique and generally follows the steps identified in Figure 1-4.

1.2.3 Machine Learning

Machine learning is a set of artificial intelligence techniques and algorithms designed to extract patterns from large datasets, without being explicitly programmed. Typically, machine learning algorithms find similarities between group of items (classification and clustering) or find relationships between variables (correlations, associations). Some of the common machine learning types are shown in Figure 1-5 below.

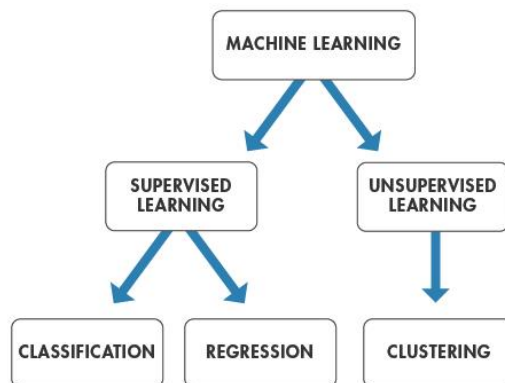


Figure 1-5: Machine Learning Types (Mathworks, 2018)

This author views PM as a form of unsupervised machine learning, creating clusters of similar items on data that has not been previously labelled or categorised.

1.2.4 Data Visualisation

Data visualisation facilitates the communication of complex information using graphical representations. Its aims to display the data in a compact, accurate, unbiased form. It converts large datasets into visually comprehensible formats such as histograms, plots, and information graphics. Stephen Few (2004) enumerated eight types of quantitative relationships, including frequency distributions, nominal comparison, correlations, and ranking and identified the optimal format for their graphical representation. Since then, many new visualisations have emerged including bubble graphs, heat maps etc. to visualise more complex data, many of which are used in this research (see Section 4.1.6).

1.2.5 Process-Oriented Data Science

Process-oriented data science (PODS) is an emerging research area bridging traditional process analysis and data-centric analysis. PODS studies the sequences of events in processes and is not solely focussed on outcome measures or the results of data mining experiments. Timestamped, case-oriented, event data is the main source of information for PODS. The principal data mining technique in use in PODS is PM.

1.2.6 Process Mining

Process Mining (PM) is the collection of techniques and algorithms applied to event data with the objective of discovering, checking and enhancing process. It is an emerging data mining technique aiming to extract high level knowledge from low level data. PM has been positioned in the field of business process management, business intelligence, and lean technologies by van der Aalst (2016, p. 44) and by Schrijvers et al (2012). It is seen as bridging the gap between traditional model-based process analysis and data-centric analysis such as data mining (Mans, et al., 2015, p. 5). It does this, first, by discovering process models from event data, i.e. time ordered data extracted from an organisation's information systems. These models are abstract representations of the essence of a process reflecting the common pathways followed and are used for many purposes within organisations including configuration, specification, documentation and verification of systems. They are also used to give insight and to provide a structured basis for discussion of the processes (van der Aalst, 2016, p. 29). The second main type of PM, conformance checking, establishes to what degree event logs agree with existing process models. The third type enhances models with additional information and is known as process enhancement. This environment is commonly represented as in Figure 1-6 below.

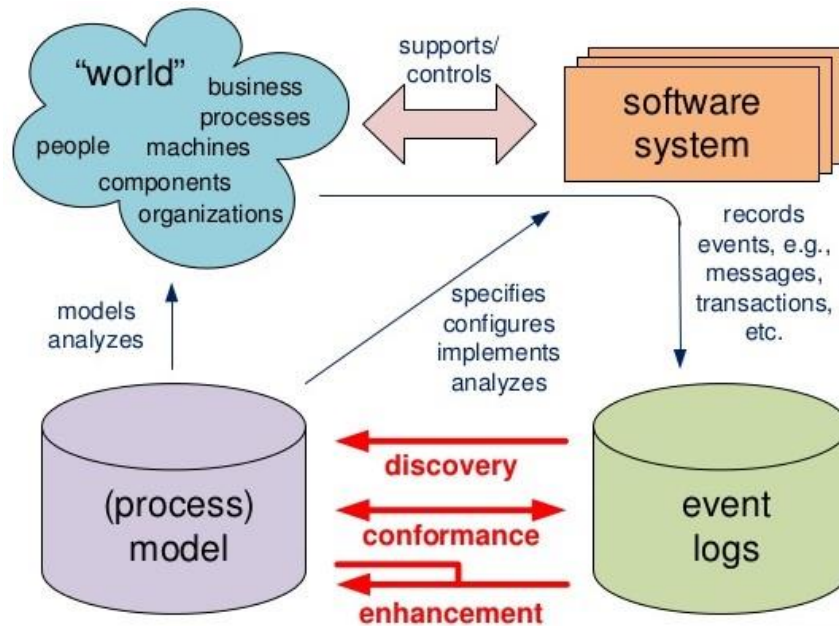


Figure 1-6: Process Mining Types and Environment (Mans, et al., 2015, p. 22)

Event data contains time-ordered lists of discrete activities or events i.e. well-defined steps in processes and an event log (EL) is a subset of this event data created for the purposes of executing a single experiment. The most basic EL contains a case-identifier, an event name and a timestamp. The quality of a process model will primarily be determined first, by whether the discovered process model generates all the behaviour in the log and second, by how close is behaviour of the discovered process to the behaviour of the original process. The comprehensibility of the discovered model and to what extent it is generally applicable are other important characteristics.

1.3 Dental Domain Terms

1.3.1 Dentistry

According to the American Dental Association (2018), dentistry is defined as the evaluation, diagnosis, prevention and/or treatment (nonsurgical, surgical or related procedures) of diseases, disorders and/or conditions of the oral cavity, maxillofacial area and/or the adjacent and associated structures and their impact on the human body; provided by a dentist, within the scope of his/her education, training and experience, in accordance with the ethics of the profession and applicable law.

In layman's terms it is the profession of caring for the human mouth, teeth and other related health matters. This manifests as creation of oral health, prevention and treatment of dental disease and restoration of damage to the teeth and the mouth. Disease of the teeth usually starts with carious lesions in the tooth surface, also known as caries which can then progress to cavities or holes in the tooth surface Disease at this level is commonly

treated through restorative measures such as fillings and in more extreme cases, crowns and bridges, implants, and extractions.

In what ways is dentistry different and similar to general healthcare?

Although relief from toothache has been mentioned in medical texts as far back as Hippocrates, dentistry and medicine have traditionally been separate occupations. During the 17th century dentistry was often carried out in the barber's chair – by the barber (Hoffmann-Axthelm, 1981, p. 161) (Ring, 1985, p. 150), while medicine was already well established as a profession at that time. Dentistry was initially seen as a primarily mechanical job, fixing and extracting diseased teeth (Hoffmann-Axthelm, 1981, p. 159) and those carrying out the activities in Germany were often known as the *Zahnbrecher* ('tooth-breakers'). In 1840, at the University of Maryland in Baltimore, the introduction of dentistry as a medical speciality was rejected and as a result, the first dental college in the world was opened – the Baltimore College of Dental Surgery.

Clearly, things have changed since these early days and there is now wide acceptance that oral health influences general health (World Dental Federation, 2016) and a concerted effort to 'put the mouth back in the body' is leading to closer ties between the professions. However, dental schools are still typically separate from medical schools, as are dental and medical hospitals. Dental insurance and medical insurance are normally separate products, as oral problems are often seen as inevitable, even if often preventable.

For the purposes of this research it is assumed that dentistry is a branch of medicine and is a form of healthcare. As with general healthcare, dentistry is delivered at three different levels. Primary care deals with common problems such as examinations, cleanings, and restorative work and is often the first point of contact for a patient. Secondary care is typically more specialised such as periodontal procedures, endodontics etc. and normally requires referral from primary care in Ireland and in the U.K. Tertiary care involves rare and complex conditions and can arise for example from trauma incidents or special-needs patients.

The classification of healthcare processes shown in Figure 1-7 is directly transferrable to dental care – the main categories of 'Non-elective care' and 'Elective care' and the subcategories of 'Emergency' (e.g. trauma), 'Urgent' (e.g. pain or abscess), 'Standard' (e.g. screening), 'Routine' (e.g. simple filling) and 'Non-Routine' (e.g. root canal treatment), are all directly applicable to dental treatment.

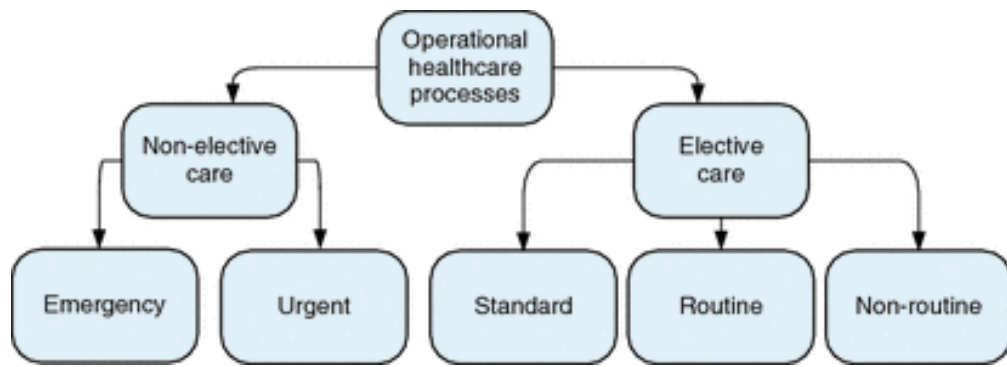


Figure 1-7: Main kinds of organisational healthcare processes (Mans, et al., 2015, p. 13)

This research assumes that dental treatment processes and healthcare processes are similar enough allowing us to use existing healthcare PM research and publications as the basis for this research.

Healthcare processes are also subject to a separate categorisation: medical treatment processes and organisation processes (Kaymak, et al., 2012; Rojas, et al., 2015), the former being clinical processes managing the patients and the later focussed on knowledge necessary to coordinate collaborating healthcare professionals and units without support for medical decision making.

1.3.2 Public Health Dentistry

While acknowledging that public health dentistry is organised differently in different countries, the American Dental Association (2018) defines it as “...*the science and art of preventing and controlling dental diseases and promoting dental health through organized community efforts. It is that form of dental practice which serves the community as a patient rather than the individual. It is concerned with the dental health education of the public, with applied dental research, and with the administration of group dental care programs as well as the prevention and control of dental diseases on a community basis*”. Dental public health has also been defined “*as the science and practice of preventing oral diseases, promoting oral health and improving quality of life through the organised efforts of society*” (Daly, et al., 2013). It is concerned with promoting oral health of the population as a whole by, diagnoses of the population’s oral health problems, to identify the cause of these problems and planning effective interventions to target identified problems leading to action at the community level. The World Health Organisation (WHO) recognises the importance of public health intervention against early childhood caries (ECC) (Phantumvanit, et al., 2018).

Public health dentistry contrasts with private dental practice where private practices are often owned by a dentist, a group of dentists, or a corporate body for the purposes of

delivering dental services to individuals. In Ireland, a child's first contact with dental services will often be through the public health school screening program. In the U.K., private practices are the first point of contact for patients who require dental treatment or oral health maintenance.

1.3.3 Oral health

The World Dental Federation (2016) defines oral health as being “... *multi-faceted and includes the ability to speak, smile, smell, taste, touch, chew, swallow and convey a range of emotions through facial expressions with confidence and without pain, discomfort and disease of the craniofacial complex*”. The World Dental Federation (FDI) definition proposes a common understanding of oral health in order to: clearly position oral health within general health, demonstrate that oral health affects general health, raise awareness of the different dimensions of oral health and how they shift and change over time and empower people by acknowledging how values, perceptions and expectations impact oral health outcomes.

1.3.4 How are oral health outcomes measured?

There are many established measures of oral health and the suitability and availability of some of these for our research were considered e.g. DMFT, Quality of Life (QoL), International Caries Detection and Assessment System (ICDAS <https://www.iccms-web.com/>). Potential quality outcomes were also proposed in the Steele Report (NHS England, 2009, p. 66) e.g. the increase or decrease in the rate of restoration and the rate of antibiotic prescription. Significant Caries Index (Sic Index) (Brathall, 2000) is another option to measure disease where the attention is focussed on those individuals with the highest caries scores in the population. It is a recognition of the high number of individuals with no detected disease and the resulting skewedness of DMFT to 0. For clarity, a DMFT score of 0 means that none of a person's 32 permanent teeth are decayed, missing, or filled i.e. it is the 'perfect' score in terms of caries. However, it does not take account of tooth loss due to other reasons such as trauma or periodontal disease.

For this research, the only criterion in deciding which oral health outcome to use was a pragmatic assessment of what information is present in the EHR to help assess oral health. The results of that assessment showed that the EHR had no information on QoL and the recorded caries information was insufficiently detailed for an ICDAS assessment. While the necessary data was present in the EHR to calculate DMFT, again, insufficient data was present to calculate its more detailed variants. Accordingly, DMFT was selected as

the outcome measure for this research. Other more detailed indices such as ICDAS should be considered in future EHR designs as they give further insight into the degree of disease present.

DMFT, D_{3c}MFT, D_{3vc}MFT

‘DMFT’ is a measure of tooth decay in permanent teeth and has three components. D refers to the number of decayed teeth where caries is to dentinal threshold as is more accurately represented as D₃. DMFT is used throughout this thesis as shorthand for D₃MFT. M refers to the number of teeth missing due to decay. F refers to the number of teeth filled due to decay. The use of capital letters indicates that the index applies to permanent teeth only. ‘dmft’ is the same measure but applied to primary teeth only. A DMFT score of 0 means none of the 32 permanent teeth are diseased, missing or filled due to decay. If a patient had 1 tooth extracted for decay, 1 filling due to decay, and 1 cavity, their DMFT score would be 3.

While DMFT has been in use for over 60 years (Broadbent & Thomson, 2005) as an index of oral health, it has well documented shortcomings such as its failure to recognise the presence of non-cavitated lesions and the fact that caries is a continuum rather than a present-absent dichotomy (Lewsey & Thomson, 2004). There are also many factors influencing the development of caries in individuals as shown in Figure 1-8 below (Selwitz, et al., 2007; O’Mullane, et al., 2016; Petersen, 2008). Many of these are personal factors such as smoking, oral hygiene and socio-economic status while others are related to the presence of fluoridation of water supplies etc. The World Health Organisation (WHO) examination criteria dictate that only dental caries at the cavitation level are recorded. This is known as D_{3c}MFT with the ‘3’ indicating that the caries is recorded at the dentinal level and the ‘C’ indicates that the lesion is cavitated (Whelton, et al., 2006). Acknowledging that dental caries is a disease of stages and acknowledging that increased access to dental services could give misleading D_{3c}MFT readings, a further refined measure, D_{3vc}MFT, incorporates visible but not cavitated lesions. However, DMFT remains widely used and accepted because of its usefulness and the need to make historical comparisons (Lewsey & Thomson, 2004). From the perspective of EHR data requirements, less information is needed to estimate DMFT than more sophisticated indices such as ICDAS. The mechanism used to calculate DMFT/dmft from the data extract is detailed in Appendix 10.9.

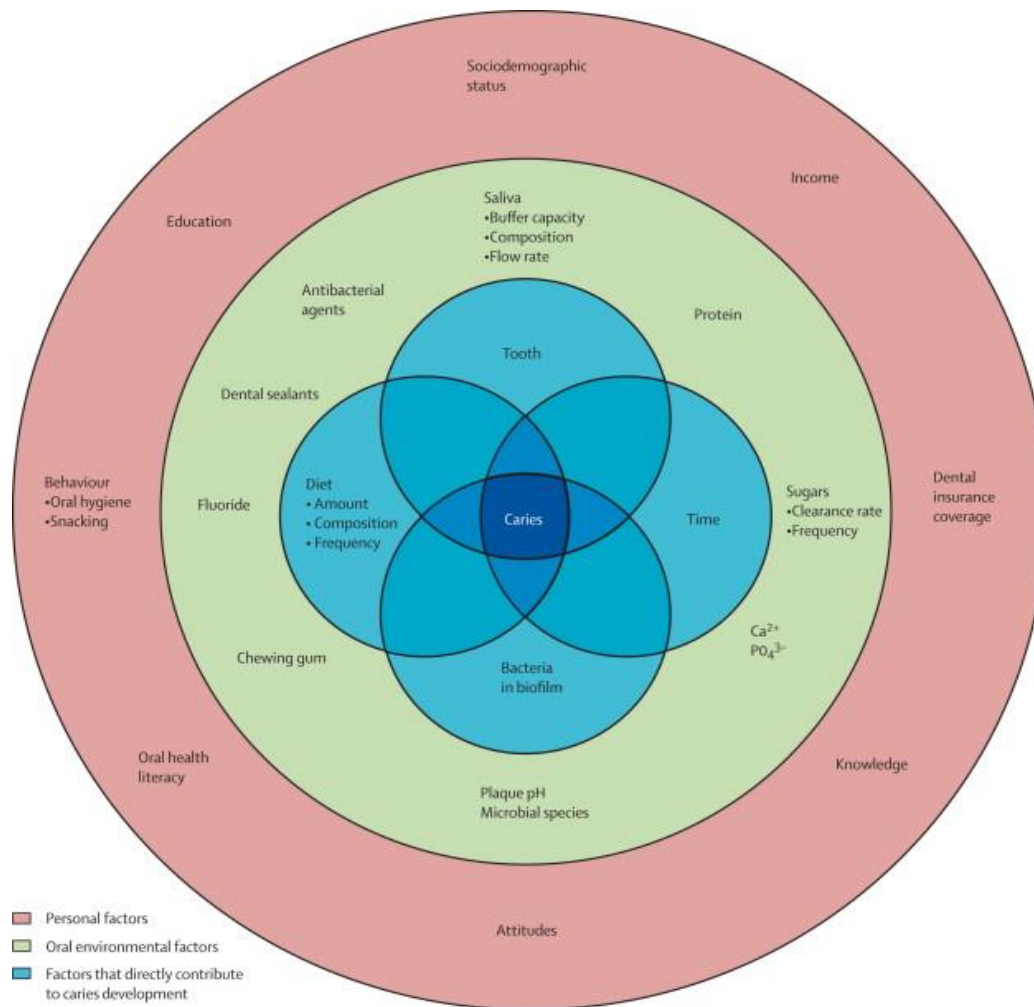


Figure 1-8: Factors involved in caries development (Selwitz, et al., 2007)

1.3.5 Care Pathways and Clinical Guidelines

Schrijvers et al. (2012) defined care pathways as ‘...a methodology for the mutual decision making and organisation of care for a well-defined group of patients during a well-defined period’. They detail essential steps in the care of patients with specific problems. According to Schrijvers et al., care pathways have their roots in established management theories such as the Critical Path Method, Lean Engineering, and Six Sigma with the goal of improving quality while reducing duration times and error-risk, reducing treatment variations, and reducing costs. While care pathways were originally introduced in the U.S. to standardise processes and reduce costs, other countries are using them mainly to achieve improvements in quality of care, and in the U.K. their use aims to achieve continuity of care across care settings and disciplines (Harris & Bridgman, 2010). Care pathways are often confused with care protocols and Harris & Bridgman state the term ‘care pathway’ denotes a distinctive type of clinical guideline, specifying each step in the care process, rather than stating broad principles that practitioners should follow. Care pathways are also seen as offering a structured means of implementing evidence-

based clinical guidelines though the development of local protocols (Campbell, et al., 1998). Examples of care pathways in dentistry in the U.K. are to be found in the Steele Report (NHS England, 2009) and in the NICE guidelines (National Institute for Health and Care Excellence, 2018).

Evidence-based clinical guidelines (CGs) are systematically developed statements containing recommendations for the care of individuals by healthcare professionals that are based on the highest quality scientific evidence available (Irish Oral Health Services Guideline Initiative, 2012). CGs are specific recommendations on how to diagnose and treat a medical condition to ensure that patients receive appropriate treatment and care. They summarise the current medical knowledge and give specific recommendations based on this knowledge. For example, clinical guidelines published in Ireland describe the ideal process for applying fissure sealants and are to be found in the Irish Oral Health Services Guideline Initiative (2010) and for providing an oral health assessment for school-aged children in a second publication (2012).

For the purposes of this research, care pathways are viewed as guiding the overall treatment process of the patients whereas clinical guidelines are viewed as focussed on specific dental treatments. In this research, we will investigate process mining's potential to generate process models that can be compared to the referenced care pathways and clinical guidelines.

1.3.6 Oral Health Strategy and Policies

Oral health policies take many forms. These include population-wide oral health promotion measures, policies addressing social determinants, route epidemiological data collection and the integration of oral diseases in policies addressing non-communicable diseases and general health (World Dental Federation (FDI), 2017). However, measuring the effectiveness of policy decisions and strategies is difficult. Daly et al. (2013) state that the evaluation of population-based prevention is particularly difficult to undertake, especially measuring success by examining changing patterns of disease. They propose that other types of evaluation such as the success of the process and investigating how many people participated in the screening program may be useful. It is hoped that the research in this thesis contributes to establishing methods of evaluating such processes and programs.

1.3.7 School Dental Screenings

Dental screening generally refers to the brief oral examination of children, usually in the school setting, in order to identify those with obvious treatment needs (Irish Oral Health Services Guideline Initiative, 2012). In Ireland school dental screenings are provided by the public health service as detailed in Chapter 4. In the UK's National Health Service (NHS), dental screening's role developed from securing treatment in times of widespread decay, to identifying children not in receipt of regular dental care, early detection of disease, and stimulation of registration with a dentist (Milsom, et al., 2008).

Why do Screenings? Are they Useful?

Internationally, the WHO argues that school-based oral health promotion is effective and efficient (World Health Organisation, 2003, p. 17). Screenings enable early detection and timely interventions against oral diseases and conditions leading to substantial cost savings. They must not necessarily be carried out by dentist or dental auxiliaries (World Health Organisation, 2003, p. 45). This latter point does not reach the recommended best practice of carrying out examinations in dental clinics as proposed in the Irish Oral Health Services Guideline Initiative (2012, p. 18). Hebbal & Nagarajappa (2004) found that screening increases follow-up visits. There is significant debate on this issue and another study by Milsom et al. (2006) found this not to be the case.

The function of examinations can include primary preventive measures e.g. oral health advice, application of fluoride gels or varnishes as well as secondary preventive measures e.g. limiting the progression of oral diseases (Riley, et al., 2013). A preventive program based on caries risk and recall intervals was shown to reduce initial caries lesions in children (Abanto, et al., 2014).

Clinic based screening is seen as the gold standard with school-based screening more likely to only identify children with more advanced caries (Irish Oral Health Services Guideline Initiative, 2012).

The usefulness of school dental screenings in improving dental attendance rates or reducing disease levels has been questioned by Milsom et al. (2006) where the authors tested three models of screening and a control, on a population of 13,000 children. They found no significant difference in caries reduction in either the deciduous or permanent teeth, nor did they find any significant difference in the secondary outcome measures, prevalence of sepsis, gross plaque, calculus or trauma. While they did acknowledge the short timeframe of the study, they also found no significant difference in dental attendance between the groups in the four-month period following the screening date.

Further work suggested that, notwithstanding considerable U.K. government support since 1918, no scientific evidence exists that school dental screening leads to improvements in health for the individual children or for the child population as a whole (Milsom, et al., 2008) and in fact, such screening exacerbated social division. While scrapping of school dental screening is not being advocated-for in those publications, the authors encourage development of clear objectives for such screenings and for scientific evaluation of the data available from those countries with such programs.

In contrast, a retrospective cohort study investigating the effects of dental recall visit intervals on the oral health of Irish school children by Brody (2016) concluded that children having one oral health assessment in 2nd class (Age ~ 7-9) had significantly higher levels of tooth decay at 6th class assessment (Age ~ 11-13) than children who received an additional oral health assessment in 4th class (Age ~ 9-11). The children receiving only two assessments were also found more likely to have attended for an emergency visit for pain in a permanent tooth in the period between assessments.

There are varying opinions on the effectiveness of school dental screenings and this research addresses aspects of this in the following sections and again in Section 7.4.

When should screenings be done?

As permanent molars account for at least 80% of the caries in children's permanent teeth in Ireland, the age at emergence of the first and second permanent molars is a key milestone for oral health assessment. The second key milestone is the emergence of the permanent maxillary canines (Irish Oral Health Services Guideline Initiative, 2012, p. 18). The guideline goes on to suggest that the periods between the ages 5-7 and 11-14 are the most crucial for regular assessment to prevent and treat caries and monitor oral health development.

How often should dental recalls be done?

The ideal interval for recalling dental patients is also an active discussion. The Cochrane Database Systematic Review on Recall intervals for oral health in primary care patients. Riley et al. (2013) looked at the evidence around varying recall intervals' effects on oral health and resources. This review updated earlier work by Beirne et al. (2005; 2007) and confirmed the original work's position, namely that there is a very low-quality body of evidence which is insufficient to draw any conclusions on the effects of altering the recall interval between dental check-ups. Further, they recommended that high quality Randomised Control Trials (RCTs) be carried out to address this question. Abanto et al.

(2014) found that each follow-up visit attended reduced new initial lesions in children although they acknowledge that these patients were also receiving oral health and dietary advice during their visits which may have impacted their findings.

Although the U.K. NHS does not explicitly recommend a recall interval, its remuneration structure supports six-monthly checks. In their systematic review, Davenport et al. (2003) found no existing high-quality evidence to support or refute six-monthly recall intervals in adults or children. They identify risks of lengthening recall intervals as moving away from the preventive paradigm and consequently more serious sequelae of caries, e.g. infection or extraction, as well as reduced contact with patients and accordingly, a loss of opportunity to encourage better oral hygiene and treatments. They identified possible advantages of lengthening the recall interval as, reducing inappropriate treatment and reduction in costs. The review noted the heterogeneous nature of the previous work and the difficulty comparing the studies.

A risk-based maximum recall interval of 12 months for patients is recommended in the NICE Clinical Guideline (National Institute for Health and Care Excellence, 2004). This may be reviewed subject to the outcome of the ongoing INTERVAL (Investigation of NICE Technologies for Enabling Risk-Variable-Adjusted-Length Dental Recalls Trial) Dental Recalls Trial expected in 2019 (BioMed Central Ltd, 2018). This risk-based approach, with a suggested maximum interval of 12 months for children in Ireland is supported by clinical guidelines (Irish Oral Health Services Guideline Initiative, 2012). Discovering care pathways around these principles and investigating the potential application of PM and data analysis of dental EHR data to answer related questions forms the basis of much of this research.

1.3.8 Initial Exams, School Screenings, Recalls, Recall Intervals & Check-ups.

This research treats ‘school screening’ as analogous to a ‘recall visit’ which was defined as ‘*the planned, unprecipitated return of a patient who, when last seen was in good oral health*’ (Royal College 1997, as cited in Riley et al. (2013), when a ‘recall examination’, ‘routine dental check-up’, or ‘oral health review’ may be carried out (Riley, et al., 2013). The recall interval is the time between recall examinations and is usually specified in months or years. The policy governing the frequency at which school screenings are carried out is then directly related to the recall interval as presumably, a policy dictating 3 screenings in the primary school setting will have a shorter recall interval between screenings than a policy dictating 2 screenings. In the research’s data, detailed in Chapter 5, the treatment item known as ‘Initial Exam’ in the research dataset is a ‘school

screening’. This assumption facilitates thinking about the effects that varying recall intervals can have on treatment processes and oral health outcomes in a similar way to considering the effects of varying screening frequency. The author does not believe that this is a high-risk assumption as it is not central to the aims but rather opens the door to the technologies being used for either scenario.

It is clearly arguable that screenings and recalls are not the same thing and this research is not proposing that ‘school screenings’ and ‘recalls’ are identical, rather, treating school screening frequency and recalls as equivalent benefits this research as it allows it to demonstrate how the techniques and technologies can link and contribute to the wider debates on care pathways in dentistry and the ongoing debates on recall intervals as ‘recalls’ and ‘recall intervals’ are the standard terms in use.

1.3.9 Care Pathways for School Dental Screenings and Recall Intervals

Decisions regarding optimal recall intervals is one of the key questions raised by dental care pathways research. There is a clear move towards risk-based treatment and recall intervals (NHS England, 2009, p. 46; NHS, 2012, p. 16; National Institute for Health and Care Excellence, 2018). School screening recall intervals and similar strategy and policy questions such as these will be examined in the context of data mining and PM.

1.4 Linking Process Mining to Care Pathways and Clinical Guidelines

Care Pathways and Clinical Guidelines in Dentistry

Care pathways, clinical guidelines and process-oriented approaches to the delivery and assessment of dental care need to be supported by technologies which facilitate process-oriented data science and analysis. PM is one such approach and the work in this thesis applies it and supporting technologies to a data extract from a dental EHR to assess its applicability and usefulness.

Efficiencies in healthcare can be gained by analysing care pathways and processes and by applying operations research techniques, workflow analysis, and other process re-engineering techniques to optimise the delivery of services. The research in this thesis shows how process discovery from EHR data can produce process models helping assess the delivery of dental service according to these ideal care pathways. Daly et al. (2013) state that evaluation of population-based preventive measure is difficult and assessing the success of the process can be a valid alternative. Mans et al. (2015, p. 3) conclude that the traditional methods of gathering the information required for such analyses by observation and interview are costly and flawed due to their subjective nature and further

claim that objective suggestions for improving processes can be readily gained from event log data. They maintain that because healthcare processes require flexibility and ad-hoc decisions, therefore, rigorous workflow management, business process management and business process reengineering techniques cannot be applied. There are also other problems unique to healthcare such as data quality issues, process complexity and organisational issues within healthcare bodies.

The use of process modelling and care pathways is well established in healthcare. The U.K.'s NHS Modernisation Agency applied process modelling to the health sector and it is now in widespread use there (Harris & Bridgman, 2010).

Care Pathway Initiatives in Dentistry

The use of care pathways in dentistry is well established and important recommendations are to be found in U.K. NHS strategy publications such as (NHS England, 2009, p. 45) also known as the Steele Report. This report recommends that NHS primary care dentistry be staged around a care pathway, with features including urgent and continuing care, formal oral health assessment, disease prevention and advanced restorative care for the purposes of continuity of the relationship between patients and dentist with recall intervals as a key element and using oral health as the outcome measure.

Care pathways are also a key part of U.K. dental contract reform and have received widespread support amongst pilot practices and patients (NHS, 2012, p. 5). The care pathways proposed by the NHS (2012) are based on the Steele Report. Here, the four main causes of poor oral health; dental caries, periodontal disease, tooth surface loss, and soft tissue conditions result in a risk status being applied to the patients and recall intervals and interventions being decided thereon. In the U.K, the NICE Guidelines on Oral Health (National Institute for Health and Care Excellence, 2018) operate on similar principles, providing a process model or flow-chart structure to guide dental professional in their care delivery. These publications are significant showing the commitment of the U.K.'s public health service to care pathways.

The annual oral health assessment proposed in the Irish Oral Health Services Guideline Initiative (2012, pp. 6,7) as the best practice approach promoting, protecting and maintaining the oral health of Ireland's 5-7 year-olds is summarised as: examination and risk assessment, oral health instruction and, if high caries risk, administration of protective measures such as fluoride varnish and fissure sealant. The Irish Oral Health Services Guideline Initiative (2010, p. 6) presents a clinical guideline portraying the Fissure Sealant Cycle in the form of a process flowchart.

1.5 Structure of the Thesis

Chapter 2 provides the literature review of PM in dentistry and defines a vocabulary for this research's PM. This research is cross-disciplinary and for this reason, literature is explored in several chapters, as close to its point of use as possible. Nine previous literature reviews were identified and used to frame dental PM within the larger healthcare PM area. Also, in Chapter 2 is an explanation of the basic PM terminology in use in this research and relevant literature. Many of the basic technology definitions and dental terms and their associated literature are introduced in Chapter 1 and developed further in the specific validating experiments in Chapter 7. Chapter 3 details the aims and objectives of the research. Chapter 4 introduces and profiles the research data and concludes with a description of the data pipeline and system architecture. Chapter 5 applies an anonymisation framework and introduces a data quality framework and its application to the research data and concludes with details of the data transforms necessary for the research. Chapter 6 examines the existing PM project methods, analyses their strengths and weaknesses and documents a synthesised method for applying PM to this dental EHR data. Chapter 7 details the PM experiments using the methodology from Chapter 6. Chapter 8 discusses the implementation of the methodology and its validation experiments and Chapter 9 forms the conclusions of the research.

2 Literature Review

The rationale for this literature review was to find the existing published work applying process-oriented data science to dentistry and to establish to what extent this technology had been applied to EHRs and large public health datasets. The literature search criteria included PM in primary care, and PM in public health. To enhance general applicability and interdisciplinary impact of this work, the general literature on PM will be also be referred to in this research.

Existing work in the area of dental informatics, which is the application of health information technology and information science to healthcare delivery (American Dental Association, 2018), was also included in the search. Informatics is a research discipline aimed at uncovering fundamental principles and methods relating to information and computers and, while primarily focussed on the dental domain, the search was cognisant of Schleyer's (2003) hints on the dangers of discipline-based informatics areas such as nursing informatics and dental informatics. He suggests that an excessive number of boundaries between specialised application areas may have the effect of 'balkanizing' informatics and he encouraged broad and inter-disciplinary collaboration between the specialist communities as the best way to develop discipline-specific solutions. Schleyer reinforced the opinion that informatics benefits from interdisciplinary collaboration as the RQs tend to be complex and use scientific methods from several areas, primarily information science, computer science, cognitive science, and telecommunications.

The review establishes what dental questions have been investigated using data mining and PM? what methods were used in applying PM to dentistry? what did they find out? what PM has been carried out in public health and primary care? and where are the gaps and research opportunities in dental PM?

2.1 Previous Literature Reviews & Related work

The relatively new data mining technique of PM, although still niche, has much to offer for broad information systems audiences, offering potential for increasing efficiency and effectiveness of services (Thiede & Fuerstenau, 2016). PM has already been effectively applied to many areas of industry, business (van der Aalst, 2011) and healthcare (Rojas, et al., 2016) including specialist healthcare areas such as stroke-care (Mans, et al., 2008), diabetes (Fernandez-LLatas, et al., 2015), and oncology (Kurniati, et al., 2016).

There have been several previous reviews of the use of PM in healthcare. In the first review Kaymak et al. (2012) identified the inability of the available PM algorithms to analyse healthcare process and they pointed to the need for PM algorithms incorporating

medical knowledge and the need for pre-processing the clinical data using medical knowledge e.g. reducing data granularity to improve the resulting models. Through analysis of ten available PM healthcare publications, they also identified that medical practitioners may be pursuing multiple goals in a process and the PM algorithms need to be aware of this to produce useful results. Their findings were largely supported by Yang & Su (2014) suggesting that PM algorithms are not efficient enough to deal with unstructured processes. They had analysed 37 studies of PM in healthcare with the goal of clinical pathway design, control, and evaluation and improvement. They noted that medical processes are more complex than business processes, being dynamic and unstructured. They point out that the existing algorithms only consider the event name and starting time – not the outcome.

Mans et al. (2008) found that the heuristic miner produced incomprehensibly complex models when applied to hospital stroke healthcare data due to disease and patient variants. The term commonly used to describe such models is ‘spaghetti’ models. They used pre-processing techniques on the event data for example seeking higher level events to represent lower level activities. They also proposed use of simplification techniques such as clustering and the specialised search algorithms as approaches to simplify the models. These algorithms will be detailed in Section 2.2. Mans et al. (2013) examined 37 process discovery publications and 7 conformance-based papers in the context of their proposed healthcare reference model and concluded that, as a rule, the existing body of work underutilised the available data and would benefit from using such a reference model to enhance the value of their work.

Rojas et al. (2015) completed an overview of the main approaches using PM in healthcare and introduced the main challenges encountered in previous work. These challenges included data access, data quality, integration and pre-processing as well as the incorporation of medical knowledge in the algorithms. The comprehensive literature review carried out by Rojas et al. (2016) built on their earlier work and categorised the published work by process-type, data types (sources), frequently posed questions, PM perspectives, tools used, methodologies, implementation strategies, analysis strategies and geographical and medical fields. This review revealed 74 PM healthcare papers. These papers were sourced from web searches and the healthcare PM repository (www.processmining.org). The review included journal articles, conference presentations, postgraduate and doctoral theses and a specialist book. In their systematic mappings of PM studies in healthcare, Erdogan & Tarhan (2016), (2018) found the field

of PM in healthcare to be rapidly growing despite the healthcare data and technique related challenges. They identified 172 studies in the area of PM in healthcare.

In a recent review of PM in primary care, Williams et al. (2018) confirmed that little research existed in the area of PM in primary care and suggested that this is indicative of challenges to be overcome in this area and that future work should look to identify and resolve these problems, though they offer no insight as to what these problems might be.

2.2 Process Mining Tools, Discovery Algorithms and Techniques

2.2.1 Introduction

Many software tools are available to facilitate PM such as ProM, Disco, Celonis, Interstage Business Process Manager, Rapid Miner, and ProMiner. According to Mans et al. (2013), ProM, which is an open source solution, has become the *de facto* standard for PM in research and is used in all the PM dental research literature. Although ProM offers a wide variety of PM techniques and algorithms and is an open framework environment allowing the development of plug-ins by researchers, a brief functional analysis of the available products in the literature would have been useful. Disco, a commercial product, has a more intuitive interface and would be more appropriate in some scenarios e.g. where the user has limited PM experience.

PM algorithms are specialised data analysis techniques designed to examine the EL and to produce a process model representative of the EL's contents. These are often classified in three groups; **deterministic**, **heuristic** and **genetic** algorithms (Gehrke & Werner, 2013). Some of the commonly used PM algorithms are the **Alpha Miner**, **Heuristic Miner**, **Fuzzy Miner**, **Inductive visual Miner**, **Genetic Process Mining**, **Region-based process mining**. Deterministic Algorithms produce defined and reproducible results. They are based on the ordering relationships between events. The Alpha Miner and its variants are deterministic algorithms. Heuristic algorithms incorporate the frequency of occurrence of events and can discover short sequences of events. The resulting process models reflect frequency of occurrence of traces and accordingly can eliminate 'noise' and rarely occurring events and traces if required. The Heuristic Miner is an example of this type. Genetic Algorithms much more resource intensive, generating large numbers of possible process models before deciding on the optimum. Typically, they follow the four steps; initialisation, selection, reproduction and termination, iteratively improving the final model over several generations. The AGNEs Miner (Goedertier, et al., 2009) is another algorithm facilitating the inclusion of negative events. A brief description of these follows.

2.2.2 The Alpha Miner

The Alpha algorithm produces a petri-net (place-transition) from a sequence of events. It does this by examining causal relationships between tasks. It takes an event log (or workflow log) W and a set of possible events T as inputs. It assumes that the log is complete with respect to all binary sequences and contains no noise (De Weerd, et al., 2012). In its basic form, it has several limitations and it has been enhanced as the Alpha+, Alpha++ and Alpha# models. The Alpha miner is mainly of theoretical interest and too simple to apply to real-life logs. It builds a model based on local relations between activities. It cannot deal with noise. Silent steps, non-local free-choice constructs, and duplicate steps (local loops) cannot be discovered. Short loops can be dealt with by the Alpha+ algorithm and Alpha++ can detect non-free choice constructs. Its strength is that it is a simple algorithm containing the basic PM ideas and concepts and can be formalised in a short form. It is, however, not robust and unsuitable for real world event logs.

2.2.3 Heuristic Miner

The Heuristic Miner was developed to address many of the problems of the Alpha Miner and can deal with noise and exceptions. It is especially suited to a real-life setting (De Weerd, et al., 2012). It outputs a heuristic net which can be converted to a Petri net which in turn can be formally analysed using the process-quality metrics. It is generally useful with real-life data containing ‘not too many’ different events. It is an extension of the Alpha Algorithm and can discover short loops and non-local dependencies. It has a noise threshold parameter setting making it suitable for a real-world setting. It applies frequency information to three types of relationships between activities in an event log; direct dependency, concurrency, directedly-connectedness. It derives XOR and AND connectors from dependency relations and can exclude exceptional behaviour and noise by leaving out edges. It lacks the capability of detecting duplicate activities. As with the Alpha & Alpha++, it builds a model based on local relations between activities.

2.2.4 Fuzzy Miner

This technique addresses some of the problems of large numbers of activities and highly unstructured behaviours. It employs an adaptive simplification and visualisation technique. It outputs a fuzzy model. It can simplify the process model at a desired level of abstraction and uses significance/correlation metrics to do this. It can hide less important activities in clusters and builds a model based on a global approach looking at the whole event log. This tool aims to emphasize graphically the most relevant behaviour,

by calculating the relevance of activities and their relations. Two metrics are used to present this. First, 'significance' measures the frequency of occurrences of events in the log, and second, 'correlation', determines how closely related two events that follow each other are, so that events highly related can be so represented in the model. It has limited ability to define choices and to define parallelism of events.

In both ProM and Disco, it is presented with an interface where the settings can be configured and their effect on the model can be seen immediately. The widths of the edges between the nodes is proportional to their importance (i.e. absolute frequency) and the darker edges indicate a higher level of correlation between the nodes i.e. their tendency to follow one another (Mans, 2011). The Fuzzy Miner is also capable of animating and replaying the log on the model. This gives a rapid, intuitive understanding of the process and quickly shows heavily executed paths and bottlenecks. Shortcomings of the models generated are that the model is without clear semantics which cannot be converted to other models. Due to this, the formal metrics commonly used to evaluate process models i.e. fitness, precision, simplicity and generalisability cannot be applied to the model.

2.3 Data Mining in Dentistry

Analysis using data mining tools has been previously applied in dentistry. Gansky (2003) applied knowledge discovery and data mining to a Rochester caries study. Classic Regression, Artificial Neural Networks (ANN) and Classification and Regression Tree (CART) caries prediction models were compared in this research. He used visualisations such as Area under the curve (AUC), CART, cumulative capture response graph and Receiver Operating Characteristic curves (ROC) to communicate the findings. He identified overfitting as a major cause of unreliable models and pinpointed data quality as an ongoing issue. Hsin-Fang (2013) applied similar methods and random forest algorithms to a large multidimensional dataset to identify the factors associated with untreated dental decay in childhood. Using multivariate logistic regression analysis, Nomura et al. (2004) concluded that cariogenic bacteria were the most important risk factor for dental caries in Japanese preschool children. Several other examples of data mining and analysis exist in dentistry: Kaplan-Meier survival curves were used to compare the longevity of restorative materials in varying circumstances (Kakilehto, et al., 2009) and apicectomy outcomes (Raedel, et al., 2015), To construct a dental caries prediction model by data mining, Tamaki et al. (2009) applied a balancing technique to the conventional neural network analysis, logistic regression analysis and decision analysis. This sought to ameliorate the skewedness of dental caries distributions, Work

by Bokhari et al. (2015) builds a model for classifying dental patients based on the importance of attributes and their relationships, using unsupervised classification and k-means, Choudhary & Bajaj (2015) applied a cross validation classification technique to assist automated prediction of root canal treatments to conclude that patient age was the most important decision attribute, and Cosgun et al. (2015) applied data clustering techniques to dental health centres in Turkey according to the services they offer.

2.4 Review Method

2.4.1 Search Process

The search terms used were (“process mining”) AND (“dentistry” OR “dental” OR “oral health” OR “oral disease”). The initial Google Scholar search was conducted through Harzing’s Publish or Perish (Harzing.com, 2018). The initial set of results (125 sources) was sorted by title and exported to Excel. Then the following source-based steps were taken by the author and reviewed by two supervisors. Adding the term ‘primary care’ and ‘public health’ to the OR clause yielded no additional results.

Further databases selected for searching were OVID Medliner, Pubmed, ACM DL, the Dental Informatics Online Community (DIOC) repository, and the Processmining.org repository. No additional articles were found. No results were found using the search terms (“process mining”) AND (“public health”). 1998 was the year the term ‘process mining’ was coined by Van der Aalst (1998). De Weerd et al. (2012) and Schimm (2003) attributed the foundational approaches to Agrawal et al. (1998), Cook & Wolf (1998) and Datta (1998). Previously used terms included, ‘workflow mining’, ‘pathway mining’ and ‘process analysis’. Use of these additional terms had no effect on the search results.

The 125 results were filtered as follows: remove non-English alphabet titles, remove non-English titles, remove nonsensical titles (e.g. author’s names etc.), remove duplicates (by title), remove titles relating to mining (of raw materials etc.), remove articles without a valid author, year, source, publisher, remove obviously irrelevant articles, remove non-healthcare or dental articles, remove fraud detection articles. On the basis of the title 40 articles were excluded. Abstract-based checking followed on the remaining 85 articles, and 33 additional results were identified as clearly not relevant in this phase. The author reviewed all 52 papers select for full text review and the 2nd two supervisors reviewed the abstracts of same and all agreed on the relevant dental PM literature. Full-text checking showed 28 were not dental nor healthcare related; 8 were fraud detection articles and 8 were pure data mining articles, leaving 9 articles on data mining, 7 on PM and 3 dedicated

to dental PM. An ancestor search presented no additional sources. The three dental publications identified are also those identified by Rojas et al (2016).

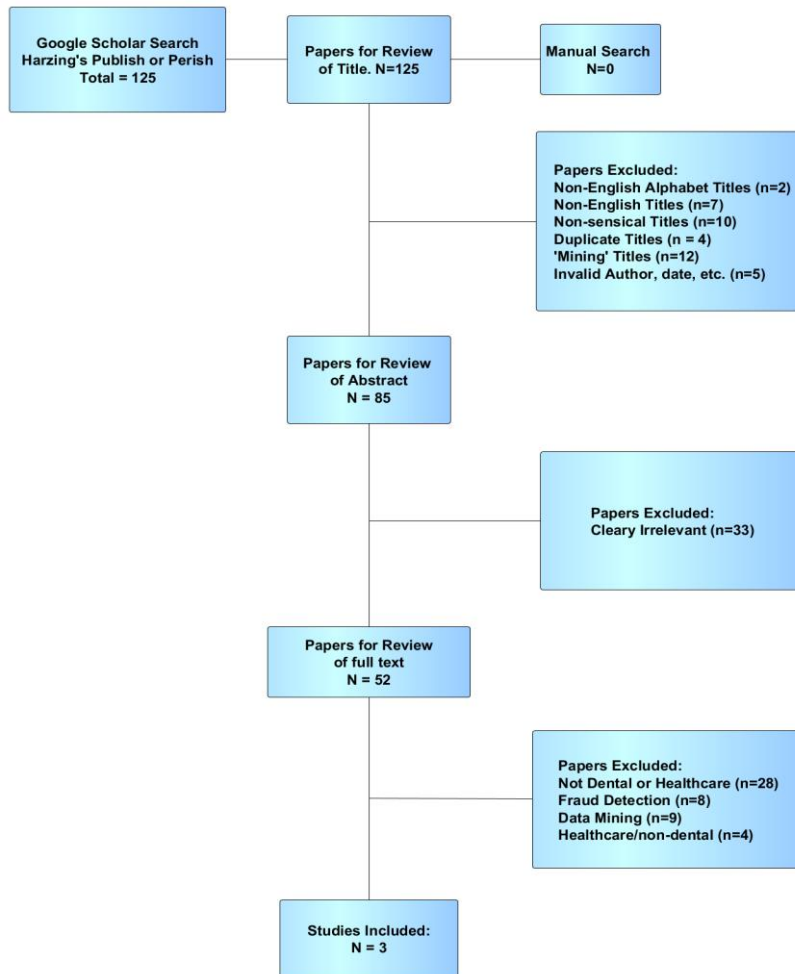


Figure 2-1: Literature search results and removal criteria

2.4.2 Quality Assessment

The quality of the search process was enhanced by implementing a series of measures. The initial search was undertaken by the author in October/November 2016 using Google Scholar in Incognito mode, through Harzing's Publish or Perish, to avoid previous searches introducing bias and to ensure consistent results. The author reviewed all 52 papers select for full text review and the 2nd two supervisors/ reviewed the abstracts of same and all agreed on the relevant dental PM literature. All dental PM literature originated in the Technische Universitaet Eindhoven, one of the main global PM centres. Both 2nd supervisors were of the opinion that the corpus of literature was small and it was agreed to supplement this review with a summary of the findings of previous reviews of healthcare process mining and dental data mining as presented in Sections 2.1 and 2.3 above. No dental process mining publications additional to those identified by Rojas et al. (2016) were found nor any relating to public health datasets. This reflects the emerging

nature of the PM discipline and proved useful in confirming the need for this research to be undertaken. The final follow-up search in September 2018 revealed no new publications relating to dentistry however, additional health literature review had been published in the interim. All supervisors reviewed and advised on all phases of the study.

2.5 Review Results

2.5.1 Introduction

Three dental PM articles were identified in the search process – all with the same combination of authors. These journal articles were ‘*A process-oriented methodology for evaluating the impact of IT: A proposal and an application in healthcare* (Mans, et al., 2013)’ and ‘*Is Your Upgrade Worth it? Process Mining can tell*’ (van Genuchten, et al., 2014). The third journal article, ‘*Mining processes in Dentistry*’ (Mans, et al., 2012), focusses on dental implants, a high-end recent innovation in dentistry. The publications emanated from a year-long research process that investigated to what extent workflow technologies could be used to help make the transition from analog dentistry to digital dentistry. Digital dentistry is the term used to describe the computer-based technological advances being applied in the delivery of dental care. Digital technologies such as dental imaging systems, x-rays, scanners replacing conventional dental impressions, digital placement software for dental implants, Computer Aided Design/Computer Aided Manufacturing (CAD/CAM) for crown manufacturing, digital printing, practice management systems all come under the digital dentistry umbrella (BDA, 2018).

While two of these journal articles were ostensibly about dentistry, it is more accurate to say that dentistry was used as a case study for their main research objectives, evaluating IT investments.

The review in this research looks at the existing literature relating to PM in dentistry, the literature relating to PM in public health and the literature relevant to PM’s potential to deliver worthwhile and novel insights to dental public health. An overview of the current work in these areas will be provided with the ultimate objective of identifying research gaps and opportunities to build on this existing work. To avoid confusion, the three publications, (Mans, et al., 2013), (Mans, et al., 2012) and (van Genuchten, et al., 2014) are referred to as (1), (2), and (3) below.

2.5.1.1 Mining Processes in Dentistry (Mans, et al., 2012) (1)

This work introduces the application of PM to digital dentistry and how this leads to digital islands in what the authors describe as, the predominantly analog world of

dentistry. The idea of cross-organisational PM is introduced by describing the role of the dental lab in the value stream. The paper used an explorative approach and the main PM perspectives were introduced: control flow, organisational, and performance. Their main conclusion is that PM is a useful tool for gaining a deep understanding of the dental processes and that workflow management technology is needed to make the introduction of digital dentistry a success. The main steps in their PM method is shown in Figure 2-2.

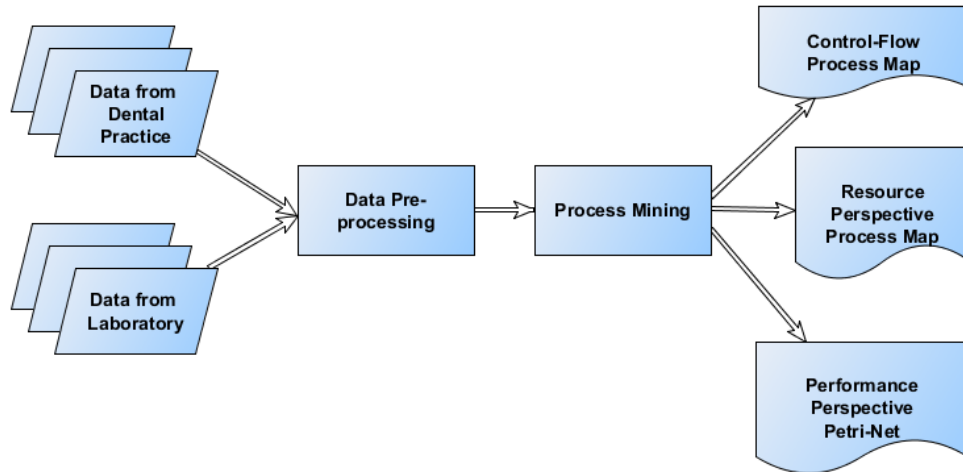


Figure 2-2: Process Mining in Dentistry adapted from Mans et al. (2012)

2.5.1.2 A process-oriented methodology for evaluating the impact of IT (Mans, et al., 2013) (2)

This work also develops the idea of digital dentistry and proposes a methodology using PM in combination with discrete event simulation to assist with the evaluation of proposed Information Technology (IT) innovations ahead of implementation. Their proposed methodology is compared to the L*life-Cycle as detailed in the Process Mining Manifesto (IEEE, 2011) and the framework detailed by Zhou & Piramuthu (2010). They conclude that these existing methods focus on the analysis of an existing business system. They aim to neither forecast the effects of a change nor to evaluate a change within these processes and to address this shortcoming, they propose the steps in Figure 2-3 below.

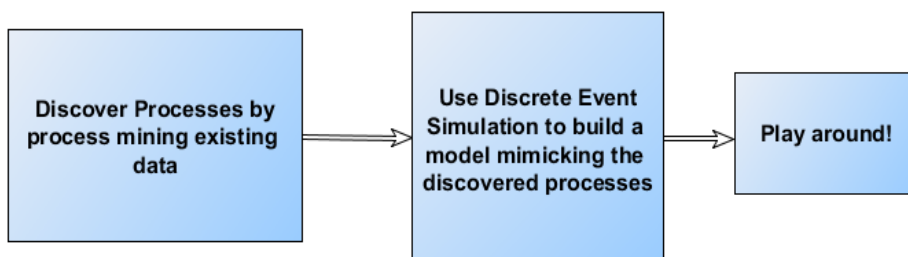


Figure 2-3. Evaluating the impact of IT using Process Mining adapted from Mans et al. (2013)

The discovery of models using PM accelerated the development of a simulation model. Key Performance Indicators (KPIs) are identified and used to quantify and evaluate the process changes, including patient throughput time, required dentist time and required lab-technician time. The results are validated using statistical processes.

The method is evaluated using a dental case study, the implant value chain. PM is used to get a detailed quantitative understanding of the process as-is. This includes the process from the making of dental impressions using the traditional impression tray through to the production of the restoration using conventional techniques. This process is then compared to the ‘digital’ process which utilises intra-oral scanning to produce a ‘digital’ impression and the Computer Aided Design/Computer Aided Manufacturing (CAD/CAM) production of the final restoration. The KPIs identified include patient throughput time, required dentist time and required lab-technician time.

Omitted from their journal article for brevity was a precursor available at <http://bpmcentre.org> (Mans, et al., 2013) and providing some additional insight to the application of the methodology to the crown process.

2.5.1.3 Is your upgrade worth it? Process mining can tell (van Genuchten, et al., 2014)

This work (3) proposes the use of PM to demonstrate that upgrading to new software releases provides quantifiable benefits to users i.e. end-users, software suppliers and researchers. Applying PM to digital dentistry was part the investigation of how workflow technologies could help transition from analog to digital dentistry. The writers propose that software suppliers would apply PM to quantifiably assess the benefits of a software upgrade to their customers. They proposed a 5-step methodology for this as in Figure 2-4.

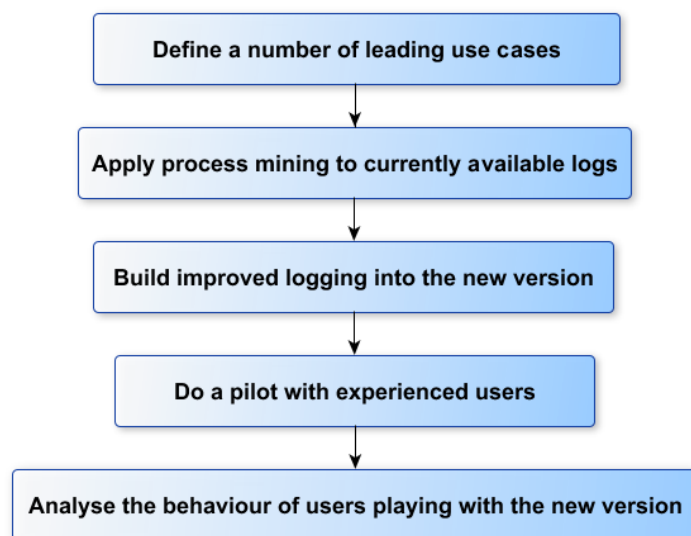


Figure 2-4: Is your upgrade worth it? adapted from van Genuchten et al. (2014)

They also identified pre-requisites to the successful implementation of PM in this scenario including accessibility of data on high-frequency processes and comparable use cases and stakeholder consent and participation. The use case was a dental design software package consisting of three parts: intra-oral scanning, the design of the dental element and CAD/CAM or 3-Dimensional (3D) printing of the dental element. PM was applied to the 2nd phase, the design of the dental element, producing Petri-net models showing control-flow, resource and performance perspectives. The comparison of 900 designs made with the existing software to 500 designs made with the newer version concluded that the new software led to an 11% reduction in design time for the end-users. The study also uncovered the importance of bug-fixing over feature development to the end-user, which proved valuable to the vendor and its engineers and concluded that presentation of these figures to clients would provide a persuasive argument to upgrade.

2.5.2 Thematic Analysis

The underlying theme of the articles is the transition from analog dentistry to digital dentistry and the impact that this has on workflows and the organisations and stakeholders. Publications 2 & 3 focus on the impact of IT innovations on the business processes. Both use dentistry as a case study to demonstrate their methodologies. Publication 1 addresses the applicability of PM to dentistry and focusses on the dental treatment, implant & crown, to demonstrate this.

To do a thematic analysis of the available literature, close reference is made to a comprehensive review of the literature on PM in healthcare by Rojas, et al. (2016), who provided a concise series of themes to analyse and assess literature on the topic.

Process types are categorised as medical treatment processes and organisational processes, with medical processes subdivided into elective and emergency.

Four sources of data for PM were identified: administrative systems, clinical support systems, healthcare logistics systems and medical devices. Typical PM questions were uncovered: What happened? Why did it happen? What will happen? and What is the best that can happen? Three main types or perspectives of PM are identified: process discovery, conformance checking and process enhancement. This analysis can be extended through: organisational mining, simulation model construction, model extension and repair, prediction, and recommendations based on history.

The PM tools are listed as ProM, Disco, RapidProM and Rapidminer. Several of the papers used PM methodologies e.g. the L*Life-cycle and Sequence Clustering Analysis.

Implementation strategies were categorised into three types: direct implementation where PM is carried out directly on a set of historical data, semi-automated implementation where the extraction of data and the creation of an event log is a bespoke operation and finally an integrated suite implementation for specific environments.

Analysis strategies were categorised according to the degree of sophistication of the techniques used in that phase of the PM study. Further themes established by their study were Geographical Analysis and Medical Fields.

Consistent with this approach, the dental PM literature has been analysed in reference to the applicable general PM literature and is summarised in Table 2-1 below.

A detailed critique of the PM in dentistry literature using this structured analysis is in Appendix 0 and could be repeated when additional literature becomes available.

Table 2-1: Summary of dental process mining literature analysis

Themes	1 Mans et al. (2012)	2 Mans et al. (2013)	3 van Genuchten et al. (2014)
Process Types	Elective treatment/ organisational process	Business processes	Business processes
Data Types (Sources)	Administration and clinical support data	Administration and clinical support data	Administration and clinical support data
Frequently posed questions	What happened? Why did it happen?	What happened? Why did it happen? What will happen? What is the best that can happen?	What happened? Why did it happen? What will happen? What is the best that can happen?
Process Mining Perspectives	Control Flow Performance Organisational/ Resource	Control Flow Performance Resource	Control Flow Performance Resource
Process Mining Tools	ProM	ProM	ProM
Techniques and Algorithms	Heuristics Miner Social Network Miner Petri-net Performance-analysis with Petri-net	High Level Petri-nets Dotted Chart Non-specific regarding algorithms	Not Specified (Petri-nets shown)
Methodologies	Non-Specific	Add to existing methodologies - Discrete Event Simulation	Non-Specific
Implementation Strategies	Direct Implementation	Direct Implementation	Direct Implementation
Analysis Strategies	Advanced Strategy with new Plug-in and Ontological input	Advanced strategy incorporating Discrete Event Simulation	Non-Specific
Geographical Analysis	Europe/Netherlands	Europe/Netherlands	Europe/Netherlands
Medical Fields	Dentistry	Dentistry	Dentistry

2.5.3 Thematic Summary and general discussion

2.5.3.1 Key findings of previous papers

The previous work in dental PM provides valuable insights on how this technology can be applied in private clinical practice. Clear examples of the discovery of process details using the control-flow perspective are demonstrated in all the literature. Issues such as cross organisational mining are addressed (Mans, et al., 2012). Additional valuable demonstrations of examining processes from a resource and performance perspective are presented by Mans, et al (2012; 2013). Suggestions as to how the effects of technology upgrades are presented (van Genuchten, et al., 2014; Mans, et al., 2013) and these provide interesting templates for how strategy or policy initiatives could be assessed, and this is of particular interest in this research.

Publication 1's main conclusion is that PM is a useful tool for gaining a deep understanding of the dental processes and that workflow management technology is needed to make the introduction of digital dentistry a success. Publication 2's main contribution is the development of a combined approach, using PM and discrete event simulation to allow for evaluation of an IT initiative in clinical dental practice. Publication 3 concludes that software suppliers would benefit from using PM to demonstrate to customers that an upgrade will be worthwhile. All articles resulted from a year-long research effort to establish how workflow analysis and technologies could aid in the transition from analog to digital dentistry in private clinical practice.

2.5.3.2 Limitations of previous papers

Their work focussed on implants and crowns, in a private dental practice and not on the general context of dentistry. By focussing on a single treatment process, implants and crowns, the work has limited generalisable value to dentistry, above and beyond other studies on specific medical treatments. In their research, they suggest that this technology will lead to the discovery of how dental processes are executed in reality. It is claimed that PM offers a less subjective version of events than the more traditional ways of investigating business processes (Mans, et al., 2015, p. 3). It is suggested that interviews, for example, have the potential to deliver 'highly subjective information' (Mans, et al., 2012). The article neglects to mention established methods to address the biases inherent in interviewing and other techniques as demonstrated by Chenail (2011) and Pannucci & Wilkins (2010).

The authors provided us no structured method to verify that the discovered processes are truly representative of what is happening in the real world. While the article refers to the

writers having validated the resulting discovered processes with the owner of the process, there is no detail as to how this was achieved and inevitably there is some subjectivity in this process and reintroduces the subjectivity biases that the authors suggested would be removed by using PM technologies. Related to this, there is no structured method to ensure that all the relevant steps in the treatment process have been recorded in the organisation's information technology systems in the first instance.

This article goes on to analyse a real case of the diagnosis and placement of implants and the final restoration, including the activity of the dental lab producing the crown. The article chose to view PM in dentistry at the micro-level, focussing on a single treatment, the implant/crown process. High-end treatments such as implants are the reserve of the few and gains in efficiency in these are unlikely to impact public health, treatment availability or outcomes. The value of choosing a high-end, very expensive treatment for such a study would appear to be limited to gaining efficiencies within the single process itself and not generally applicable to dental care processes.

In publication (1) with unfiltered data, the Heuristic miner produced a complex, spaghetti-like process model. In the methodology section, they describe a process of consolidating event names and the use of a new ProM plug-in to effect this. It is unclear whether the plug-in is exclusive to dentistry. They also speak about mapping event-names to 'subjects' though there is no additional information on these 'subjects'. It is unclear whether the research used any standard diagnostic or treatment codes such as ICD 9/10 or SNODENT in this phase. The authors also manually linked patient information from two disparate systems - the dental practice and the laboratory. This seemed to be on a 'best-guess' basis and no detailed method was described. This may have been error prone. Again, the authors refer to the validation of the discovered processes with the owners of the process without any methodology or data. A structured approach would lead to more reproducible research.

In my view, a major shortcoming in the work is the lack of a predefined structured method although in publication (2) Mans et al. (2013) referred to existing methods but pointed out that none aimed to evaluate changes within the process. They then aimed to develop a method appropriate to PM combined with discrete event simulation. All three research efforts could have benefited from applying the discipline inherent in the above methodologies.

In the conclusion of publication (1), the authors claim that several innovative methodological and technical steps were taken in the research. They are however not specific about this and it would benefit from clarification. They also introduce new

material on the increasing need for workflow management technologies and constraints which could have been introduced and contextualised earlier in the article.

2.5.3.3 Implications for future research and direction for my research

PM is an emerging technology in the field of data mining. This review investigated to what extent it has been applied to the dental domain. Only three publications referred to the dental domain. Two of these publications used dentistry as a case study to evaluate IT innovations and the third focussed solely on dental implants & crowns. There was no prior publication on PM in dental public health and no publications for public health.

Examining the current literature showed that applying state of the art PM techniques to dentistry can add significantly to the existing work in PM done in the Technical University in Eindhoven (<http://processmining.org/>). Application of a structured method such as PM² would force a disciplined consideration of project objectives and RQs and the suitability of the existing information systems to answer these questions. While ProM may have become the *de facto* standard for PM research, the attributes of all tools also should be considered for other scenarios where usability or user-friendliness may be factors. The data extraction and pre-processing phases require close attention as extraction and filtering techniques can lead to a loss of accuracy in the data analysis. The choice of techniques and algorithms deserves careful consideration as they each bring their own advantages and limitations as seen above. Also, the assessment of the quality of the PM results is a crucial step in the process. The quality attributes of fitness and accuracy conflict with each other and this requires close attention.

All of this points to a large opportunity to researchers who can access large data extracts from EHRs or public health datasets with a view to analysing dental treatment processes, dental care pathways and dental care processes. The existing literature focusses on the single treatments of implants or crowns with a view to achieving efficiencies in specific treatments rather than the objective of improving the care pathways followed by patients and this goal can be more easily achieved with large-scale datasets and a thorough methodological approach.

Broad Public Health Focus

By focussing on the high-end procedures of crowns and implants, the existing research ignored the benefits that public dental services could derive from PM. The existing research's value is primarily in provide specific valuable insights in the control flow of the implant/crown process and assessing resource usage and performance. In our view,

it would also be useful to focus on the common procedures such as examinations, fillings, fissure sealants etc. - procedures that make up the bulk of public health population level activities. Even a minor improvement in these commonly executed processes could have significant impacts both financially and on outcomes. Conformance analysis of the processes with established standards such as those in the Steele Report as shown in Section 7.1.4 and established evidence based clinical guidelines (Irish Oral Health Services Guideline Initiative, 2012) would provide valuable feedback to policy makers and other stakeholders. It is clear from the existing literature that little or no research is being done at the public health level in dentistry. PM research at this macro level would bring many of the issues mentioned above into sharper focus.

2.6 This Research's Process Mining Vocabulary

2.6.1 Introduction

PM has developed a terminology or vocabulary of its own. To understand the PM technology and how it is applied and evaluated, it is important for this new vocabulary to be clear and unambiguous. PM is an emerging data analysis technology and as with all data science, it incorporates skills from many disciplines and overlaps with these disciplines in places (Mans, et al., 2015, p. 4). For the purposes of this research, the author has attempted to list some of the terms commonly used, and to explain or define them as they are used in this research. Any words in bold are further defined or expanded on later in this chapter. The aim is to force clarity about the objects that exist in this research and how they are related to each other and how this research fits in the wider discipline of PM. Clarification of these terms is helpful within the PM community but more importantly, when communicating our findings to healthcare domain experts who may have prior exposure to these terms in different contexts and with different meanings.

PM aims to extract high level knowledge from low level data. It aims to do this by discovering process models from event logs (**Process Discovery**), checking the conformance of **event logs** with existing **process models** (**Conformance Checking**), and enhancing models with additional information (**Process Enhancement**). This is confirmed by the IEEE task force on Process Mining having stated that PM is not limited to process discovery but includes other dimensions including conformance checking, performance diagnosis, organisational mining, prediction etc. They identify the key requirement to be that:

“analysis is based on ‘facts’ from an event log and that process models whether discovered or modelled, play a role in this”. (IEEE CIS Task Force on Process Mining, 2010)

There are demonstrable ambiguities in the use of several of the common PM terms and the objective here is to provide a vocabulary for PM in this research that is clear and stable. To provide additional structure and clarity to this research’s vocabulary, the author has divided the vocabulary into three sections; Data-level, approaches-level and model quality-level. First, data-level terms describe the data terms used in PM. Second, the approaches-level describe types of PM, PM perspectives and PM objectives. Third, as we are not formally evaluating model quality, the model quality-level terms section has been omitted from this thesis.

2.6.2 Process Mining ‘Data-level’ Vocabulary

2.6.2.1 Event data

Event data are an extract from an organisation’s information systems, suitable for PM. They contain time-ordered lists of discrete **activities or events** i.e. well-defined steps in **processes**. Event data is the raw data needed for all of the PM discovery and conformance checking techniques and event data should be as “raw” as possible (van der Aalst, 2013). It is a subset of organisations’ data systems extracted for the purposes of PM. The Event Data could be extracted from multiple IT systems in an organisation e.g. it could consist of data from a customer relationship management system (CRM) and an accounting or purchasing system. Sufficient data will usually exist in event data for multiple event logs and studies. Event data will often be divided into subsets either by time, event classes, or other criteria. These subsets are called Event Logs

2.6.2.2 Event log

An event log (EL) is a subset of **event data** created to execute a single experiment. An EL might be created to examine the events that are related to a specific result e.g. an event log might contain all of the events that proceed the extraction of a diseased tooth. This EL might then contain events such as x-rays, fillings, and dressings for many extractions. This event log would often then be analysed to find the most common pathways, to analyse resource usage and similar questions.

It contains time-ordered lists of discrete **activities or events** i.e. well-defined steps in **processes**. The most basic EL contains a **case-identifier**, an **event name** and a **timestamp**. Bozkaya et al. (2009) refer to ELs or audit trails typically existing in information systems supporting business processes. An EL typically contains information about the start &

completion of process tasks together with related context data (e.g. actors and resources) and timestamps (Rovani, et al., 2015). This introduces another word to describe an ‘event’ i.e. ‘process task’. This author believes this additional term is unnecessary. “An event log can be viewed as a set of traces (also known as cases, or in the emergency room, episodes), each containing all of the activities executed for a particular process instance” (Rojas, et al., 2017). There can be many ELs extracted from a single instance of event data.

2.6.2.3 Activity log

An activity log is a partial description of an EL. It is a list of the distinct processes in an EL along with their frequency and the events contained in each one (Verbeek & van der Aalst, 2015). Given the definitions used here, this could also be applied to ‘cases’ instead of ‘processes’ and for this research’s purposes will be so applied. In the case of the event log containing tooth extraction events as described above, some of the extracted teeth may have had no preceding events, some may have had a tooth dressing only whereas others may have had prior restorations, x-rays etc. Whereas the EL would contain each of the events, the activity log would only contain a list of the unique pathways and their frequency e.g. The sequence “Filling, tooth-extraction” occurs 100 times.

2.6.2.4 Pathway (path)

A pathway is a set of broadly similar traces or ‘variants of processes’. They do not have to be identical but should have enough similarity to merit grouping together for analysis or discussion. Referring to the example above, the existence or not of a prior x-ray may not be considered significant to the research question and therefore both process variants would be included in the pathway. The term ‘Groups’ is also used by Rojas et al. (2017). In the PM healthcare context Yang & Su (2014) cite clinical pathway as a structured, multidisciplinary, patient care plan in which diagnostic and therapeutic interventions performed by physicians, nurses, and other staff for a diagnosis or procedure, sequenced on a timeline.

2.6.2.5 Process

This is very similar to a pathway. It is a grouping of similar sets of activities or events within a problem domain. In this research, pathway, path and process are synonymous.

2.6.2.6 Subprocess

A subprocess is a distinguishable part of a process. This might be a specific pattern of events within an overall process and might be given a name in order to simplify the process model. A process can be decomposed into subprocesses (Rojas, et al., 2017). Decomposing a process into subprocesses can be useful for simplifying complex process where a number of events or steps can be represented by a single subprocess.

2.6.2.7 Trace & Case

A trace or case is the complete, specific sequence of events, as recorded in the EL for a single experience of the process of interest typically, in healthcare, by a particular patient. van der Aalst (2013) described a trace as the lifecycle of a particular case in terms of the activities executed. It is an instance of a process or “process-instance”. Following van der Aalst (2013), the events belonging to a trace or case are ordered and can be seen as one run of the process. Bozkaya et al. (2009) also appear to suggest ‘trail’ or ‘run’ as alternatives. These authors consider these additional terms to be unnecessary.

2.6.2.8 Variant

The phrase process variant is used to describe the set of traces which follow identical sequences of events and, of course, where traces follow a unique sequence of events they can also be described as variants.

2.6.2.9 Classifier

A classifier is the name given to a trace, e.g. the trace occurring most often in an EL might be known by the classifier, ‘Most-common-trace’. It is a convenient way of referring to a trace using a name instead of a series of events e.g. ‘The Swedish Trace’ instead of ‘ABBAABBACDEFG’. In a dental scenario, the ideal sequence for a school dental service might be ‘Examination’, ‘Oral Health Instruction’, and ‘Apply Fissure Sealants’. This sequence might be given the name like ‘Ideal School Service’.

2.6.2.10 Episode

The term ‘episode’ arose in the context of emergency room PM. ‘*Episodes can be clustered into groups*’ (Rojas, et al., 2017). This seems to be the same as ‘case’.

2.6.2.11 Event

An event is a well-defined step in a **process**. An event is an instance of an **event class**, In PM it is also known as an **activity** or process-step. An event can only belong to one **case**. Each event in such a log refers to an activity, i.e. a well-defined step in some process and is related to a particular **case** i.e. a **process instance**. The events belonging to a **case** are ordered and can be seen as one instance of the **process** (van der Aalst, 2013). Events in the example above would be the process steps preceding the extraction of a specific tooth i.e. the specific ‘filling’ carried out or the specific ‘dressing’ applied to the tooth.

2.6.2.12 Event class

An event class is a distinct event within an event log i.e. a type of event. There can be multiple instances of an event class in an event log or even within a **case**. In the example, this would be ‘filling’ or ‘dressing’ but does not refer to a specific occurrence of these.

2.6.2.13 Activity

In PM an event is also known as an activity or process-step. It is synonymous with event in this research. Bozkaya et al. (2009) suggest that an activity can have multiple events. Although the authors are vague, it appears that an activity would be e.g. ‘process an insurance claim’, their use of the term ‘activity’ is ambiguous and confusing in this context. ‘Pathway’, ‘processes’ or ‘trace’ might be more appropriate in that case. In summary, this thesis considers event, activity and process-step to be synonymous .

2.6.2.14 Timestamp

A timestamp is a record of the time that an event took place. This is an essential element of the EL, allowing us to create time-ordered sequences of events. Timestamps can be of varying levels of detail e.g. date alone, date & time to minute level, second level etc.

2.6.2.15 Process model

This is an abstract representation of the essence of a process reflecting the common pathways. Models may be descriptive – describing a process – or prescriptive, enforcing a particular way of working. This model is often presented as a petri-net, BPMN model or similar. Figure 2-5 below shows a sample dental process model in petri-net format.

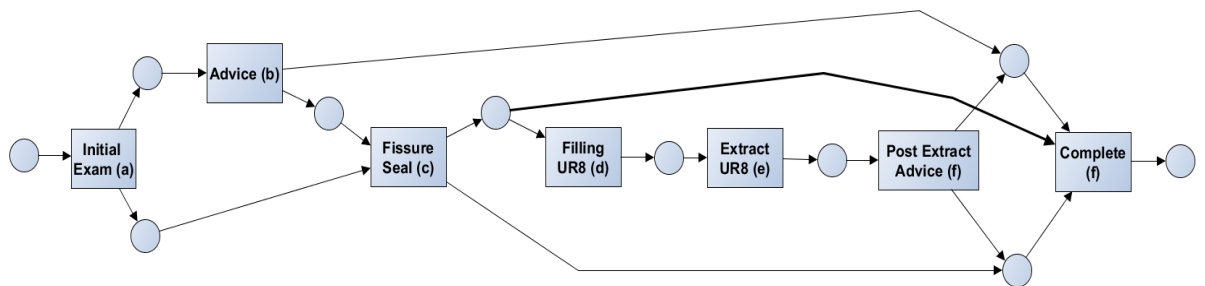


Figure 2-5: Example Process Model

Models can be described as either *de facto* or *de jure* depending on their origin. *De facto* models are based on facts and in the PM world, this means that they originate from an event log i.e. from a record of facts about historical events (van der Aalst, 2011). *De jure* models originate in laws, rules, guidelines, standard operating procedures, and such like.

2.6.2.16 Putting these terms together.

Taking the above definitions and linking them: a case is made up of a sequence of specific events from an EL. Each of these events belongs to a class of events. A case’s trace is its unique sequence of events and there may be multiple cases in the EL with the same sequence of events. All of these cases would be of a particular trace, named by a classifier. Processes, pathways or paths are groups of broadly similar traces. They may be grouped

together to facilitate common analysis or discussion for the purposes of an experiment. Each EL can consist of many different processes. ELs are subsets extracted from event data for the purposes of specific experiments. The event data are the data extracted from the organisations information systems for the purposes of PM. This can be roughly represented in the onion format in Figure 2-6 below. It could also be represented in a hierarchical or entity relationship format. In our example above, the organisation's data would be all of the data contained in the dental service's information technology systems. The event data would be the extract from these systems potentially capable of addressing multiple questions. The event logs would be subsets of the event data to address a single question such as the process leading to extraction. Processes, pathways and paths would be all of the cases that are similar to each other and where the differences are not significant. Traces and Activity logs help describe the types of processes a case might experience and might also have information about how frequently each variation exists in the log. Each tooth extracted under general anaesthetic is a case and each step in each case is an event, process step, or an activity.

The vocabulary model in Figure 2-6 below is developed from the terms defined above and represents the commonly used terms in describing the data used in PM and their relationship to each other. More specifically, it demonstrates how this research's author views each term to be a subset of the larger term, or to be an instance of the larger term or to be equivalent to a term at the same level (in the same ellipse).

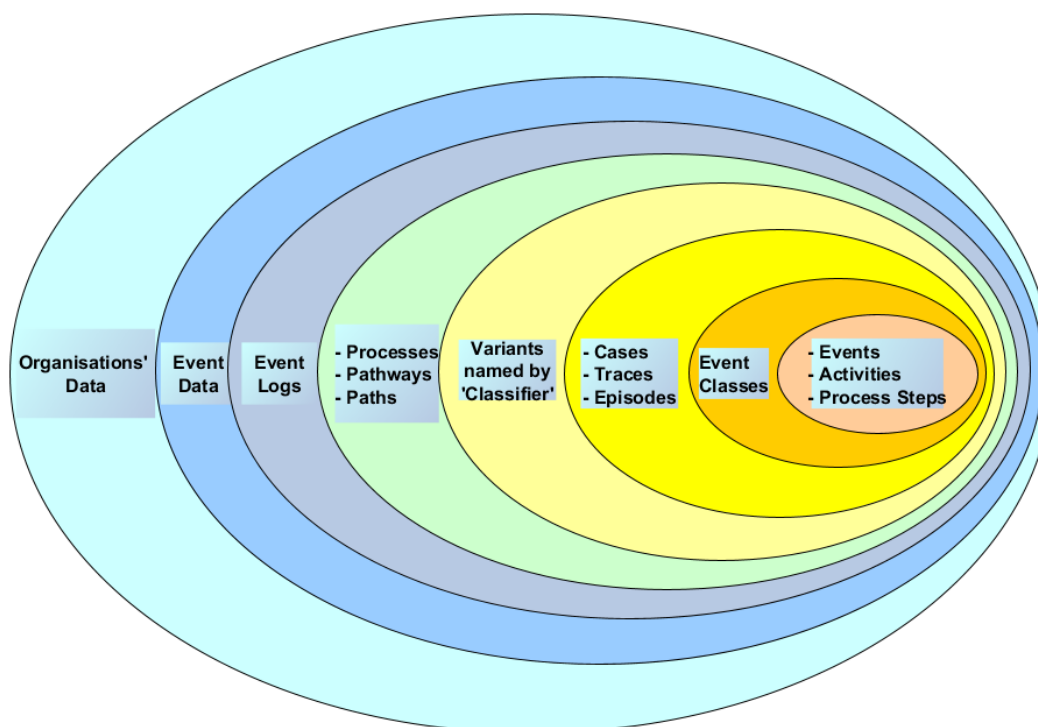


Figure 2-6: Process Mining Data-level Vocabulary Model

2.6.2.17 A little more information about event logs...

While an event log (EL) is a multiset of cases and a subset of event data, it is important to recognise that it is likely only a sample of the traces possible in real-life. This is a potential significant bias in all research involving event logs and PM. This bias is a phenomenon known as the assumption that ‘What you see is all there is’ (WHSIATI) (Kahneman, 2012). ‘Log Completeness’ describes the degree to which an EL is representative of the possible sequences of events in the real world.

ELs have several properties that provide an overview of their contents.

- A visual overview of an EL can be seen using a dotted chart.
- Number of event classes. This is the number of distinct events in the EL.
- Number of cases.
- Number of distinct cases (traces).
- Level of Detail – average number of event classes per trace.
- Structure - amount of observed behaviour compared to the amount of theoretically possible behaviour. A low value here is challenging for PM because of the difficulty in representing unstructured behaviour.
- Mean Affinity - has similarities to Structure and represents the mean relative overlap of direct following relations between each two traces in the event log. A low value again makes process discovery difficult - indicating diverse behaviour.
- Does it contain loops, skips, non-free choices, duplicates?

2.6.3 Process Mining ‘Approach-level’ Vocabulary: Types and Perspectives

Existing Literature

This research author’s view is that the literature is ambiguous and unclear on the distinction between PM ‘type’ and ‘perspective’, and they have overlapping definitions and descriptions. In their publication ‘Process mining in Healthcare’ (2015, p. 5), Mans et al. identify three ‘types’ of PM; discovery, performance and enhancement. If the EL has extra information, then ‘*we can learn additional perspectives and enrich the model*’. Specifically, they enumerate the organisational perspective, the case perspective and the time perspective. The control flow perspective was added by van der Aalst (2016, p. 34) who stated that these perspectives are orthogonal to the PM types. According to Rojas et al. (2016) there are four PM perspectives; control flow, performance, conformance, and organisational. The Process Mining Manifesto says PM includes automated process discovery, conformance checking, performance, enhancement, social network and organisational mining, construction of simulation models, model extension, model repair,

case prediction and history-based recommendations (IEEE, 2011). Weijters et al. (2006) list three PM perspectives; process, organisational, and case. Rebuge & Ferreira (2012) list four perspectives, control-flow, organisational, data, and performance. Their definition of the data perspective is unclear. Bozkaya et al. (2009)'s outcomes cover the control flow perspective, the performance perspective, and the organisational perspective. The existing literature and publications, though using varying terminology, agree that the ordering of activities from the EL and the presentation of this information in the form of a process model is a key type of PM. Their varying terms, Control Flow, Discovery, Automated Process Discovery and Process can all easily be understood by using the term 'Process Discovery'. The varying approaches are summarised in Table 2-2 below.

Table 2-2: Summary of Literature's PM 'types' and 'perspectives'

	PM Types			<i>If the Event log has extra resource information, the PM Perspectives to the right may be available</i>	PM Perspectives			
Process Mining in Healthcare (Mans, et al., 2015) (Types)	Automated Process Discovery.	Conformance 'Monitoring Deviations', Model Repair	Enhancement 'Extend' and 'Improve'		Organisationa 1	Performance Time perspective	Case	
Process Mining in Healthcare (Rojas, et al., 2016) (Perspectives)	'Control Flow' Ordering of Activities	Conformance			Organisationa 1	Performance		
Process Mining Manifesto (IEEE, 2011) (Types)	Discovery	Model Repair	Model Extension,		Social Network/ Organisationa 1		Case Prediction	Simulations Recommendation s
Process Mining with the Heuristics Miner Algorithm (Weijters, et al., 2006) (Perspectives)	Process Discovery	Conformance Checking			Organisationa 1		Case	
Process Diagnostics (Bozkaya, et al., 2009) (Outcomes)	Control Flow				Organisationa 1	Performance		
Business Process Analysis (Rebuge & Ferreira, 2012) (Perspectives)								
Process Mining: Data Science in Action (van der Aalst, 2016)	Discovery	Conformance	Enhancement		Organisationa 1	Performance	Case	Control-Flow
Consolidated Terms for this research (Types)	Process Discovery	Conformance Checking	Model Enhancement		Organisationa 1	Performance	Case	Control-Flow

This researcher considers van der Aalst's (2016, p. 34) use of the terms to be the most intuitive and therefore uses these as the basis for the consolidated terms for use in this research as presented in the last row Table 2-2 above. A brief description of each follows.

2.6.3.1 How this research views 'types' and 'perspectives'

Using van der Aalst's idea that PM perspectives are 'orthogonal' to the PM types, this could be represented in a 2-D grid format, where every PM exercise can be categorised using two words, one from each of the following 2 sets {Discovery, Conformance, Enhancement} and {Control Flow, Organisational, Cases/Data, Performance/Time}.

Table 2-3: Process Mining Types and Perspectives

	Discovery	Conformance	Enhancement
Control Flow Perspective	1	5	9
Organisational Perspective	2	6	10
Case/Data Perspective	3	7	11
Performance/Time Perspective	4	8	12

2.6.3.2 Process Discovery

Referring to Table 2-3 above, 1 through 4 are the Process Discovery perspectives.

Process discovery is used when there is no existing (a priori) process model. A process model is 'discovered' from an event-log using specialised **PM discovery algorithms and techniques**. This type of PM focuses on the ordering of activities or events and presents the discovered model as a **petri-net**, an Event Driven Process Chain (**EPC**), in Business Process Model Notation (**BPMN**) model or similar. The model shows the control flow of the process with the events ordered by their timestamp.

1. This is a process discovery exercise using the control flow perspective i.e. discovering processes by finding the sequence of events in the process.
2. This is a process discovery exercise using the organisational perspective i.e. we are trying to discover processes where the important thing is establishing the structure of the organisation by classifying people into roles and units, or by creating the social network. This approach is not in use in this research.
3. This is a process discovery exercise using the case/data perspective i.e. we are trying to discover processes where the interesting thing is the properties of the data associated with sequence of events or steps in the process. An example in use in this research is where we look at the oral health outcome DMFT at the time that a dental examination event occurs.

4. This is a process discovery exercise using the performance/time perspective i.e. we are trying to discover processes where the important thing is the sequence of events and the durations, times-between, and volumes are the important factors.

2.6.3.3 Process Conformance

Referring to Table 2-3 above, 5 through 8 are the Process Conformance perspectives. This is where an existing process model is compared to event log of the same process to check if reality (i.e. the event log) conforms to the model and vice versa. This is often termed ‘model-alignment’. Comparing an EL to an existing model is known as conformance checking (CC). This existing model can be an output from process discovery above or a process model from a different source although some literature suggests that CC only compares a process model with its corresponding EL (Van der Aalst, 2015). He continues to propose three primary use cases for CC: auditing and compliance, evaluating process discovery algorithms, and conformance to specification of software and services. Evaluating process discovery algorithms suggests that the process model used in CC must have been directly created from the EL against which its conformance is being checked. The data in the EL are then compared with the existing model and **model discrepancies and deviations** are identified and analysed. “Do the model and the log *conform* to each other?” CC seeks to identify discrepancies between the model and the EL and to quantify these discrepancies with metrics, (Rozinat & van der Aalst, 2008), also known as ‘business alignment’. Can each case in the EL be replayed on the process definition? (Bozkaya, et al., 2009). The most important requirement for conformance is ‘fitness’. The other important requirements are precision, simplicity and generalisation. Conformance can be viewed from two points of view; local conformance which checks for deviations at specific nodes and global conformance measures measuring the overall relationship between the model and the log. Rojas et al. (2017) define conformance checking as ‘...based on comparing a process model with an event log to verify whether the process is executed according to that model’.

This should not be confused with compliance checking which deals with the adherence of a process to internal or external rules e.g. the requirement that at least two people are involved in a process to reduce the opportunities for fraud and error, also known as the ‘4-eyes principle’ (Gehrke & Werner, 2013). Wil van der Aalst (2015) considers compliance to be a use case of CC. In contrast, Rovani et al. (2015) propose a PM methodology ‘to check the compliance of the clinical guidelines (the *de jure* model) against the actual clinical practice, recorded as an event log’ and to ‘check the

conformance of the de jure model, which encodes the medical guide-lines, against the actual process executions, which are recorded in logging data'.

This research takes the broader view and CC is understood to be the checking the conformance of an event log against a process model irrespective of the model's origin. This approach is supported by recent publications where CC is comparing a discovered model to a reference model (Bloemen, et al., 2018; Burattin, et al., 2018).

5. This is a process conformance checking exercise using the control flow perspective i.e. to what extent the processes model agrees with the sequence of events or steps present in the event log is being checked.

6. This is a process conformance checking exercise using the organisational perspective i.e. to what extent the model of the organisation agrees with the organisational information available in the event log is being checked.

7. This is a process conformance checking exercise using the case/data perspective i.e. to what extent the case/date information in the existing model agrees with the case/data information available in the event log is being checked.

8. This is a process conformance checking exercise using the performance/time perspective i.e. to what extent the model performance information agrees with the performance information available in the event log is being checked.

2.6.3.4 Process Enhancement

Referring to Table 2-3 above, 9 through 12 are the Process Enhancement perspectives. This is where an existing process model is enhanced or extended to include additional information from the event log. These perspectives are not used in this research.

9. Process enhancement using the control-flow perspective. This is sometime known as 'process repair' where additional control flow information is added to the existing model to make it more accurately reflect reality.

10. Enhancing existing model with organisational information from the event log. Enhancing existing model with case/data information from the event log.

11. Case mining is concerned with the properties of cases. Cases can be characterised by their path in the process or by the resources used by the case. When additional data corresponding to cases and events exists, the cases can also be characterised by the values of their corresponding data elements. For example, if a case represents a specific treatment of patients in a hospital, it might be interesting to know the differences in throughput times between smokers and non-smokers (Weijters, et al., 2006). This research has outcome information (DMFT) associated with the 'Initial Exam' events. This

facilitates observation of how the outcome measure evolves with time and with the sequence of treatments received.

12. Enhancing existing model with performance information from the EL. This requires an *a priori* model and extends it with information about times between events, event durations etc. improving the performance of the existing model.

2.6.4 Deciding which PM Algorithm and Process Mining Software to use

From the outset it was difficult to decide on the appropriate PM algorithm. Information on which PM discovery algorithm to use varies from the statement that the Heuristic Miner algorithm is especially suited in a real-life setting by De Weerd et al. (2012) to Fluxicon's PM tips recommending Heuristics miner, Fuzzy miner, and Multi Phase miner (Fluxicon, 2017). Detailed assessments of PM discovery algorithms by De Weerd et al. (2012) and Wang (2013) proposing a framework for efficient selection of PM algorithms using 48 model characteristics provide further guidance. Weber et al. (2012) produced a framework for the analysis of PM algorithms viewing the mining algorithms as learning the distributions of processes over traces. Relevant to healthcare research, Rojas et al. (2016) found that the most commonly used PM algorithms in healthcare were Fuzzy Miner, Heuristic Miner, and trace clustering.

The approach to this decision considered the above reviews and, in addition, used empirical testing of the available algorithms using this research's specific research data to help identify appropriate techniques. Multiple tests were executed with Disco (Version 2.2.1) and ProM (versions 6.6, Revision 28643 and 6.7 Revision, 35885) using sample event logs to evaluate the Alpha, Fuzzy, Heuristic, and Inductive miners.

The priority in selecting the PM algorithm and technology for this research was that the models must be recognisable and comprehensible to dental experts and that they must demonstrate to them the potential for actionable insights to be generated by the technologies. While acknowledging the importance of formal model quality metrics, it was not the intention of this research to formally analyse process models. This encouraged consideration of informal models produced by the Fuzzy Miner and the Heuristic algorithm with the Fuzzy Miner ultimately being preferred due to its more comprehensible and recognisable results.

The choice of which software technology to use came down to a direct choice between ProM and Disco. ProM, which is an open source solution, has become the *de facto* standard for PM in research and is widely used in the PM research literature as seen in the literature review. ProM offers a wide variety of PM techniques and algorithms and

has the advantage of being an open framework environment allowing the development of plug-ins by researchers. Disco is a commercial product spin-out from Technische Universiteit Eindhoven, the home of ProM and, certainly in its initial phases, shared common personnel with ProM. In the interim, an alternate commercial product, Celonis, has significantly increased its presence in the market and currently boasts ‘the godfather of process mining’, Professor Wil van der Aalst, as a board member and chief scientific advisor. However, Celonis was not assessed for suitability in this research although future research would certainly include it in the technology assessment phase.

When comparing use of the Fuzzy Miner with Disco or ProM, Disco had distinct advantages. Disco utilises a single PM algorithm, the Fuzzy Miner, and accordingly delivers a cleaner, less cluttered, and ultimately more efficient user interface. Its user interface only must facilitate use of this single algorithm and its specific input parameters. Disco supported, as input, the .csv format required for this research. Its smart log import could automatically assign Case, Activity and Timestamp to the imported columns. This was a significant time-saver over similar exercises in ProM. Furthermore, it could combine multiple columns to create Activities. Also, its graphical outputs must deal with the characteristics of the Fuzzy Miner outputs alone. It also supported .png output of the process models. The quality of these was superior to that from ProM. It allowed easy generation and switching between two views of the models, frequency and performance. The frequency view shows both the case frequency and the absolute frequency of occurrence of individual events. It also explicitly enumerates the frequency with which sequence-pairs occurred in a dataset. The produced process model also shows this information visually by thickening and darkening the arcs between the events in accordance with their frequency of use and similarly darkening the most frequently visited events. When using the performance view of the process model, the time between events is directly shown and the user can choose between total time between events, mean time between events or maximum time between events. Its main process-model user-interface screen has two user-controlled functions facilitating simplification of the default process models.

These are sliders controlling the percentage of activities and paths visible in the generated model. Both control the percentages of the total paths and activities present in the event log that are included in the final model allowing rapid simplification of the model if required. This can help mitigate the spaghetti effect as detailed above. When the EL is initially imported, Disco assesses the size and complexity of the EL and selects a value for both the percentage of activities and the percentage of paths to be displayed. It is not

documented what the algorithm's criteria are for these settings, but it would appear to be guided by efforts to create comprehensible models within the constraints of viewing on a computer monitor. The user can then adjust these percentages up and down if required. In some of the ELs in this research, the default paths-percentage was <1% and this often resulted in valuable information such as the split between 'Prevention' and 'Restorative' after 'Initial Exam 1' not being evident in the models. Setting the paths-percentage value to 1 or 2% resolved this, but it highlights the need to be vigilant with the PM technologies. It was also necessary when using the product to record the application settings to ensure reproducible experiments. The product also has extensive filtering functionality although this was not availed of in this research due to the preference to enhance reproducibility by doing all filtering in the data transforms.

This is not to say that much of this functionality is not also available in ProM, rather, once the decision to use the Fuzzy miner was made, the accessibility of Disco, the tailoring of its interfaces to the Fuzzy Miner, its ease of use, and the high quality of the output graphics made the decision to use it for this research easy.

3 Research Aims, Objectives, and Research Questions

3.1 Aims

The literature review has shown that while healthcare generally has been subject to much process-oriented data analysis, dentistry has been lacking equivalent attention and the dental community was largely unaware of the potential of PM. No literature referred to the application of PM to dental public health datasets. This research addresses that gap by applying data mining, process mining, and data visualisation to an extract from a public health school dental service database. It aims to provide new insights into the variable care pathways leading to different outcomes. Additionally, it investigates the feasibility of using PM for conformance and compliance purposes by comparing the discovered process models with established care pathways and clinical guidelines.

Much of the published literature on PM has lacked the necessary methodological rigour. Very few have a strong published method and do not incorporate the complexities of handling EHR data in a research environment. This work takes a rigorous approach to data handling and data provenance. This work aims to present this methodology in a structure that provides useful guidance to future applications of PM to oral health data.

This research is primarily about developing new approaches to engineering data and presenting it in a novel way for the benefit of policy planning and decision making by healthcare staff. Although the work develops and illustrates a new approach to analysing large electronic health databases and shows how these data can be presented, the drawing of conclusions about the data itself is beyond the scope of this thesis.

In summary, this research aims to apply PM to an oral health dataset, illustrating the value of the data in the dental repository, and demonstrating how it can be presented in a useful and actionable manner to address public health questions. A subsidiary aim is to document and present the rigorous methodology used in this research in a structure that provides useful guidance to future applications of PM to oral health datasets.

3.2 Objectives

Several steps are required to deliver the aims of this thesis and can be presented as the following objectives:

1. Review relevant dental and healthcare literature establishing the state-of-the-art.
2. Apply a rigorous and documented methodology in this research.

3. Evaluate existing PM methods relative to this research's aims
4. Extend existing PM methods as necessary to achieve the aims of this research.
5. Process mine the research data addressing interesting and relevant questions
6. Based on the experience of the research, identify desirable dental EHR data to facilitate effective dental PM.
7. Identify challenges encountered when applying PM to routine dentistry data and how these can be overcome.

Achieving these aims and objectives is achieved through the rigorous preparatory approach taken and the methodology applied to answering the following detailed research questions.

3.3 Research Questions

As identified in Section 1.2.6 and by Mans, et al. (2015, p. 22), the three types of PM are discovery, conformance checking and process enhancement. The following RQs address the first two of these. Process enhancement is beyond the scope of this research as the researcher did not have ongoing access to the operational dental service.

Research Question 1: *Can PM discover care pathways, from a dental EHR?*

Research Question 2: *Can PM help assess compliance of real-world processes with recommended care pathways and clinical guidelines?*

The discovered *de facto* process being recorded in the EHR are compared to the ideal or *de jure* process. This will assess the usefulness of PM by comparing discovered care pathways with recommended care pathways and clinical guidelines. The research examines the suitability of the research data (or similar data) for assessing compliance with care pathways set down in the Steele Report (NHS England, 2009), dental contract reform (NHS, 2012) and the Irish Oral Health Services Guideline Initiative (2010).

Research Question 3: *Can PM discover dental care pathways associated with a specific outcome – e.g. extraction under general anaesthetic?*

Research Question 4: *Is PM and PM4D capable of assessing the impact of policy changes on service delivery and oral health outcomes, from the dental EHR, using the following two examples?*

- Can analysis of the EHR be used to evaluate the impact of ‘frequency of school screening’ policies on oral health outcomes?
- Can analysis of the EHR assess the impact of ‘age at first school screening’ on oral health outcomes?

Research Question 5: *What Data is Needed in an EHR for Effective PM?*

This question aims to identify data that would be needed in an EHR if applying the new PM approach to discover dental care pathways and facilitate the evaluation of policy implementation i.e. RQs 1 through 4. The objective of this exercise is to enhance the initial BridgesPM1 data model, which describes the research dataset, with additional desirable entities and attributes identified from existing standards in the literature. Experience gained in the course of this research should provide further information on additional desirable entities and attributes.

Research Question 6: *When applying PM to routine dentistry data, what challenges does one encounter and how can these be overcome?*

Identify the challenges encountered when applying PM to routine dentistry data. The challenges identified in previous PM healthcare research are briefly reviewed to provide a framework for those encountered in this research. Other challenges experienced in this research will also be considered and proposals for overcoming these challenges will be considered. In particular, data access and communication between researchers and data owners, data quality, and process model quality will be considered.

4 The Data and the Research Data Environment

4.1 Research Data Description

4.1.1 Introduction

The research data is an extract from Bridges, a single-centre relational database containing information relating to patients and their dental treatment in the Health Service Executive (South), Ireland. The database contains information resulting from the school screenings and subsequent treatment of children between 2000 and 2014 as mandated under the legislation (Government of Ireland, 1953; Government of Ireland, 2000). The core role of the School Dental Program is shown in Figure 4-1 below.

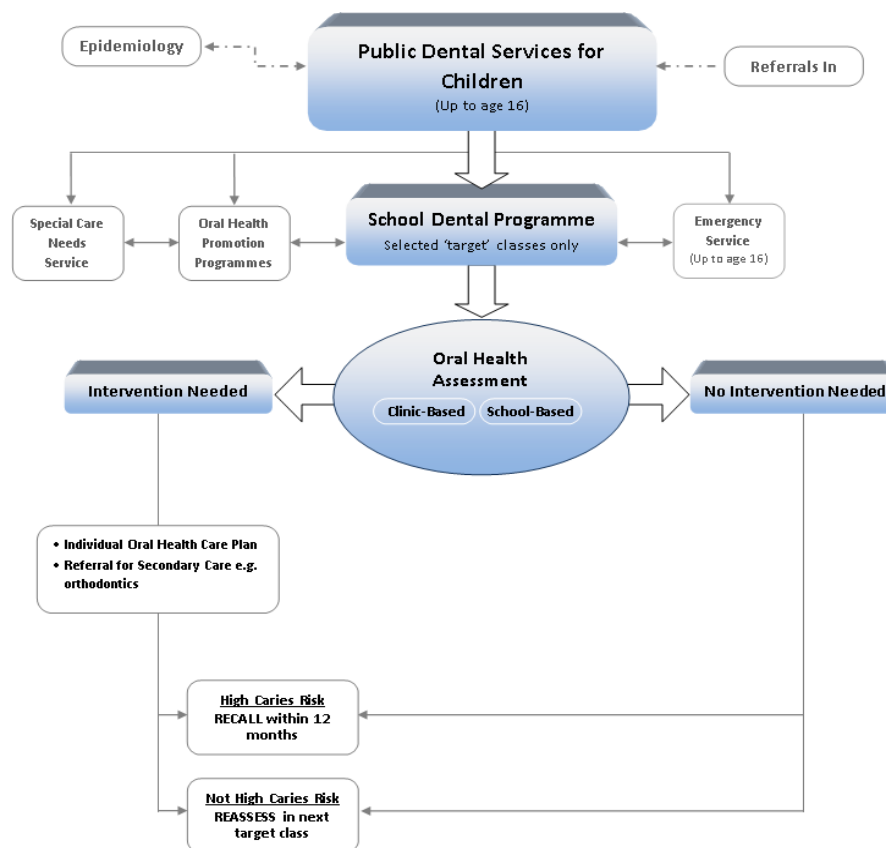


Figure 4-1: Organisation of Public Dental Services for children in Ireland. (Irish Oral Health Services Guideline Initiative, 2012, p. 12)

The database also contains information on emergency visits to the dental service and some data on special-needs adults. Data includes appointment history, attendance records, demographic data, medical history, clinical charting, notes, treatment plans and dental health status measures and is a by-product of the operational activity. The database facilitates access to patients' dental health status through DMF measures and KPI queries and supports research projects in these areas. Supported research includes the Irish Health Research Board funded Fluoride and Caring for Children's' Teeth (FACCT) (CARG/2012/34), a project which evaluates the impact of policy changes in 2002 and

2007 on children's oral health and Mapping the Divide (MTD) (HRA_HSR/2012/25), a project which analyses the distributions of oral healthcare services in relation to population and oral disease levels in children. The Bridges database is notable because the dataset spans 15 years of dental school screenings and resultant treatments and contains clinical, administrative, oral health outcome, and KPI data.

In this chapter, the EHR's use in dental clinics is outlined. The data acquisition process is described including the author's relationship to the data. The data extract itself, known as BridgesPM1, is described and profiled in some detail.

4.1.2 The Bridges EHR Application in General Use

The Bridges EHR application was used by clinical staff at chairside and also by administrative staff. Each patient's demographic information was registered using the Client Manager module. A medical questionnaire was completed and parental consent for examination and treatment obtained, the contents of which were sometimes transcribed into Bridges and sometimes scanned images were stored with the patient record. A sample capture screens is included for reference in Appendix 10.7. A dental charting was then usually carried out for each patient. This was recorded on a specialised screen representing the commonly used 'odontogram' format for recording details about individual teeth, their surfaces, diagnoses such as cavities, fissure sealant required and treatments such as fillings, extractions etc. A sample Bridges odontogram is shown in the top left panel of Figure 4-2 below demonstrating some typical charted conditions including amalgam fillings, missing teeth, cavities etc. Commonly used materials, conditions and surface combinations were available to users from the panels on the right-hand side of the figure. A list of proposed and completed treatments is displayed beneath the odontogram. These treatments could be automatically generated by the act of charting conditions or alternatively, could be manually entered by the users.

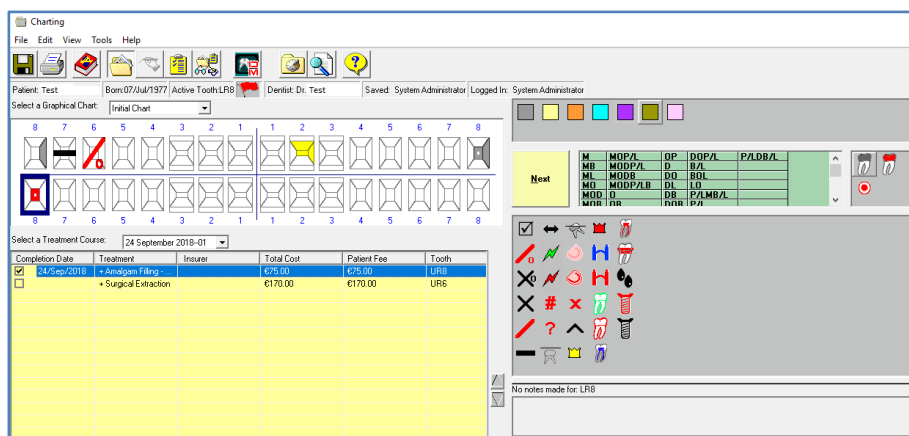


Figure 4-2: Sample Bridges Odontogram

This charting module generates the core of the patients' clinical records. Pre-existing conditions and diagnoses are recorded here along with planned and proposed treatments. The charting on the odontogram provides the basis for calculating the patients' oral health outcomes, DMFT/dmft, in this research. The treatments list, along with the patients' appointment history provide the basis for the treatment process models in this research.

4.1.3 The Author's Relationship to the Bridges Application and the Data

The author is uniquely positioned to provide insight in many areas of this research's data due to his intimate involvement with the Bridges application software and its implementation in the HSE. The author founded the software company responsible for the development and implementation of the EHR. He led the design, programming and implementation of the initial versions of the software and of the later enterprise versions in conjunction with senior clinicians in the HSE South. He is intimately familiar with the underlying data structures, the user interface and the protocols and training in the implementation of the software in over 250 clinics. He has executed multiple data extractions and anonymisation processes for research projects and academic theses over the last decade and provided the data extract instructions for this research's data extract, in compliance with the data controller's conditions for allowing him access the data.

Due to his relationship to the software, the author has a potential bias to protect the reputation of the software developers and other stakeholders. The author is alert to not allow this potential bias take effect and the potential bias risk is reduced by the fact the Bridges product is no longer actively marketed.

4.1.4 Data Acquisition Process

The University of Leeds School of Dentistry and Leeds Institute of Data Analytics (LIDA) and The Oral Health Research Services Centre in University College Cork (OHRSC) collaborated in the process of obtaining ethics clearance in Ireland and in securing permission from the data-owner and controller, the HSE, to extract, anonymise and process the data for research purposes. The matter was referred to the Office of the Data Protection Commissioner (DPC) in Ireland for an opinion and access to the data was ultimately granted by the HSE's Primary Care Research Committee (PCRC).

The dataset was extracted by a staff member of the Bridges and was anonymised by him. As the data had been acquired during routine school screenings and treatments, there were no requirements on dentists and ancillary staff in the data acquisition process. The pre-

processing and anonymization exercises were also cost neutral to the data controller. The extract, known as BridgesPM1, contains integrated, de-identified and comprehensive clinical and administrative dental data for persons under the care of the HSE South in the timeframe 2000-2014. The data is not currently openly accessible and is only released for this specific research. Efforts will be made to have the data openly available under a data-use agreement for the purposes of reproducing clinical studies and perhaps for further use in the areas such as academic research, similar to MIMIC (Johnson, et al., 2016).

Full ethical approval was granted by the Clinical Research Ethics Committee of the Cork University Teaching Hospitals (CREC) on August 2nd2016 (Reference: OHSRC00516). The research was approved by the Primary Care Research Committee (PCRC) at its meeting on 17th January 2017 with conditions that the researcher not be involved in the data anonymisation process and with the understanding that the PCRC protocol requires that the PCRC will have sight of any draft report prior to publication and that their opinion will be considered in relation to the publication. The requirement for individual consent was waived as all protected health information was de-identified. There was no requirement for additional ethical approval from the University of Leeds, School of Dentistry. The ethical approval documents are included in Appendix 10.4.

4.1.5 The Data Extract (BridgesPM1)

The original Bridges EHR application database contained 199 user defined tables. Many of these 199 tables relate to application logic, application settings, user settings, payments, waiting-lists, inventory, insurers etc. and were not useful for this research. A subset data extract was designed to fulfil the research requirements. Two main areas were central to the research. First, information regarding treatment events and appointments were required to allow creation of treatment process models and maps. Second, clinical and demographic information were required to profile the dataset, create cohorts, and calculate health outcomes. The code underlying the creation of the BridgesPM1 database is available on Code CD (5). Data anonymization is further detailed in Section 5.2. BridgesPM1 initially had little aggregated data, just raw events avoiding a shortcoming of traditional data warehouses, where events are aggregated into quantitative data, thus hampering process analysis (van der Aalst, 2016, p. 162). BridgesPM1 is now described.

4.1.5.1 Classes of Data

BridgesPM1 contains data associated with 231,760 distinct patients, primarily school-going children undergoing school screenings and special-needs adults with visits between

2000 and 2014. Further data available in the BridgesPM1 database includes time stamped events, treatment items, appointment details as well as dictionary-type data such as nationality, treating clinic and region. A summary list of data classes and details is found in Table 4-1, details in Appendix 10.3. BridgesPM1 is a relational database consisting of 20 tables. Tables are linked by identifiers having the suffix “ID” e.g. PMClient.PMClientID refers to a unique patient. The attribute prefix ‘PM’ indicates that this is the anonymised ID, not the original ID. PMClientID has been propagated to other tables and enforcing anonymity throughout the database.

The data tables closely represent those in the original database. No transformations beyond the anonymization process had been made at this stage. Primary identifiers such as name and address and other data such as free-text notes were not included in the BridgesPM1 dataset. Twelve of the twenty tables contain individual-level data and the remaining eight contain reference data. It would have been possible to merge the reference data tables into the individual-level tables. However, file sizes would have increased significantly, and performance could have been adversely affected.

Table 4-1: BridgesPM1 Data Classes

Class of Data	Description	No. of Rows	Size
PMClients	Client Demographics	231,760	37 Mb
PMTreatments	Treatment Event Description	3,169,864	1.44 Gb
PMTreatmentCourses	Treatment Course Identifiers	285,518	27 Mb
PMCharts	Chart ID and DMF measure	1,016,197	145 Mb
PMTooth	Tooth Description	32,219,452	3.7 Gb
PMToothPart	Tooth Part Description	16,649,791	4.2 Gb
PMCondition	Tooth Condition Description	32,291,681	8 Gb
PMAppointments	Appointment Time and Duration	1,760,923	376 Mb
PMAttendances	Attendance History	5,516,738	1.16 Gb
PMQuestionnaire	Medical Questionnaire Identifier	332,600	43 Mb
PMQuestionAnswers	Medical Questionnaire Answers	9,754,820	2 Gb
PMQuestions	Medical Questionnaire Questions	16,912	37 Mb
(D)PMToothType	Dictionary of Tooth Parts	52	5 kb
(D)PMToothPartType	Dictionary of Tooth Part Types	45	8 kb
(D)PMConditionType	Dictionary of Condition Types	56	8 kb
(D)PMNationality	Dictionary of Nationalities	25	2 kb
(D)PMClinic	Dictionary of Clinic Names	41	7 kb
(D)PMRegion	Dictionary of Region Names	5	1 kb
(D)PMAppointmentStatus	Dictionary of Appointment Status	9	1 kb
(D)PMAppointmentType	Dictionary of Appointment Types	11	2 kb

4.1.5.2 Entity Relationships between Data Classes

The BridgesPM1 database centres on patients and treatments (PMClients & PMTreatments) with individual clients (patients) having a one-to-many relationship to charts, treatment courses, medical questionnaires, and appointments. Patient charting, treatment course, and appointment information are stored in related tables, PMCharts, PMTreatmentCourses and PMAppointments. Additional dictionary and detail information is stored in further tables. An overview of the tables, fields, and relationships is provided in a summary entity relationship diagram (Chen, 1975)) in Figure 4-3 below.

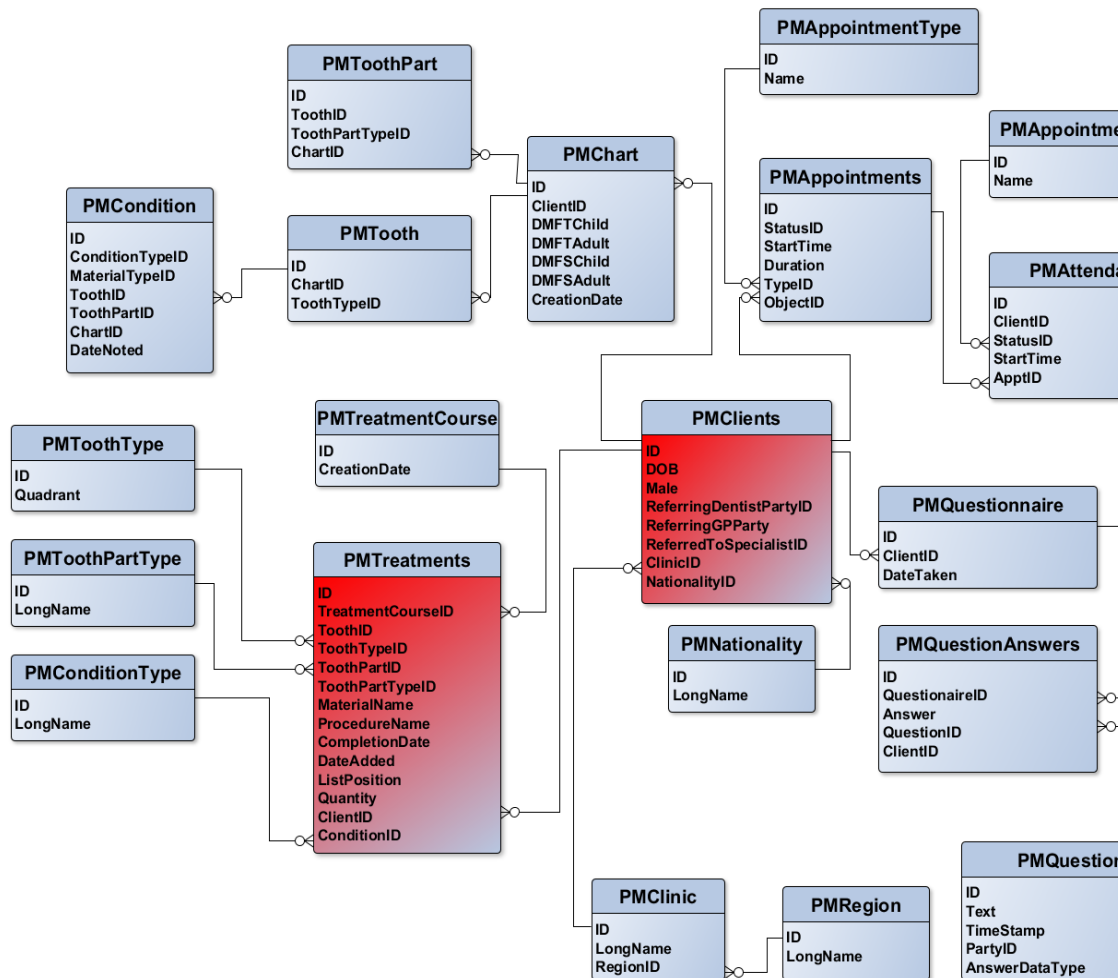


Figure 4-3: Entity Relationship Overview of the BridgesPM1 database

The author had intimate knowledge of the underlying Bridges EHR database structure and accordingly could guide the personnel creating the extract as to what data was required. Simultaneously, the author had knowledge of the data requirements of PM research and used the Healthcare Reference Model (Mans, et al., 2015, p. 27) as a guide to creating the data extract. Section 7.6.5 outlines how the BridgesPM1 extract compares to the PM data model standard, the Healthcare Reference Model.

4.1.5.3 Technical Validation of the extract

Best practice for scientific computing was followed wherever possible e.g. the FAIR principles of: findable, accessible, interoperable and reusable (Wilson, et al., 2014). Although quality issues with the data were tracked, no data extract iterations with improvements were possible due to resource limitations i.e. it was not feasible to request any improved data extract based on our new-found data quality information. BridgesPM1 was provided by the data-owner as a collection of comma-separated value (csv) files along with scripts for importing the data into SQL Server.

4.1.6 Profiling the BridgesPM1 Dataset

Weiskopf & Weng (2013) state that gaining an overview of the dataset is a valuable exercise although it is often overlooked by researchers. O'Neil & Schutt (2014, p. 29) in the data science domain, call this exploratory data analysis, or “*making plots and building intuition for our dataset*”. This is done by plotting histograms, summary statistics and scatterplots to get an intuitive feel for the data.

How does this relate to the research questions?

This research contains a strong element of exploratory data analysis (EDA) both in the data preparation and in the process modelling phase and involves building models from reality which involves many steps including building intuition for the data (O'Neil & Schutt, 2014, p. 29). EDA is a critical part of the data science process (O'Neil & Schutt, 2014, p. 34) and some of the reasons to do it are: gaining intuition about the data; making comparisons between distributions; checking scales and formats; identifying outliers and missing data and summarising the data. This research approach, as introduced by Tukey (1977), is akin to detective work where exploration and gaining knowledge of the data in as many ways as possible is encouraged with the overriding theme of avoiding confirmation bias i.e. the tendency to favour data supporting the predetermined hypothesis. EDA can also be characterised by a goal-oriented approach of detecting clusters and relationships. Yu (2017) proposes a goal-oriented taxonomy of EDA, finding clusters, screening variables out of many relationships and discovering patterns and relationships. According to Yu, with the advent of high-powered computing and large datasets, these methods have come to be known as data mining. O'Neil recommends the plotting of histograms and scatterplots to get a feel for the data and describes the basic tools as plots graphs and summary statistics and the method as systematically going through the data and plotting distributions of the variable and their relationships. What distinguishes the EDA approach from the classical approach to statistics is an emphasis

on graphical techniques to gain insight as opposed to the classical approach of quantitative tests. Quoting (O'Neil & Schutt, 2014, p. 37) *“It’s been a disservice to analysts and data scientists that EDA has not been enforced as a critical part of the process of working with the data.”* and this can be seen in the central role it plays in the data science process as represented in Figure 1-1.

When creating a data profile, some questions can be easily answered with simple query scripts, accessing only one or two tables, and may not require formal recording of the query code due to its simplicity. Some of these will include:

- What data exists?
- Listing the entities and attributes.
- Determining the size and scale of our data, Counts of each entity.
- Breaking down entities by key attributes.

More difficult questions require creation of additional aggregation, calculated values or outcome tables. These profiles may require temporary data-structures, cursors, and formalised procedures and functions such as those available in SQL Server and Python/Jupyter. Examples include:

- Generation of histograms and distributions of entities and attributes.
- Creation of cohorts with specific characteristics.
- Creation of health outcomes and other calculated values.

Most of the following profile figures were generated within Jupyter Notebooks (See Supplemental Material). More complex queries were formalised in procedures and functions within SQL Server and called from the Jupyter/Python Notebooks. This facilitated simpler, more systematic and structured approaches to query construction begging the question, why not do everything in Jupyter/Python? The author took a pragmatic approach to this aspect of the research and proceeded as far as possible in the SQL environment as he had prior advanced skills in SQL. As it became obvious that Python/Jupyter offered significant benefits in data manipulation, statistics, and visualisation, more complex work was then completed in the Jupyter environment, before finally folding some of the prior work into the Python environment. In the following sections, profiles of varying complexity are presented.

4.1.6.1 PMClients

Much of the patients’ demographic details was omitted from the data extract to reduce reidentification risk. All direct identifiers such as names, addresses, identification numbers, photos, scanned letters, and free-text notes were removed from the client

records when the data was extracted from the EHR. Date-of-birth is a key attribute as it defines many of our cohorts in the validating questions in Chapter 3.

Clients' dates of birth histogram

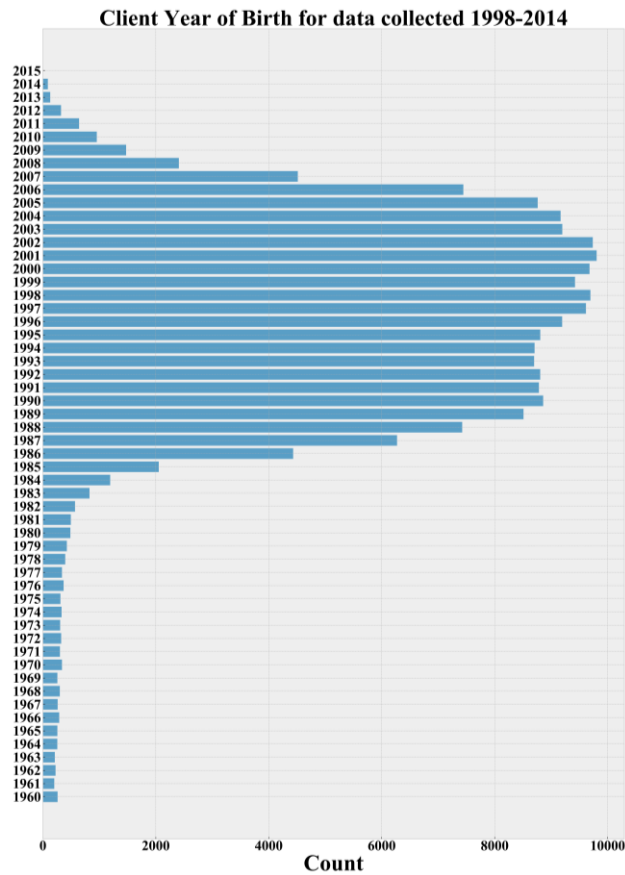


Figure 4-4: Clients' year of birth histogram for data collected 1998-2014

The Bridges dental EHR came into use initially in 1998. After a successful pilot it was expanded to all clinics in the HSE-South in 2000 and was fully operational by 2001. The service is aimed at primary-school-going children and also caters for a small number of adults and special-needs clients, explaining the distribution tail back to 1960. Although there is variation in the age groups of children targeted by the public dental service, most of the focus is on children in 6th class. Children in 1st or 2nd class are also frequently targeted to facilitate preventive care for their newly erupted permanent teeth. This focus is reflected in the histogram which shows a relatively large number of children born in 1986 who would have been aged 12 in 1998 when the first pilot system commenced. Visualising the dataset in this manner contributes to the research questions by exposing the long tail of dates of birth, identifying the existence of adults in the dataset and thereby empowering the researcher to eliminate these individuals in all queries if appropriate.

Client Nationalities

Each client registered should have had their nationality noted. A simple query counting the occurrences of each nationality showed it not being consistently entered by the application users. Accordingly, it would be invalid to use ‘Nationality’ as a criterion in cohort selection. This graphic is restricted to nationalities with a count > 100. The issue has been logged in the DataQualityIssuesRegister. (See Appendix 10.17)

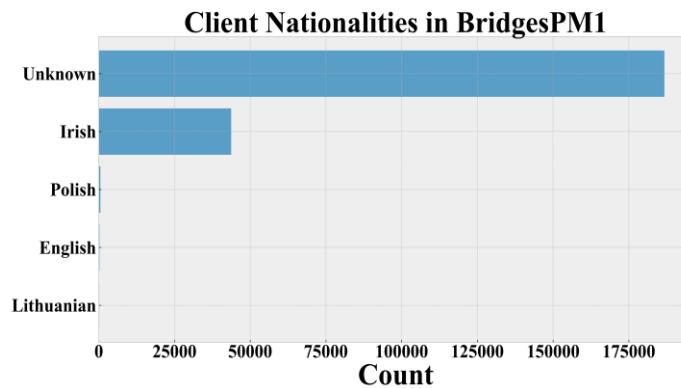


Figure 4-5: Clients' Nationalities

Visualising the dataset in this manner contributes to the RQs by exposing the data quality issue here and ensuring that nationality is not considered to be a factor in any of the RQs.

4.1.6.2 PMTreatments

The PMTreatments class along with PMClients, provide the core data for creation of PM Event Logs (ELs). PMTreatments contains the treatments (events) patient (case) received. The PMTreatments table contains 3,169,864 entries with 9,287 distinct procedure names. The large number of procedure names is due to editing of the core procedure name by the Bridges EHR application users permitted in the early years of the application’s usage. Of the 9,287 distinct procedure names, over 8,000 appeared only once, were highly specific, and often inappropriate as procedure names. They often contained more information than would normally be in a procedure names and should arguably have been in the patient notes. These variations were introduced by the application users and arguably, they should not have had the facility to change the procedure name in the user interface. This functionality was subsequently removed. This could ultimately have been prevented and accordingly, has been treated as a data quality issue (Issue 15). Using all of these would inevitably lead to the spaghetti models as found by Mans et al. (2008). As in their work, the mapping techniques used are a pre-processing transform on the event data with the objective of eliminating rarely occurring names. For example, seeking higher level events to represent lower level activities has a similar ultimate effect as clustering techniques as used in the fuzzy miner when it is necessary to simplify process models to make them comprehensible and useful to domain experts.

Only 363 procedure names had 10 or more instances and only 142 of the 9,287 distinct procedure names appear more than 100 times in the table. This research focuses exclusively on these 142. Within the 142, further simplification is possible, and the additional mappings are detailed in Appendix 10.2.

Treatment Counts

The procedure counts recorded in the EHR dataset is shown in Figure 4-6 below. This shows the most common item to be Fissure Sealants with just under 700,000 instances, over 300,000 screenings (Initial Exam) and over 200,000 amalgam fillings.

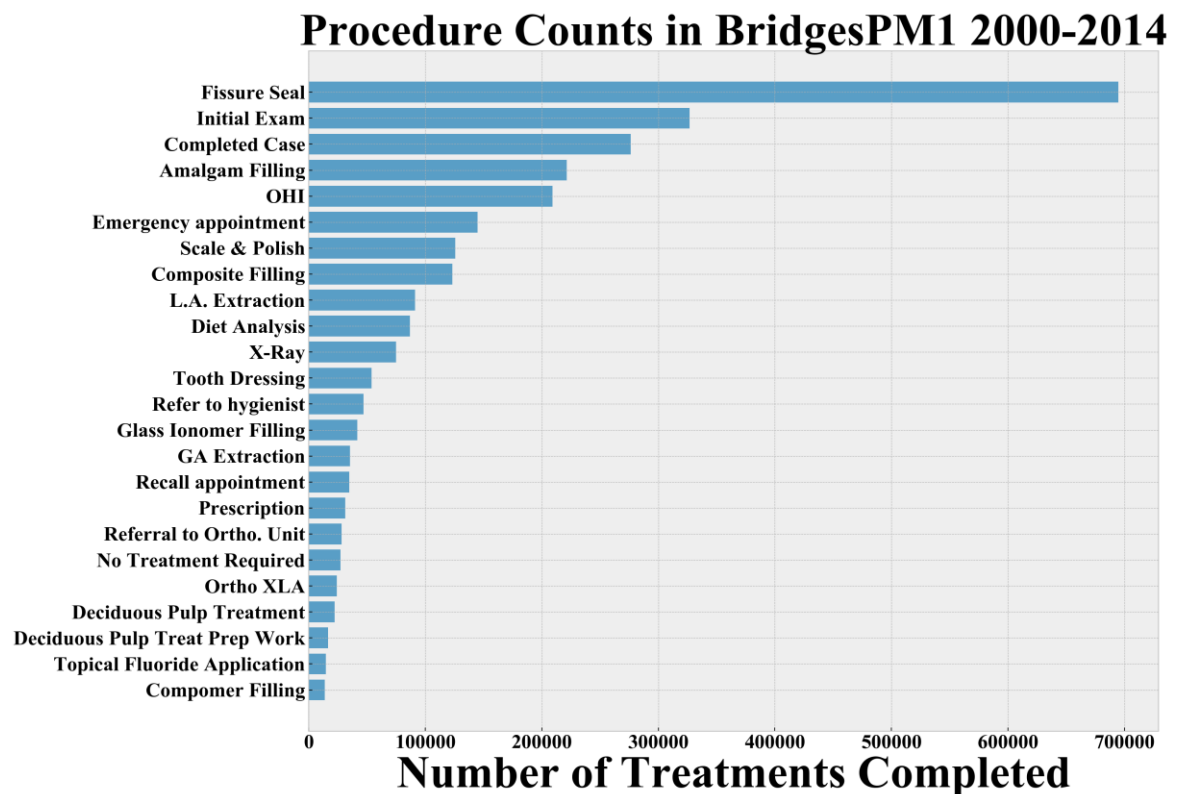


Figure 4-6: Procedure counts 1998-2014

Visualising the dataset in this manner contributes to the research questions giving an intuitive feel for the treatments delivered by the service. Data quality issues relating to this element arose later in the research. For example, the ‘Initial Exam’ procedure was sometimes inappropriately applied. More sophisticated visualisations at this point could have uncovered this and led to efficiencies further down the research pipeline. This visualisation also reveals the low level of topical fluoride application relative to the large number of fissure sealants.

DMFT Distribution

This histogram shows the distribution of DMFT values of those patients receiving their first oral examination in 2007. This shows just the DMFT values for the dataset although the dmft (for deciduous teeth) could also have been calculated from the data. This would be of interest for future research as past caries history is the best predictor of future caries risk and there is a relationship between the health of a primary dentition and a permanent dentition. If using dmft, calculating the dmft on the C, D, and E primary teeth would help eliminate the incorrect 'Missing' values due to the natural loss of deciduous teeth.

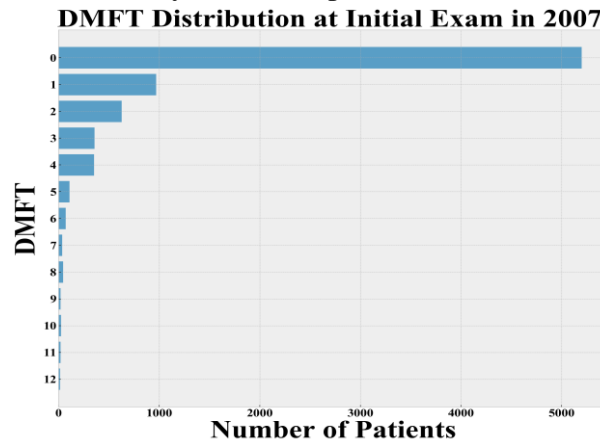


Figure 4-7: Distribution of DMFT values for patients examined in 2007

This distribution gives a good indication of how DMFT is skewed to zero and would suggest caution when using parametric statistical tests. While it is known from the next section that most children receive their first examination at ages 7,8, or 9, it would be nonetheless useful to have incorporated age into the visualisation.

Patient Age when receiving treatment

This histogram shows the distribution of treatment-counts over age and demonstrates the expected spikes at age 8/9 and 12/13 when the school screenings are often scheduled. It includes the full range of treatments shown in Figure 4-6 above and the less common procedures not included in that figure.

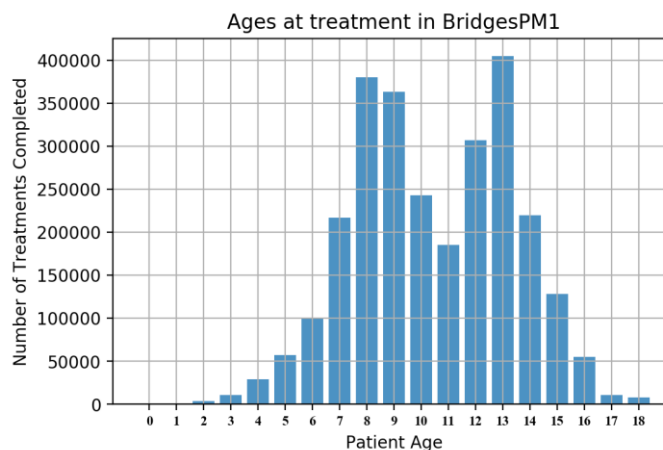


Figure 4-8: Ages when treatment received for data collected 2000-2015

This distribution contributes to research questions by confirming that the majority of treatments are generally carried out at the ages suggested by the published situation analysis (UCC/HRB, 2005/6).

Ages at first school-screening (Initial Exam)

One overall histogram and one histogram is shown for each region here. The period shown 2006-2015 was selected. Use of the EHR commenced in 2000 and many of those children showing up in the from the period 2000-2006 already had their first screening before the introduction of EHR and would possibly be incorrectly included in the count if the period before 2006 was used. The period 2006-2015 showed a notable decrease in the proportion of children receiving their first screening from the period 2000-2015 and is likely the more accurate. This contributes to the research question by showing why the EHR should be let run for a period before the data can be considered good data. The EHR must “ramp-up” and reach a steady-state to be capable of delivering good data (Kennebeck, et al., 2012; Ward, et al., 2014).

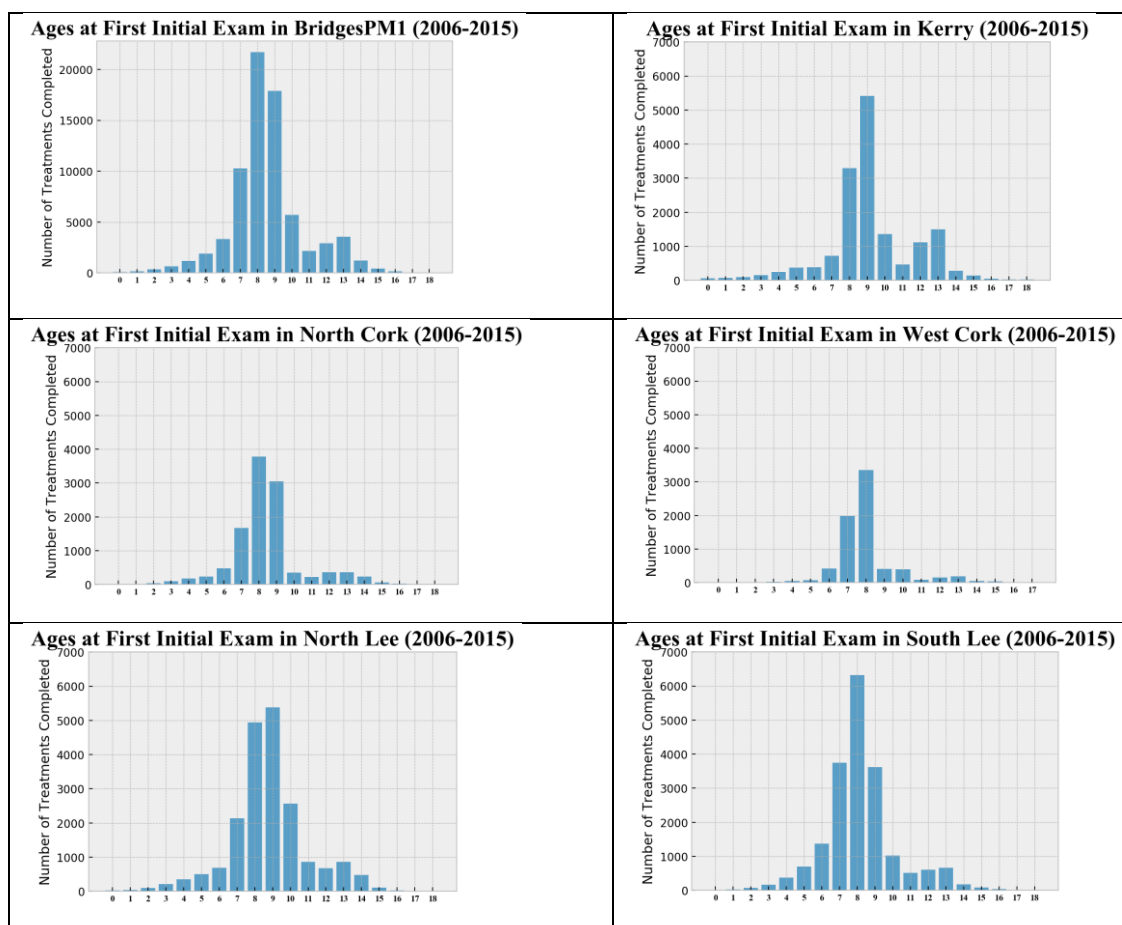


Figure 4-9: Ages at first School Screening Histograms

4.1.6.3 Geo-mapping of population DMFT

This basic geo-map in Figure 4.10 shows the locations of high DMFT values (>3) in the HSE South. The population density is represented in Figure 4-11 (Central Statistics Office (Ireland), 2012). This representation is of limited value and in its current form it is only presenting an overview of the main population centres. Anonymisation precluded the

inclusion of individual client addresses in the data extract therefore the location of each ‘dot’ is based on the GPS location of the clinic attended by the patient. Jittering was used to scatter the instances around each single clinic location and give an impression of the density of occurrences. However, with additional under-laid information, e.g. the population density, the fluoridation status or socio-economic status, such geo-maps can provide valuable visual overviews of the health status of regions and communities.

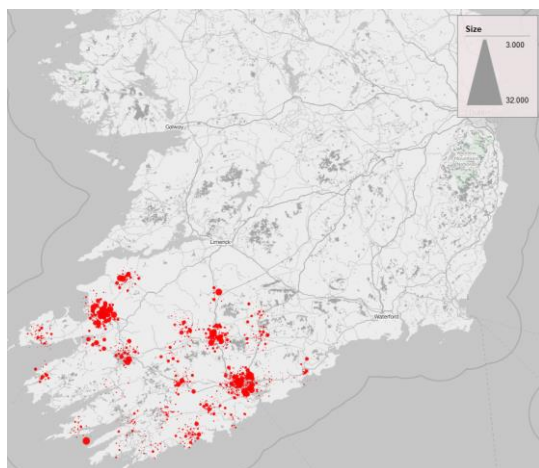


Figure 4-10: Geo-map of HSE South high DMFT values (>3).

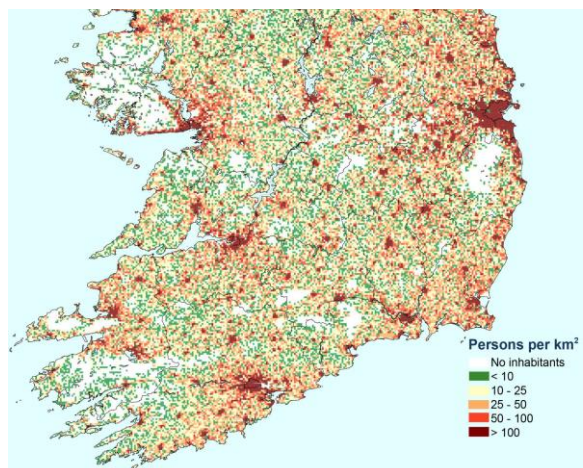


Figure 4-11: Population Density (Central Statistics Office (Ireland), 2012, p. 12)

4.1.6.4 Medical Questionnaires

Medical questionnaires are filled out routinely by patients, normally in advance of examination and treatment. In the early stages of the implementation of the Bridges application in the HSE, practitioners had the option of designing their own questionnaires, however, this was subsequently changed, and all questionnaires were standardised. There were 28 Yes/No type questions and a free text area on the questionnaire. The full list of questions is included in Appendix 10.8. Due to re-identification risks, the free text is not part of the data extract. As part of the data profiling and exploratory data analysis, the 10 most commonly positively answered questions were extracted and matched to the DMFT outcome if there was a charting on the same date that the questionnaire was taken. If there was no contemporaneous examination and charting the questionnaire was ignored for this profile. The summary data is presented in Table 4-2 below and charted in Figure 4-12.

Table 4-2: Recorded Medical Conditions and DMFT Distribution

	D₃MFT 0 (%)	D₃MFT 1 (%)	D₃MFT 2 (%)	D₃MFT 3 (%)	D₃MFT 4 (%)	D₃MFT 5 (%)	D₃MFT 5+ (%)
Population (N=130226)	59.14	13.53	9.66	5.96	5.09	2.33	4.28
No Conditions (N=55937)	55.74	13.97	10.42	6.62	5.70	2.72	4.83
Any Medical Condition (N=63617)	58.44	13.90	9.76	6.03	5.09	2.33	4.46
Asthma etc (N=17552)	53.98	15.06	11.03	6.75	5.46	2.75	4.99
Hay Fever etc (N=14952)	54.98	15.03	10.82	6.65	5.58	2.51	4.42
Pills/Drugs (N=10797)	53.48	15.10	10.95	6.89	5.75	2.62	5.21
Cold Sores (N=10591)	45.77	15.96	11.89	8.06	7.12	3.41	7.79
Under Treatment (N=6968)	53.34	14.67	11.17	6.79	5.84	2.99	5.21
Any Illness (N=6593)	53.09	14.82	11.04	6.98	6.25	2.75	5.08
Allergies etc (N=7603)	54.25	14.93	10.89	6.27	5.79	2.78	5.09
Heart Murmur (N=4037)	54.82	14.74	11.17	6.29	5.67	2.65	4.66

This is summary raw data and has not been adjusted for any confounding factors such as age etc. It shows that those patients with no medical conditions marked in the medical questionnaire also had the highest percentage of DMFT=0 values at 55.74 percent. This group also had the lowest DMFT>5 at 4.83%. Patients registering ‘Cold Sores’ had the lowest DMFT=0 values at 45.77%, with all of the others in the low/mid 50’s.

Age is, without doubt, a confounding factor here and a potentially powerful next step would be to complete an age-specific analysis and these methods could illustrate the need for a greater emphasis on prevention for children with medical issues. The probability of diagnosis of a medical condition and the chances of having caries both increase with time meaning that the older children get, the more likely they are to report at least one diagnosed medical condition and also to have caries. For example, both DMFT and exposure to the virus causing cold sores cannot decrease over time and accordingly, DMFT should increase and colds sores become more prevalent with increasing age. These profiles could easily be developed to address age and other confounding factors. However, detailed analysis is not the intention of the data profiling and exploratory data phase, rather to gain familiarity with the general data properties. Exploratory analysis such as this helps give an intuitive feel for the data and can sometimes reveal unexpected hypothesis generating results appropriate for more detailed future analysis.

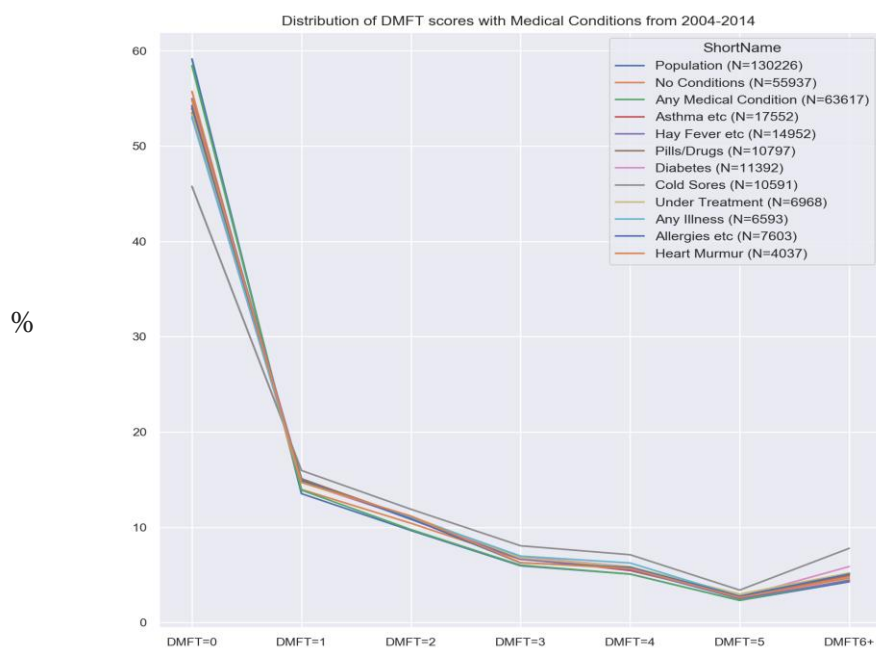


Figure 4-12: Recorded Medical Conditions and DMFT Distribution

It should be noted that the ‘Diabetes’ values in this figure are not completely relevant to the patient because the number included anyone in the patient’s family with diabetes.

4.1.6.5 Visualisation of DMFT Profile

These stacked charts give an overview of DMFT for the complete dataset from 2000 to 2014, overall and by region. The calculation is done individually for each year. Every charting completed in each year has its DMFT calculated and the first charting for each patient is used in the calculation. Hence, the same patient could appear in the data for multiple years, but only once for any given year.

This calculation shows a steady decrease in DMFT values since approximately 2005. This is confirmed by the regional breakdown. The higher DMFT values in the Kerry region are thought to be due to lack of fluoridation in many of the water supplies (Whelton, et al., 2017). The oral health measure DMFT, is particularly suited to presentation in a stacked-chart format as this allows presentation of the three components (Decayed, Missing, Filled) with different colours. It is helpful to consider the ‘Age at first screening/examination’ in Figure 4-9 when comparing the level of DMFT among regions in Figure 4-13. Because DMFT is a cumulative score, the modal age distribution at first screening/examination will be a contributing factor to the mean DMFT calculated for the region. For example, in Kerry the distribution is bimodal with a modal age at first screening of 9 years with a second mode at 13 years, suggesting less exposure to early clinic based preventive care and a median age of 9 years. The mean DMFT for Kerry is considerably higher than that for other areas. The other area with a modal age of 9 years

at first examination us North Lee, however there is less polarisation towards age 9 and, unlike North Cork, the proportion of children screened at age 9 is not substantially greater than the proportion first screened at age 8. Furthermore, North Lee incorporates a large urban cohesively fluoridated region whereas the Kerry region covers a wide area which includes many rural areas without fluoridation. Such confounding issues (age and fluoridation status) must be considered when interpreting the variation in the distribution of mean DMFT among the areas.

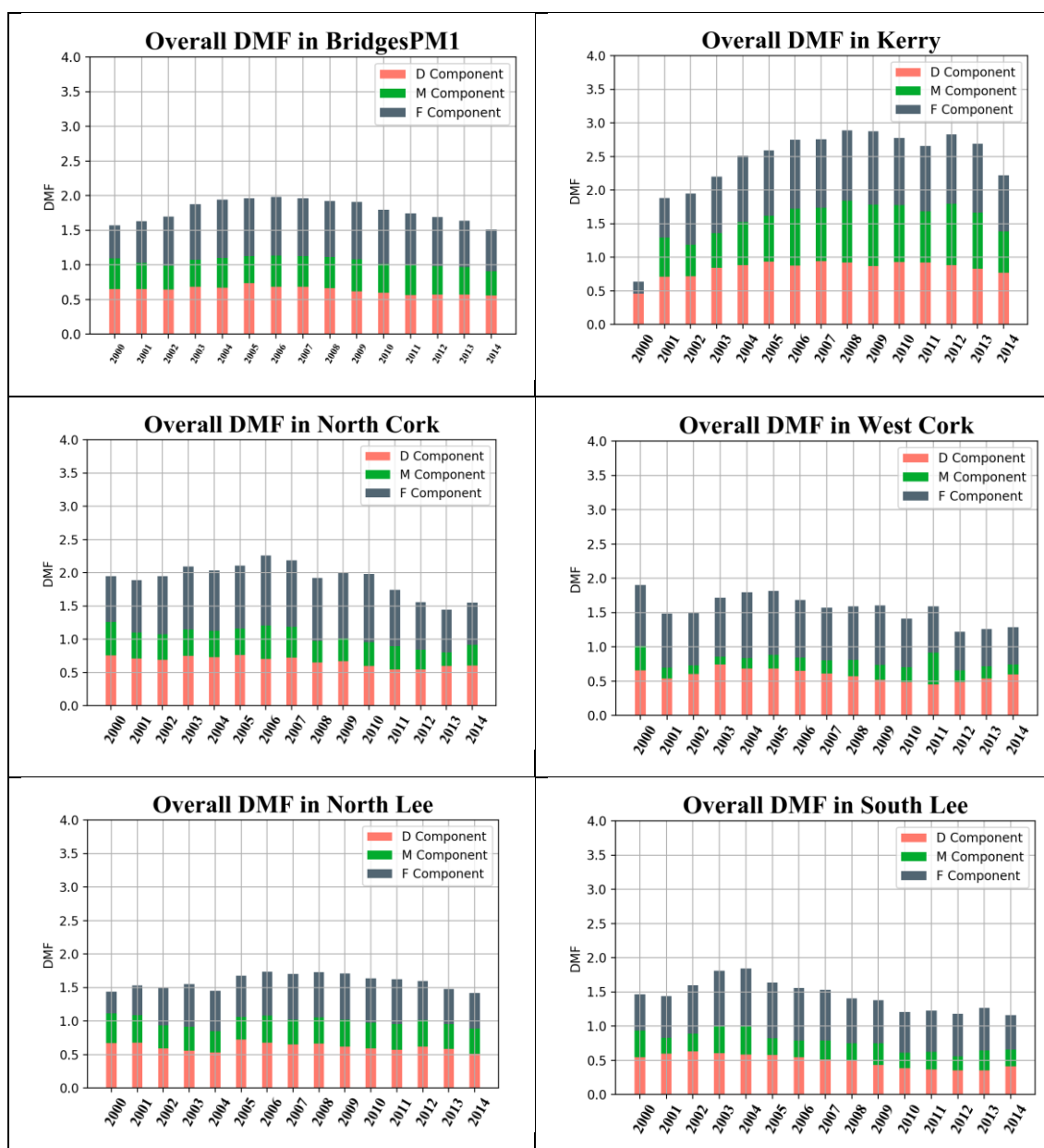


Figure 4-13: Overall DMFT Values (2000-2014)

4.1.6.6 Visualisation of DMFT Profile by Tooth

The charts in Figure 4-14 show the contribution of each tooth to the overall DMFT for two sample years, 2005 and 2014. This shows, as expected, that DMFT is concentrated on the teeth numbers 6 & 7 and shows a decrease in DMFT values from 2005 to 2014.

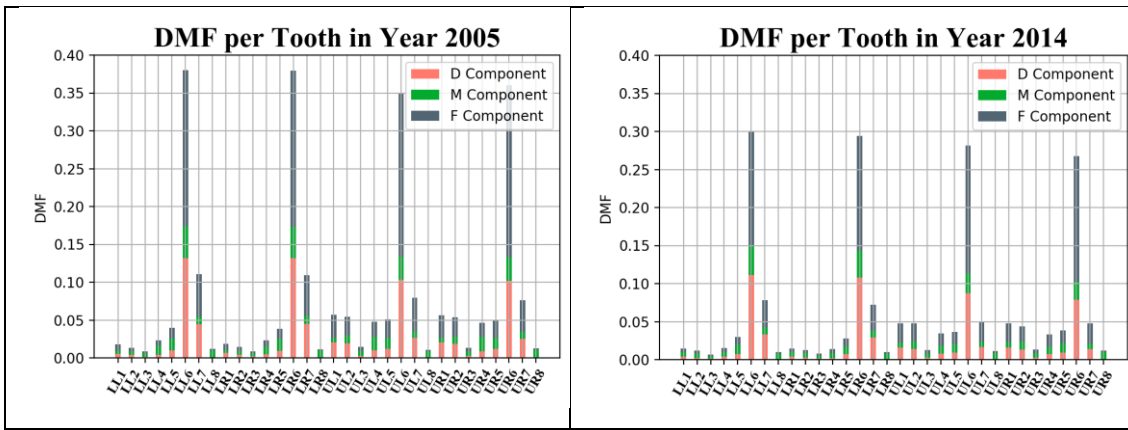


Figure 4-14: DMFT Values, by tooth number for two sample years 2005 & 2014

The following four charts in Figure 4-15 focus on the four most affected teeth, UR6, UL6, LR6, and LL6 and show the aggregated DMFT contribution of these teeth from 2000 to 2015. The DMFT value shows a steady decrease from approximately 2005; assuming the absence of demographic shifts, changes in the water fluoridation or changes in target groups for examination, these data are likely to give a reliable indicator of trends. However, these assumptions should be tested and further analysis to control for confounding due to age, water fluoridation and socio-economic status (SES) would be needed before drawing definitive conclusions.

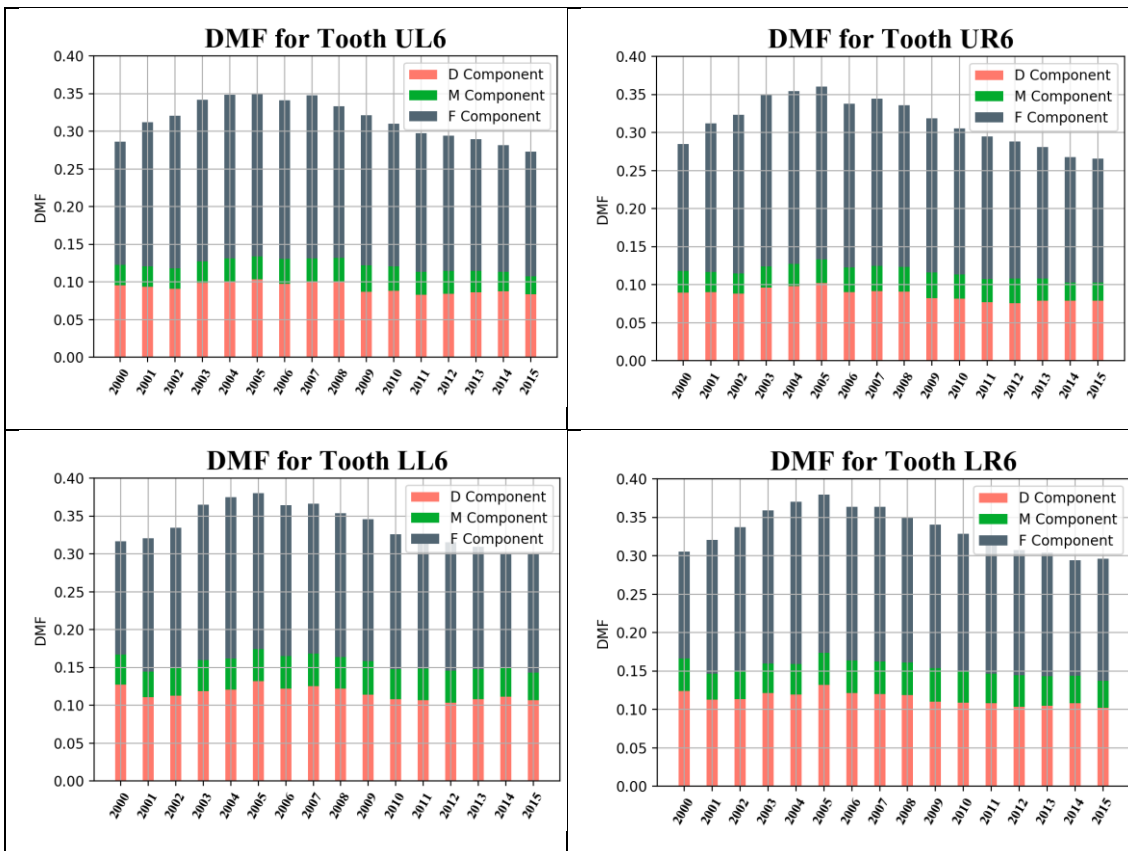


Figure 4-15: DMFT Values for the 6s, for the years 2000-2015

4.1.6.7 Heat map visualisation of DMFT by region

A heat map is a two-dimensional representation of data in which values are represented by colours. In this example, low DMFT values are displayed in blue and high values displayed in red with the legend displayed to the right of the heat map. It provides a quick overview of DMFT ‘hotspots’ but, as with geo-maps, would benefit from lower granularity of locations and incorporation of information on other relevant variables such as fluoridation status, socio-economic status, age, population size etc.

Each area has two rows in this heat map. The first row includes all patients with all DMFT values at the baseline or initial screening examination and the second row shows only those with DMFT=0 at the baseline or initial screening examination, both groups having their first examination in 2007. The first column contains the number of patients targeted for initial school screenings in schoolyear 2007. The cohort details are as follows:

- Screening (Initial Exam) carried out between September 1st of target year (2007) and 31st August of 2008. This was the first screening for that patient.
- The patient was aged 7, 8 or 9 at the time of the screening.
- The data quality was acceptable.
- The first row having initial DMFT = 0, the second having all initial DMFT values.

For example, in the first row, first column, the DMFT is 0.9 and n=1090. This means that 1090 patients complied with the criteria above in Kerry and all DMFT values, averaging at 0.9. In the second row, first column, the DMFT is 0.01 and n=664. This means that 664 patients complied with the criteria above in Kerry and each had a DMFT value of practically 0. Each subsequent column then represents the members of the first column seen in each of the following 5 years. For example, in the first row, second column (Year2DMFT), the DMFT is 0.88 and n=68. This means that, of the original 1090 seen in the first column, 68 were seen in the following year. Continuing through the columns, each n value represents the number of the original 1090 who were seen in the subsequent years and their average DMFT. It shows how these groups’ average DMFT values developed over the next 5 years and highlights the trajectory for those who were free of dentine caries at the first examination.

It should be pointed out that children are not seen annually systematically and may be seen in any combination of school years with first class (age approximately 7) and sixth class (approximately 12 years) being the most common combinations and some regions seeing children on one other occasion in the intervening period (See Table 7-1 for policy details). The heat map follows children seen for the first time in 2007. The data need to be interpreted with caution as children at high risk of developing decay or children being

monitored for orthodontic reasons may be recalled more frequently for more intensive care which may confound the data. However, such policies are likely to be duplicated in each region which increases the validity of comparison among regions.

The mean DMFT figures reported at the different time points in this heat map have higher values than those reported in epidemiological studies. For example, the mean DMFT at age 12 in 2014, from the FACCT study is 0.8 in fluoridated areas in Cork and Kerry and 1.4 in non-fluoridated areas of Cork and Kerry (Whelton, et al., 2017). In comparison, the DMFT figures seem high in year 6 (2012). There are several potential explanations for this. All the children examined in years 2-6 are a subset of those examined in Year 1. In general, high caries risk children are recalled on an annual basis, thus high risk or high caries children are over-represented in the heat map. Furthermore, data collected in a clinical setting include more disease than those collected in an epidemiological setting because of better lighting, examination conditions, ability to dry the teeth and the use of radiographs. The epidemiological examination seeks to record a stage of caries progression (e.g. clearly into dentine) whereas the clinical examination seeks to capture the full extent of disease i.e. all stages of caries progression. Therefore, the DMFT might be expected to be higher from a clinical examination than from an epidemiological survey.

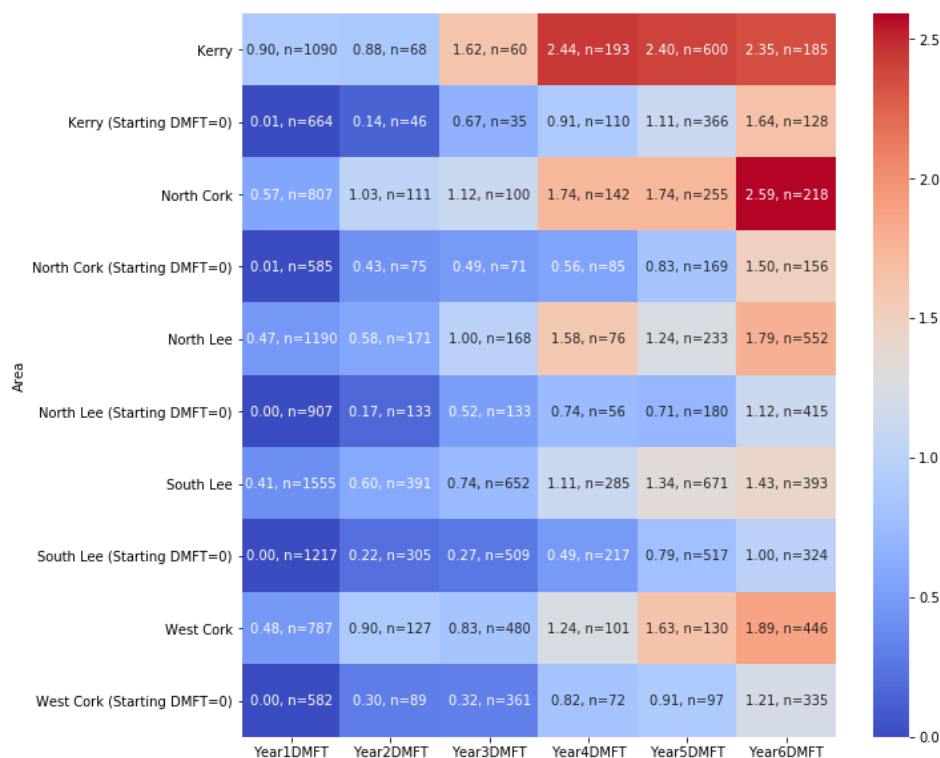


Figure 4-16: DMFT Heat map, All Areas, All DMFT values, starting 2007

These profiles of DMFT over time and by region are independent of the research questions but contribute to the research by illustrating the differences by geographical

area possibly reflecting the importance of fluoridation status on oral health outcomes. This and the other intuitions and sanity-checks gained through these types of exploratory data analysis provided the research with a valuable feel for the data and confirms the importance placed on this phase by O’Neil & Schutt (2014), Tukey (1977) and others.

4.2 Data Pipeline Environment

A range of technologies are used in the data pipeline environment in this research.

The HSE/Bridges dental EHR application used a SQL Server 2008 database and the anonymisation process was executed with Transaction-SQL Scripts (T-SQL). The data was exported with SQL Server Integration Services (SSIS) scripts. The data transfer took place by physical, personal delivery by the author on flash-drives using Bitlocker encryption. Reloading at destination was executed with SSIS scripts to SQL Server 2017. DQ assessment and transforms were executed with T-SQL Scripts, as were data extractions, cohort creation and event log creation. SSIS export scripts were used to generate the event log CSV files. Data profiling was carried out with T-SQL Scripts and Python within the Jupyter Notebooks Module and Anaconda Integrated Development Environment. Data mining for the primary purposes of data-profiling was carried out with T-SQL and Python using specialised packages such as Pandas, Matplotlib, SciPy, NumPy, Seaborn, Cufflinks, Orange3, with the Jupyter Notebooks module. PM was primarily executed with Disco using CSV files and XES Event logs. ProM was initially used to evaluate some of the algorithms e.g. Alpha, Fuzzy, Heuristic, Inductive miner etc. Statistical analysis was carried out using Python.

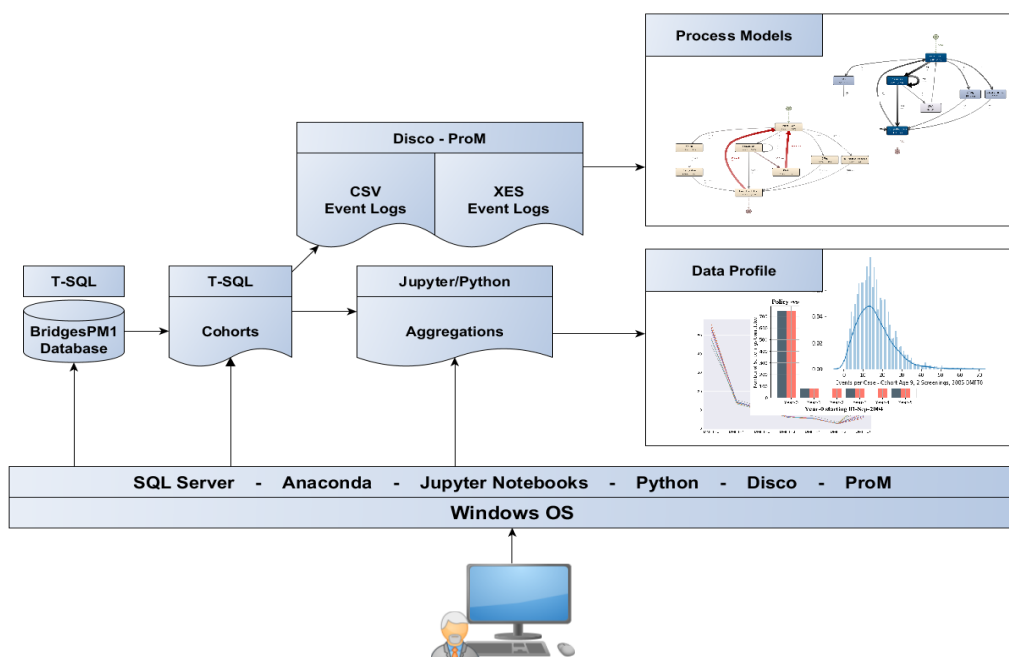


Figure 4-17: Data Pipeline Environment

There were two principal data-streams. The first prepared and profiled the data and generated the outcomes to the validating experiments. The second process-mined the data. The streams had common elements in data extracting, loading and DQ handling. Once the cohorts for the experiments were created, the PM and outcome generating streams diverged. The pipelines are represented in Figure 4-17.

4.3 System Environment & Architecture

Referring to the PM environment in Figure 1-6, our research environment can be represented as in Figure 4-18 below. This summarises the position of PM in an organisation. Applying this model to our research, the “world” is the Health Service Executive using Bridges as its Dental EHR software system. Bridges records details of the patients’ dental attendances and treatments which is extracted into attendance and treatment ELs. The HSE/Bridges produces event data as a by-product of operations. PM techniques are then applied to this event data using either discovery of processes, or compliance of the organisations activities with an established standard. This research does not employ process model enhancement techniques.

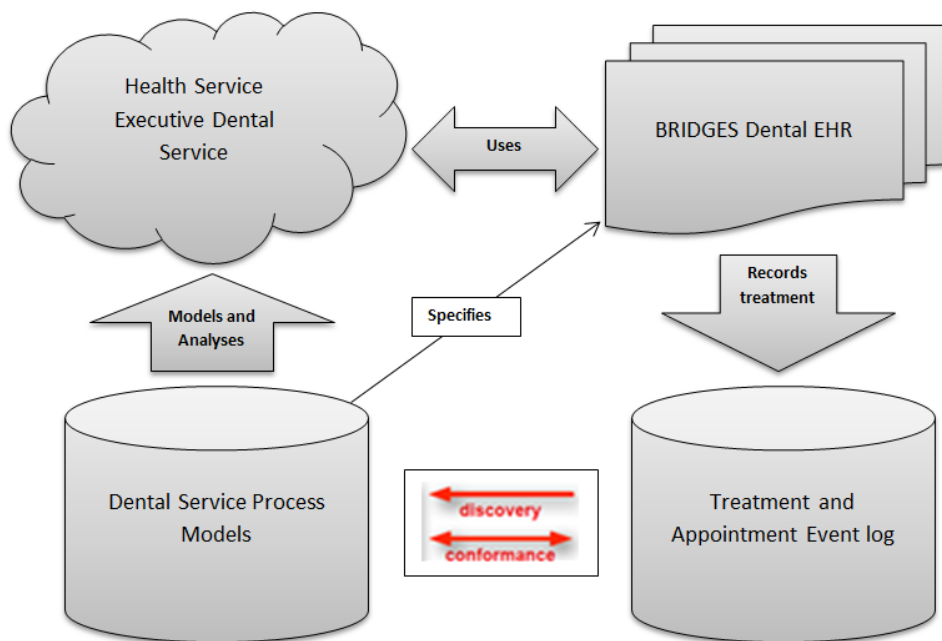


Figure 4-18: Process Mining Environment (Adapted from Mans, et al. (2015, p. 22))

The architecture in use in this research is represented in Figure 4-19. The initial database component shows the extract from the live database to the research environment (BridgesPM1). The ontological component i.e. SNOMED/SNODENT exists independently of this research. Domain expertise was used to confirm ontological mappings and for general advice. One of the most complex components of the research

architecture is the pre-processing component. The scripted interface to the data was specific to this research but could be adapted for other research work as described in the Code Reuse Guide (Appendix 10.11).

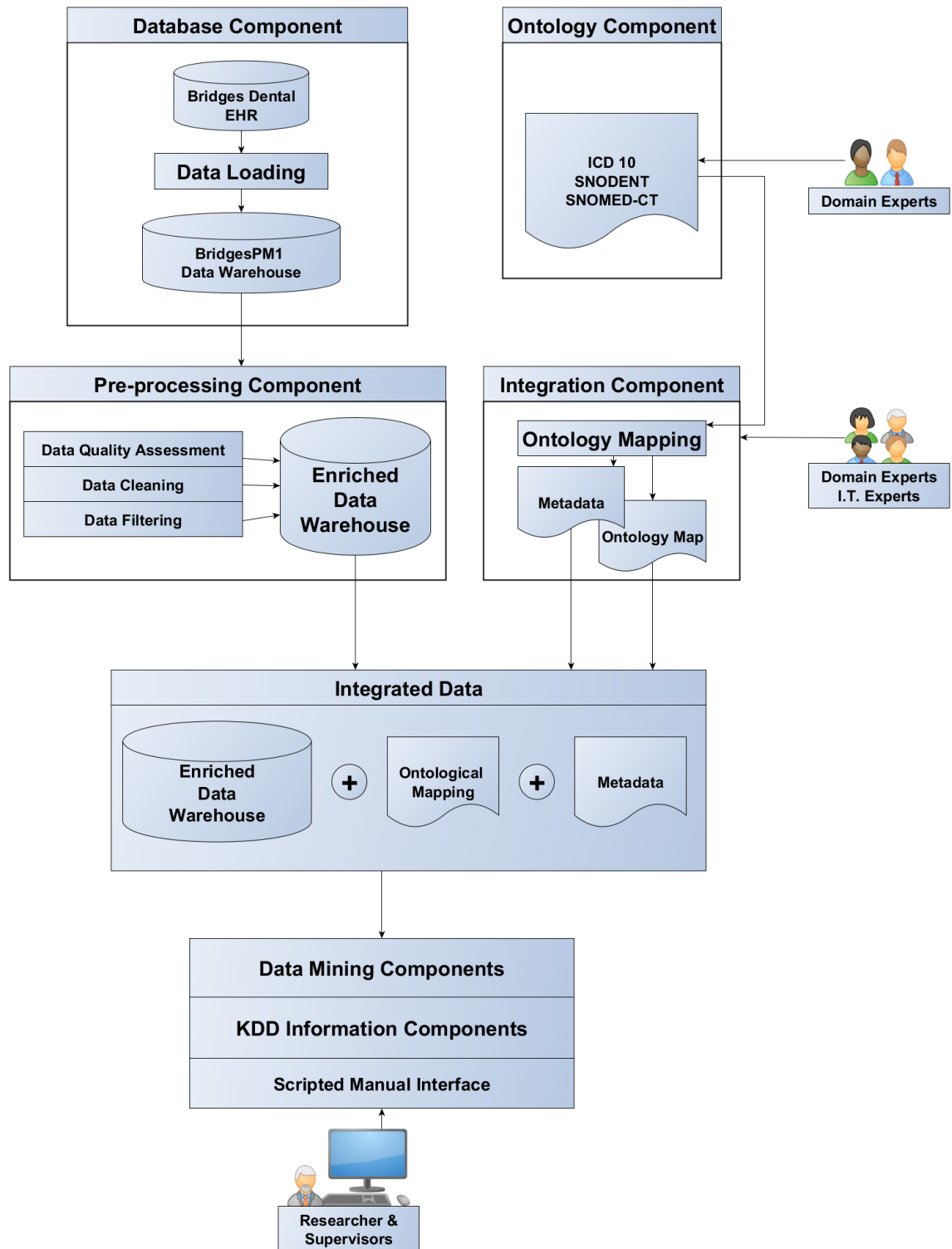


Figure 4-19: System Architecture representation adapted from Santos, et al. (2013, p. 275)

5 Challenges When Applying PM to Routine Dentistry Data

5.1 Introduction

The Process Mining Manifesto (IEEE, 2011, p. 10) enumerates some of the challenges facing the technology. Finding and merging data, sometimes from multiple organisations, is identified as one issue. Data may be distributed over multiple sources and may need to be merged to form the complete picture of the process. These data may be incomplete, suffer from noise, differing levels of granularity, context-specific variations, and other quality issues. Event logs vary significantly in complexity. Their overall size can be problematic, as well as internal characteristics such as the number of cases, the number of unique cases, the number of events per case and the fact that what exists in the event log is not necessarily all that exists in the real world. A process does not always remain constant over the course of a PM analysis and this problem is known as concept drift, although this is a confusing title. The lack of representative benchmarks for comparing the many PM techniques and algorithms remains a challenge and problems balancing the quality criteria of fitness, precision, simplicity and generalisation reflects this lack of practical guidelines for PM applications. In the interim, a comprehensive bench marking framework (CoBeFra) to carry out conformance checking has been developed by vanden Broucke et al. (2013). Additionally, a model for comparing PM techniques has been developed by Weber et al. (2012). The manifesto also identifies the positioning of PM in the world of operations research as an important challenge. As an emerging technology, it will benefit from combining itself with other modelling technologies such as simulation, lean value stream mapping, and data visualisation. This issue to has been addressed to some degree in the interim (Schrijvers, et al., 2012; van der Aalst, 2016, p. 46). This integration with other technologies will help address some of the other challenges such as improving comprehensibility and usability for non-experts.

Rojas et al. (2015) completed an overview of the main approaches using PM in healthcare and introduced the main challenges encountered in previous work. These challenges included data access, data quality, integration and pre-processing as well as the incorporation of medical knowledge in the algorithms.

Rehse & Fettke (2018) state that evaluation of PM results must be complete, relevant, sound, and reproducible to a degree producing scientifically substantial results and suggest that the validity, reliability, and credibility of published results in this area are potentially threatened by incomplete evaluations. They detail six categories of ‘process-mining crimes’: using the wrong evaluation data, misleading quality assessment, scientific inaccuracies, incomplete evaluations, improper comparison of evaluation

results, and missing information. While undoubtedly of value, their assessment is primarily based on evaluation of PM techniques under the headings of fitness, precision, simplicity, generalisation, and computational efficiency. This restricts their analysis to techniques producing formal models such as petri-nets and is inapplicable to evaluation of this research's results.

Incomprehensible models as a result of the ad-hoc, flexible, and the dynamic nature of healthcare processes is a central problem for PM in healthcare. Many approaches to this issue have been taken. Decomposing PM ELs into collections of smaller ELs, each containing fewer activities was proposed by Verbeek & van der Aalst (2015) to address their assertion that PM algorithms scale badly with increasing numbers of activities. Event abstraction using supervised learning techniques was proposed by Tax et al. (2016) leading to more comprehensible process models. Clustering cases having similar properties was addressed by Mans et al., (2008) who also used the abilities of the Fuzzy Miner to reduce the complexity of the discovered process models as does van der Aalst (2016, p. 417). Rovani et al. (2015) propose a declarative approach to PM acknowledging healthcare's complex, unpredictable processes requiring flexibility in its delivery and linking the 'spaghetti effect' to the explicit representation of all possible paths in a highly complex, dynamic environment.

The experience of this research matches the above in many respects. Data access, data quality and process model quality, in particular spaghetti-type modes, were the three key challenges encountered. We will look at each of these in turn and detail how this research sought to overcome them.

5.2 Data Access

A key challenge to this research was securing access to the research dataset. Understandably, given the richness of the dataset, the data owner had concerns about its release and required the researcher to provide assurances about the security and anonymity of the dataset. Ultimately, the research proposal was referred to the Data Protection Commissioner (DPC) in Ireland (<https://dataprotection.ie/>) for an opinion. The DPC deemed the research to be exempt from the legislation and the data-owner subsequently granted access to the data. Notwithstanding this, and considering that anonymity of data is not cut-and-dried, the Anonymisation Decision Framework (Elliot, et al., 2016) as published by the UK Anonymisation Network, was applied to this research with the aim to demonstrate that a robust data governance procedure was followed when managing the data in this research.

5.2.1 The Anonymisation Decision Framework Method

The ADF was developed to address a need for a practical guide to anonymisation that gives operational advice, while being less technical than the statistics and computer science literature. The ADF requires the user to understand how a data privacy breach might occur, understand the consequences of a breach and to reduce the risk to a negligible level i.e. to a level that a reasonable man would ignore. It is intended for those needing to anonymise their data with confidence, usually in order to share it. It consists of an assessment and management of reidentification risk. Following its steps should include reference to all the components of the ADF: the data, other external data sources, legitimate data use and potential misuse, governance practices, and legal, ethical and ongoing responsibilities (Preface to (Elliot, et al., 2016)).

The ADF consists of ten components incorporating three different activities: understanding the data situation (Points 1-5), disclosure risk assessment and control (6&7) and impact management (8-10). A summary of each step follows.

1. **Describe your data situation.** This describes the relationship between the data and the environment. This relationship maybe static or dynamic i.e. the data may stay in one environment or may move between differing environments.
2. **Understand your legal responsibilities.** This requires the researcher to understand their role in the data environment and their responsibilities in each environment.
3. **Know your data.** Identify the data's properties. Who are the subjects? What are the data types? Does the data include personal identifiers?
4. **Understand the use case.** Why is the data being released? Who will access the data? How will those accessing the data use it?
5. **Meet your ethical obligations.** Identify your obligations and implement good governance structures to achieve and manage these.
6. **Identify the processes you will need to assess disclosure risk.** Should the data be released? How much disclosure control should be applied? What is the optimum means for releasing the data?
7. **Identify the disclosure control processes that are relevant to your data situation.** What are the processes available to change the data or change the data situation to reduce disclosure risk?
8. **Identify who your stakeholders are and plan how you will communicate.** Build trust with the stakeholders through good communication strategies.

9. **Plan what happens next once you have shared or released the data.** Plan for handling the data set in the light of technology advances and increasing risk due to the ever-increasing number of available datasets. Do not adopt a release and forget approach.
10. **Plan what you will do if things go wrong.** Put a robust audit trail and trained personnel in place to help manage a breach.

The framework also enumerates the five principles upon which the ADF is founded:

1. You cannot decide if data are safe to share/release or not by looking at the data alone.
2. But you still need to look at the data.
3. Anonymisation is a process to produce safe data but it only makes sense if what you are producing is safe useful data.
4. Zero risk is not a realistic possibility if you are to produce useful data.
5. The measures you put in place to manage risk should be proportional to the risk and its likely impact.

Before extraction of the data into the BridgesPM1 database, it was de-identified cognisant of the Anonymisation Decision making Framework (ADF) (Elliot, et al., 2016).

The ADF steps as applied in this research are addressed in turn and detailed in Section 10.21.

5.2.2 Discussion of the ADF

The ADF provided this researcher with a strong basis to appeal for access to the research data. It provided a strong, defensible structure with which to assess the data environment, assess the risk of a data breach and manage such a breach should one occur. The framework's strength was its end-to-end approach to this issue – from deciding whether or not to release the data through to communicating and managing a breach. This had the effect of making anonymisation and information governance an ongoing process. It also had the strength of being pitched at a level accessible to many of the stakeholders such as data-owners and researchers and assisted communication between these parties.

5.3 Data Quality Management in this Research

5.3.1 Introduction - Using EHR Data in Research

Electronic Health Record (EHR) systems are now well established in many countries and healthcare settings. The importance of the secondary use of EHR data for research is widely recognized. Reliable research demands data of good quality or, at least, data of a

known quality and without this, research results are impossible to evaluate. Robust data provenance and data of acceptable and known quality must become the norm.

The use of big data and secondary use of EHR data for healthcare research is gathering momentum and is supported by business (McKinsey Global Institute, 2011), health authorities and governments (The Parliamentary Office of Science and Technology, 2017; Wilson, et al., 2016; European Commission, 2014, pp. 5-9). Many benefits have been identified. Danciu et al. (2014) identify rapid cohort identification, quality of care assessment, comparative effectiveness research, data privacy and de/re-identification as some of the areas where access to clinical data can aid researchers. Syndromic surveillance, public health, research and quality improvement were identified by Anker et al. (2011) and Botsis et al. (2010). There is a growing body of literature that uses data derived from EHRs to inform health research. There is also a growing, but noticeably smaller, body of literature on the underlying data quality (DQ) problems inherent in using EHRs as a research data source. Frameworks such as those proposed by Weiskopf and Weng (2013) and Kahn et al. (2016) can address the huge scope for non-random human error across multiple dimensions and can be used to categorize the dimensions of EHR DQ helping identify strategies for mitigation. Their adoption in EHR research is urgent. Notwithstanding existing work in EHR DQ, questions remain on the appropriate use of routinely collected health service data for research purposes. Some suggest that data should only be used for the purposes for which it was collected (van der Lei as cited by Weiskopf and Weng (2013)). According to Schmier et al. (2005), clinical decisions often take priority over data collection and measures must be taken to validate any data collected in a clinical setting. Other limitations of data collected in a clinical setting being used for research include, representational bias, clinician-related biases regarding missing data and outcomes, non-standardisation of data entry, data redundancy, inaccuracy, restriction to retrospective study, and difficulties extracting data (Song, et al., 2013). Weiskopf and Weng (2013) suggest that there is no 'absolute' DQ measure, 'fitness for purpose' being the appropriate criterion i.e. the data must be of sufficient quality to answer the RQs being asked. The data are of sufficient quality when they serve the needs of a given user pursuing specific goals. However, understanding the clinical significance of the data and the way they are coded in the clinical setting is a major and necessary task (Danciu, et al., 2014). They also note that many Enterprise Data Warehouses are designed to support business intelligence goals and not for research. Researchers are often unaware of the complexity of clinical data systems and of the provenance of the data, hence, the

creation of the optimal dataset often requires several iterations between clinical users, software developers and database administrators.

Anker et al. (2011) identified root causes for some DQ issues in the secondary use of data created for project management of EHR implementation as:

- Differential incentives for the recording of data i.e. data tended to be more accurately recorded if needed for contractual or financial purposes.
- Flexibility in software systems allowing multiple ways of doing the same task.
- Variability in documentation practices between personnel.
- Variability in the use of standardised vocabulary and changes in procedures and electronic system configuration over time.

Botsis et al. (2010) also identified missing, inaccurate, and inconsistent data issues in their study of pancreatic cancer data. This was due to information fragmentation in the healthcare system and poor documentation of critical information. Inaccuracies were also caused by poor granularity of diagnosis terms or incorrect use of the terms. Inconsistencies arose due to different data sources in the EHR and inconsistent use between clinicians. They also proposed some solutions involving formal information exchange mechanisms, clinical registries and personal health records as well as the sharing of effective strategies for secondary use of healthcare data.

EHR data quality can also be viewed through the lens of compliance with and use of standards. SNODENT provides a useful reference for standard diagnostic and procedure nomenclatures. Standards for Electronic Dental Record System design in ANSI/ADA Standard No. 1067-2013 provides useful guidance.

There is an urgent need for PM to focus on techniques addressing DQ problems (Bose, et al., 2013). Secondary use of EHR data for research demands validated, systematic methods of EHR DQ assessment (Weiskopf & Weng, 2013). These authors encourage systematic logging techniques and the development of repair and analysis techniques to improve the quality of the ELs and consequently, improving the outputs of PM exercises. To further this, a method for enumerating and managing DQ issues in research using EHR data is proposed. Further, a method to pre-process the research data is proposed, both marking the data if its quality is compromised and mitigating the DQ issues if possible. PM in healthcare is especially challenging because care patterns vary widely between patients, health care professionals, and organizations and the reliance of the method on the completeness of time stamped ELs adds additional requirements for measurable DQ. As with other forms of data mining, systematic logging and repair techniques are important, as is the need for transparency around data cleaning and checking steps.

5.3.2 This Research's Data Quality

An initial assessment of the dataset revealed many potential DQ issues arising from differing sources – from the developers of the application, the users, the data extraction process, and potentially from the research itself. Previous work using similar data from the dental EHR highlighted some quality issues, for example, inconsistencies in recording fluoridation status, trauma status and gender. This research also benefitted from the author's intimate familiarity with the EHR, its design and its day-to-day usage, allowing a birds-eye overview of possible sources of DQ issues and their impact. The author co-designed the underlying data structures and much of the user interface as well as implementing the EHR application in the clinical setting. He defined the research dataset for extraction and executed the technical data transformations within the research. Accordingly, the author was ideally positioned to identify potential DQ issues arising through all the phases of the data's existence. Classifying and managing the numerous issues remained problematic as it became apparent that they arose from various sources e.g. application users; could affect the data at different levels e.g. row or field level; and were identified by various means. Further, the impact of a data issue was dependent on the RQ or experiment e.g. date-of-birth was essential for some queries and irrelevant for others. The author chose to examine these issues in a structured manner and to document and audit every change or transformation made to the data, whether such a transformation was to address a DQ issue or to enrich the data for analysis purposes.

5.3.3 This Research's Data Quality Framework

The complexity of the DQ issues was such that it necessitated a formal framework i.e. the care pathway data quality framework (CP-DQF) for managing and, if possible, mitigating these data issues. The framework facilitated the systematic identification, recording, managing and, in some cases, mitigating of the quality issues. It also facilitated reporting of the issues and their scale. A database of potential DQ issues was established, both from the author's own experiences with the application development and with the data itself and from the existing published literature on DQ and forms another output from this phase. This proved to be a valuable and productive undertaking and demonstrated that formal DQ assessment is an essential step in research using EHR data. The framework developed has the potential to be generalised to other research using EHR data and the author believes that the framework and the list of discovered DQ issues can assist other researchers to discover, manage and mitigate the DQ issues in their own work. It provides

a valuable, timesaving, pragmatic starting point for other researchers undertaking research using EHR data.

The details of the Data Quality Framework that addresses the specific needs of PM of Care Pathways (CP-DQF) based on the PM and EHR DQ literature are presented in Appendix 10.19.

5.3.4 Improving Data Quality

Weiskopf & Weng (2013) made several recommendations to improve DQ when using EHR data for research. They encourage the use of systematic methods to assess the quality of an EHR-derived dataset for a given research task. Our research is following such a systematic approach. They suggest the use of a consistent taxonomy for DQ assessment. This research seeks to build upon their suggested categories to reflect the specifics of this dental research. Integrating DQ work from other fields is another suggestion from their work. This research has included an extensive data profile presented in a Jupyter Python Notebook (See Supplemental Material) satisfying their suggestion to include distributions, summary statistics and histograms in publications. A complete log of the data cleaning and transforms is included allowing full replication of the research if required.

Twelve guidelines for logging with the aim to improve data quality were proposed by van der Aalst (2016). These are tabulated below along with this research's approach to each of the guidelines.

	12 Logging Guidelines	This research's approach
GL1	<i>Reference and attribute names should have clear semantics, i.e., they should have the same meaning for all people involved in creating and analysing event data</i>	This research employed mapping techniques on the raw data to achieve this. See Section 10.2 for details.
GL2	<i>There should be a structured and managed collection of reference and attribute names</i>	This research mapping treatments to SNOMED-CT concepts to achieve this. See Section 10.2 for details.
GL3	<i>References should be stable (e.g., identifiers should not be reused or rely on the context).</i>	All identifiers used are GUIDs (Globally Unique Identifiers)
GL4	<i>Attribute values should be as precise as possible. If the value does not have the desired precision, then this should be indicated explicitly (e.g., through a qualifier).</i>	This issue arose with the treatment attribute, CompletionDate where date without time was recorded. This was addressed as a data transform in Section 10.18.3
GL5	<i>Uncertainty with respect to the occurrence of the event or its references or attributes should be captured through appropriate qualifiers.</i>	The dental EHR users were motivated to accurately record treatment events and appointment events. No audit of the accuracy was possible in this study.

GL6	<i>Events should be at least partially ordered. The ordering of events may be stored explicitly (e.g., using a list) or implicitly through an attribute denoting the event's timestamp.</i>	This issue was present in our research data and was addressed as a data transform in Section 10.18.3.
GL7	<i>If possible, also store transactional information about the event (start, complete, abort, schedule, assign, suspend, resume, withdraw, etc.).</i>	The existence of treatment courses could be considered analogous to transactions but was not formally addressed in the EHR in the research extract.
GL8	<i>Perform regularly automated consistency and correctness checks to ensure the syntactical correctness of the event log.</i>	The event log for this research was a once-off extract and a full data quality analysis was executed. See Section 5.3.
GL9	<i>Ensure comparability of event logs over time and different groups of cases or process variants</i>	As the event data spans 15 years, this clearly required attention and is discussed as a data quality issue in Section 5.3..
GL10	<i>Do not aggregate events in the event log used as input for the analysis process.</i>	This was fully complied with in the research. All aggregations were done in the analysis phase
GL11	<i>Do not remove events and ensure provenance. Reproducibility is key for PM</i>	All data transforms were logged in Section 10.18
GL12	<i>Ensure privacy without losing meaningful correlations.</i>	This is a trade-off situation and much personal information about the clients was not included in the event data to reduce the risk of re-identification of individuals as detailed in Section 4.1.6.1

Table 5-1: Event Logging Guidelines, adapted from van der Aalst (2016, p. 152))

5.3.5 Data Transforms

After the assessment of the DQ, the data was pre-processed to facilitate answering the RQs. Changes made to the data are documented in this section. Weiskopf & Weng (2013) have pointed out that, like data-profiling, this step is often missing from research documentation and publications. Data transformations are necessary to streamline the data for analysis, to eliminate rarely occurring data or noise. Data transformations may also be necessary to make future queries comprehensible. Such transforms may also facilitate enhanced performance of queries. Complex queries with multiple joins or subqueries may take too long to execute or may not be practicable with the available computing resources. The transformations applied to the BridgesPM1 data are detailed in Appendix 10.18.

5.4 Process Model Quality

5.4.1 Spaghetti models

Who says they are a problem?

Incomprehensible models as a result of the ad-hoc, flexible, and dynamic nature of healthcare processes is a central challenge for PM in healthcare. Fernandez-Llataz, et al. (2015) suggest that it is the main problem facing PM technologies. The term commonly used to describe these models is ‘spaghetti’ models symbolising their unstructured nature. On the other hand, structured predictable models are often known as ‘lasagna’ models. Healthcare’s complex, unpredictable processes are acknowledged by Rovani et al. (2015) and they link the spaghetti-effect to case heterogeneity and to the explicit representation of all possible paths in a highly complex, dynamic environment. Even though the notion of a process exists in these environments, the actors deviate from it to accommodate the needs of the case in hand i.e. real-life business processes are not strictly enforced by their supporting information systems (citation Trace Clustering in Process Mining). It is known that many high-tech systems produce logs of very fine granularity lead to spaghetti-like process models. However, these spaghetti models also provide important insights about the process and often indicate that it is driven by the experience and intuition of service providers and often incorporate trial-and-error, rules-of-thumb and qualitative information (citation Process mining: discovering and improving Spaghetti and Lasagna processes) and while challenging for PM, can provide substantial benefits.

In healthcare, Mans et al. (2008) found that the heuristic miner produced such models when applied to hospital stroke healthcare and attributed this to disease and patient variants. The dental PM literature also encountered spaghetti-type process models.

What are they suggesting be done about it?

Several approaches to alleviating this issue have been taken. Decomposing PM ELs into collections of smaller ELs, each containing fewer activities, was proposed by Verbeek & van der Aalst (2015) in an attempt to address PM algorithms scaling badly with increasing numbers of activities. Event abstraction using supervised learning techniques was proposed by Tax et al. (2016) leading to more comprehensible process models. Clustering cases having similar properties was addressed by Mans et al., (2008) who also used the abilities of the Fuzzy Miner to reduce the complexity of the discovered process models as does van der Aalst (2016, p. 417). Trace clustering has been shown to be effective i.e. partition event logs into subsets of homogeneous cases (deLeoni, et al., 2016).

Higher levels of abstraction can be achieved using ontologies (Pedrinaci & Dominique, 2007) e.g. SNODENT. Mans et al used pre-processing techniques on the event data for example seeking higher level events to represent lower level activities. They also proposed use of simplification techniques such as clustering and the specialised search algorithms as approaches to simplify the models.

Fernandez-Llatas, et al. (2015) referred to a number of strategies for mitigating the spaghetti effect: PM algorithms that make simple, less dense models, the use of Activity Based PM incorporating the results of the activities into the maps, time abstractions grouping large numbers of similar transitions or arcs together, rendering algorithms providing additional visual cues to the important paths and events, filtering algorithms and clustering techniques, and improved navigations apps.

In the dental PM literature, to arrive at a comprehensible model, the dental researchers applied a strategy where only events that occurred in more than 10% of the process instances were included. This is a type of slice-and-dice filtering. Unfortunately, no discussion was held on the value of the discarded data. Perhaps the deviant processes are also interesting, and it is certainly worth consideration. There is no analysis as to what information was lost in this process, nor its value. It would be essential to assess the omitted information with the help of domain experts. In publication (1) with unfiltered data, the Heuristic miner produced a complex, spaghetti-like process model. In the methodology section, they describe a process of consolidating event names and the use of a new ProM plug-in to effect this. It is unclear whether the plug-in is exclusive to dentistry. They also speak about mapping event-names to 'subjects' though there is no additional information on these 'subjects'. It is unclear whether the research used any standard diagnostic or treatment codes such as ICD 9/10 or SNODENT in this phase.

How does this research manage spaghetti models?

This research took a number of steps to reduce this problem. Putting it simply, the process models must be presented in a way that they are useful to the users. The models must be legible if on paper i.e. the nodes and font sizes readable and the arcs distinct and distinguishable from each other. On screen, there is some additional flexibility as zoom features are often available. The limitations of presenting healthcare process models on paper or small screen formats are clear. In this research it proved difficult to interpret more than 30 different event types (nodes) with 60 connections (arcs) on an A3 sheet. Resolution and font-size limitations make the model details impossible to read in printed

formats irrespective of their substantive content. Viewing these models on screen, within Disco or as exported vector-graphic (.png) files resolves this matter to a degree in that the vector graphic file allows zooming in or magnification without pixilation or loss of definition of the image. However, for the purposes of a printed thesis, 30 nodes was an approximate but practical upper limit and all complete models presented in the thesis are guided by this limit. For the purposes of dealing with more involved models, the relevant excerpts of the complete model are presented.

Whether a process model is recognisable or comprehensible also has a more subjective dimension and is dependent on the person viewing the model. Is it a dentist, a process analyst, or a lay person? What previous exposure to process models have they had? How much domain expertise do they have? These are all relevant questions when assessing recognisability and comprehensibility. For the purposes of this research, recognisability and comprehensibility were assessed utilising a convenience sample of dentists and process miners. Discovered process models were viewed by supervisors, colleagues and presented at several local and international research conferences and meetings where they were subjected to scrutiny and debate. This feedback led to some additional adjustments in the presentation of the models, particularly in removing less important details and focussing on the core issues in the models.

In an ideal experiment, a representative sample of dentists would be presented with process models of various familiar scenarios and their ability to recognise and comprehend the models accurately and in a timely manner would be recorded to give a more scientific assessment of the models quality characteristics. This was not feasible in this research due to time and resource constraints but would merit consideration for future work. Several steps were taken to reduce models to this level.

Higher Level Abstraction of Events

This is a pre-processing step. In summary, one-off and rare events were removed from the event logs. As detailed in Appendix 10.2 and 10.18.1 only 142 of the 9,287 distinct procedure names (events) appeared more than 100 times in the events table of over 3 million events. This research focused exclusively on these 142. Within these 142, further simplification was possible by mapping similar events to a single event e.g. ‘1 Surface Amalgam Filling’ and ‘2 Surface Amalgam Filling’ were mapped to ‘Amalgam Filling’. Further abstraction to SNOMED terms was also carried out but did not prove useful at the practical PM level because the SNOMED terms were often very descriptive and lengthy e.g. the SNOMED Concept Name for ‘Amalgam Filling’ was ‘Insertion of

amalgam restoration into tooth (procedure)’. This was a constant issue with SNOMED and made the resulting models illegible. While the mapping to SNOMED was retained and is available in Appendix 10.2, its use in the PM phase was abandoned.

Seeking higher level events to represent lower level activities has a similar ultimate effect as clustering techniques as used in the fuzzy miner when it is necessary to simplify process models to make them comprehensible and useful to domain experts.

Using Disco

The choice of Disco as this research’s PM tool is detailed in Section 2.6.4 and is key to dealing with the issue of spaghetti models. Disco uses the Fuzzy Miner algorithm for process discovery and aims to balance the four quality criteria of fitness, precision, generalisation, and simplicity. Two user-controlled functions controlling the percentage of activities and paths visible in the generated model facilitate their simplification and help reduce the spaghetti effect. When the EL is initially imported, Disco assesses the size and complexity of the EL and selects a value for both the percentage of activities and the percentage of paths to be displayed. It is not documented what the algorithm’s criteria are for these settings, but it would appear to be guided by efforts to create initial comprehensible models within the constraints of viewing on a computer monitor. The user can then adjust these percentages up and down if required. The product also has extensive filtering functionality although this was not availed of in this research due to the preference to enhance reproducibility by doing all filtering in the data transforms.

5.4.2 How was ‘recognisable’ and ‘comprehensible’ assessed?

As stated above, the priority in selecting the PM algorithm and technology for this research was that the models must be recognisable and comprehensible to process mining and dental experts. The limitations of presenting healthcare process models on paper or small screen formats are clear. From a practical perspective for example, it is difficult to interpret more than 30 different event types (nodes) with 60 connections (arcs) on an A3 sheet. Resolution and font-size limitations make the model details impossible to read in printed formats irrespective of their substantive content. Viewing these models on screen, within Disco or as exported vector-graphic (.png) files resolves this matter to a degree in that the vector graphic file allows zooming in or magnification without pixilation or loss of definition of the image. However, for the purposes of a printed thesis, 30 nodes appears to be an approximate practical upper limit and all complete models presented in the thesis

are guided by this limit. For the purposes of dealing with more involved models, the relevant excerpts of the complete model are presented.

Whether a process model is recognisable or comprehensible also has a more subjective dimension and is dependent on the person viewing the model. Is it a dentist, a process analyst, or a lay person? What previous exposure to process models have they had? How much domain expertise do they have? These are all relevant questions when assessing recognisability and comprehensibility. For the purposes of this research, recognisability and comprehensibility were assessed utilising a convenience sample of dentists and process miners. Discovered process models were viewed by supervisors, colleagues and presented at several local and international research conferences and meetings where they were subjected to scrutiny and debate. This feedback led to some additional adjustments in the presentation of the models, particularly in removing less important details and focussing on the core issues in the models.

In an ideal experiment, a representative sample of dentists would be presented with process models of various familiar scenarios and their ability to recognise and comprehend the models accurately and in a timely manner would be recorded to give a more scientific assessment of the models quality characteristics. This was not feasible in this research due to time and resource constraints but would merit consideration for future work.

6 Methodology

6.1 Introduction

This chapter details the strategy and methodology used to achieve the aims of this research. First, it places the research in the broader research landscape with reference to research philosophies, approaches, strategies, methodology, time horizons, data collection techniques, and analysis procedures. It then looks at the research from the perspective of using EHR data for research and specifically, for trying to answer public health questions. It looks at existing PM methods and synthesizes these different perspectives into the methodology used to achieve the aims of this research. For convenience, this methodology is named PM4D (Process Mining for Dentistry).

6.2 How Process Mining in Dentistry Fits in the Research Landscape.

6.2.1 The Theory

The fundamental theory of this research is the belief that man plus machine is greater than man alone - that information technology is a useful addition to the workplace and can assist in many areas including clinical decision making. The theory assumes that Electronic Health Records (EHRs), with their strengths and weaknesses, are a useful source of information for research. The theory proposes that data extracted from EHRs can be used to evaluate public health policy and strategy decisions in the dentistry domain. This evaluation will be based on the oral health outcomes and the treatment processes experienced by the patients in the EHR and as such it is a retrospective cohort study. The theory assumes that the EHR can deliver valid oral health outcomes. The theory proposes that the EHR can deliver valid process models of the patient care pathways. The research can also be viewed as deductive theory testing in that someone else's theories are being used and operationalised by measuring the concepts from their theories and oral health outcomes measured from the EHR. The research will add knowledge to the field of dental informatics.

6.2.2 The Research Philosophy

'The research philosophy you adopt can be thought of as your assumptions about the way in which you view the world' (Saunders, et al., 2012, p. 128) and these assumptions define the research strategy and methods. This ultimately affects our understanding and interpretation of the research. There are two major ways of thinking about research philosophy, ontology and epistemology.

The research onion analogy by Saunders et al. (2012), as adapted by the University of Derby (2018) provides the canvas to position this research within the research methods landscape.

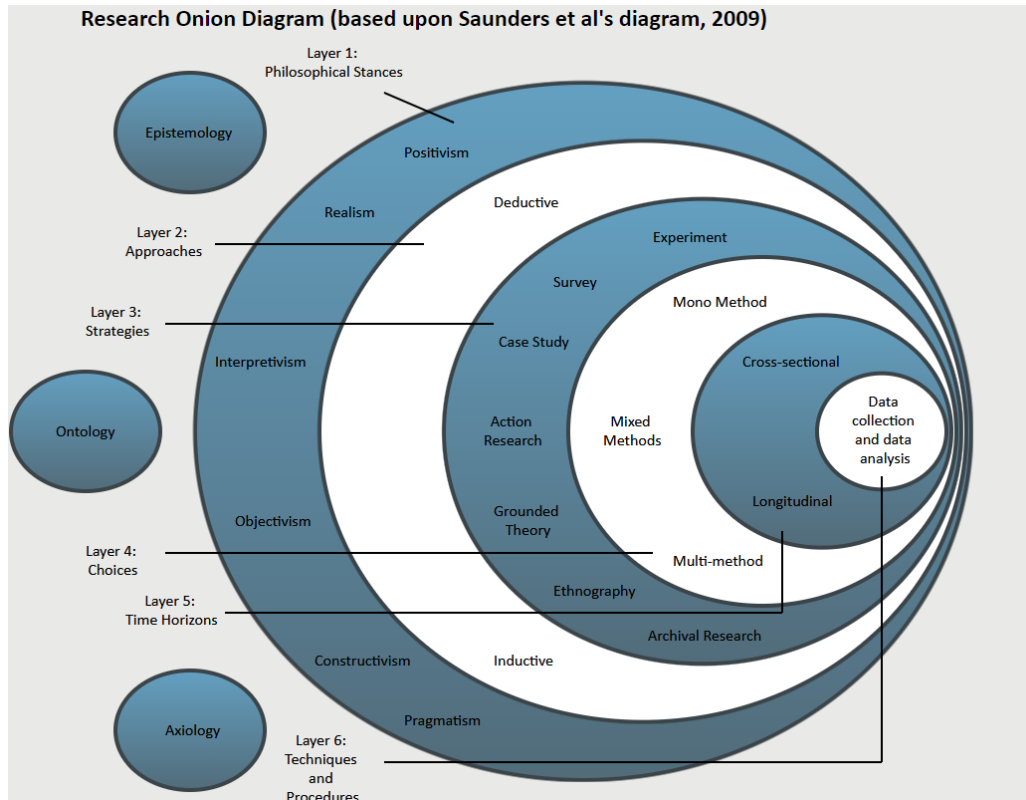


Figure 6-1: The Research Onion (University of Derby, 2018)

The author's ontological and epistemological views shaped his approach to this research. The outer layer of the research onion refers to the philosophical approach to research - to the ontological perspective and whether there is an objective reality. The author's view is that there is an objective observable reality and he will search for regularities and causal relationships in the acquired data. This positivist approach is qualified by suggesting that the dataset may have been influenced by the social actors in the system and accordingly, a philosophically realistic approach in data quality assessment and analysis may also be appropriate.

The author's epistemological stance describes how this research can come to knowledge given the ontology. The research includes a commitment to accurately record methods and findings i.e. how the results, findings and conclusions were arrived at. The research takes an attitude of scepticism to both the data and the data and PM methods to ensure that the results are defensible and uses the most credible sources of knowledge located. Authoritarian knowledge is used in the literature review and to establish the background for this research and efforts are made to spot the ontological, and epistemological

positions of the authors in the literature. Empirical knowledge from the EHR data abstract is used with the intention of creating new logical knowledge by applying experimental techniques, analysis, and reasoning to this data. The epistemological approach to data quality helps to define the relationship between the data and the actual existing phenomena in the real world/ Critical realism as applied to the data quality is beneficial. In general, this research leads the author to the positivist epistemological approach, albeit with elements of critical realism.

The axiology is to undertake the research in a value-free way. The author has committed to stay separate from the things being studied, to leave his personal beliefs behind and acknowledge, deal with, and control his biases as far as is possible. This stance helps remove bias from both the research reality and the authors conditioned reality. It will help clarify questions of the type, “Why is a hamburger called a hamburger, but a cheeseburger called a cheeseburger?”

6.2.3 The Research Approach

The next layer is whether a deductive or inductive approach should be taken. In using the EHR data to investigate specific questions relevant to dentistry, the author is using a largely positivist philosophy. There is an objective reality to be measured, and the outcome of an intervention can be predicted, a hypothesis established and tested. That is how the author approached gaining knowledge about the phenomenon being tested and this is a deductive approach.

When describing PM4D (Process Mining for Dentistry), a more inductive approach was taken. PM is a relatively new technology. Its application to the dentistry domain and public health is also new. This newness offered opportunities to address issues such as data quality, ontologies, EHR usage in research, and others from a new and fresh perspective and accordingly, it was not obvious from the start how the methodology would unfold and develop. As the author understands it, this is an inductive approach to methodology, and it is represented in the commonly used Figure 6-2 below.

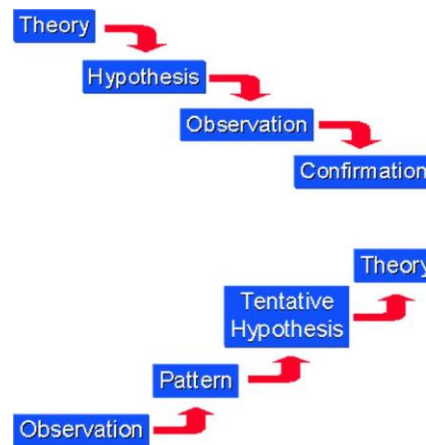


Figure 6-2: Deduction (top down) & Induction (bottom up) approaches to research

6.2.4 The Methodological Choice

Understanding PM4D requires looking at the research from several perspectives. Framing this research within the research onion is complicated by the fact that the data is already collected, and this existing data forms the opportunity for this research. Approaching the research onion using a strictly outside-in approach is somewhat misleading because of that. Specifically, no decision regarding data collection is necessary or possible. The existing data spans 15 years of school screenings and hence offers longitudinal studies as the obvious choice. Experimental strategies also appear applicable. The research uses data extracted from dental EHRs and uses theoretical constructs from the area of “secondary use of EHR data” as part of the methodology. Using EHR data for public health research offers specific issues for consideration and also influences the methodology used. Data mining is the general area of data science in use in this research and PM is the specific technology being applied. PM has its own established methods and constitutes the main body of PM4D. General experiment methods and specific PM experiment methods are then utilised within PM4D to answer the validating RQs.

6.2.5 The Research Strategy

Given the author’s positivist philosophy and proceeding by deductive reasoning with pre-existing data, experimental research design is the obvious strategy with a primarily quantitative approach.

6.2.6 The Time Horizon

The time horizon is longitudinal, and the data collected is from an archived source allowing statistical analysis if appropriate.

6.2.7 Data Collection

The data collection is detailed in Section 4.1.

6.2.8 Data Analysis

Analysis of the data is detailed in Chapters 4 and 7.

6.3 Process Mining Project Methods

This is a description of the existing PM methods and their strengths and weaknesses, which elements of these have been chosen for this research, and why they were chosen. The PM4D research methodology is based on 6 established formal PM methods. First,

the Process Diagnostics Method (PDM) (Bozkaya, et al., 2009) which addresses the complexity of healthcare processes, Business Process Analysis in Healthcare Environments Methodology (Rebuge & Ferreira, 2012) building on the PDM above, the L* life-cycle method as detailed in the Process Mining Manifesto (IEEE, 2011), the PM² method (van Eck, et al., 2015), The ClearPath Method (Johnson, et al., 2018), and finally and a Question-Driven Methodology for Analyzing Emergency Room Processes using Process Mining (Rojas, et al., 2017). An outline of each is described below. PM4D identifies and tries to address gaps and limitations in these existing methods identified by this research i.e. secondary use of dental EHR data in a research environment. Our systematic examination of the existing methods in Sections 6.3 led us to the methodology documented in Section 6.4.

6.3.1 The Process Diagnostics Method (PDM) (Bozkaya, et al., 2009)

6.3.1.1 Introduction

This method performs process diagnostics using PM and proposed five phases: Log preparation, Log Inspection, Control Flow Analysis, Performance Analysis and Transfer Results as shown in Figure 6-3 below.

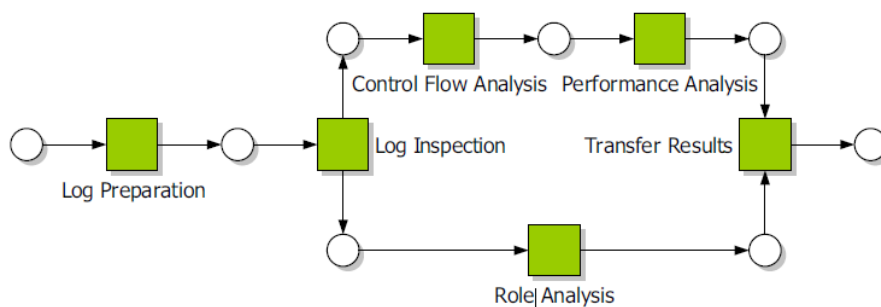


Figure 6-3: Phases of the methodology (Bozkaya, et al., 2009, p. 23)

This is an early method constructed with the intention to make PM a repeatable service in the circumstances where the parties had no prior knowledge, i.e. the event-log is presented to the process miner as a *fait accompli*. Also, the process miner has no domain specific knowledge and no role in analysing the results of the diagnostics. The steps involve pre-processing to appropriate datatypes, profiling of the log, conformance-checking against known organizational process if any exists, and process discovery utilizing the fuzzy miner algorithm. This is followed by replaying the log on the discovered model i.e. performance checking, to find bottlenecks for example and role-analysis to establish ‘*who does what?*’. In their case study, presenting the results to the process owner and the accompanying discussion with the process miner was a key step in helping the client interpret the outcomes.

6.3.1.2 Strengths of the Method.

- Asks the question “does the information system really reflect the state of affairs of the business process?” not the common PM mantra that the information system reflects how the process is ‘actually’ executed, (Rebuge & Ferreira, 2012; Rovani, et al., 2015).
- Even without domain knowledge the outputs were recognizable to the domain experts.
- The method can provide an overview of the organisation’s processes quickly.

6.3.1.3 Limitations of the Method

- It is a quick method intended to give a broad overview without much detail.
- Snapshots of processes are limited and need enhancing by domain experts.
- As the EL is presented to the researcher (miner) as a *fait accompli*, the researcher has limited facility to assess the data provenance or quality.
- Due to the process miner’s lack of domain-specific knowledge, there is also little capacity for making common-sense or obviously helpful adjustments to the log.
- The terminology used in the introduction to describe events, traces, activities, cases is inconsistent with much of the literature and introduces terms not seen elsewhere e.g. ‘trail’ and ‘run’. This is understandable as this was the first published PM method.
- The method assumes that the EL is readily available in the information systems and the section on pre-processing is vague and not reproducible. Log inspection results in incomplete cases being removed resulting in a log ready for ‘Control Flow Analysis’. This is insufficient and other measures may be necessary at this point to prepare the log for process discovery such as removal of invalid data or data of inadequate quality.
- They use the terms ‘conformance’ and ‘compliance’ interchangeably.
- Noise and infrequent behaviour are treated as if they are the same and simply removed from the log to facilitate creation of simpler models. This is insufficiently dealt with as the removed information may be important. Such information may reflect exceptional behaviour necessary for a particular patient group.

6.3.2 Business Process Analysis in Healthcare Environments (Rebuge & Ferreira, 2012)

6.3.2.1 Introduction

The PDM above was extended by Rebuge & Ferreira (2012) to deal with the highly dynamic, highly complex, multi-disciplinary, and ad-hoc nature of healthcare processes which can result in incomprehensible process models. This method applies PM techniques leading to the identification of regular behaviour, process variants, and exceptional medical cases. An additional ‘clustering’ step was incorporated using Microsoft SSAS Sequence Clustering, to identify regular behaviour and group similar processes together thereby improving the readability of the resulting models. This additional step is shown in Figure 6-4.

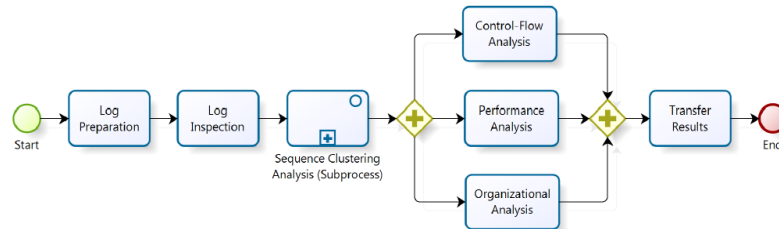


Figure 6-4: Proposed Method for BPA in healthcare (Rebuge & Ferreira, 2012, p. 107)

This method proposes running a sequence clustering algorithm to discover patterns of behaviour, infrequent behaviour, and process variants. The regular behaviour is established by identifying the clusters with the highest support. The regular behaviour is then identified by the examining the Markov chain associated with the cluster which gives probabilities of specific events following each other. A Markov chain is a model of a sequence of possible events in which the probability of each event occurring depends only on the state existing in the previous event. There may be several clusters with high support and unpicking this may require domain expertise. Clusters with lower support are then categorized as ‘process variants’ and those with least support as ‘infrequent behaviour’.

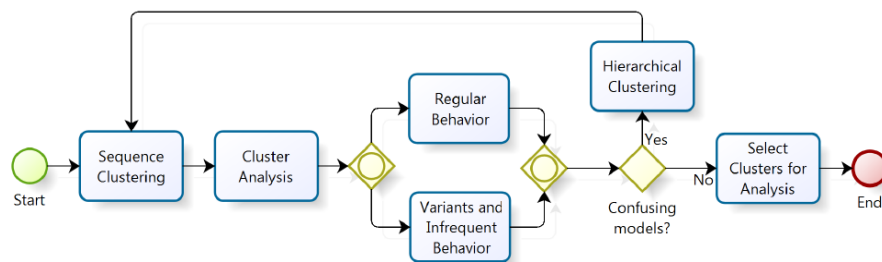


Figure 6-5: The Sequence Clustering Analysis subprocess (Rebuge & Ferreira, 2012, p. 108)

To distinguish these effectively, the authors indicate the importance of careful inspection here as ‘infrequent behaviour’ may give insight into special patient groups and clinical situations. If the model remains incomprehensible, then an additional step of ‘hierarchical clustering’ is applied to further simplify the output.

6.3.2.2 Strengths of the approach

- This approach has a clearly articulated understanding of the difficulties of modelling healthcare processes and proposes addressing the issue of incomprehensibility with clustering techniques.

6.3.2.3 Limitations of the approach.

- The PM studio used appears to be only applicable in their specific case-study which uses an in-house developed EHR.
- The format of the event-logs is MXML which has now been replaced with the XES format. This limits the general usability of the work, though the implementation of data mining techniques and ideas remain useful.
- Another PM perspective, ‘data’, described here as ‘...related to the data objects that serve as input and output for the activities in a case’ is inadequately explained.

6.3.3 L*Life-cycle process mining method (IEEE, 2011)

6.3.3.1 Introduction

The IEEE Task Force on Process Mining issued a declaration of its principles and intentions in the form of a manifesto with the objective of promoting the development and use of PM as a management tool (IEEE, 2011). PM is positioned as an “enabling technology” for management approaches such as Continuous Process Improvement (CPI), Business Process Improvement (BPI), Total Quality Management (TQM), and Six Sigma. Additionally, this author suggests that PM is complementary to the lean approach and the Value Stream Mapping (VSM) technique. These techniques aim to improve operational performance in organisations. Other organisational objectives such as compliance and conformance can also be progressed by PM. The five stages of the L*life-cycle PM project method are summarised below.

Stage 0 is the planning and justification phase. This involves investigation of the domain to establish what process-related information the stakeholders require. Which of the PM perspectives (discovery, conformance, and enhancement) will be employed? Are the available information systems ‘process aware’? Are the available data sources clinical, administrative or healthcare support systems? What type of PM project is this i.e. is it curiosity, question or goal driven? What are the desired data and target dataset? What data is available to us? What questions can be answered? What value can be added?

Stage 1 translates into the aims, objectives and the RQs. Stage 1 produces process models. KPIs and handmade or *de jure* models may have emerged from existing documentation, domain experts and stakeholders describing best practices.

Stage 2 involves the construction of the event log and linking it to the control flow model. One of the guiding principles of the manifesto states that log extraction should be driven by questions. This principle also serves to minimise the data requested and push PM away from discovery science towards a more traditional scientific method.

Stage 3 Additional information is now incorporated to extend the model from stage 2, e.g. timestamps and calculated durations could estimate wait times and throughput.

Stage 4 Here, knowledge from historical PM is used to monitor and control currently running cases. This could be used to predict outcomes, throughputs and to flag deviations and adverse events.

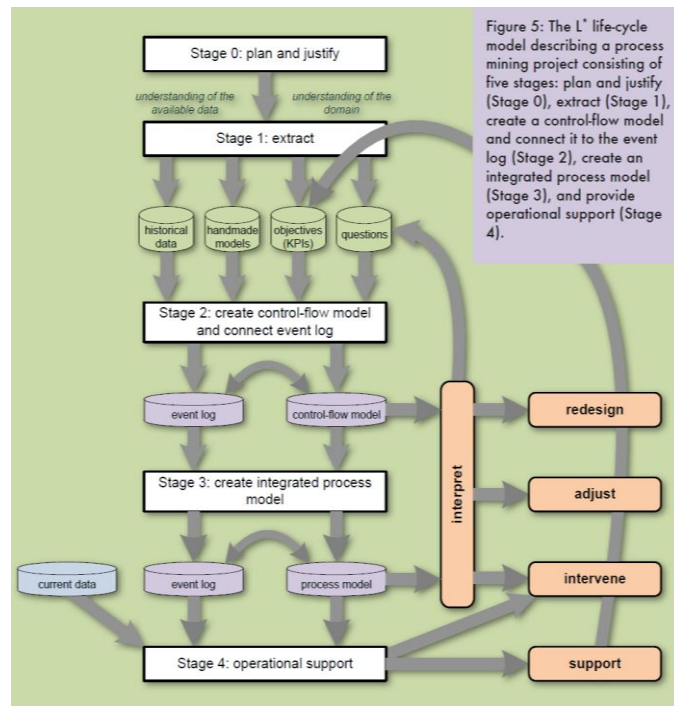


Figure 6-6: L* life-cycle methodology (IEEE, 2011)

6.3.3.2 Strengths of the approach

- It expands the previous methods to the planning and justification of the PM exercise. This is made imperative by asserting that PM should be driven by RQs.
- The authors introduce the last stage of a PM project where the models can be used in operational support. It is explicit in articulating this use of the process models to feedback into the 'Extract' phase.
- They introduce the idea of the artefacts produced in the 'Extract' stage. This should be expanded to all stages of the project, in particular in research environments, producing a thorough audit trail and supporting research reproducibility.

6.3.3.3 Limitations of the approach

- The method does not accommodate the complexity inherent in the healthcare domain and its processes, possibly because it was designed for structured processes aimed at producing a single integrated process model (van Eck, et al., 2015).
- To this author it appears to be a general method approach without the detail necessary for application in a specific research domain.

6.3.4 The PM² method (van Eck, et al., 2015)

6.3.4.1 Introduction

PM² was developed in response to the high-level nature of the previous methods, and to address some of the limitations of PDM and L*life-cycle method described above. The authors identified that the scope of PDM was limited, covering only a small number of PM techniques. The other major limitation identified was the emphasis on avoiding the use of domain knowledge during the analysis which makes it less useful for larger, more complex projects. They critique the L* life-cycle method in that, while being broader than PDM, it was primarily designed for the analysis of structured processes and aimed at discovering a single integrated process model. They state that neither method explicitly encourages iterative analysis, and both could benefit from additional practical guidelines for inexperienced practitioners. The phases and input/outputs in PM² are summarised in Figure 6-7. For each phase of the PM² method, inputs and outputs are clearly defined as are concrete steps to be performed, referred to as activities. The goals of a PM project can be very concrete such as achieving a 10% cost reduction for a particular process or more abstract such as obtaining valuable insights regarding the performance of several processes. Through PM² these goals are translated into concrete RQs which are iteratively refined and answered.

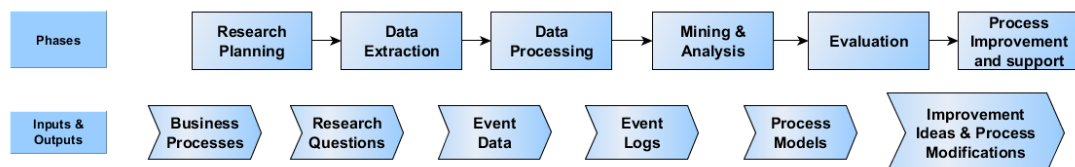


Figure 6-7: PM² Method Steps

6.3.4.2 Strengths of the method

- It provides significantly more guidance to process miners. While its phases are similar to those of the previous methods, more implementation detail is provided.
- It is explicit in defining the inputs and outputs of each phase. This is useful when starting a PM project encouraging reproducibility and creation of an audit trail.
- This is the first method to explicitly introduce the importance of data quality in the event data and to suggest that this will likely affect the outcomes of the project.

6.3.4.3 Limitations of the method

- It does not incorporate issues specific to healthcare processes.

- It does not incorporate issues specific to research.

6.3.5 Declarative Process Mining in Healthcare (Rovani, et al., 2015)

6.3.5.1 Introduction

While clinical guidelines aim to improve the healthcare delivery process, there are often good reasons to deviate from them. Healthcare is a complex, unpredictable process requiring flexibility in its delivery. The authors link the ‘spaghetti effect’ or unreadability of healthcare process models to the explicit representation of all possible paths in a highly complex, dynamic environment. As an alternative they propose using declarative models where the models are expressed as a series of constraints. They propose a method to check the *de jure* model against the actual clinical practice and adjust the *de jure* model to reflect the actual clinical practice, leading to a *de facto* model. Their proposed method in Figure 6-8 provides useful additional methodological steps to achieve this.

6.3.5.2 Strengths of the method

- Their understanding of the complexity of healthcare processes and the necessity for flexibility in their execution and the knock-on effect of making the procedural models incomprehensible. They claim that the proposed ‘declarative’ approach defines a process as a series of constraints and is more compact and understandable.
- Their method splits the data into ‘training’ and ‘test’ in the traditional machine learning method facilitating cross-validation. This encourages quality assessment of the models and as in data-mining, should enhance repaired model accuracy.
- Incorporates domain expertise in the model repair phase.

6.3.5.3 Limitations of the method:

- The results are based on a single case-study
- Though the text says that the clinical guidelines are updated if found that the actual execution of the process is in fact the correct process, the authors did not indicate this learning feedback in the method diagram.
- Their approach assumes that the EL represents how a process is ‘actually’ executed. It is more correct that the EL is an accurate representation of IS records.
- It requires familiarity with the Declare (Linear-temporal-logic) e language, somewhat mitigated as there is a Declare Miner ProM plug-in as well as Analyzer and Checker to execute conformance checking. Requires familiarity with the Declare Designer to create the original *de jure* model. This is acknowledged in their conclusions.

- Each discrepancy between the *de jure* model and the log must be explained to decide whether to ‘repair’ the *de jure* model and this is a resource intensive activity
- It may be problematic if the domain expertise’s availability is limited.

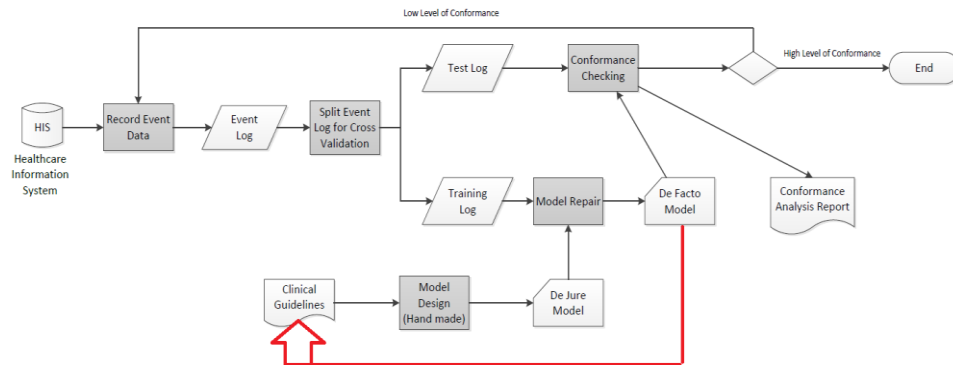


Figure 6-8: Method for the analysis of medical treatment processes (Rovani, et al., 2015)

6.3.6 Question-Driven Methodology for Analyzing Emergency Room Processes Using Process Mining (Rojas, et al., 2017)

6.3.6.1 Introduction

The aim of the work was to create a methodology for answering frequently posed questions in emergency room management using PM. The authors identified the need for data reference models to identify and manage the information necessary to answer the questions. They also identified the need to reduce the complexity of resultant process models by asking specific questions and finally they identified the need to apply PM in flexible environments. The method provides detailed activities, descriptions and guidelines in six main stages: data extraction, event log creation, filtering, data analytics, PM, and results evaluation stages.

6.3.6.2 Strengths of the method

- Creation of question classifications.
- Proposed data reference model to guide the data extraction from the HIS.
- Creation of a question driven methodology specific for emergency rooms.
- Focus on data quality.

6.3.6.3 Limitations of the method

- Lack of outcomes-based question classification. This could have been used in addition to the ‘episode’, ‘triage’ etc. classifications.
- Could have considered use of ontologies such as SNOMED in Activity 1.3

- This method suggests managing filtering of the EL in the PM application, Disco. Given that this research is focussed on specific questions, the filtering could also be managed at the creation of the EL and it is this author's opinion that the research is more reproducible using this method. For example, the next version of Disco might use an enhanced algorithm producing different results. However, the rationale could be that the filtering of the EL at the creation stage reduces flexibility in the PM stage.

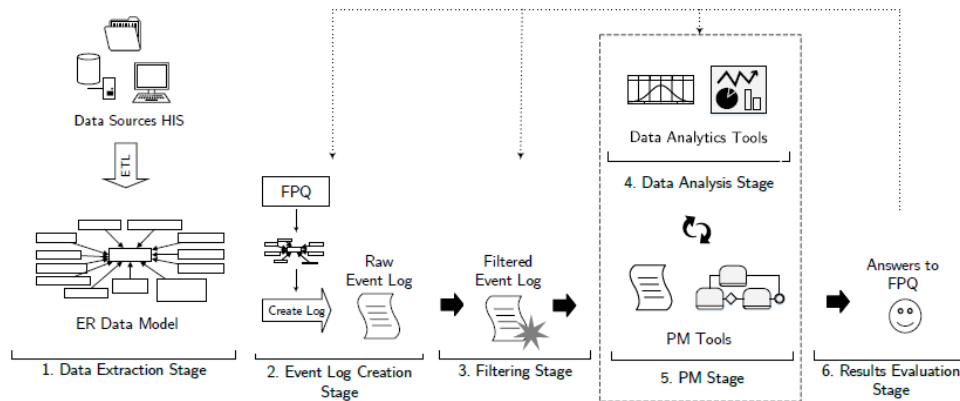


Figure 6-9: Question-Driven Methodology for Analyzing Emergency Room Process Using Process Mining (Rojas, et al., 2017)

6.3.7 The ClearPath Method for Care Pathway Process Mining and Simulation (Johnson, et al., 2018)

6.3.7.1 Introduction

The ClearPath Method also builds on the PM² method and adds a simulation phase to help communicate care pathways to stakeholders and explore what-if options to facilitate improvement of these pathways. The method also addresses issues of poor data quality and supports rich stakeholder engagement. The authors emphasise the involvement of domain experts in iteratively refining PM efforts and how simulation models have previously been effective in motivating that engagement.

The method uses NETIMIS (www.netimis.co.uk), a cloud-based online service used to manage and create models of care pathways as runnable simulations, with nodes representing events and pathways animated with moving tokens representing patients. The simulator requires no patient-level data as the tokens are randomised with population attributes. Users can interact with the simulation using a series of features such as zoom and inspect. Iterations of models can be run side-by-side allowing easy comparison. The method acknowledges the importance of formal data quality management and auditing of data extract and transform activities. It advocates for an agile approach to produce simulation models using an iterative approach supported by software tools, traditional

academic research methods and traditional business process analysis. The ClearPath method follows the six stages as in PM² : planning, extraction, data processing, mining and analysis, evaluation, and process improvement and support. Simulation is introduced in Stage 5 where mined models are recreated by hand in NETIMIS and enhanced with information from the business process analysis actions. Evaluation takes place in conjunction with the Clinical Review Board (CRB) to view and evaluate candidate NETIMIS models. Once accepted, models can be published on NETIMIS and shared with other organisations.

Uniquely, the ClearPath method utilises an evidence template to create the early-stage models. This evidence-base is based on references to source material in the literature which is then improved with reference to the CRB.

The method is then illustrated using three case studies. The case studies highlight problems common in modelling healthcare processes including the lack of sufficiently detailed information recorded in the EHRs, difficulty in extracting detail capable of providing rich insights and spaghetti-models.

6.3.7.2 Strengths of the method

- Inclusion of simulation adds a valuable ‘what-if’ dimension to the process models.
- Three case-studies greatly enhance the validity of the method. The case-studies add generally to PM knowledge by exposing data issues arising therein.
- It requires no patient-level data easing ethics and data protection issues.
- It assesses data quality issues directly using the data quality framework.
- The method utilises an evidence-base from literature.
- Use of standards such as SNOMED-CT and data models for aggregation.
- It has an iterative improvement approach supported by a clinical review board.

6.3.7.3 Limitations of the method

- The method requires knowledge of simulation and the NETIMIS product
- It is heavily dependent on the CRB whose availability will typically be limited.

Process Mining Methods Comparative Summary					
Process Diagnostics Method (PDM)	PDM (& sequence clustering)	L*Lifecycle	Declarative Process Mining	PM ² (ClearPath additions)	Question-Driven
(Bozkaya, et al., 2009)	(Rebuge & Ferreira, 2012) Extends (Bozkaya, et al., 2009)	(IEEE, 2011)	(Rovani, et al., 2015)	(van Eck, et al., 2015) (Johnson, et al., 2018)	(Rojas, et al., 2017)
Log Preparation: Identification of information within organisation's IS Identification of events and activities Clarification of timestamps	Data Gathering: Created a subset of the IT systems to a new database Created a 'Medtrix' Process Mining Studio with architecture Specified a RQ regarding patient handovers	Stage 0: Plan the project Outputs: Understanding data leading to... Stage 1: Extract Understand the available data leading to... Extract event data from systems Understand the domain leading to... Extract models, objectives and questions from domain experts and management Outputs are... historical data, handmade models, objectives, questions	New Step before Extraction: Model Design (<i>de jure</i>) Create process model based on clinical guidelines This is done by hand Use declarative language Record Event Data/ Event Log	Phase 1: Planning Set up project Determine RQs Set goals: Improve processes, check conformance Time-Boxed, pre-booked meetings with CRB Phase 2: Extraction Determine scope (attributes, granularity, timeframe) Extract Event Data Transfer process knowledge (from business experts to analysts) Development of evidence-base	1: Data Extraction Stage Identify available data Verify timestamp Name events Create specific fields Verify Data Quality
Log Inspection: Create overview log statistics, No of cases, roles etc., no of different events Sizes of processes, max/min events per case Filter to remove incomplete cases				Phase 3: Data Processing Creating views (ELs) Aggregating Events (Using SNOMED-CT) Enriching Logs Filtering Logs	2: Event Log Creation Identify data to answer specific question Create Event Log Include specific data for each event
Control flow analysis: Generate model using discovery Compare to pre-existing model i.e. do conformance checking	Additional Step before Control flow analysis - Sequence Clustering Analysis: Find patterns in the EL	Stage 2: Create control flow model and connect event log Use automated process discovery techniques	Additional Step before Control flow analysis - Split EL to cross validate Create test log Create training log (TrL)	Phase 4: Mining and Analysis Process Discovery Conformance Checking Enhancement	3: Filtering Stage Basic filtering Clinical Filtering Question-driven filtering 4: Analysis Stage(DA)

Filter log using Pareto principles to get high frequency sequences	Provides insight into regular/infrequent behaviour Simplify maps clustering similar processes Critiques PM algorithms	Model may trigger redesign or adjust Filter and adapt EL using the model	Create model based on TrL Repair <i>de jure</i> model – becomes <i>de facto</i> model	Process Analytics (using other Data Mining techniques)	Select DA techniques Statistical Analysis DataMining Analysis
Performance Analysis: Use dotted chart to compare processes and throughput times Are there bottlenecks? Compare performance of different sequences.	Performance Analysis: Use petri-net to compare processes and throughput times	Stage 3: Create integrated process model Extend the control flow model perspectives e.g. data, time, resources May answer more Quests May trigger more actions	Conformance checking Using the <i>de facto</i> model and the test log. If High conf. then End If Low, regenerate EL Feedback from <i>de facto</i> model to CGs Report	In Phase 4 Hand-made models using NETIMIS	5:Process Mining Stage(PM) Identify PM tools Process Discovery Conformance analysis Performance analysis Organisational analysis Question specific analysis Iterate on DA & PM
Role Analysis: Who executes what activities? Create a role-activity matrix Discover specialist. Create a social network	Organisational Perspective: Create a social network to track deviations.	Also executed in Stage 3:		In Phase 4 Phase 5: Evaluation Diagnose Verify and Validate Iterative process using Clinical Review Board	
Transfer results: Discuss outputs with client/domain experts		Stage 5: Operational support Models may be used for operational support		Phase 6: Process improvement and Support Implement Improvement Supporting Operations Model acceptance by CRB and publishing	6: Results Evaluation Stage Identify experts Define feedback instruments Obtain Feedback

Table 6-1: Process Mining Methods Summary

6.4 Extending the Existing Methods for Dentistry Research

6.4.1 Introduction

In the dental PM literature review, no formal use of existing PM methods was evident. As this research developed, new phases and steps emerged necessary to execute PM in a dental healthcare research setting adding to those in the methods reviewed above. Some of these additional phases and steps were essential to dental research and others significantly aided the process. For example, ethical considerations are an essential step in dental research whereas the use of a dental data reference model is helpful in defining the required data. While it could be argued that some of these steps are common to all research, they are mostly omitted from published PM literature. They constitute an essential part of this research's methodology. This research's methodology, known as PM4D for convenience, borrows heavily from the methods outlined above, primarily the later methods, PM², ClearPath and the Question-Driven methods. This was an iterative process and many of the steps were revisited, reordered, and optimised. A limitation of the PM4D steps is that they can appear to be strictly sequential and linear. The reality was somewhat less clear-cut, and the methodology required some flexibility from the author before it was finally documented.

6.4.2 Key points on this Methodology

- For convenience the methodology used in this research is named PM4D.
- PM4D tells the story of this research.
- The PM² and the Question Driven methods described above provided the starting structure for PM4D.
- Additional steps were added to meet the needs of dental PM research.
- PM4D consists of 13 steps. Steps 1 through 8 are general preparatory steps followed in this research. These are, in many cases, unique to the experience of this research and although they might provide useful guidance for future dental PM research, they are not necessarily easily transferrable.

1. Plan
2. Assess the available data
3. Get appropriate research permissions
4. Prepare and document the research environment
5. Data Extraction
6. Data Pre-processing
7. Assessing data quality

8. Create data description and profile
 - Steps 9 and 10 are the preparatory steps necessary to define the specific research questions and prepare the data to execute the process mining experiments.
9. Define detailed RQs
10. Data processing to facilitate answering RQs
 - Step 11 is the key PM step and is based on the Question Driven method as summarised in Figure 6-9).
11. Mining & Analysis
 - Step 12 is the Evaluation and Discussion of the results.
 - The additional Step 13, Process Improvement and Support, in the summary in Table 6-2 is the logical and desirable next step, present in some of the existing methods, but not formally executed in this research.

The following section describes the steps, their strengths and limitations in more detail.

6.4.3 PM4D Methodological Steps and Critique

6.4.3.1 Plan

This research applied process-oriented data science to a large clinical dental EHR extract and aimed to provide new insights into the variable pathways leading to different outcomes. According to the Process Mining Manifesto (IEEE, 2011, p. 7) and Rojas, et al. (2017), PM should ideally be driven by RQs and this should guide the extraction of meaningful event data. This research set out with the broad aim as outlined above and without clearly defined RQs. This reflected the emerging nature of the process mining technology, the exploratory approach taken to the data, and the lack of previous applications to large dental datasets. It was also a reflection of the challenges of assessing the ability of PM and EHR data to answer specific clinical questions. For example, it was intended to examine the effects of applying fissure sealants on treatment process and oral health outcomes, but this was not possible due to difficulties defining cohorts having received (or not received) the treatment. However, this broader aim led to a more comprehensive dataset than might have otherwise resulted and opened the door to a more exploratory approach in the research.

The required minimum dataset for execution of PM is case identifier, event and timestamp. Considering a patient as a case, treatments as events, and time of treatment as timestamp, one would expect these basic and minimum data elements would be available from every EHR. However, this research considered the full dataset available from the EHR and attempted to maximise its utility through expanding the dataset facilitating a

more exploratory approach. This step is closely linked to the log preparation stage of the PDM (Bozkaya, et al., 2009), the data gathering stage introduced by Rebuge & Ferreira (2012), Stage 0 of the L* Lifecycle (IEEE, 2011), the planning phase of PM² (van Eck, et al., 2015) and the ClearPath methods (Johnson, et al., 2018), and the data extraction and event log creation phases of the question driven method (Rojas, et al., 2017).

A clear shortcoming of this research is the lack of a formal clinical review board as incorporated in the ClearPath method, which would have a crucial role in development of RQs and in the assessment of the available data. It is clear that each PM research project will have its own aims and RQs and accordingly its own data requirements and available data sources. Hence, it is not feasible to have a strict cookbook approach to the planning phase.

6.4.3.2 Assess the available data

Two standards relevant to PM of dental EHR data were identified: the PM healthcare reference model (Mans, et al., 2015) and the ANSI EHR standard (American National Standard/American Dental Association, 2013, pp. 27-52). These standards provided the basis for first, assessing the ‘completeness’ of the available dataset and second, for making recommendations for an ‘ideal’ dataset. Both these standards, while useful, had limitations in this application. The healthcare reference model (HRM) is generated from the information systems of several hospitals without any specified dental service and many components were not relevant to the dental service on which this research is based. The ANSI standard was functional in its definition and did not have specific data definitions or a data dictionary.

Bearing this in mind, the available data was compared to the standards and a gap analysis was completed. This step positioned the dataset within the proposed standards and produced a generalisable benefit in informing future dental EHR designers wishing to accommodate process and data mining. While useful, it is limited by the ‘unknowability’ of other potential RQs. There are many areas of specialism in dentistry not considered (endodontics, orthodontics etc.) each having their own specialist data requirements to assess process and outcomes. Only when the RQs are finalised could an ideal dataset be described. The existing methods did not reference either of these standards. The ANSI standard is specific to dental EHRs and was not relevant to the existing methods. However, the HRM could have been referenced in the methods based on healthcare processes. This part of the research is detailed in Chapter 7.

6.4.3.3 Get Appropriate Research Permissions

A step not specifically developed in the existing methods is that of acquiring the appropriate permissions for the research. In this research it was a complex and time-consuming phase. At this stage, the research aims and objectives were clear as was the data required to answer them. This allowed a concrete application for data access to stakeholders: data-owners and controllers, ethics committees, and software suppliers whose assistance was required in the extraction process. This phase in a research scenario involved completing application forms and satisfying stakeholder's requirements regarding the proposed use of the data. Ethics clearance documentation and permissions were received in return. This phase required ethical clearance from University College Cork followed by agreement from the Primary Care Research Committee and the agreement of the Principal Dental Officer where the EHR was in use. The request was referred to the Office of the Data Protection Commissioner for an opinion which returned supportive of the research. This part of the research is detailed in Section 4.1.4.

6.4.3.4 Preparing and Documenting the Research Environment

In advance of receiving data, preparations were made for the research environment. Commitments were given to data owners regarding the security and handling of the data. Complying with this included describing the hardware and software architecture and generating documents such as data and anonymisation management plans describing the data situation and the data flow around the organisations involved. Further consideration was given to data protection issues at this point.

A reproducible research document was drafted, detailing what efforts can be made to have the research data placed in an accessible repository when the current research is completed. This will also facilitate verification of the research results and, with the permissions of the data-owners and controllers, the use of the data for further research. This part of the research is detailed in Sections 4.2, 4.3, and 5.2.

6.4.3.5 Data Extraction

The primary data extraction culminated in CSV files being presented to the researcher. Several steps were undertaken here. Extraction code was written and executed. Many of these steps were specific to the Bridges EHR, were conducted without significant input from the researcher, and are included here primarily to indicate the outputs and resulting artefacts as presented in Table 6-2 below.

6.4.3.6 Data Pre-processing

This consists primarily of the anonymisation process in which the researcher was not involved and as above, is included here primarily to indicate the outputs and resulting artefacts as presented in Table 6-2 below. An anonymisation plan and anonymisation code was created and executed resulting in anonymised data. Documents such as the Anonymisation Standard Planning Record (See Appendix 10.6) were created in this phase. Many of these steps were specific to the Bridges EHR, were conducted without significant input from the researcher, and are included here primarily to indicate the outputs and resulting artefacts as presented in Table 6-2 below.

This anonymisation phase, though clearly necessary, has a generally negative impact on the value of the data as information such as the location of the individuals is removed at this point. It is also clear that if the data owner was seeking to answer the research questions in-house, this step may be unnecessary and could result in higher quality research data.

6.4.3.7 Assess Data Quality

After receiving the anonymised data, it was loaded into the research environment using data load scripts. There were several intertwined steps following this phase. Data and metadata were assessed for quality referencing the RQs, and data of inadequate quality was marked as such. Some data quality (DQ) issues found at this stage disqualified the data from all research e.g. data integrity issues while other DQ issues only disqualified the data for a specific RQ. The complexity of the DQ issues necessitated the development of a data quality framework as detailed in Section 10.19.3.

6.4.3.8 Create Data Description and Profile

After DQ assessment and transformation, the data was described and profiled. This resulted in documents such as entity relationship diagrams, data descriptions, data profile documents, and DQ documents.

6.4.3.9 Define detailed RQs

Next, detailed RQs, hypotheses, and experiments are defined. This part of the research is detailed in Chapter 3.

6.4.3.10 Data processing to facilitate answering RQs

This can be divided into two parts: general data processing with transforms to facilitate the RQs, and data processing to facilitate answering specific RQs

Part 1: General data processing and further transforms to facilitate the RQs are executed resulting in additional data tables and fields, including aggregated and calculated data. These transforms are detailed in Appendix 10.18.

Part 2: Data processing to facilitate answering specific RQs. This has two phases: cohort definition and creation, and event log creation.

Phase 1: Cohort Definition and Creation

- First the Cohort must be defined in terms of selection criteria resulting in a list of patients who fulfil the criteria to have their record included in the experiment. This step is akin to temporal electronic phenotyping (Hripcsak & Albers, 2013) (Liu, et al., 2015)
- A table, CohortX, containing the ID for each patient in the cohort is created in the BridgesPM1 database, where X is the name-summary of the RQ.

Phase 2: Event Log Creation

- All relevant treatment process events experienced by these patients are then extracted and exported to a csv/txt file called RQn.txt where n is the RQ number. The minimum required data elements to carry out this experiment were ClientID (Case), ProcedureName (Event), and CompletionDate of each treatment event (Timestamp). This is similar to the Filtering Stage described by Rojas et al. (2017) here whose method proposes using basic, clinical and question-driven filtering from within the PM tools to create the EL and these filters are normally included in the PM tools. In this research most of the filtering was carried out here i.e. when the cohorts were defined and when it was decided which events to include in the EL. This was done primarily for ease of auditing and reproducibility. It is difficult to capture filter settings from the PM tools as they are often set interactively in the user interface. The author is not aware of a facility for capturing these settings along with the process model output. Therefore, the filtering is captured in the SQL files creating the cohorts and selecting the events for processing. By filtering at the event-log creation stage, some flexibility and agility at the point of use of the PM tools is lost, but the SQL scripts used facilitate easy editing and recreation of the cohorts and events when required.

- The csv/txt file is then converted to an XES formatted EL within Disco using standard Disco functionality.
- To provide an overview of the EL, from information available within the Disco application, the fundamental statistics around cases and events are then summarised. This gives information on the proportion of variants and non-unique pathways in the event log and is useful for gaining intuition and understanding of the data.

6.4.3.11 Data Analysis and Process Mining

Part 1: Data Analysis

Analyse the data supplemental to the PM analysis. This is akin to the data profiling already executed in this research but is specific to the experiment with the aim of discovering different patterns and knowledge on data contained in the event logs. Rojas, et al. (2017) characterise this as: selecting statistical analysis and data mining techniques and tools that are then used to characterize an event log, identifying the frequency of activities, the distribution of cases over time, and variants of process execution, among others.

Part 2: Process Mining

This is followed by the process mining using Disco. The reasons for the decision to utilise Disco is detailed in Section 2.6.4. Analysis of results follows.

If using the complete dataset yields an incomprehensible spaghetti model as is often the case with healthcare processes then, consistent with this research's strategy to carry out all filtering at the cohort and event log creation phase, the event log will be regenerated omitting less frequently occurring events. Test-runs are to be carried out using various thresholds for inclusion of an event in the EL, cognisant of our guidelines for legibility and comprehensibility in Section 5.4.2 until the cohort yields acceptable process models. The software's frequency model was set up with 'Case Frequency' as the primary metric and 'Absolute Frequency' as the secondary metric. Case frequency indicates the number of distinct patients who experienced an event (i.e. received a treatment). Absolute frequency refers to the number of times an event occurred and hence, a patient could appear in this count multiple times. The performance model set up with 'Mean Duration' as the primary statistic and case frequency as the secondary metric.

6.4.3.12 Process improvement

Process improvement is not addressed in this research. However, they are included in the methodology for completeness.

6.4.3.13 Support

Process improvement and support are areas not addressed in this research. However, they are included in the methodology for completeness.

The 13 steps in PM4D, a listing of inputs and outputs and their locations in the thesis are shown in Table 6-2 below. Those steps that are new or add significantly to the existing methods are marked with an asterisk and printed in red.

Table 6-2: Extended Methodology Steps

Methodology Step	Inputs	Outputs	Where are these outputs in this thesis?
Plan	This research is the application of Process-Oriented data science to a large clinical dental EHR extract. Research aims are the creation of a robust methodology to achieve this and its validation. The processes examined are closely tied to the aim/objective of improving Dental Public Health. Data Mining Literature Review e.g. various algorithms, discovery and conformance.	Overall RQs. Ideal Dataset description. Minimum Dataset Description The resulting general RQs will generate an ideal dataset which are compared/mapped to the Healthcare Reference Model and the ANSI Standard	Chapter 3 Figure 4-3, Section 7.6.6 Case, Event, Timestamp
* Assess data. Data Model and Mapping. Map 'ideal' and 'available' data to Healthcare Reference Model and ANSI Standard	Ideal Dataset, Minimum Dataset Healthcare Reference Model (HRM) Dental EHR Standards (ANSI 1937 2013)	Mapping Document Compare Datasets to the HRM. Gap analysis of both	Appendix 10.2 Section 7.6.5 Not formally done
* Get Appropriate research permissions	Application Forms	Ethics Clearance Data Controller Permission Software Supplier IP Agreement Other Governing Documents	Appendix 10.4 & Code CD(20) Appendix 10.4 Appendix 10.5
* Prepare research environment	Hardware & Software Architecture Software Installation Security & Integrity Applying e.g. UK Data Archive data life-cycle, create a plan.	Environment description Reproducible Research Document Data Management Plan (UoL requirement) Data Protection Plan	Section 4.3 Appendix 10.1 Not formally required as data anonymised. See ADF in Section 5.2
Extraction	Data extraction to address Overall RQs: This ideal dataset description will be compared/mapped to the available dataset (Bridges). This output is Bridges-PM1)	Data Extraction Script SQL Server Services Integration package to export to CSV files Event Data	Code CD (1) Code CD (3) CSV files not available due to data-owner restrictions
* Pre-Process	Anonymise Transfer to research location	Anonymisation Standard Planning Record	Appendix 10.6

	Load data into research environment	Anonymisation Plan Data Anonymisation Script Data Anonymisation Checklist Anonymised event data SQL Server Data Load scripts	See ADF in Section 5.2 Not Included to protect anonymisation process (2) CSV files not available due to data-owner restrictions (4) Code CD (5)
* Assess Data Quality	Apply the Care Pathways Data Quality Framework to manage and mitigate data quality issues	Quality Assessed Data (Event Data) List of Data Quality issues Data quality report	SQL Server database See Section 10.17 See Section 10.17
* Create Data Description & Profile	Describe Data and Metadata Profile Data - Document general information about the event data: Volumes, Variety, demographics	Data Description Entity Relationship Diagram Data Profile Jupyter Profiling Notebook	See Chapter 4.1 See Figure 4-3, See Chapter 4 Code CD (9)
* Define detailed RQs Apply Policy & Strategy approach	The RQs defined here should be detailed enough to create a testable hypothesis and to define cohorts for testing Initiate an Experiment Design Document	RQs Hypotheses Experiment Design Documents	See Chapter 3 See Chapter 3 Not Completed
Data Processing (Transforms) Additional data tables and fields may be created in this phase including aggregated and calculated data. * Apply Policy & Strategy approach	Experiment Design Documents Event Data Each experiment design will result in an event log	Data Transforms Code Data Transforms Description Cohort Creation SQL scripts Event Logs Experiment Documentation (partial) Calculate Cohort outcomes SQL Server Services Integration package to export to CSV files CSV/txt Event log files	Code CD (6) See Section 5.3.5 Code CD (7) & (8) Code CD (7) & (8) Not Completed See Sections 7.3.3.11 Code CD (7) & (8) Code CD (7) & (8)
Mining & Analysis * Apply Policy & Strategy approach	Cohorts' outcomes Event Logs	Jupyter Analysis Notebook Process Models Analysis/Critique of our techniques & algorithms Experiment Documentation (Cont.)	Code CD (9) Code CD (7) & (8)
Evaluation	Process Models Domain Expertise Bias Assessment	Improvement Ideas Experiment Documentation (partial for iterations) & later complete documents	
Process Improvement and support Not executed in this research	Improvement Ideas	Process Modifications	Not executed in this research

6.4.4 Policy and Strategy Questions Methodological Approach (RQ4)

An additional dimension of this research is the use of EHRs to investigate the effects of, or to evaluate a policy or strategy decision. During this research a secondary approach emerged to formalise this. This outlines more a way of thinking about answering policy and strategy questions from EHR data rather than a strict method. A summary is proposed in Table 6-3 below.

Steps 1 through 8 of PM4D are carried out as before.

Step 9 of PM4D is replaced by Steps 9.1 through 9.8 below.

Step 10 of PM4D is similar to Step 10 below.

Step 11 of PM4D is similar to Step 11 below.

Table 6-3: Policy and Strategy Questions Methodology

Step No.	General experiment method for using EHR data to evaluation a policy, strategy or decision.	Specific Steps in this research (Sections 7.3.3.1 and Sections 7.4.3.1)
9.1	Identify a situation that represents a policy or strategy change or decision of interest	Varying strategies in HSE South on school dental screenings
9.2	Assemble evidence/documentation this policy happened in fact, and is recorded in the EHR.	Situation Analysis (UCC/HRB, 2005/6)
9.3	Is the policy/strategy visible in the EHR?	See Screening analysis below
9.4	Does the EHR data comply with the policy? If not, can the policy be reliably simulated from the EHR data? Define how this is determined.	Partially. See Screening analysis below.(Sections 7.3.3.1 and Sections 7.4.3.1)
9.5	What are the appropriate outcomes to measure the effects of a policy/strategy?	DMFT, QoL, ICDAS etc. See Background / introduction
9.6	Which of these appropriate outcomes are available from the EHR?	DMFT.
9.7	With the objective of ensuring cohorts are from a level playing field, identify potential exposures, outcomes, confounders and mediators and mitigate if possible.	Establish the baseline DMFT (2007) for these cohorts. These should be Caries-Free (DMFT=0).
9.8	Develop the specific RQs around the policy/strategy, answerable with the EHR data	Is there a different health outcome or treatment process for the patients subject to the policies?
10	General data processing as in Step 10, Part 1 of PM4D. Define cohorts on all sides of the policy/strategy or decision & Create Event Logs as in Step 10, Part 2 of PM4D.	See sample Cohort Selection Code in Appendix 10.12
11(a)	Results: Establish the outcomes for these cohorts with Data Analysis as in Step 11, Part 1 of PM4D	See DMFT outcomes in Sections 7.3.3.11
11(b)	Results/Discussion: Are the outcomes different for the cohorts?	See Outcomes analysis in Sections 7.3.3 and 7.4.3.
12(a)	Results: Establish the treatment process model experienced by the cohorts as in Step 11, Part 2 of PM4D	See Process mining outputs in Sections 7.3.3.12 and 7.4.3.12
12(b)	Results/Discussion: Are the treatment process models of adequate quality? If not, Iterate to Step 10. Are the treatment processes different for the cohorts?	See discussion section in 7.3.3 and 7.4.3.

This can be summarised as in Figure 6-10 below.

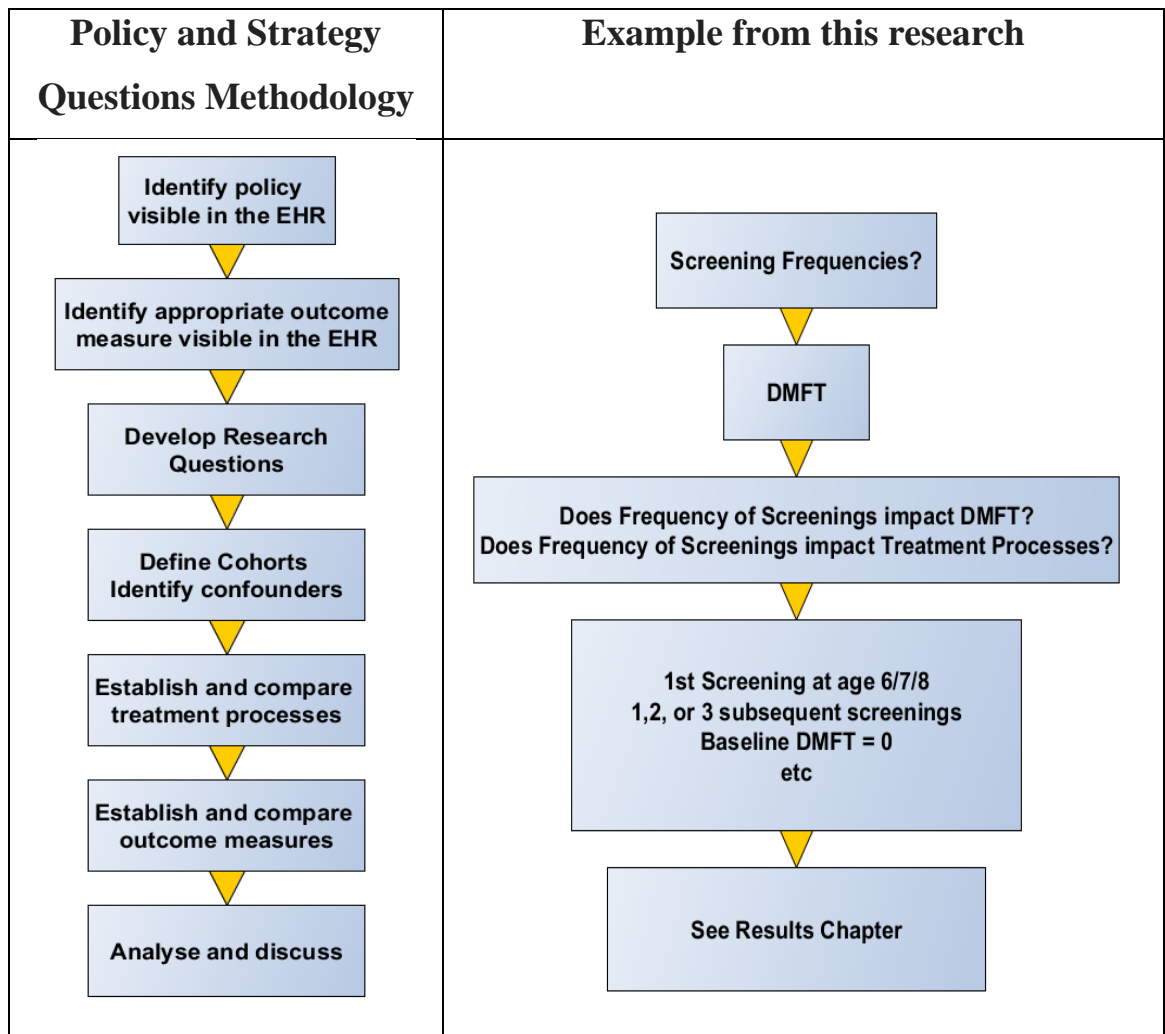


Figure 6-10: Policy and Strategy Questions Methodology with Example

6.4.5 Conclusion

The published PM methods were reviewed and analysed for their strengths and limitations and how they could be applied to this research. The requirements of PM a dental EHR in a research environment necessitated additional steps and the experience of this research is described in Section 6.4. PM4D identifies in detail the inputs and outputs for each step, the artefacts created and where they are to be found in this research and provides a structured approach and checklist for future research in this area. It is important to note that the steps were not necessarily executed in a strict sequence and there were several iterations to fully document the method. A secondary approach was necessary for assessing the impact of strategy or policy changes in an EHR presented in Figure 6-10 and Table 6-3.

With the methodology in hand, the research questions became the next focus.

7 Validation of the Methodology: Experiments and Results

7.1 Introduction

Validating examples were required to answer research questions 1 through 4.

- Can PM help assess compliance with recommended care pathways and clinical guidelines (CGs)? This addresses RQ1 and RQ2.
- Can PM establish the treatment pathway preceding a specific outcome? This addresses RQ3.
- Can analysis of the EHR assess ‘frequency of school screening’ policies? This addresses RQ4.
- Can analysis of the EHR assess the impact of ‘age at first school screening’? This addresses RQ4.

7.1.1 Assessing Compliance with Care Pathways and Clinical Guidelines

7.1.2 Introduction and Aims

Can PM help assess compliance of real-world processes with recommended care pathways and CGs? It aims first to establish if PM can discover pathways from dental EHR data addressing RQ1. Second, it aims to see if those discovered pathways are comparable with established care pathways and CGs addressing RQ2.

In summary, this investigates PM’s potential to assess compliance of the real-world *de facto* processes in our research dataset with established *de jure* processes from the literature. The main objective is to discover the treatment processes experienced by the cohort, present them in a comprehensible format and thereby get an overview of PM’s abilities with the dataset.

First, the care pathway proposed in the Steele report (NHS England, 2009) is considered. Then, two examples from the Irish Oral Health Services Guideline Initiatives (2010; 2012) are considered. For various reasons, these *de jure* processes being used in this research have not been implemented in the HSE and the following examples are therefore hypothetical in nature. However, they serve to investigate whether or not the research data is at the correct granularity, i.e. the correct level of detail to facilitate such analyses. The experiments also have the potential to indicate to what degree the service follows the guideline and, if the guidelines were to be implemented in the future, would serve as an initial benchmark. If the data proves to be suitable for PM then this should encourage the collection of data with tooth details and treatment detail as was done pre-2006 in the NHS, as opposed to the aggregated band-level data currently being collected and would facilitate such PM research.

7.1.3 Success Criteria

The success criteria for this question are twofold:

First, the research data must be suitable for creating PM event logs and producing models recognisable and comprehensible to our PM and dental domain experts. Second, it is desirable that the models be comparable with the established care pathways and clinical guidelines allowing insight as to the degree of compliance of the *de-facto* models.

7.1.4 Steele report (NHS England, 2009)

This pathway starts with a new patient visiting the dentist. They are attending either for routine care or for urgent care requiring pain relief. After pain relief is administered, ‘Urgent’ patients are encouraged to undergo an assessment of oral health, thereby joining the same pathway as the ‘Routine’ patients. After the oral health assessment, patients receive disease management, routine management and risk-based recalls. This pathway is also the gateway to advanced dental care and is shown in Figure 7-1 below.

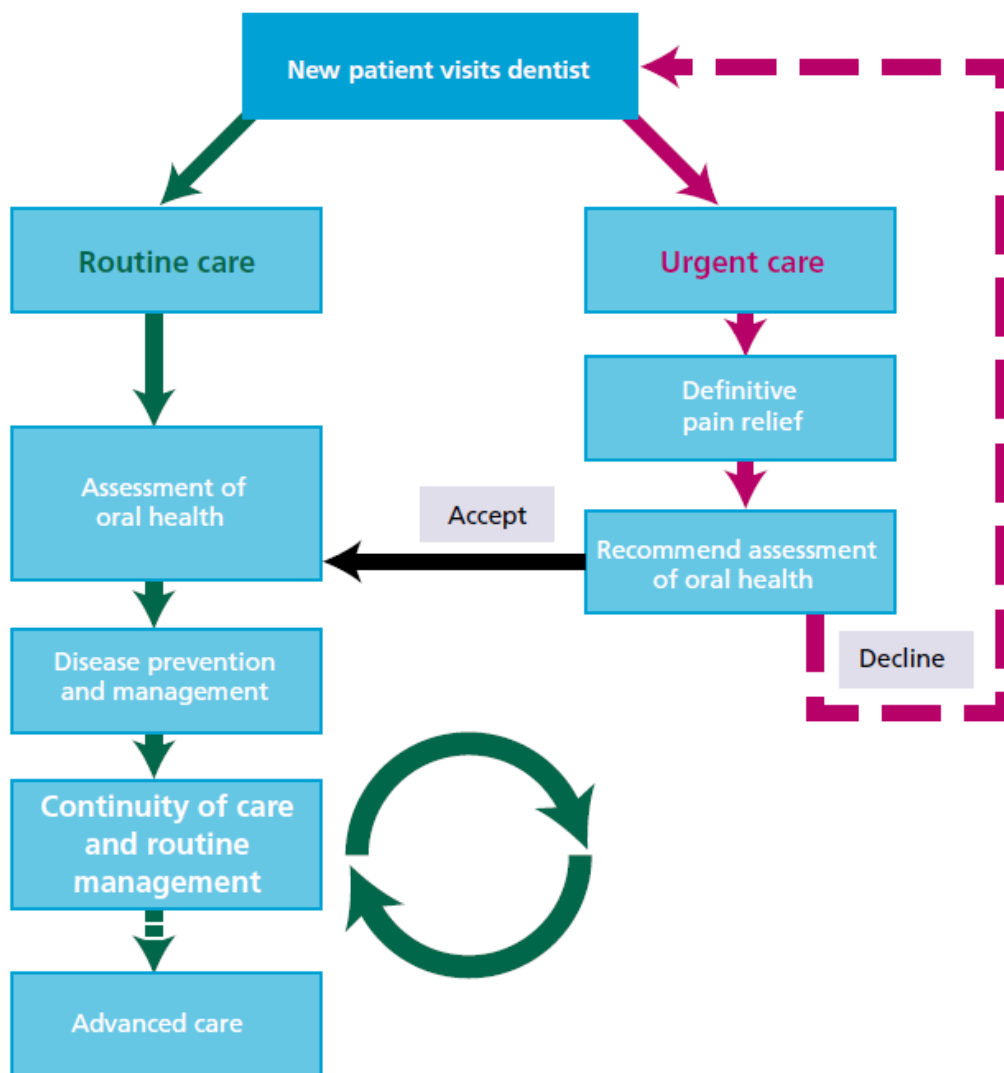


Figure 7-1: Proposed Dental Care Pathway (NHS England, 2009, p. 45)

7.1.4.1 Process Mining Method

Following the PM methodology steps from Section 6.4.3.

Steps 1 through 8 are the general preparatory steps followed in this research, common to all RQs and have been completed earlier in the research as detailed in Section 6.4.3.

Step 9, defining the detailed RQs 1&2, is addressed in Chapter 3.

Step 10, Part 1, General data processing to facilitate RQs is in Appendix 10.18.

Step 10, Part 2, RQ-specific data processing is now addressed.

Phase 1: Cohort Creation

One week of activity was used to demonstrate the comparability of the EHR data and PM outputs to the *de jure* pathway. The cohort was defined as follows:

It was the patient's first visit to the dental service. They received either an Initial Exam (Routine) or Emergency Appointment (Urgent). This took place between 1st Sept 2007 and 5th Sept 2007. Data quality was OK.

A table called CohortCarePathway with the ID for each of these patients was created.

Phase 2: Event Log Creation

- All subsequent treatment process events experienced by these patients were then extracted and exported to a csv/txt file. The minimum required data elements to carry out this experiment were ClientID (Case), ProcedureName (Event), and CompletionDate of each treatment event (Timestamp).
- Convert the csv/txt file to an XES formatted EL using Disco functionality.
- To provide an overview of the EL, the fundamental statistics around cases and events are summarised in Figure 7-2 . Notable is the high proportion of variants, typical of healthcare processes, known for their flexible nature. Of the 88 cases, 86 followed unique pathways.

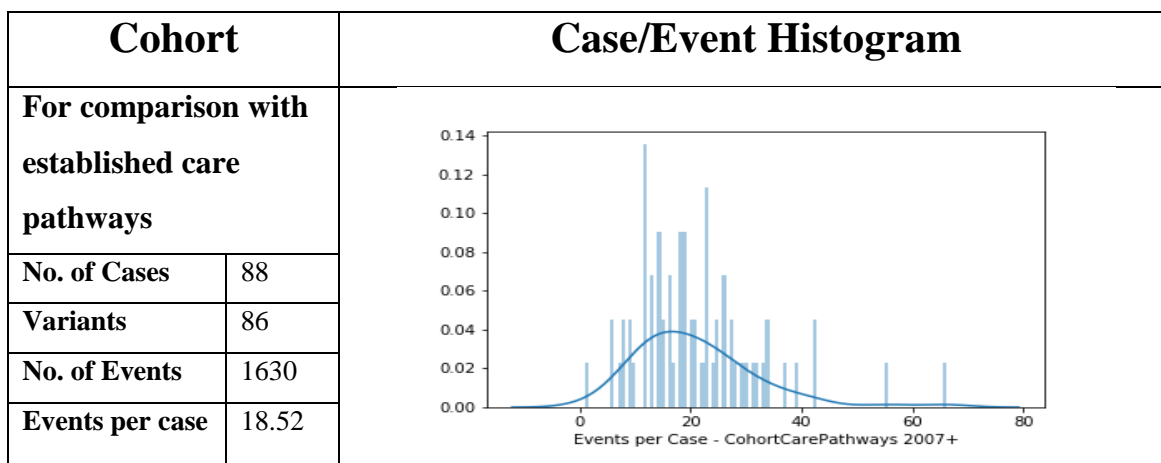


Figure 7-2: Event Log Characteristics (Care Pathway Compliance)

Step 11: Data Analysis and Process Mining

Part 1: Data Analysis

No supplementary data analysis was necessary for this experiment.

Part 2: Process Mining results and output

Using the complete dataset yielded an incomprehensible spaghetti model as would be expected with the high number of variants. Consistent with this research's strategy to carry out all filtering at the cohort and event log creation phase, the event log was re-generated omitting less frequently occurring events. A number of tests were carried out using various thresholds for inclusion of an event. This was done cognisant of our guidelines for legibility and comprehensibility (See Section 5.4.2). Restricting the EL to events occurring more than 20 times for the cohort yielded process models within the guidelines.

This model with both 'Routine' and 'Urgent' is shown in Figure 7-3 below. Viewing sample 'Urgent' and 'Routine' patients in isolation gives us a concise view of the different paths being followed as presented in Figure 7-4 & Figure 7-5.

The darker coloured boxes (events) indicate higher frequency of execution of these procedures and the heavier arrows indicate the most travelled pathways. The larger font number within the box indicates the number of patients (cases) receiving the treatment and the smaller font number within the box indicating the number of times the treatment was executed reflecting that a patient may receive a treatment on multiple occasions.

The default settings for this PM tool aim to present the main features of the dataset i.e. the most frequent activities and the most frequent paths.

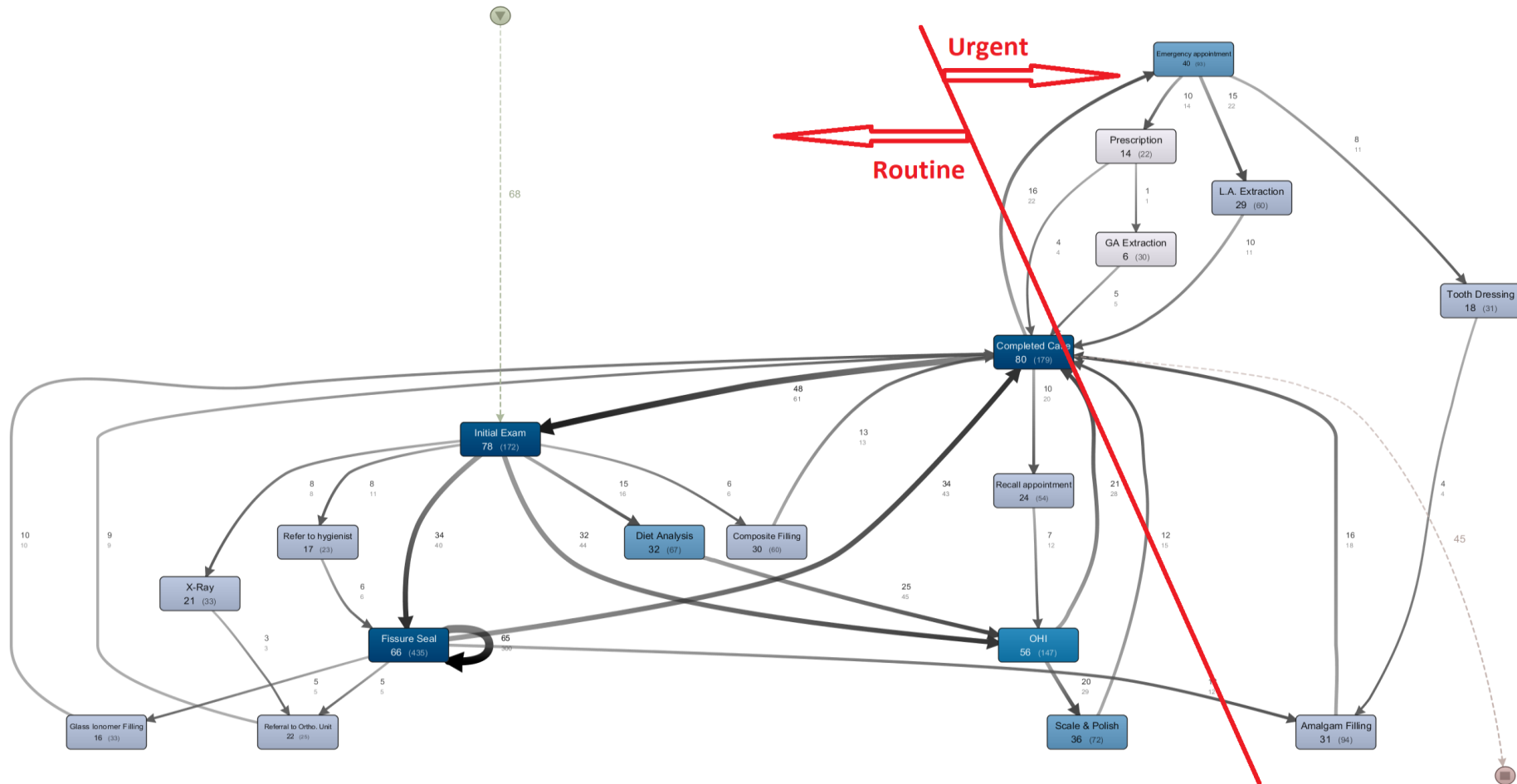


Figure 7-3: Routine and Urgent Care Pathway generated from a single week of data.

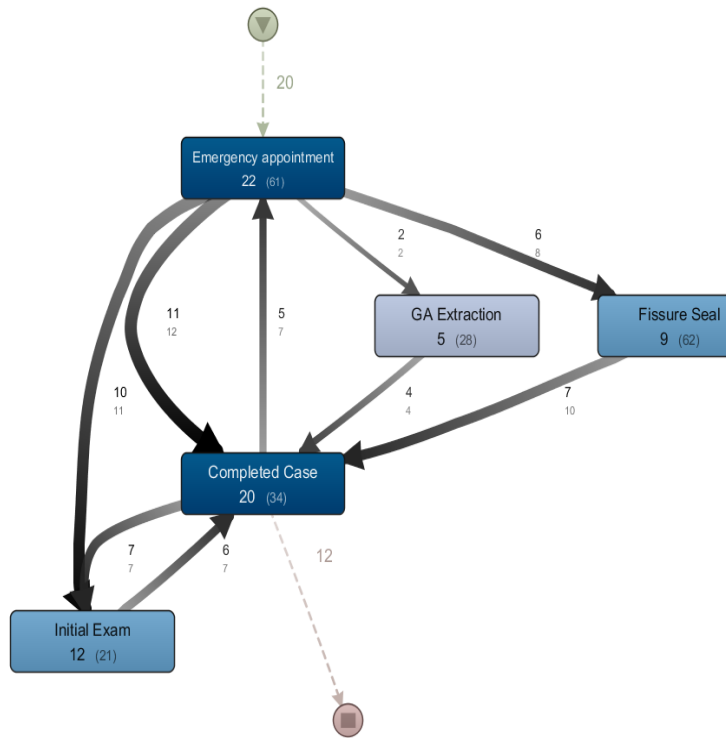


Figure 7-4: Urgent Care Pathway generated from a single week of data.

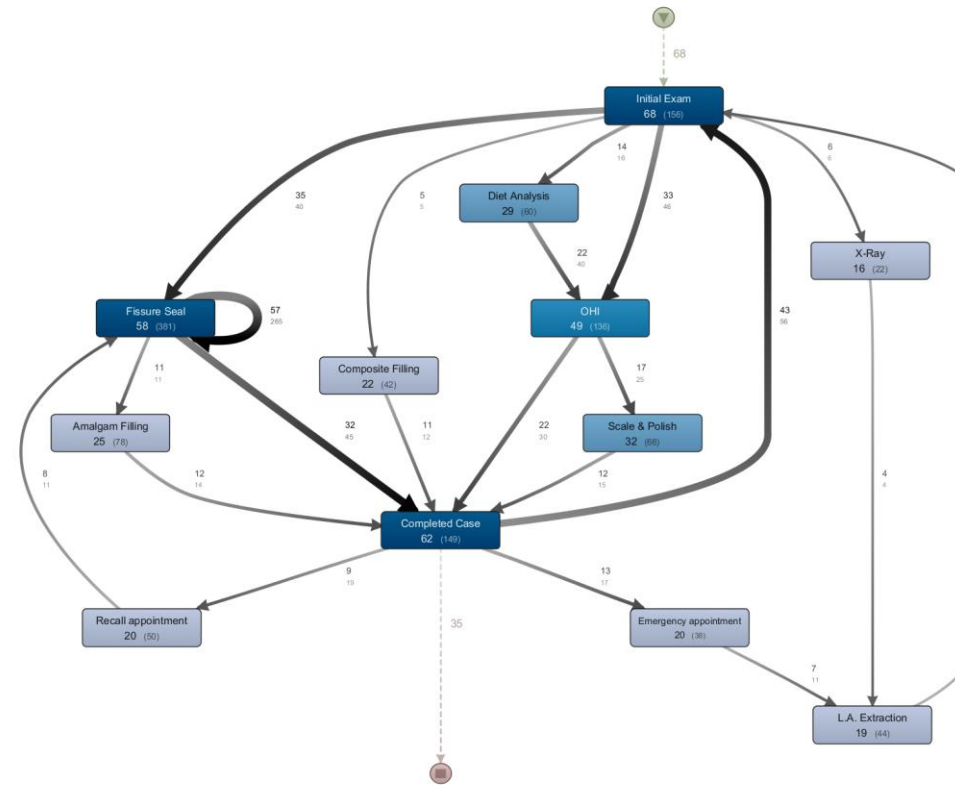


Figure 7-5: Routine Care Pathway generated from a single week of data.

Step 12: Evaluation and Discussion

The overall model (Figure 7-3) generated had the following key characteristics:

- It showed an obvious demarcation between paths followed by patients whose initial visit was ‘Routine’ and those whose initial visit was ‘Urgent’.
- Patients presenting as an emergency had predominantly restorative treatments such as fillings, extractions, and dressings.
- Patients presenting for a routine examination had predominantly preventive events such as fissure sealants, oral health instruction etc.

The models showed that 66% of patients presented for ‘Routine’ treatment and the balance for ‘Urgent’ treatment. From the ‘Urgent’ pathway, over 50% subsequently re-joined the ‘Routine’ pathway having a routine initial examination. This is shown in Figure 7-4 where it can be seen that 12 of the 22 patients presenting for an emergency appointment had a subsequent initial examination. This is not obvious from Figure 7-3 as it is impossible to see whether those having a routine examination had entered the service through the emergency pathway, proving the necessity to present all three views.

Although the cohort was generated from a single week of initial visits, the event log contained all of this cohort’s subsequent encounters with the service and could therefore identify those who re-attended the HSE dental service in the routine stream at a later date. It is also possible that members of this cohort could have had additional dental treatments outside the HSE service i.e. in private dental practices.

Addressing the RQs:

Can PM discover care pathways from the dental EHR?

Addressing the RQs involves several practical questions. Is the data in the EHR comparable with the care pathway? i.e. are the treatment events recorded in the EHR similar to the steps indicated by the care pathway? Do they use a similar terminology? Are they at a similar granularity or degree of detail?

Answering these questions for this research dataset: The treatment events recorded in the EHR have timestamps and can be ordered into a process model. Using the complete dataset generated from the cohort’s EL, an incomprehensible, spaghetti-type model was generated. Therefore, the data was restricted to events occurring more than 20 times for the cohort. This threshold yielded models that fitted our criteria for comprehensibility as detailed in Section 5.4.2 and shown in Figure 7-3, Figure 7-4, and Figure 7-5 above. Even though these models have been simplified as detailed above, Figure 7-3, the most complex

model, is barely legible on A4 paper. When represented as portable graphics network (.png) format file, all models can be enlarged on-screen or for printing as needed without loss of definition.

The events in the models are similar to the steps in the care pathway. ‘Initial Exam’ and ‘Emergency Appointment’ can be associated with ‘Routine’ and ‘Urgent’. While the granularity is finer in the EHR data, the similarities between the terms is apparent. The preponderance of preventive measures such as ‘Oral Health Instruction’(OHI), ‘Fissure Sealant’, ‘Dietary Analysis’ in the path followed by ‘Initial Exam’ i.e. routine patients is in contrast to the predominantly ‘pain-relief’ measures such as prescriptions, fillings, and extractions for the emergency patients. The research data appear generally suitable for PM are recognisable and comprehensible to our PM and dental domain experts. They are legible on paper and on the computer screen? The nodes & arcs can be identified, isolated, and understood. The spaghetti-type model can be simplified for comprehensibility.

Does PM the EHR data produce a useful process model? Are the pathways similar to the recommended pathway? Are the relationships between events similar?

It can be seen that the technologies used, when tailored to the research data, produced models that were comprehensible and legible as in Section 5.4.2. The pathways in the models corresponded closely with the recommended pathway from the Steele Report. While it is unclear whether there is a target for desirable proportions of ‘Routine’ and ‘Urgent’ presentations, it is easy to calculate these from the discovered process models and to identify deviations from the ideal process if required. These technologies offer insights unique to PM, namely, the ordering of events into comprehensible process models which identify the frequencies of use of the pathways and facilitate comparison with ideal models. PM can tell us if patients are following this care pathway and show that over 50% of emergency patients re-joining the routine care pathway as hoped.

This shows the plausibility of using such technologies to monitor compliance of real-world activities with the desired care pathway. Such models could be of value in planning and monitoring care pathways towards improved oral health outcomes.

There a number of important limitations in this experiment.

- This care pathway has not been implemented as policy at this dataset’s source. The objective of this experiment is to assess PM’s potential to evaluate compliance with such pathways using EHR data with similar characteristics to this research’s dataset.

- It is possible that the patients arriving for an emergency appointment also had an Initial Exam incorrectly registered on that day giving a misleadingly high percentage of patients re-joining the 'Routine' pathway. This may explain the discrepancy in numbers attending in Figure 7-3, Figure 7-4, and Figure 7-5. This is a DQ issue in Section 10.17.
- It is possible that the selection of dates in September i.e. the start of the school year, could have an impact on the numbers presenting for emergency appointments.
- It is possible that members of this cohort could have had additional dental treatments outside the HSE service i.e. in private dental practices.
- A shortcoming of the Fuzzy Miner used to create the process models is that the formal measures of process model quality i.e. fitness, precision, simplicity, and generalisability, are not calculable on fuzzy models. These models are a best-guess with an emphasis on graphically emphasising the most relevant behaviour, by calculating the relevance of activities and their relations.
- A shortcoming of the process models presented here is that the sum of the numbers of cases on the arcs is sometimes less than the number of cases in the originating node. This can be confusing but is a direct result of the fuzzy miner eliminating infrequent paths or noise. This enhances the comprehensibility of the models. This can also manifest as the number of cases in the originating node being more than the sum of the subsequent nodes. It is of course possible that small numbers of important cases are omitted in this fashion and accordingly, caution should be exercised when interpreting such models.

7.1.5 Irish Oral Health Services Guideline Initiative (2012, pp. 6,7)

The proposed best practice approach for promoting, protecting, and maintaining the oral health of school-aged children in Ireland is shown in Figure 7-6 below. Most of the distinct steps in the model are present as treatment items in the BridgesPM1 data extract: a medical questionnaire is completed for each patient, a clinical examination (initial exam) is carried out, caries risk assessment functionality, though available in the application software, did not go through the process of adoption; caries prevention instruction (oral health instruction) exists as a treatment event as do fluoride varnish application, glass ionomer, fissure sealant and recall. The caries risk assessment tool captures the variables which help to categorise a patient's caries risk profile. A clinician's judgement without the use of a tool is also valid and is a widely practiced approach. However, such a tool serves as a reminder to the clinician to consider all the most relevant variables, and although helpful it is not an essential requirement for categorising patients according to caries risk.

The decision point ‘Moisture control adequate’ is not a treatment event in this research’s dataset, however it may be present in free text notes not extracted for this research. The dataset is at the correct granularity level to assess compliance with this guideline and if the ‘Caries Risk Assessment’ functionality in the Dental EHR had been implemented and if it were possible to deduce if moisture control was adequate then, compliance checking with this level of guideline would be feasible using this research’s EHR data extract. Notwithstanding that there are some steps in the *de jure* process unavailable in the research data, it is nonetheless clear that the data is at the appropriate level of detail to facilitate comparison to the guideline. Given that the guideline was not implemented, the author has not created cohorts and process models in this case.

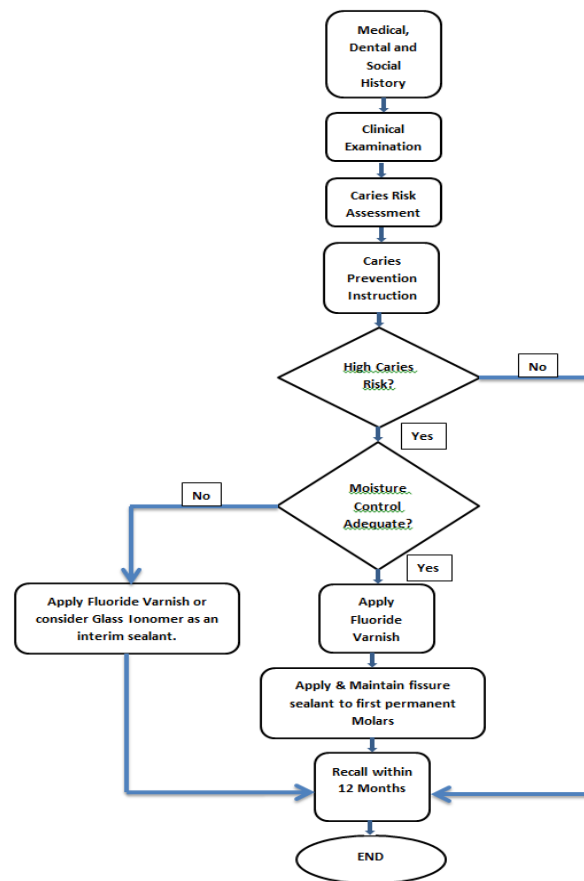


Figure 7-6: Oral Health Assessment Program Proposal (adapted from Irish Oral Health Services Guideline Initiative (2012, pp. 6, 7))

7.1.6 Fissure Sealant Cycle (Irish Oral Health Services Guideline Initiative, 2010)

The high level of detail present in the guideline for the Fissure Sealant Cycle (Figure 7-7) presents some additional challenges for PM of the EHR data. As in Section 7.1.5 above, this guideline requires a caries risk assessment, and this is followed by several additional decision points. While some of the details required to follow the guideline are collected by default during the oral examination, e.g. Sealed/ Sound/ Demineralisation/ Suspicious,

many other points in the guideline require additional documentation at the time the fissure sealant is being carried out e.g. Sealant intact? Adequate Moisture Control, Caries into dentine, X-ray required, and choices of treatment. While the guideline is clearly valuable in defining the process itself, for PM to be of value in assessing whether it is being complied with, much of this detailed information would have to be explicitly documented for each tooth assessed for fissure sealant. While the clinician may be processing the clinical clues or information in their own mind, much of the detail may not be documented at a tooth level. It is unclear whether collecting this additional detail would be practical or not. The benefit of collecting this level of detail on an ongoing basis is questionable. It appears that this would be a time-consuming requirement, could slow down practitioners and increase appointment times for patients. A time-and-motion study or similar method could give a clear indication of these effects. Without this information, such detailed data might be better collected on an occasional basis perhaps as an audit tool or to address specific research questions. Given that the guideline was not implemented, the author has not created cohorts and process models in this case.

7.1.7 Conclusions

Can PM discover pathways from dental EHR data addressing RQ1. If so, are these discovered pathways comparable with established care pathways and CGs?

This experiment has shown that PM is capable of producing process models from the dental EHR data, that are recognisable to dental domain experts and PM experts and comply with the requirements of comprehensibility in Section 5.4.2.

PM showed us to what extent patients are following the care pathway from the Steele Report and also showed that over 50% of emergency patients re-joining the routine care pathway as hoped. From the explorations in Sections 7.1.5 and 7.1.6, it can be seen that the level of detail of data contained in the BridgesPM1 dataset is at a higher level than that required for comparison with these Fissure Sealant Clinical Guidelines but it would appear to be at an appropriate level for The Oral Health Assessment Program Proposal.

In conclusion, this experiment showed PM's ability to discover care pathways from the research data and that these models are comparable with established care-pathways and CGs. This shows the potential for using such technologies to monitor compliance of real-world activities with the desired care pathway. It is also worth considering that such models would be of value in planning and monitoring care pathways towards improved oral health outcomes.

Fissure Sealant Cycle

The use of pit and fissure sealants for high caries risk individuals or groups should form part of an overall caries preventive programme, which includes advice on home care, with a focus on twice-daily tooth brushing with fluoride toothpaste containing at least 1,000 ppm fluoride and appropriate dietary advice. Maintenance of fissure sealants is important to ensure their continued effectiveness, and sealant integrity can be assessed at recall. It is recommended that the recall interval for high caries risk children should not exceed 12 months.⁶⁰

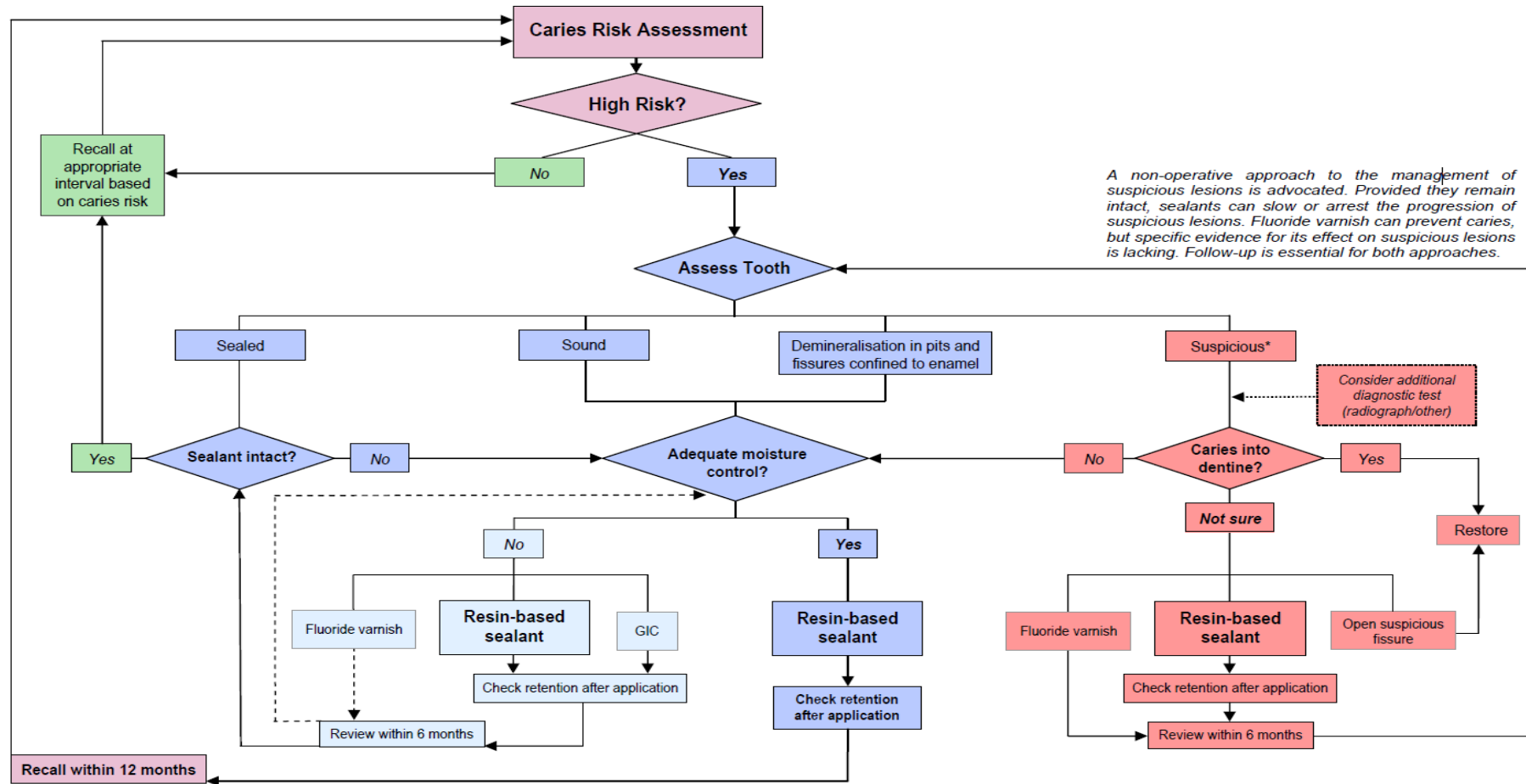


Figure 7-7: Fissure Sealant Cycle (Irish Oral Health Services Guideline Initiative, 2010, p. 6)

7.2 Establishing the Treatment Pathway for a Specific Outcome

7.2.1 Introduction and Aims

Can PM establish the pathway preceding extraction under general anaesthetic (GA)?

Sometimes it may be of interest to simply establish the steps surrounding an event of interest. This experiment studies GA extractions (GAX) aiming first to establish if PM can discover treatment pathways preceding then from dental EHR data and addressing RQ3. Second, it aims to see if those discovered pathways yield useful insights.

This is an interesting topic as it is an expensive, resource-intensive intervention, traumatic for patients and should be avoided if possible. In 2015/16 approximately 43,700 children were admitted to hospital in England for the treatment of dental caries and in most cases for the extraction of multiple teeth, at a cost of £30m (Knapp, et al., 2017), although it unclear if these were in-patients or day-cases. The numbers in Ireland are less clear with the Irish Dental Association (1977) claiming that 10,000 were admitted to hospital for GAX and reports (RTE, 2015) from the Department of Health suggesting that the figure was around 3,600 per year. Proportionally, this would be in line with the English figures. Undoubtedly it is a financial burden on the health services and additionally, the procedure carries risks of morbidity, particularly nausea, pain and bleeding, and occasionally mortality. It is also a traumatic experience for the child and family (Knapp, et al., 2017). It has been suggested that these numbers could be reduced if children were seen earlier and more frequently by dental professionals for prevention and early intervention. Studying the events preceding GAX has the potential to inform these debates.

7.2.2 Success Criteria

The success criteria for this question are twofold:

First, the research data must be suitable for creating PM ELs and producing models recognisable and comprehensible to our PM and dental domain experts. Second, it is desirable that the process models deliver insights on the events leading to a GAX.

7.2.3 Methodology

Following the PM methodology steps from Section 6.4.3.

Steps 1 through 8 are the general preparatory steps followed in this research, common to all RQs and have been completed earlier in the research as detailed in Section 6.4.3.

Step 9, defining the detailed RQ, is addressed in Chapter 3.

Step 10, Part 1, General data processing to facilitate answering RQs is in Appendix 10.18.

Step 10, Part 2, Data processing to facilitate answering RQ3 is now addressed.

Phase 1: Cohort Creation

- Teeth extracted under GA between 1-Jan-2004 and 1-Jan-2014 were selected with a PM ‘case’ being the combination of ClientID and ToothType. Other treatment events in the database which may be related to GAX, Specifically, treatment items ‘Refer for general anaesthetic’, ‘Refer for oral surgery’, and ‘Surgical extraction’ are present in significant numbers in that timeframe. Their relationship to GAX has not been investigated. No distinction was made between permanent and deciduous teeth.
- Data quality was OK.
- A table, CohortGA, containing case IDs was created in BridgesPM1.

Phase 2: Event Log Creation

- All subsequent treatment process events experienced by these teeth were then exported to a csv/txt file. The minimum dataset was the combination of ClientID and ToothType (Case), ProcedureName (Event), CompletionDate of treatment (Timestamp).
- The csv/txt file was converted to an XES format EL using Disco functionality.

Step 11: Data Analysis and Process Mining

Part 1: Data Analysis

The profile of the dataset’s GAX and prescriptions in that timeframe is presented in Figure 7-8, Figure 7-9, and Figure 7-10 below. It shows the rate of GAX peaking at ages 5 & 6 and prescriptions peaking at 7 & 8. Where the prescriptions could be associated with a patient who ultimately received a GAX, the ages again peaked at 5 & 6, which would be expected. The dataset has no detail on the direct reasons for the prescription.

Part 2: Process Mining results and output

The objective is to discover the treatment processes experienced by the cohort, present them in a comprehensible format and get an overview of PM’s abilities with the dataset. Using the complete dataset yielded an incomprehensible spaghetti model as would be expected with the high number of variants. Consistent with this research’s strategy to carry out all filtering at the cohort and event log creation phase, the event log was re-generated omitting less frequently occurring events, cognisant of our guidelines for legibility and comprehensibility in Section 5.4.2.

A number of tests were carried out using various frequency thresholds. Restricting the EL to events occurring more than 20 times for the cohort yielded process models in Figure 7-11 - Figure 7-13 below.

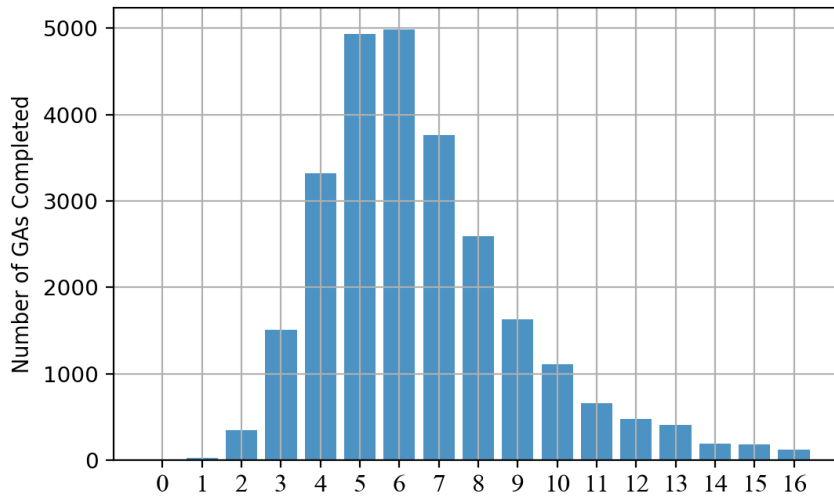


Figure 7-8: Number of GA Extractions by age (2004-2014)

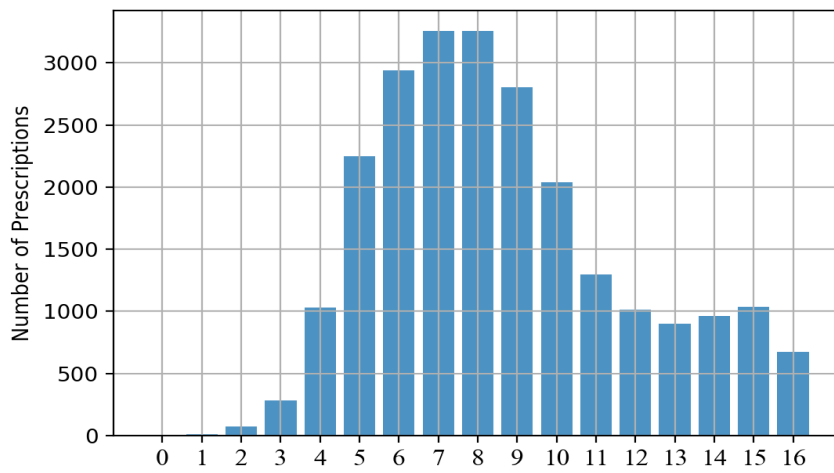


Figure 7-9: Number of Prescriptions by age (2004-2014)

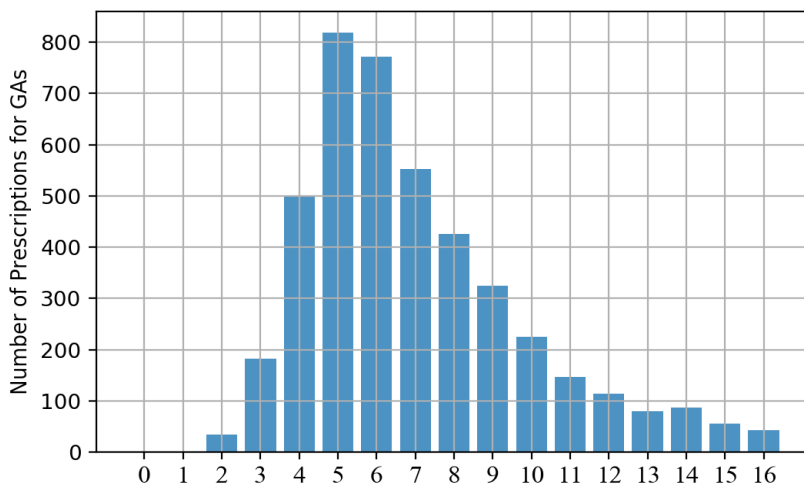


Figure 7-10: Number of Prescriptions by age - followed by GA (2004-2014)

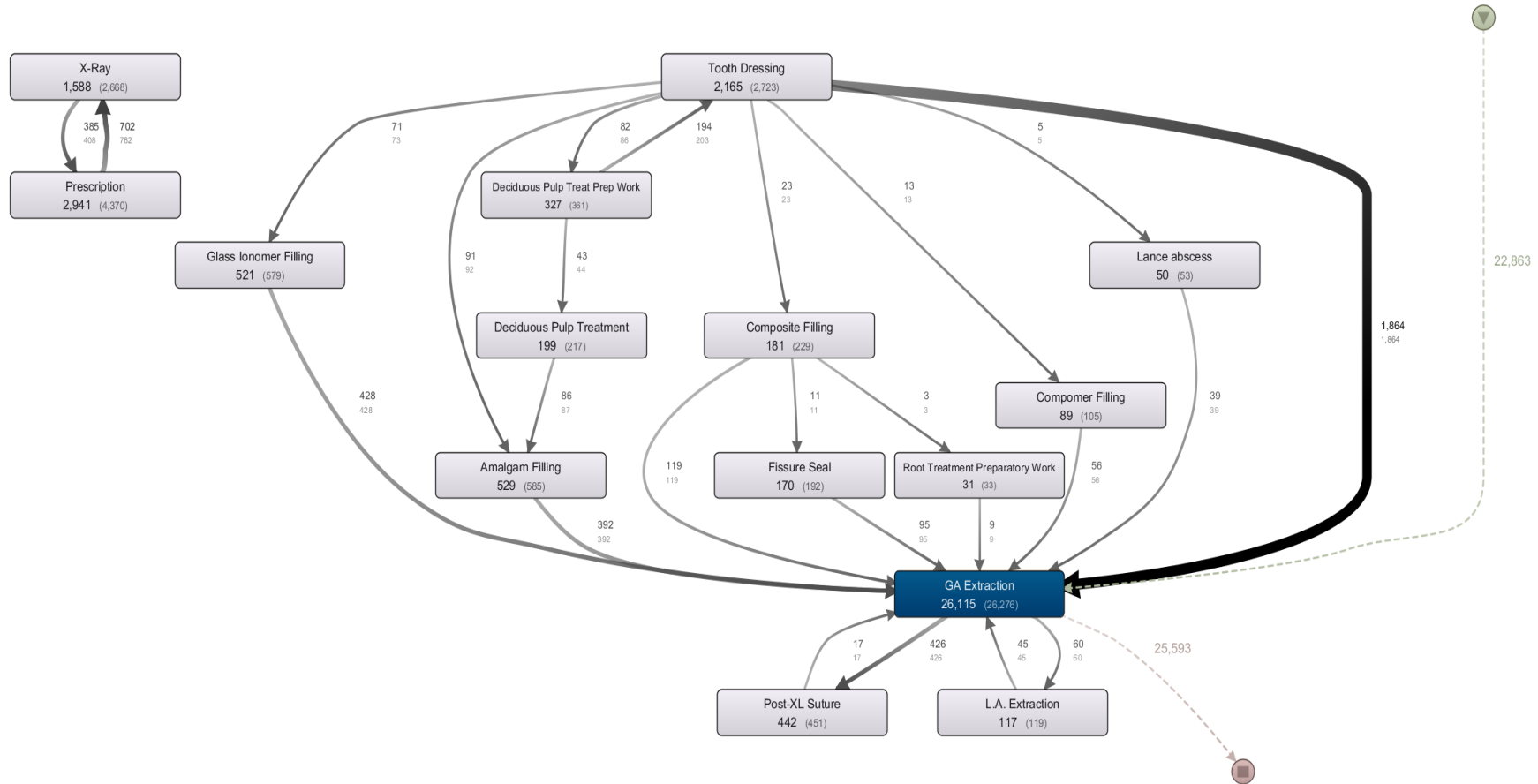


Figure 7-11: Process mining frequency analysis of General Anaesthetic Extractions. Temporal sequence for teeth extracted under general anaesthetic between 2004 and 2014 and all preceding events.

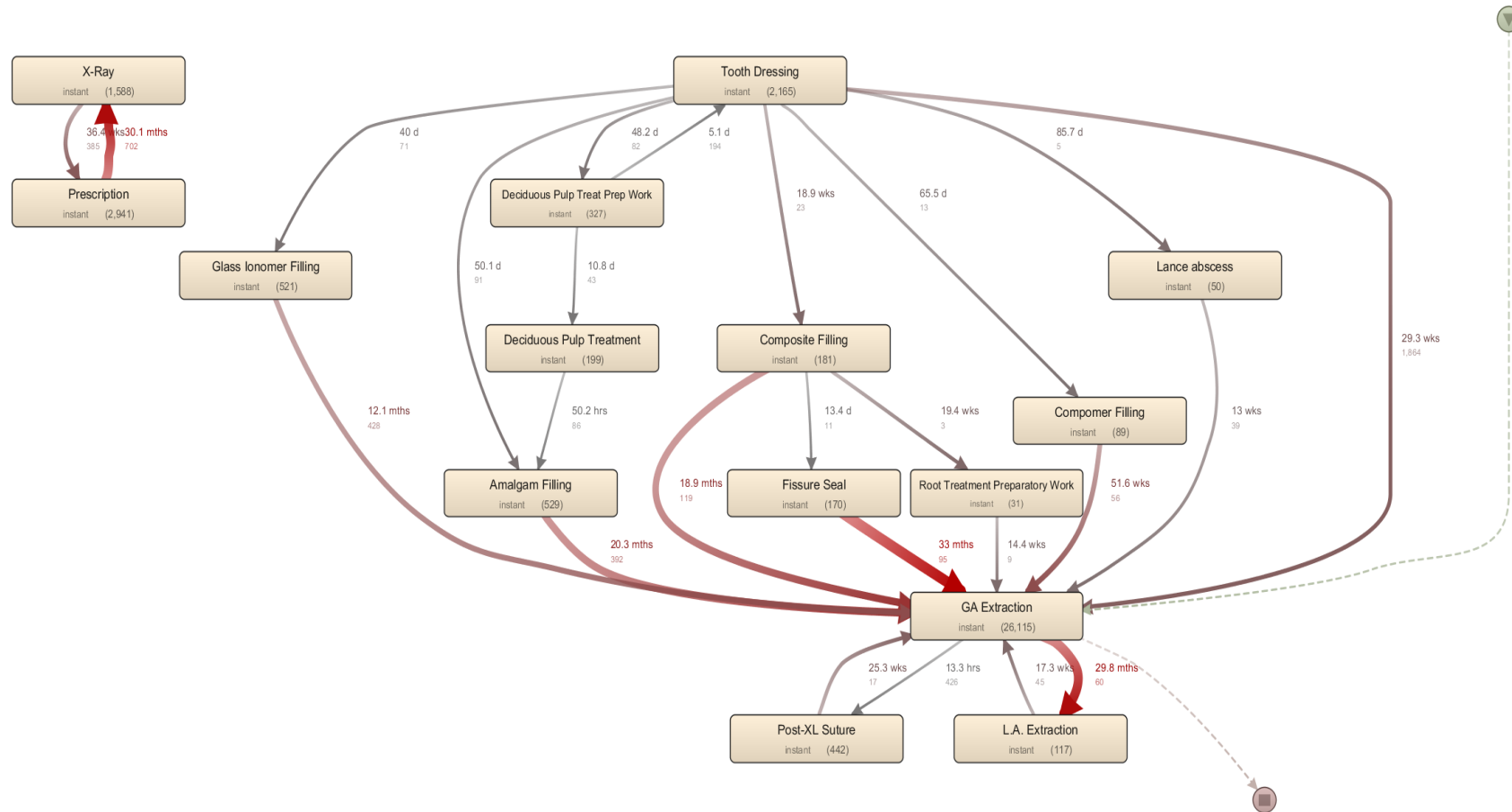


Figure 7-12: Process Mining performance analysis of General Anaesthetic Extraction. Temporal sequence for teeth extracted under general anaesthetic between 2004 and 2014 and all preceding events.

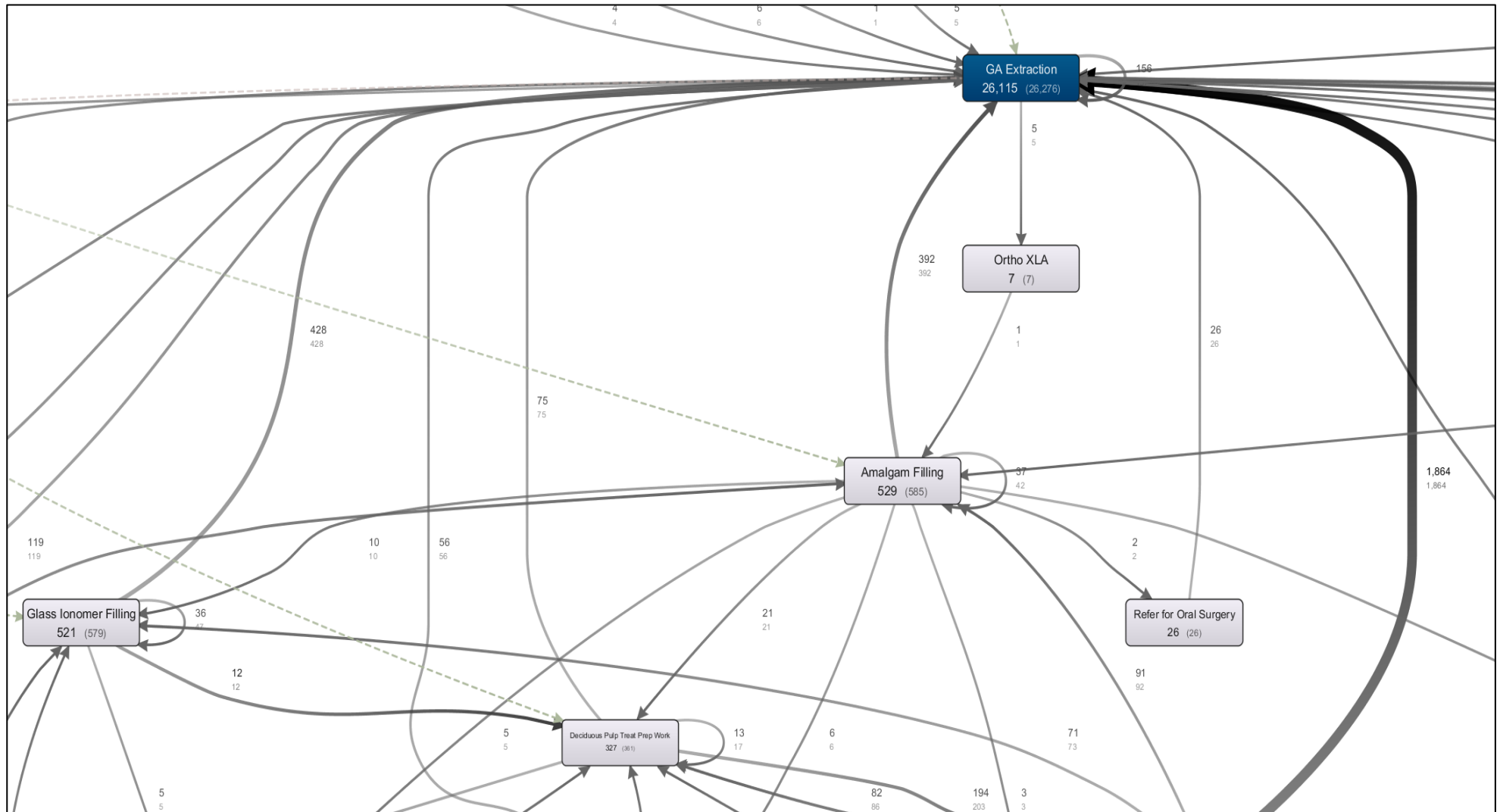


Figure 7-13: Detail of paths taken between Amalgam Filling and GA Extraction

Step 12: Evaluation and Discussion

Can PM discover the treatment pathway for a specific outcome e.g. before GAx?

Key features of the process models:

- 26,115 teeth experienced GA extraction (GAx) in this 10-year time-period
- Over 9% (2,165 instances) of the teeth received a tooth dressing before GAx
- Average time between tooth dressing and GAx was 6 months indicating a service under pressure

Figure 7-11 showed that 26,115 teeth experienced GAx in this 10-year time-period and of these 22,863 had GA without any intermediate intervention. These are represented as the light arrow going directly to 'GA Extraction' on the right-hand side of the process model. Over 9% (2,165 instances) of the teeth received a tooth dressing before GAx and, over 7% (1,864 instances) had a prior tooth dressing and no other treatment before GAx. Several restoration events were sometimes in evidence. Of the total number of extracted teeth, 529 teeth received an amalgam filling on average 20.3 months prior to extraction and 392 went directly to GAx without any intervening treatment. The remaining 137 took an alternate path to GAx that is not shown on the process model. Increasing the detail showed these less travelled paths but made the overall model difficult to comprehend. Figure 7-13 shows detail for the paths taken after 'Amalgam Filling' are shown. Of the 137 that did not go directly from Amalgam Filling to GAx, 21 had a Deciduous Pulp Treatment, 10 had a Glass Ionomer Filling and two were referred for oral surgery. On the left of the model, prescriptions and x-rays administered to these patients can be seen. Performance analysis in Figure 7-12 revealed that the average time between tooth dressing and GA extraction was 29.3 weeks. This translates to a six-month waiting-time between the tooth dressing and ultimate extraction of the tooth, suggesting a service that is under strain providing such emergency treatments.

Suggestions for further and better use of these models.

- The utility of these models would be increased if they were generated for comparator groups, e.g. to compare the impact of varying policies for service delivery or compare the outcomes of use of amalgam vs composite material in an age-standardised cohort. The approach has the potential to extract value from the dataset for planning and evaluation of services based on real life processes pathways and outcomes.

- Other clinical applications might include investigation of alternative treatments or comparison of materials and drugs. Such an approach would require random allocation of patients within a prospective study design.
- PM in prospective studies would provide more in-depth evaluation of the comparative effectiveness of an intervention than simple endpoint comparison.
- If answering more specific question was of interest to a researcher, the event log can be tailored to that question providing more detail or zooming in on part of the process.

Limitations of these models

Not incorporating the characteristics of the patient is a shortcoming e.g. the age of the child, previous oral examinations, or treatments of other teeth, no distinction was made between permanent and deciduous teeth which would possibly have different pathways. There is no information on the date of decision of the necessity for the GAx and this could give additional valuable waiting-time detail if available.

There is a slight discrepancy between the total number of GAx and the sum of the teeth in the paths followed in the process model. This is due to the exclusion of very unusual paths and events from the model in order to enhance its comprehensibility. It is also notable that a small number of teeth also are marked as having had a local extraction in addition to the GAx. This clearly cannot be the case and is most likely a data recording error where a tooth number is incorrectly identified.

7.2.4 Conclusions

GAx were studied aiming to establish if PM can discover treatment pathways preceding them from the dental EHR data. It also aims to see if those pathways yield useful insights. In this case the mapping is retrospective and involves a look back before GAx. It is clear that PM technology can show the process of treatment leading to the GAx outcome. The generated models were comprehensible and recognisable to our domain experts and, as in this case, can usefully demonstrate the pathways followed and the waiting times between events. This showed a waiting time of 6 months between a tooth dressing and GAx in many cases. This is a valuable insight to the process. PM could also be valuable in showing the effects or improvements that the addition of resources to a service would have on the outcome and on the intermediate steps leading to that outcome. Furthermore, the technology could be used in practice-based clinical trials using patient randomisation to show the outcomes of treatments or perhaps the outcomes of using certain materials and the intermediate steps.

7.3 Assessing the Impact of ‘frequency of screening’ Policies

7.3.1 Introduction and Aims

Can analysis of the EHR assess the impact of ‘frequency of screening’ policies?

This question aims first to examine whether the oral health outcomes and treatment processes vary between the cohorts that received screenings according to varying policies, delivering insights, and addressing RQ4. There are several sub-aims or objectives to achieve this: establish if the EHR can distinguish between cohorts, establish if the research data can show oral health outcomes for the cohorts, and establish if PM can discover treatment pathways followed by these cohorts.

Exploring this question using EHR data is technically more complex and challenging than those in Sections 7.1.1 and 7.2 and requires application of the Policy and Strategy Questions Methodological Approach from Section 6.4.4 and the steps in Table 6-3.

7.3.2 Success Criteria

The success criteria for this question are as follows:

First, the research data must be suitable for creating cohorts representing the groups receiving various numbers of screenings. It must also be suitable for creating PM event logs and producing models of adequate quality, i.e. recognisable and comprehensible to our PM and dental domain experts. Second, it is desirable that the analysis and process models are shown capable of delivering insights on the significance or otherwise of the frequency at which school screenings are delivered.

7.3.3 Methodology

Introduction

Following the PM methodology steps from Section 6.4.3.

Steps 1 through 8 are the general preparatory steps followed in this research, common to all RQs and have been completed earlier in the research as detailed in Section 6.4.3.

Step 9 utilises the Policy and Strategy Questions Methodology detailed in Section 6.4.4.

7.3.3.1 Step 9.1 - Identify a situation representing a policy or a strategy change .

The first situation looked at is the area of school dental screenings. The preventive value of school screenings and questions around the optimal age for administering such screenings as well as the optimal frequency or recall intervals of such screenings have long been debated. There has been no definitive answer to these questions. Historical EHR data may offer insight into these questions. This RQ profiles a public health database

and produces visualisations of the data showing the process of dental care received by cohorts whose screening history adhered to the policies described in the situation analysis (UCC/HRB, 2005/6) ('Situation Analysis'). Data mining and PM were applied to the dataset to examine their potential for demonstrating how the oral health outcomes and treatment paths followed by these cohorts can be compared and contrasted.

7.3.3.2 Step 9.2 - Assemble evidence of this policy in the EHR.

The primary evidence that this policy/strategy existed is in the findings of the Situation Analysis. An aim of that study was to establish the practice in the Irish Public Dental Health Service in three areas: school dental screening, strategies to prevent caries in high risk children, and the use of topical fluorides in caries prevention. The study established that although there was a targeted approach to dental screening in primary school in almost all areas, there was a wide variation in the programmes, practices and policies and there was considerable uncertainty about the relative effectiveness of these variations. These variations revolve around the choice of which class should be targeted to maximise protection of first permanent molars through examination and preventive therapies such as fissure sealants. The five areas for which EHR data is available took a targeted approach described as in Table 7-1 below.

The data-profile distributions in Sections 0 and 0 look at the EHR data against the findings of the Situation Analysis. The Leyden Report (Department of Health (Ireland), 1988, p. 24) had envisioned an ideal scenario with eligible children receiving annual screenings. However, they acknowledged that the service was resource constrained and would accordingly focus on screening 1st and 6th classes and treating children at high risk in these classes. The National Health Strategy (Department of Health (Ireland), 1994, p. 54) recommended improvements in the school dental services to ensure the systematic screening of children in three classes in primary and post primary schools.

Table 7-1: Classes targeted by area, 2005 (from (UCC/HRB, 2005/6)

Area	Number of classes targeted	Classes Targeted
Kerry	2	2, 6
North Cork	4	1, 2, 4, 6
West Cork	3	1, 3, 6
North Lee	3	1, 3, 6
South Lee	3	1, 3, 6

7.3.3.3 Step 9.3 - Establish whether the policy/strategy is visible in the EHR?

This is a technical question about the data. Questions like: Can the information being stored in the EHR be used to distinguish between groups or cohorts who were on the different sides of the policy? Is the data stored in sufficient detail?

In this case, screenings are recorded as ‘Initial Exams’(IE) in the EHR. The IE consists of entering the details of the patients dental condition in the EHR odontogram as shown in Figure 4-2, creating a list of planned treatments if appropriate, and recording the date they were carried out either in a clinical setting or sometimes in the school. The sequence of screenings received by a patient can be seen, facilitating the creation of cohorts receiving 2, 3, or 4 screenings and the intervals in-between. From this we can see that it is theoretically possible to distinguish between the various sides of this policy in the EHR.

7.3.3.4 Step 9.4 - Does the EHR data comply with the policy?

Does the EHR data demonstrate that the policy was adhered to?

To see whether the EHR data complied with the stated policy/strategy of the areas, the age-at-first-initial-exam distributions in Figure 4-9 was examined to get a feel for the data. All areas targeted 1st class except for Kerry, targeting 2nd class. This is somewhat reflected in the histograms with Kerry the only area with a spike at 9 years of age. The other areas show a spike at 8 years of year, indicating that they started screening a year earlier, except for North Lee which shows a sustained spike over 8 & 9. Although the available dataset started in the year 2000, the profile presented in Figure 4-9 starts in 2006. This gives a more accurate picture of the age at first screening as the first-screening data for 2000-2006 could have included patients who had received their first screening before the introduction of the EHR. In that case, their first screening appearing in the EHR would have incorrectly appeared to have been their first screening.

The next step was to see if the EHR data shows to what extent the stated classes were targeted as planned. To do this, a cohort that appears to have been targeted as initial school screenings was selected. The cohort is defined as follows:

- Screening (Initial Exam) carried out between September 1st of target year (2005) and 31st August of following year (2006).
- This was the first screening for that patient.
- The patient was aged 7, 8, or 9 at the time of the screening.
- The data quality was acceptable.
- Starting DMFT = 0. (DMFT is used throughout as shorthand for D₃MFT)

This identified a cohort in the schoolyear 2005/6 and the same individuals were tracked over the following 5 years to see in which year, if any, they received another screening.

These results are summarised as a level of policy attainment in Table 7-2 below. The level of attainment was calculated by looking at the number of patients seen in the first target year and calculating the percentage of these seen in subsequent target years. This percentage was then increased to reflect the number of patients, due to be seen, but not seen, in the target year, but seen in an adjacent year.

Kerry had a 71% policy attainment level in their 6th class targeted children, having an initial DMFT=0. West Cork had a higher level, with 93% policy attainment in 3rd class and 87% in 6th class. The loss to follow-up in the subsequent years can be due to issues such as the children moving out of the region. Overall, the figures are similar for those children with all DMFT values at initial screening. The shaded blue cells are the years that were not officially targeted according to Table 7-1 above. The detailed data and results for the school years 2005-2010 are in Appendix 10.13, Figure 10-6 to Figure 10-10 and Table 10-1.

Table 7-2: Level of policy attainment* (%) based on the number of patients seen in first targeted year having DMFT=0

	Kerry (2nd & 6th Class)	North Cork (1st, 2nd, 4th, 6th)	West Cork (1st, 3rd & 6th)	North Lee (1st, 3rd & 6th)	South Lee (1st, 3rd & 6th)
Number of patients seen in each area's first targeted screening year i.e. Year1, 100%	581 (2nd Class)	494 (1st Class)	458 (1st Class)	886 (1st Class)	1090 (1st Class)
Year2 % Policy Attained		14			
Year3 % Policy Attained			93	56	72
Year4 % Policy Attained		47			
Year5 % Policy Attained	71				
Year6 % Policy Attained		48	87	66	68

Table 7-3: Level of policy attainment* (%) based on the number of patients seen in first targeted year having all DMFT values

	Kerry (2nd & 6th Class)	North Cork (1st, 2nd, 4th, 6th)	West Cork (1st, 3rd & 6th)	North Lee (1st, 3rd & 6th)	South Lee (1st, 3rd & 6th)
Number of patients seen in each area's first targeted screening year representing 100%	895 (2nd Class)	737 (1st Class)	1188 (1st Class)	1392 (1st Class)	620 (1st Class)
Year2 % Policy Attained		15			
Year3 % Policy Attained			91	59	71
Year4 % Policy Attained		47			
Year5 % Policy Attained	71				
Year6 % Policy Attained		51	87	66	67

7.3.3.5 Step 9.5 - Which outcome to use to measure the effects of a policy/strategy?

DMFT is the outcome measure used in this RQ. No quality of life data is available in the EHR dataset. Insufficient information for ICDAS related assessment is available. See Section 1.3.4 for further justification of this choice.

7.3.3.6 Step 9.6 - Which outcomes are available from the EHR?

DMFT is available in the dataset.

7.3.3.7 Step 9.7 - Eliminating confounding factors in the cohorts

Potential exposures, outcomes, confounders, and mediators need to be identified in order to ensure the cohorts are comparable. Efforts should also be taken to mitigate against these factors if possible. It is clear from Figure 1-8 that there are many factors affecting the development of dental caries and accordingly there are many potential confounders. It is beyond the scope of this research to deal with these in detail. However, the requirement that all members of the cohorts had a starting DMFT=0 is a significant step taken to ensure that the cohorts are legitimately comparable. There is a very strong correlation between age and disease experience therefore, there is likely to be a difference between the demographics of the different age cohorts in relation to their caries risk and the Kerry children being a year older at baseline (i.e. Year 2 instead of Year 1) could possibly result in a higher mean DMFT for those with all DMFT values at baseline.

7.3.3.8 Step 9.8 - Develop the specific Research Questions (RQ)

Is there a difference in oral health outcome (DMFT) or treatment processes between groups receiving 2, 3, or 4 school screenings in areas where this was the stated policy?

In this research, no statistical comparison of the cohorts' outcomes is carried out. The intention of this research was to validate the methodologies developed and not to definitively answer the validating questions. Additionally, critical information relating to confounding factors such as fluoridation status and socio-economic-status was unavailable making comparison with previous studies impossible.

If it were carried out it would often take the form of a Null Hypothesis i.e.: There is no difference in oral health outcome (DMFT) or treatment processes between groups receiving 2, 3, or 4 school screenings in areas where this was the stated policy. Alternative Hypothesis: There is a difference between the DMFT values of the groups receiving 2, 3, or 4 school screenings where this was the stated policy.

7.3.3.9 Step 10, Part 1 - General data processing to facilitate answering RQs is in Appendix 10.18.

7.3.3.10 Step 10, Part 2 - Data processing to facilitate answering RQ4 is now addressed. Phase 1: Cohort Creation - Identify cohorts on different sides of the policy/strategy or decision. In this question, the cohorts were already created in Step 9.4 to establish the extent to which the EHR data complies with the policy. This identified a cohort in the schoolyear 2005/6, for each policy area, and the same individuals were tracked over the following 5 years to see in which years, if any, they received another screening

- Tables called CohortAgeX2005YScreenings containing the ID for each of these cases was created in the QL Server BridgesPM1 database where X is the patient's age (7,8,9) and Y is the number of screenings they received.

Phase 2: Event Log Creation

- All subsequent treatment process events experienced by these patients were then extracted and exported to csv/txt files. The minimum required data elements to carry out this experiment were the ClientID, ProcedureName (Event), and CompletionDate of each treatment event (Timestamp).
- The csv/txt file was then converted to an XES EL using Disco functionality.
- To get an overview of the event logs (EL), the fundamental statistics around cases and events were established. Each case represents a patient and each event represents a treatment item in the patient's EHR. Notable again is the high proportion of variants with almost 100% of the patients following unique pathways. The details are in Appendix 10.13.

7.3.3.11 Step 11, Data Analysis

Step 11(a): Establish the outcomes for the cohorts receiving 2, 3, or 4 screenings

This was executed in the Anaconda/Jupyter Notebook environment. Data analytics cells were written in Python, calling SQL Server functions with the following major steps:

This is an ad-hoc sequence of steps, specific to the RQ, and on Code CD-7.

- Create table to store aggregated data and entries in table for each area and policy.
- Calculate and save the number of patients whose screening history complied with this policy i.e. screening took place at the correct time, in the correct clinic/region; it was their first screening; they were aged 7, 8, or 9; data quality was good.

Calculate and save the number of these same patients who received a screening in each of the following 5 years. Calculate and save the number of these same patients attended for emergency treatment in each of the following 5 years.

- Estimate and save the number of these same patients who did not receive a screening in the intended policy year but did in a year adjacent to the policy screening year.
- Calculate and save DMFT mean and standard deviation for all patients in each of the 6 years.
- Calculate and save the number of patients receiving a screening in the all the intended policy years i.e. patients whose screening history complied exactly with the policy.
- Calculate and save DMFT mean and standard deviation for patients receiving a screening in the intended policy years.

These steps are summarised in Figure 7-14. Full dataset is in Appendix 10.13.

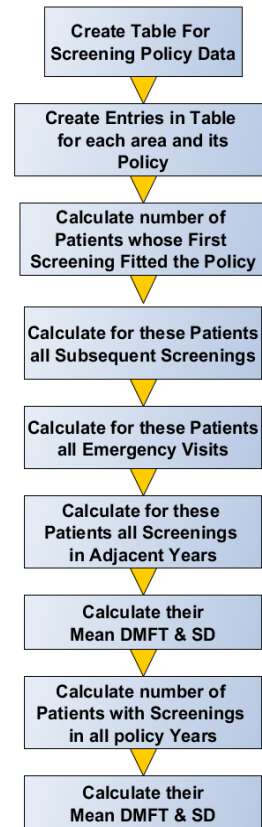


Figure 7-14: Frequency of Screening Outcomes Calculation

Table 7-4: Initial DMFT and Final DMFT for Frequency of Screenings

	Kerry	North Cork	West Cork	North Lee	South Lee
No of Screenings	2	4	3	3	3
Mean DMFT @ age 12/13 where Starting DMFT=0	1.60	1.53	1.08	1.14	0.88
Mean Starting DMFT where Starting DMFT > 0	0.73	0.72	0.47	0.52	0.39
Mean DMFT @ age 12/13 where with all starting DMFT values	2.8	2.97	1.79	1.72	3.36
Difference between baseline and age 12/13 where starting DMFT=0	1.6	1.53	1.08	1.14	0.88
Difference between baseline and age 12/13 with all starting DMFT values	2.07	2.25	1.32	1.2	2.97

One possible explanation for the higher mean DMFT for those with all DMFT values at baseline among Kerry children is that they were a year older at baseline (i.e. Year 2 instead of Year 1). It is interesting to note the difference in mean DMFT by age 12/13 for children who had DMFT=0 and those who had all DMFT values at baseline. Subtracting the baseline DMFT from final DMFT, in all areas the difference in mean DMFT is greater in the group where with all DMFT values at baseline; meaning that those children who

had some caries at the younger age developed more caries over the subsequent years than those who were free of dentine caries at the outset.

Step 11(b)

Is there a difference in oral health outcome (DMFT) or treatment processes between groups receiving 2, 3, or 4 school screenings in areas where this was the stated policy?

In this research, no statistical comparison of the cohorts' outcomes is carried out. Critical information relating to confounding factors such as fluoridation status and socio-economic-status was unavailable making comparison with previous studies impossible.

If it were carried out it would often take the form of a Null Hypothesis i.e. There is no difference in oral health outcome (DMFT) or treatment processes between groups receiving 2, 3, or 4 school screenings in areas where this was the stated policy. Alternative Hypothesis: There is a difference between the DMFT values of the groups receiving 2, 3, or 4 school screenings where this was the stated policy.

7.3.3.12 Step 12, Process Mining

Step 12(a): Establish process models for the cohorts receiving 2, 3, or 4 screenings?

From the data analysis above, marked differences between basic characteristics of the cohorts receiving 2, 3, or 4 screenings can be seen. First, in the area identified in the Situation Analysis as having 4 screenings, only 2 patients in our cohort received 4 school screenings in strict compliance with the stated policy. As this area was the sole area with 4 screenings, it would be invalid to compare this small number to other areas with 2 and 3 screenings. Second, there are marked differences in the average number of treatments received by the cohorts, ranging from 14.37 to 37.5, although the high number refers to the area with only 2 cases. The lowest number, 14.37, is associated with Kerry, the area with the minimum policy of 2 screenings, and the higher numbers (26.63, 23.65, 22.33) with the areas offering 3 screenings as policy. In PM terms this translates to a difference in the average number of events per case (treatment items per patient) and impacts the resulting process models – more events per case leading to higher complexity models.

Important is the high proportion of variants. For example, in the cohort receiving 2 screenings, of the 174 cases, all these followed unique pathways. This is typical of healthcare processes, known for their flexibility and ad-hoc nature and given the large number of different treatment possibilities. In the case of this research's data, most of the treatment names have been abstracted to 'Prevention' and 'Restorative' and it is surprising, given this higher level of abstraction, that the number of variants remains so high. This high proportion of variants can be problematic, leading to spaghetti models.

The default output from Disco for the 2-screening-cohort is in Figure 7-15. Figure 7-16 is the performance model and shows the mean time between events.

Disco allows adjustment of the amount of detail presented in the models. Showing 100% of the paths and activities results in the process model presented in Figure 10-3.

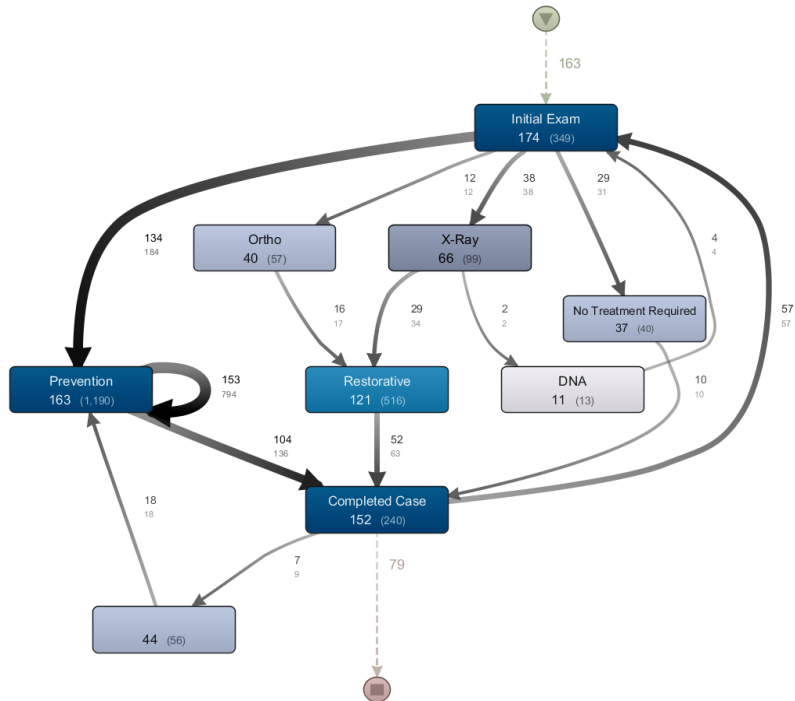


Figure 7-15: Default Frequency Model for 2 screenings. Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.

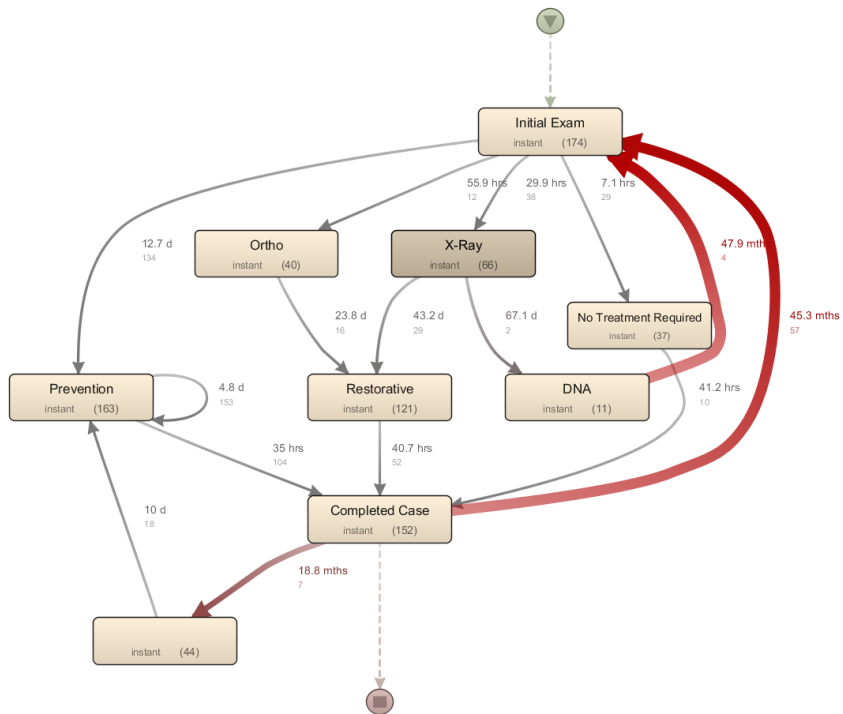


Figure 7-16: Default Performance Model for 2 screenings. Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.

Why are these default models inadequate?

These default models have significant limitations. Of interest here is the sequence of ‘Initial Exam’ (IE) events, therefore it is a requirement that the PM algorithm can distinguish the first IE from subsequent IEs for a patient. The Fuzzy miner when applied to the EHR data makes no such distinction and, as in Figure 7-15, the instances of IE events are combined in one ‘box’ as are the ‘Prevention’ and ‘Restorative’ events and it is not possible to see the sequence of screenings and treatment events received by the patients. The sequence is collapsed into a single IE and a single set of treatments.

This default output does little to portray the temporal dimension of patients’ treatment processes. In other words, it is not possible to see that these patients had 2 or 3 IEs and treatments following each of these IEs.

How was this addressed in this research?

An approach of adding ‘rank’ and DMFT to the event name was taken to address this.

The rank of an IE is its place in the sequence of IEs, i.e. the first-in-time IE has rank 1, the second IE has rank 2 etc.

To do this the following transforms were executed on the event data.

- 1) The IEs were ranked for each patient and the event name changed to reflect this e.g. if a patient had two IEs then, after the transformation, the event names were ‘Initial Exam 1’ & ‘Initial Exam 2’.
- 2) All other events were also adjusted in a similar fashion e.g. ‘Prevention’ events taking place after ‘Initial Exam 1’ were renamed ‘Prevention 1’ and ‘Prevention’ events taking place after ‘Initial Exam 2’ were renamed ‘Prevention 2’.
- 3) All events taking place prior to the first IE are marked ‘Pre’

These transforms were carried out on the data using an SQL script contained in supplemental material (Code CD 8.5). It is likely that a similar approach would be necessary for other data sets where events are repeated at intervals and the supplemental material referred to may offer some guidance for future work.

The process models were then recreated and are shown in Figure 7-17 to Figure 7-20 below. They show a comprehensible representation of the temporal sequence of events.

As Kerry was the only area with a policy of two IEs, and West Cork achieved a slightly higher percentage of patients whose sequence of screenings agreed with the policy of three screenings, this validating question focused on these two geographical areas.

The performance perspective provides a supplemental overview with a clear indication of the times between screenings evident in the models.

With the objective of incorporating outcome information in the output the following additional transform was then executed on the event data.

4) The DMFT was appended to the ranked IE events, i.e. if a patient had a DMFT of 1 at the time of the 2nd exam, after the transform, the event description was changed to ‘Initial Exam 2 DMFT=1’.

This gives an overview of the DMFT status of the cohort at the time of the IE and also has the effect of splitting them into different process streams afterwards. This is shown in Figure 10-2 to Figure 10-5 in Appendix 10.13. This transform addresses a criticism by Yang & Su (2014) where they point out that the existing algorithms only consider the event name and starting time – not the outcome. Initial inspection of the processes shows a higher percentage of the 3-screening-cohort maintaining DMFT=0 at final examination. There are two variations of adding DMFT to the event name. It can be added to all events or alternately just to the IE event. In this research, the latter was chosen as the former introduced additional complexity to the process models and added little value. Addition of the DMFT value to the event name, can lead to slight discrepancies between the frequency numbers shown on the process models using DMFT and those not using DMFT. This can occur if there is no charting available in the database to calculate the DMFT value and the consequence of this is that the case (patient) ‘drops’ from the model. This is not apparent in these models but is a notable feature nonetheless.

These process models are presented below:

Figure 7-17: Frequency model enhanced with ‘rank’ for 2 Screenings (Kerry)

Figure 7-18: Performance model enhanced with ‘rank’ for 2 Screenings (Kerry)

Figure 7-19: Frequency model enhanced with ‘rank’ for 3 Screenings (West Cork)

Figure 7-20: Performance model enhanced with ‘rank’ for 3 Screenings (West Cork)

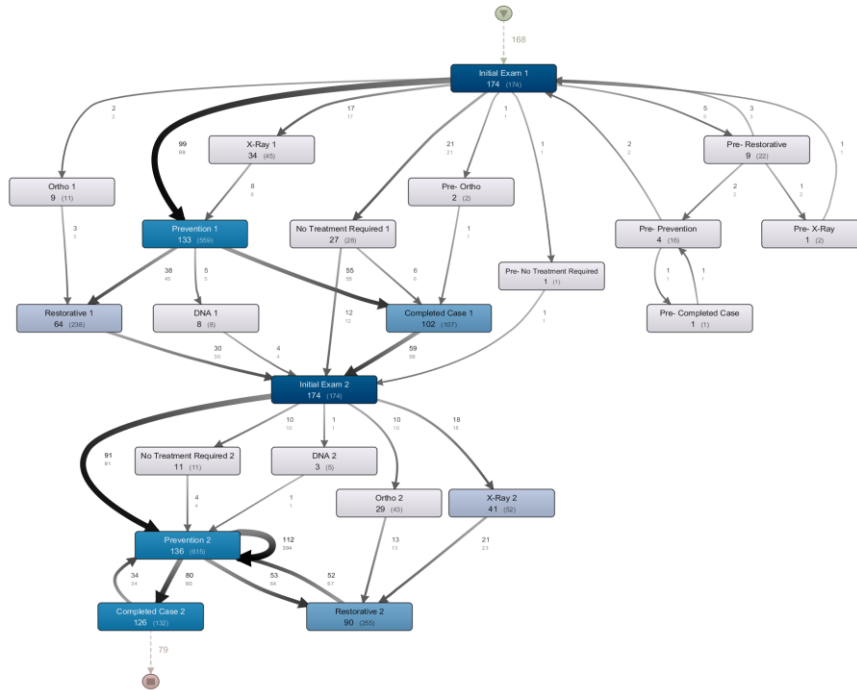


Figure 7-17: Frequency model enhanced with ‘rank’ for 2 Screenings (Kerry). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.

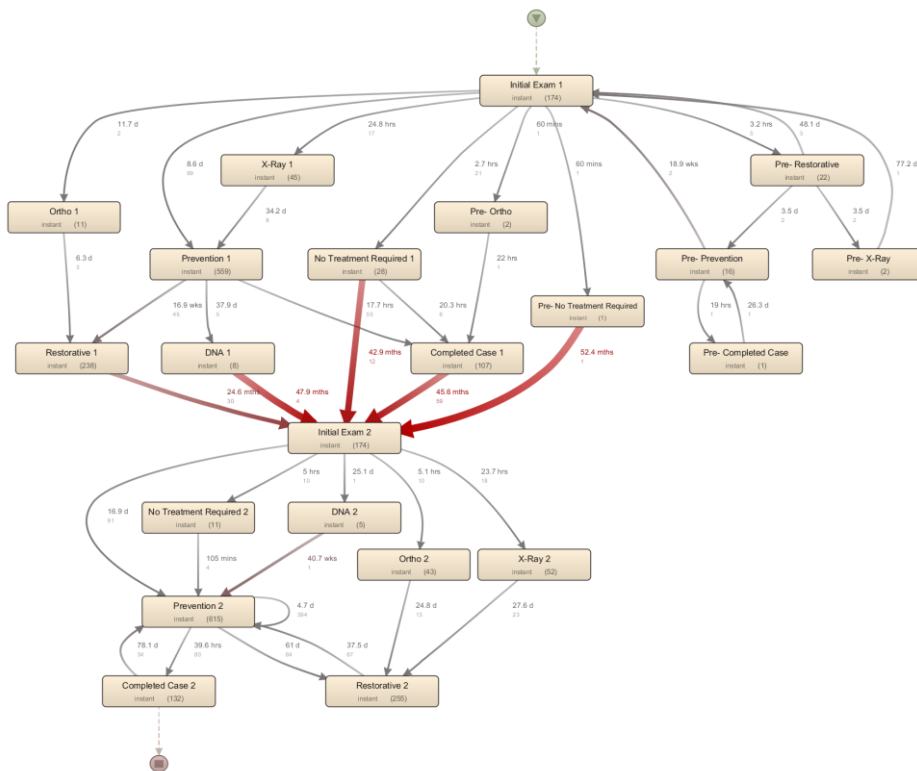


Figure 7-18: Performance model enhanced with ‘rank’ for 2 Screenings (Kerry). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.

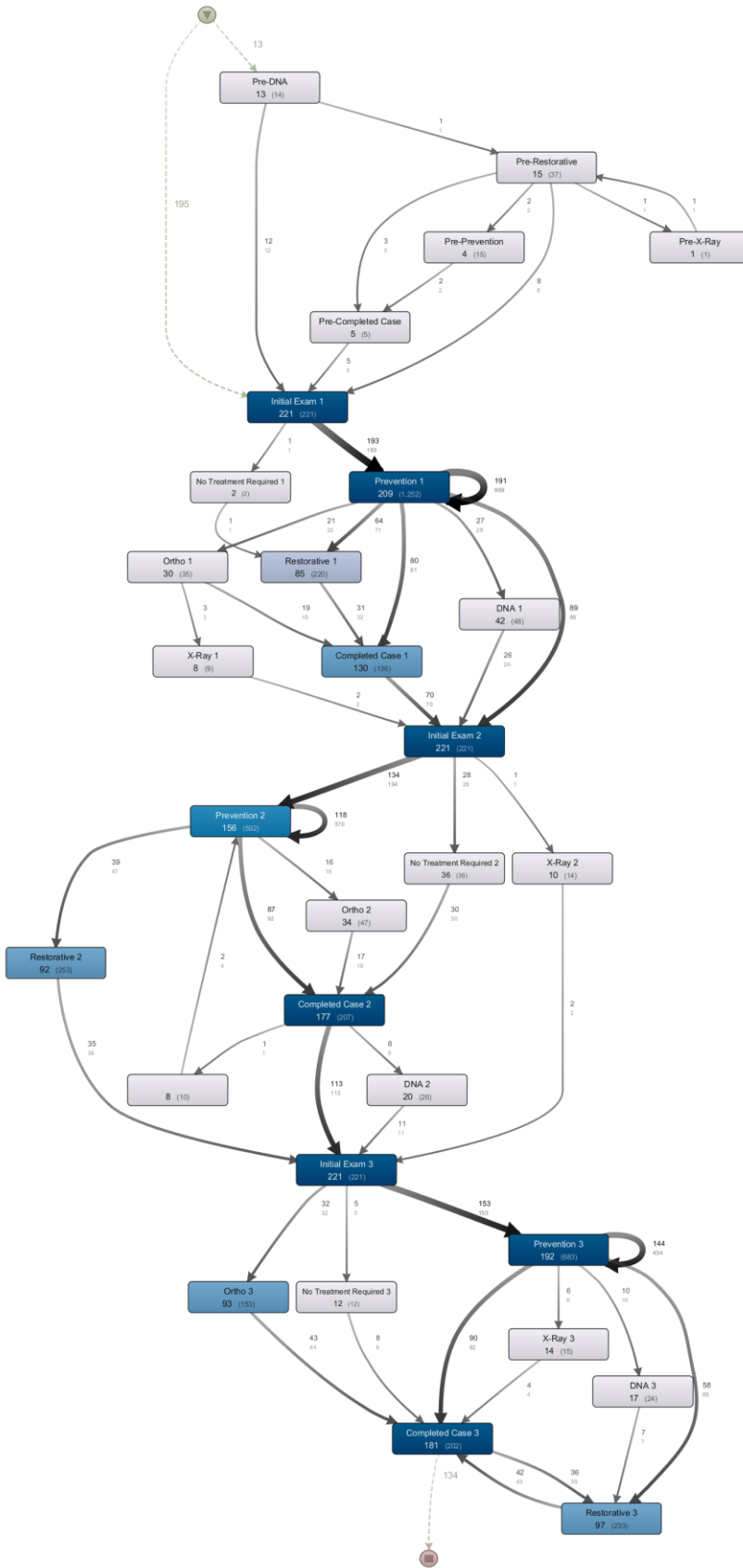


Figure 7-19: Frequency model enhanced with ‘rank’ for 3 Screenings (West Cork). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.

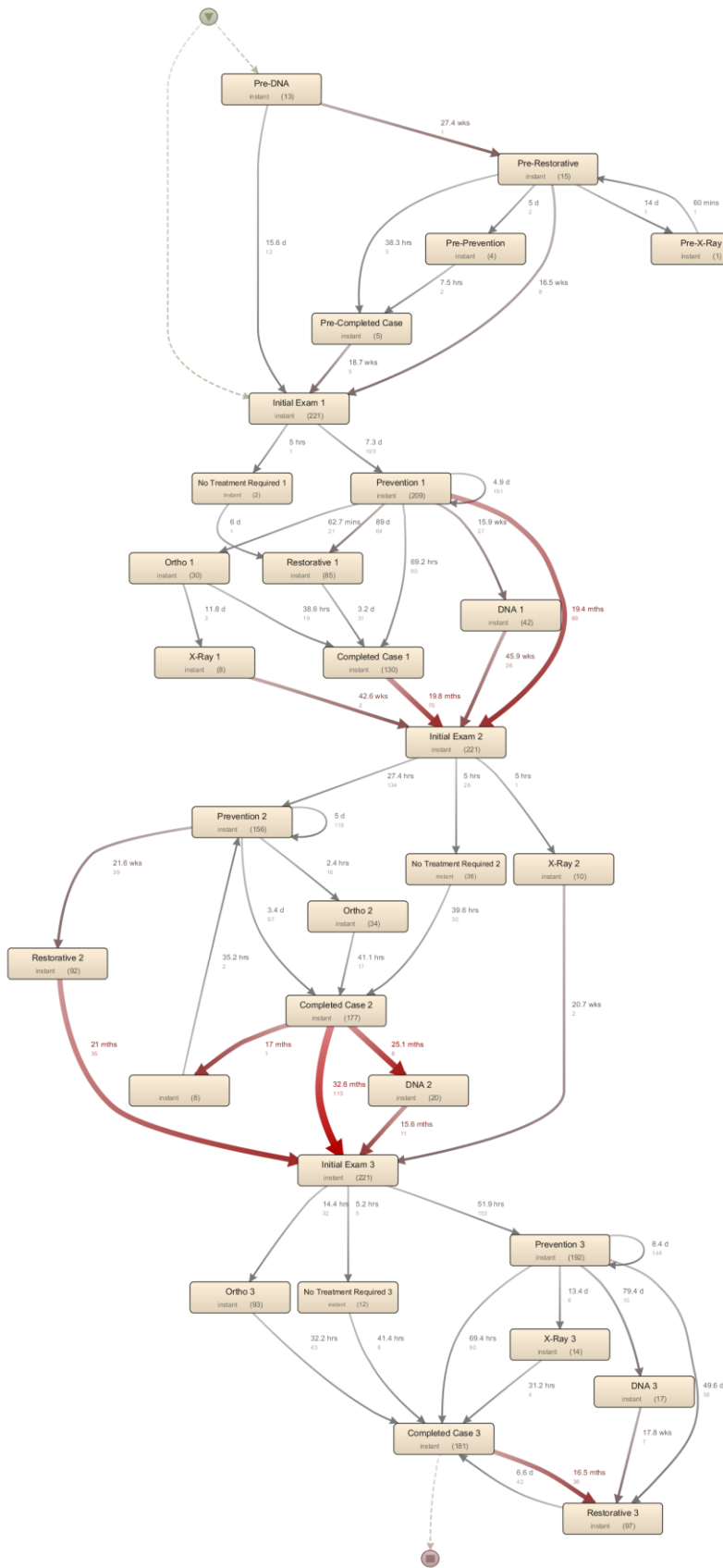


Figure 7-20: Performance model enhanced with ‘rank’ for 3 Screenings (West Cork). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.

Step 12 (b): Evaluation and Discussion of PM:

Are the treatment process models of adequate quality?

The Fuzzy Miner was used to generate the process models. Formal model quality metrics such as precision and fitness are accordingly unavailable as detailed in Section 2.6.4. However, the models generated were legible and recognisable to domain experts.

The research question asks if there is a difference in oral health outcome (DMFT) or treatment processes different for the cohorts receiving 2, 3, or 4 screenings in areas where this was the stated policy?

Three aspects of the process models are now examined:

- Commonly occurring pathways
- DMFT outcomes
- Temporal features of the pathways.

Commonly occurring pathways

First, we can identify the most commonly occurring paths. Of the 174 patients in the cohort receiving 2 screenings, 76% (133) received preventive treatments following the 1st screening and 78% (136) received preventive treatment after the 2nd screening. It is also interesting that the default process model before enhancing the event names with ‘rank’ revealed that, of the 174 patients, 94% (163) received preventive treatment after the first or the second screening. Subsequent models did not show this because the same patients could have been counted in both ‘Prevention 1’ and ‘Prevention 2’.

Thirty six percent (64) of patients received some restorative treatment between their 1st and 2nd screening and 90 received restorative treatment after their 2nd screening. 16% (27) of patients were marked with ‘No Treatment Required’ and did not receive preventive treatment. It is possible that these patients were free of dentinal caries in both deciduous and permanent teeth so they would not have required preventive treatments as they appear to be at low risk. It is also possible that these patients may have had preventive treatment outside this dental service, perhaps having received fissure sealants in private practice. Further analysis of the data could reveal additional detail regarding this. However, it was beyond the scope of this RQ. It is also interesting that the default process model in Figure 7-15, which was created before enhancing the event names with ‘rank’, revealed that of the 174 patients, 94% had received preventive treatment either after the first or the second screening. This is something that could not have been seen from the subsequent more

detailed models in Figure 7-17 and Figure 7-18, as many patients appeared in both ‘Prevention 1’ and ‘Prevention 2’ event boxes.

Of the 221 patients receiving 3 screenings, 95% (209) received preventive treatment and 38% (85) received restorative treatment before their 2nd screening. 71% (156) received preventive treatment after the 2nd screening and 42% (92) received restorative treatment before their 3rd screening. Eighty seven percent (192) received preventive treatment after the third screening and 44% (97) received restorative treatments. This information can be read from Figure 7-19 above.

DMFT Outcomes

Second, having split the patients into streams based on their DMFT at the time of screening, it was easy to examine whether the DMFT distribution at the final screening differs between those receiving 2 or 3 screenings. The results of this are presented in Table 7-5 below. The left 3 columns contain the DMFT values of the 174 patients who received 2 screenings. The right 5 columns similarly contain the DMFT values of the 221 patients who received 3 screenings. Of those that received 2 screenings, 40% still had a DMFT=0 at the time of their second and last screening whereas 60% of those who received 3 screenings remained disease free at the time of their 3rd and final screening. Also, 6% of those with 2 screenings had a final DMFT=6 or greater whereas only 1% of those receiving 3 screenings had a similar outcome.

2 Screenings 174 Patients			3 Screenings 221 Patients				
DMFT at 2nd Screening	Number of patients at 2nd Screening	%	DMFT	Number of patients at 2nd Screening	%	Number of patients at 3rd Screening	%
0	70	40	0	183	83	132	60
1	34	20	1	19	9	39	18
2	18	10	2	8	4	22	10
3	22	13	3	9	4	17	8
4	18	10	4	2	1	6	3
5	2	1	5	0	0	2	1
6+	10	6	6+	0	0	3	1

Table 7-5: DMFT Distribution at 2nd or 3rd Screening.

The interpretation of these must consider the DMFT’s confounding factors as outlined in Figure 1-8. This data carries a health warning because DMFT is correlated to socio-economic status (SES) as is dental visiting pattern. The current exercise aims to investigate the utility of PM to service planning and evaluation and a more inclusive analysis would be required to explore the impact of screenings on outcomes. For example it cannot be assumed that the process of more frequent screening in itself generates better

patient outcomes because more compliant patients, possibly with a higher SES and lower caries levels may be more likely to attend for frequent screenings whereas those with higher caries levels may be more likely to attend for symptomatic treatment only. PM could also be applied to the data to model the pathway for children whose initial appointment is for symptomatic treatment and for those who attend for 1st screening in 1st or second class (age 7-8). This approach would include 2nd and 3rd screenings in the pathway but would capture a broader spectrum of patients. Careful planning of inclusion criteria with health service administrators and clinicians would help to refine the research question and structure the model in a way that best addressed the divergent behaviours of the target population.

Temporal features of the pathways

Finally, the times between treatments can be examined from the process models. The process model for patients receiving 2 screenings in Figure 7-18 shows that the average time from a patient being classified as a ‘Completed Case’ after their 1st screening to their 2nd screening was 45.6 months and the average time from a restorative treatment to their 2nd screening was significantly shorter at 24.6 months, indicating at least some of the 30 receiving restorative work probably attended outside of the planned screening schedule. For those receiving 3 screenings, the average time from ‘Completed Case’ to the 2nd screening was 19.8 months and from there 32.6 months to the 3rd screening. It is perhaps interesting that 36 of the 221 patients received restorative treatment, an average of 16.5 months after completing their 3rd screening.

This application illustrates the potential for PM to monitor KPIs related to time to treatment or time to completion, allowing the setting and monitoring of important targets for service delivery.

7.3.4 Limitations of this experiment

The method for selecting the cohorts carries the risk of introducing an ascertainment bias to the experiment as these criteria could cause the sample to not accurately represent the intended population. For example, it is unclear what happened to the patients that had only one screening, why they had only one screening and what effect their inclusion would have had if they had had subsequent screenings.

There are other limitations in the creation and interpretation of these cohorts. The numbers in the data extracted for this experiment are not intended to represent the actual number receiving school screenings in the areas and should not be interpreted as such.

The selection criteria exclude patients on several criteria, including DQ criteria and accordingly, the numbers are less than the actual numbers seen. It is also unclear whether the population in the EHR is representative of the entire population.

It is unknown whether the patients in the EHR received treatment outside the public health system i.e. in private practice. It is possible that additional patients received their first screenings at an age other than 7, 8, or 9. Patients whose DMFT was not 0 at the time of screening were excluded and analysed separately to remove some confounding factors.

7.3.5 Conclusions

This question aims first to examine whether the oral health outcomes and treatment processes vary between the cohorts that received screenings according to varying policies. The experiment has clearly shown that cohorts can be created and oral health outcomes for the cohorts can be calculated. Finally, it has been shown that PM can discover treatment pathways followed by these cohorts fulfilling the success criteria for the experiment.

A multivariate analytical approach which could account for known confounding factors for dental caries, would be required to address the question “Is there a difference in oral health outcome (DMFT) or treatment processes between groups receiving 2, 3, or 4 school screenings in areas where this was the stated policy?” Simple comparison of outcomes for groups of children subjected to different screening frequencies in disparate geographical regions could be misleading. Critical information relating to confounding factors such as fluoridation status and socio-economic status was unavailable in this dataset, making comparison among groups unreliable. What we can say about the PM approach is that this work illustrates the feasibility of extracting valuable outcome data on the impact of the different screening frequencies if independent variables such as the child’s fluoridation status and the families socioeconomic status were collected routinely and included in the model.

Though analysis and process models have been shown capable of delivering insights on the significance of the frequency at which school screenings are delivered, the limitations of the data and the potential confounding factors dictate that any insights from this data be treated with caution.

7.4 Assessing the Impact of ‘age at first screening’ Policies

7.4.1 Introduction and Aims

Can analysis of the EHR assess the impact of ‘age at first screening’?

This question aims first to examine whether the oral health outcomes and treatment processes vary between the cohorts that received their first screenings at varying ages (7, 8, or 9), delivering insights, and addressing RQ4. There are several sub-aims or objectives to achieve this: establish if the EHR can distinguish between cohorts, establish if the research data can show oral health outcomes for the cohorts, and establish if PM can discover treatment pathways followed by these cohorts.

Exploring this question using EHR data is technically more complex and challenging than those in Sections 7.1.1 and 7.2 and requires application of the Policy and Strategy Questions Methodological Approach from Section 6.4.4 and the steps in Table 6-3.

7.4.2 Success Criteria

The success criteria for this question are similar to the previous experiment in Section 7.3.2 and PM should be shown capable of delivering insights on the significance or otherwise of the age at which first school screenings are delivered.

7.4.3 Methodology

Introduction

Following the PM methodology steps from Section 6.4.3.

Steps 1 through 8 are the general preparatory steps followed in this research, common to all RQs and have been completed earlier in the research as detailed in Section 6.4.3.

Step 9 utilises the Policy and Strategy Questions Methodology detailed in Section 6.4.4.

7.4.3.1 Step 9.1 - Identify a situation that represents a policy or strategy change

Age at first school screening is now explored. The age at emergence of the first and second permanent molars is a key milestone for oral health assessment and this question investigates whether EHR data and PM technologies can help answer the question: What is the ideal age for first school screening? The research demonstrates how data mining and PM can distinguish between the paths followed by cohorts receiving school screenings at different ages and their corresponding oral health outcomes.

7.4.3.2 Step 9.2 - Assemble evidence of this policy in the EHR.

The author is not aware of any formal policy or strategy regarding the age at which school children should receive their first screening in the HSE at the time this research's EHR data was collected. However, all patients' date-of-birth and date of first screening are

present in the dataset allowing us to investigate the usefulness of our methods in addressing the RQ.

7.4.3.3 Step 9.3 - Establish whether the policy/strategy is visible in the EHR?

This is a technical question about the data. Questions like: Can the information being stored in the EHR data structures distinguish between groups or cohorts who were on the different sides of a decision or policy?

Identical to Section 7.3.3.3 it is clear that the information being stored in the EHR data structures is capable of distinguishing between groups or cohorts receiving their first school screening with different ages.

7.4.3.4 Step 9.4 - Does the EHR data comply with the policy?

Even though there is no policy for this question, this is still relevant. If we can view the data in a way that simulates the question that we are trying to answer, it might still be possible to make some interesting findings. In this case, it may still be possible to create cohorts who had their first school screening at ages 6, 7, 8, or 9 in order to simulate the policy from the EHR data. As there was no differing policy between regions, it was possible to ignore the region and select the cohorts based on the age at first screening.

7.4.3.5 Step 9.5 - What are the appropriate outcomes to measure the effects of a policy/strategy?

Identical to Section 7.3.3.5.

7.4.3.6 Step 9.6 - Which of these appropriate outcomes are available from the EHR?

DMFT/dmft is available in the dataset

7.4.3.7 Step 9.7 - Eliminating confounding factors in the cohorts

Again, as in Section 7.3.3.7, potential exposures, outcomes, confounders, and mediators need to be identified in order to ensure the cohorts are comparable. Figure 1-8 shows many factors affecting the development of dental caries and accordingly there are many potential confounders. The requirement that all members of the cohorts had a starting DMFT=0 is a significant step taken to ensure that the cohorts are legitimately comparable. However, this itself potentially introduces a difference between the demographics of the age-groups. There is a very strong correlation between age and disease experience therefore, there is likely to be a difference between the demographics of the different age

cohorts in relation to their caries risk i.e. a 9 year old having DMFT=0 is more indicative of a low caries risk than a 6-year-old with DMFT=0 as the 6-year-old's permanent teeth are just erupting and have not yet been exposed to many of the caries risk factors.

7.4.3.8 Step 9.8 - Develop the specific Research Questions around the policy/strategy, answerable with the EHR data

Is there a difference in oral health outcome (DMFT) or treatment processes between groups age 12/13 with their first screenings at ages 6, 7, 8, or 9? In this research, no statistical comparison of the cohorts' outcomes is carried out. The intention of this research was to validate the methodologies developed and not to definitively answer the validating questions. Critical information relating to confounding factors such as fluoridation status and socio-economic-status was unavailable making comparison with previous studies impossible.

If it were carried out it would often take the form of a Null Hypothesis, i.e.: There is no difference in oral health outcome (DMFT) or treatment processes between groups age 12/13 with their first screenings at ages 6, 7, 8, or 9. Alternative Hypothesis: There is a difference in oral health outcome or treatment processes between groups age 12/13 receiving their first screenings at ages 6, 7, 8, or 9.

7.4.3.9 Step 10, Part 1 - General data processing to facilitate answering RQs is in Appendix 10.18.

7.4.3.10 Step 10, Part 2 - Data processing to facilitate answering RQ4 is now addressed.

Phase 1: Cohort Creation -

The cohort is defined as follows:

- First Screening (IE) between January 1st, 2004 and December 31st, 2008.
- The patients were aged 6, 7, 8, or 9 at the time of the screening.
- Patients received 2 or 3 screenings.
- The data quality was acceptable.
- Starting DMFT = 0 or Starting DMFT>0.

This identified 4 cohorts (ages 6, 7, 8, 9) in the schoolyear 2005/6 and the same individuals' Oral health outcome (DMFT) at ages 12/13 was assessed. This RQ uses the same methodological approach as the previous RQ. Tables called CohortAgeX2005YScreenings containing the ID for each of these cases was created in

the QL Server BridgesPM1 database where X is the patient's age (7,8,9) and Y is the number of screenings they received.

Phase 2: Event Log Creation

- This phase is identical to that in Section 7.3.3.10.

The basic characteristics of the cohorts are more similar than those in the experiment regarding frequency of screenings. The average number of treatments received by the cohorts ranges from 21.1 to 26.1. The highest number is associated with the patients receiving their first screening at age 6, and the lower numbers (21.1, 22.1, 22.9) with patients receiving their first screening at age 7, 8, and 9. In PM terms this translates directly to a similar average number of events per case.

Again, notable is the high proportion of variants with only a very small number of non-unique pathways experienced by the patients. For example, in Cohort Age 6, of the 790 cases, 788 of these followed unique pathways.

7.4.3.11 Step 11, Data Analysis

Step 11(a): Establish the outcomes for the cohorts receiving their first screening at ages 6, 7, 8, or 9.

This was executed in the Anaconda/Jupyter Notebook environment. Data analytics cells were written in Python, calling SQLServer functions with the following major steps:

This is an ad-hoc sequence of steps, specific to the RQ, and on Code CD-7.

- Create a table to store aggregated data.
- Create entry in table for each age at first screening and number of screenings received.
- Calculate and save the number of patients in each category i.e. Screening took place at the correct time, in the correct clinic/region; it was their first screening; they were aged 6-9; data quality was good.
- Calculate and save the number of these same patients who received a screening in each of the following 5 years.
- Calculate and save mean DMFT and its standard deviation for patients in each of these years. This is summarised in

Figure 7-21.

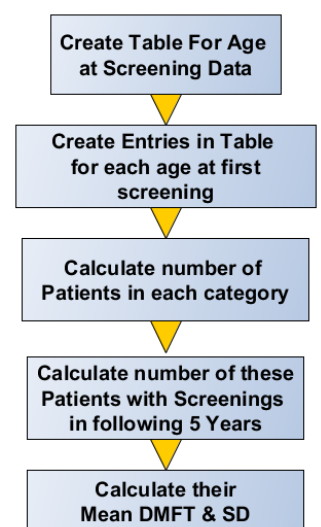


Figure 7-21: Age at First Screening, Outcomes calculation

This experiment is primarily designed to see if the age at first screening had an impact on oral health outcomes at age 12/13 or if it had an impact on the treatment processes experienced by the patients. The detailed data was extracted for cohorts aged 6, 7, 8, & 9 having DMFT=0 and is in Appendix 0, Table 10-3 (2 screenings) and Table 10-4 (3 screenings). A similar dataset having starting DMFT>0 is also in Appendix 0, in Table 10-5 (2 screenings) and Table 10-6 (3 screenings).

To clarify, two major groups were created, one with starting DMFT=0 and one with starting DMFT>0. Within each of these groups, 4 separate groups were identified, those receiving their first school screening at 6, 7, 8, or 9. Within each of these age-groups, two sub groups were identified: those receiving 2 screenings and those receiving 3 screenings, i.e. 16 cohorts in total. On average, children who are free of dentine caries at age 9 are a low-risk subset. Of those free of dentine caries at 6/7 fewer will be free of dentine caries at 8/9.

The complete visualisations of the cohort with DMFT=0 at 1st screening is represented in Appendix 0, in Figure 10-12 created from Table 10-3 and Table 10-4. The cohort with starting DMFT>0 at 1st screening is represented in in Appendix 0, Figure 10-13, created from Table 10-5 and Table 10-6. This data is summarised in Figure 7-22 & Figure 7-23.

Note: How to read the Age at Screening Profile data below.

- X axis -Age when cohort received first screening - at age 6, 7, 8, or 9.
- Patients receiving 2 Screenings - Grey Bar, 3 screenings – Orange Bar.
- Y axis - Weighted average DMFT at ages 12 or 13.

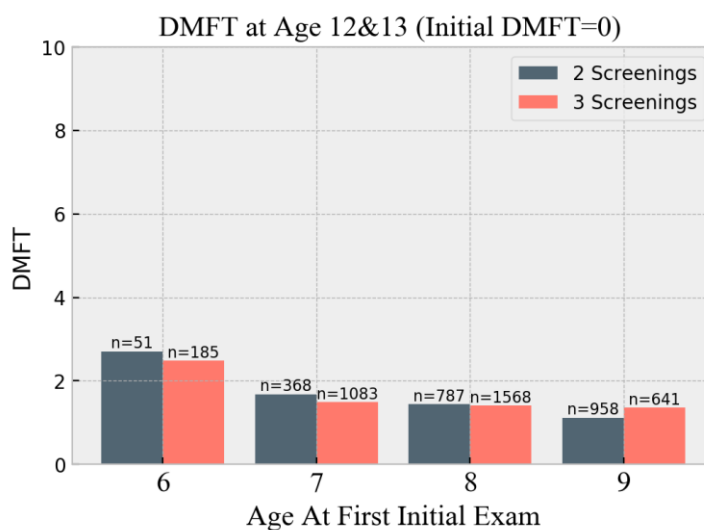


Figure 7-22: DMFT at 12/13 by age-at-first-screening for patients with a baseline DMFT=0

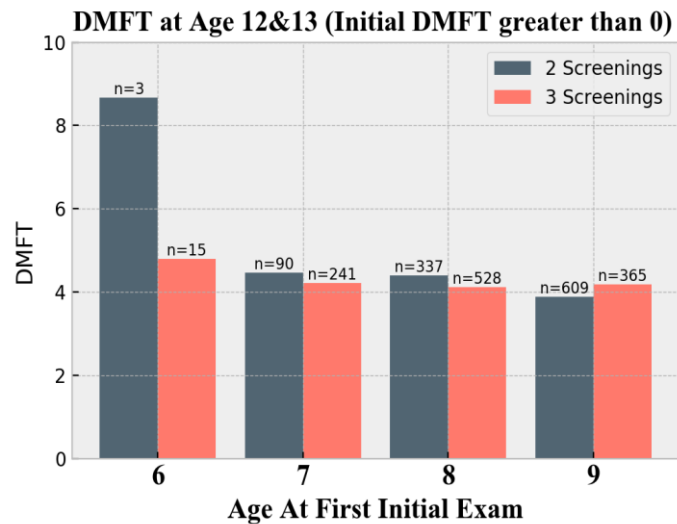


Figure 7-23: DMFT at 12/13 by age-at-first-screening for patients with a baseline DMFT>0
Step 11(b)

Are the outcomes different for cohorts receiving their first screening at age 6, 7, 8, or 9? As in the prior experiment, to support evaluation of the process models, an outcomes-based analysis of the cohorts receiving their first screening at ages 6, 7, 8, or 9 was carried out in advance of analysing the cohorts' process models. This was supported by descriptive analysis including histograms giving overviews of the activity frequencies, ages at treatment and DMFT distributions as in Section 4.1.6.

The analysis specific to this RQ described in Section 10.14 and is summarised in Figure 7-22 for those with a starting DMFT=0, and for those with DMFT>0 in Figure 7-23 above. These figures plot age at screening and DMFT outcomes for those groups receiving 2 and 3 screenings. However, although cohorts were created for patients receiving two and three screenings, this experiment concentrates on those receiving three screenings as it adequately demonstrates that PM can show the differences between the treatment processes. In any case, the analysis of outcomes showed no obvious difference between the outcomes for those receiving 2 or 3 screenings. The complete dataset is available in Appendix 0, Figure 10-12 and Figure 10-13 and in Table 10-3 to Table 10-6.

Figure 7-22 shows the DMFT at age 12 & 13 for four main cohorts, those receiving their first screening at ages 6, 7, 8, and 9 with a baseline DMFT=0. Each of these is then subdivided into those receiving two or three screenings. There is a noticeably higher DMFT for those who received their first screening at age 6 with DMFT then decreasing slightly as age-at-first-screening increased. This can be explained by the fact that free of dentine caries children at age 9 are likely to be a subset of those who are free of dentine caries at age 6 or 7. Those who are free of dentine caries at the older ages are those who

are at lower risk of caries. The lower level of incremental caries development is likely to be linked to this lower caries risk.

Similar to the previous figure, Figure 7-23 shows the DMFT at age 12 & 13 for four main cohorts, those receiving their first screening at ages 6, 7, 8, and 9. However, these patients had a starting DMFT > 0. Each of these is again subdivided into those receiving two or three screenings. There is no obvious difference in DMFT for those who received their first screening at any of the measured ages.

In this research, no statistical comparison of the cohorts' outcomes is carried out. Critical information relating to confounding factors such as fluoridation status and socio-economic-status was unavailable making comparison with previous studies impossible.

7.4.3.12 Step 12, Process Mining

Step 12(a): Establish the treatment process models experienced by the cohorts receiving their first screening at age 6, 7, 8, or 9.

The default outputs from the PM application, Disco, suffered from the same shortcomings as in the previous experiment i.e. it was impossible to discern a logical sequence from the process model because the order and sequence of screenings were collapsed into a single event. These preliminary results are excluded from this section of the thesis but are available in Appendix 10.16 for reference. To solve this issue and as in the previous experiment, the 'rank' of the event was added to the event name and the DMFT outcome was also added as described in Section 7.3.3.12.

To reduce the complexity of the models in a structured manner, an additional transform was carried out. Only 'Restorative', 'Prevention', 'Initial Exam' and 'Completed Case' were extracted into the event logs. Some of the others such as 'X-ray' and 'Ortho' could often be considered as noise and they were removed. It was difficult to see how they added any value. There is clearly a risk that such reductions could result in interesting or important paths and deviations being missed or impossible to see and they could be reintroduced if there was a specific question of interest.

In some of our ELs the default visible-path percentage was <1% and this often resulted in valuable information such as the split between 'Prevention' and 'Restorative' after 'Initial Exam 1'. Setting the paths-percentage value to 1 or 2% resolved this, but it highlights the need to be vigilant with default settings in PM technologies. To ensure that

the process models are reproducible all product settings used to generate a model must be recorded. These setting for this experiment were 50% of activities and 4.6 or 9.8% (age 6) of paths. As in the previous experiments, desirable features of the resulting process model are that the model be at least legible when printed on an A3 sheet, preferably on A4. The model should show a breakdown of 'Prevention' and 'Restorative' and all screenings. Only the key data extracted from the process models for those receiving 3 screenings are now presented. The enhanced process models are presented in Appendix 10.16.

Step 12 (b): Evaluation and Discussion of PM:

Are the treatment process models of adequate quality?

As in the previous experiments the Fuzzy Miner and Disco application is used to generate the process models. Formal model quality metrics such as precision and fitness are accordingly unavailable as detailed in Section 2.6.4. However, the models generated were legible and recognisable to domain experts.

Are the treatment processes different for the cohorts?

The research question asks if there is a difference in oral health outcome (DMFT) or treatment processes different for the cohorts receiving their first screening at age 6-9.

As in the previous experiment, three aspects of the process models are now examined: commonly occurring pathways, DMFT outcomes, Temporal features of the pathways.

Commonly occurring pathways

First, the most commonly occurring paths can be identified. Of the 790 6-year-old patients, 62% received preventive treatments following the 1st screening in comparison with 84%, 89%, and 87% of 7, 8, & 9-year-olds respectively. This lower prevention level for 6-year-olds could be due to the fact that their first permanent molars are less likely to have erupted. 65% of 6-year-olds received restorative treatment after their first screening in comparison to 46%, 40%, and 36% of 7, 8, and 9-year-olds. A possible explanation is that most caries at age 6 would be in deciduous molars as the permanent molars do not normally erupt until age 6-7. As the deciduous molars are needed to maintain space for their permanent successors erupting at age 9-12, dentists are more likely to restore decayed deciduous teeth in younger children to maintain this space for as long as possible. Restorations at this age may also be carried out to reduce pain. The older the child, the less value there is in restoring deciduous teeth as they near the time for natural exfoliation.

Table 7-6: Summary of the Age at first screening Process Model Characteristics (2004-2008), 3 Screenings, Baseline DMFT=0.

	Age 6	Age 7	Age 8	Age 9
No of Cases	790	2081	3323	1671
No of Unique Cases	788	2050	3247	1624
No of Events	20684	47735	73494	35285
% Receiving Prevention after 1st Screening (n)	62 (n=489)	84 (n=1742)	89 (n=2952)	87 (n=1450)
% Receiving Restorative after 1st Screening (n)	65 (n=512)	46 (n=957)	40 (n=1317)	36 (n=594)
% with DMFT=0 after 2nd Screening (n)	69 (n=549)	74 (n=1547)	76 (n=2513)	75 (n=1259)
% with DMFT=0 after 3rd Screening (n)	41 (n=323)	48 (n=1007)	53 (n=1761)	56 (n=929)

DMFT Outcomes

First, having split the patients into streams based on their DMFT at the time of screening, it is easy to examine whether the DMFT distribution at the final screening differs between those receiving their first screening at age 6, 7, 8, or 9. The outcomes analysis in Figure 7-22 showed little difference between the DMFT values at ages 12 & 13 for the age-groups and the process models also confirmed this.

Temporal features of the pathways

Finally, the times between treatments can be examined from the process models. Again, the temporal information in this experiment is quite mundane. However, in this experiment, it did give the opportunity to focus solely on the development of DMFT over time and this offered a different perspective on the process. The performance-based process models were generated using only the screening events and the completed case events. This resulted in a much simpler model clearly showing the time between screenings (or completed cases) and the DMFT at the time of the 2nd and 3rd screenings. Reading the time elapsed between screenings yielded the data in Table 7-7 below. For each age there are two columns. The ‘Months to 2nd screening’ column indicates the average time elapsed between the 1st screening and the 2nd screening. The ‘Months to 3rd screening’ column indicates the time elapsed between the 2nd and the 3rd screening. The DMFT column is the DMFT at the time of the screening e.g. for six-year-olds, the average time between 1st and 2nd screenings for those having a DMFT=0 at the 2nd screening is 26 months. For six-year-olds, the average time between 2nd and 3rd screenings for those having a DMFT=0 at the 3rd screening is 36 months. This dataset allows us to plot time between exams against final DMFT outcome and see if there is a trend.

The data also shows that low numbers of patients had the higher DMFT values (>5) in most cases, making the long times between their screenings less significant e.g. only a single 6-year-old had a DMFT=10 at their 2nd screening which was 80 months after their first screening (see bottom left-hand cell in Table 7-7). Of course, it could be that these ‘rare’ cases are of specific interest as they ultimately may require more extensive treatment and may be a higher burden on the service. Ignoring these unusual cases often leads to more comprehensible process models but must be done with caution as it is possible that valuable insights lie within them.

Table 7-7: Average number of months between 1st & 2nd Screening, and between 2nd & 3rd Screening related to DMFT outcome at 3rd screening, broken down by age at 1st screening

DMFT	Age 6		Age 7		Age 8		Age 9	
	Months to 2 nd Screening	Months to 3 rd Screening	2 nd	3 rd	2 nd	3 rd	2 nd	3 rd
0	26 (n=550)	36 (n=340)	23 (n=1558)	33 (n=1033)	23 (n=2538)	30 (n=1824)	21 (n=1271)	27 (n=961)
1	29 (n=101)	51 (n=116)	26 (n=292)	32 (n=393)	24 (n=410)	31 (n=593)	26 (n=213)	27 (n=297)
2	27 (n=65)	39 (n=109)	26 (n=115)	34 (n=248)	27 (n=210)	30 (n=362)	25 (n=95)	28 (n=168)
3	35 (n=39)	45 (n=85)	29 (n=68)	34 (n=161)	30 (n=94)	30 (n=229)	30 (n=41)	26 (n=97)
4	28 (n=23)	38 (n=49)	27 (n=32)	36 (n=111)	29 (n=44)	29 (n=127)	29 (n=26)	27 (n=60)
5	64 (n=5)	33 (n=36)	38 (n=7)	33 (n=58)	32 (n=9)	32 (n=76)	30 (n=8)	28 (n=38)
6	18 (n=2)	44 (n=20)	36 (n=3)	37 (n=26)	35 (n=9)	35 (n=43)	31 (n=6)	23 (n=19)
7	32 (n=1)	43 (n=16)	13 (n=1)	35 (n=19)	40 (n=2)	40 (n=19)	27 (n=2)	17 (n=12)
8	70 (n=2)	32 (n=4)	43 (n=2)	34 (n=14)	31 (n=1)	31 (n=18)	41 (n=2)	31 (n=5)
9	80 (n=1)	37 (n=7)	52 (n=1)	32 (n=10)	37 (n=1)	37 (n=8)	47 (n=5)	32 (n=6)
10	80 (n=1)	81 (n=3)	59 (n=1)	39 (n=3)	32 (n=1)	32 (n=3)	50 (n=1)	33 (n=4)

Additional temporal features were then created using the data visualisation package, Python Seaborn. Regression plots gave an overview of the relationships between variables particular useful during exploratory data analysis. Exploring data in this fashion creates intuitive knowledge of the dataset and can lead to other questions for further exploration. It does not give quantitative measures of fitness of the model. Statistical analysis could be executed with the Python Statsmodels package or similar if required. The first two models of the data plotting the frequencies of ‘time-between-screenings’ give an overall impression of the dataset. Figure 7-24 shows the data segmented by ‘Age’ and shows a much tighter range of time between screenings for the older ages. Six-year-

olds show a much wider spread of values than the older ages confirming what can be sensed from reading the data in Table 7-7. This could be anticipated as the service focuses on children to age 16 and younger children have a longer window of opportunity for recall up to age 16.

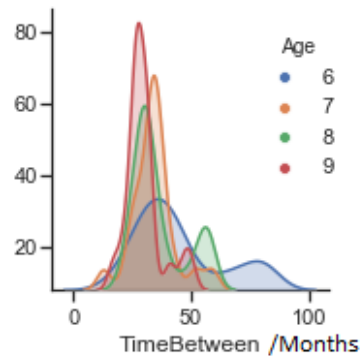


Figure 7-24: Distribution of times between screenings, by age at first screening

Viewing the same data, but segregated by times between 1st and 2nd screenings, and 2nd and 3rd screenings, shows a tighter distribution between the 2nd and 3rd screenings. Both of these figures might suggest that the 2nd interval is easier to manage as the patients are already in the system and those arriving into the system before they are officially due their first screening may be partially responsible for flattening the time between 1st and 2nd screening distribution. Alternatively it may illustrate the shorter time window for recall

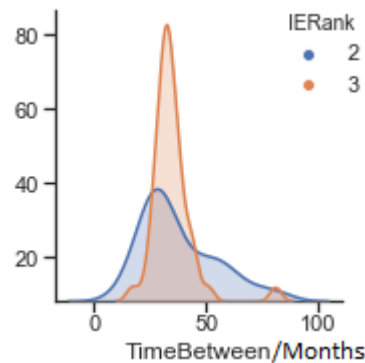


Figure 7-25: Distribution of times between screenings, segregated by times between 1st and 2nd screenings and times between 2nd and 3rd screenings.

Again, viewing the same data but incorporating the DMFT values can be used to see further variable relationships. Here, in Figure 7-26, basic linear regression models are generated using the same data. As would be expected, overall DMFT values are generally increasing with age and also increasing, albeit slightly, with an increase in the time between screenings. Again, none of these models estimate the fit, rather, act to give an intuitive understanding of the data. It is clear from Table 7-7 that many of the higher DMFT values have very low n-values which could produce misleading regression models. Further information on the reasons for longer intervals between screenings would be needed to fully interpret these findings.

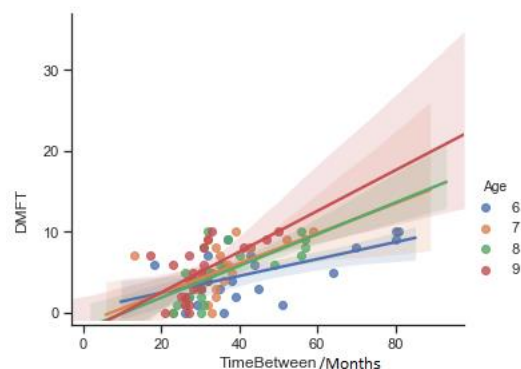


Figure 7-26: DMFT at age 12/13 & time between screenings, segregated by age.

7.4.4 Limitations of this experiment

The method for selecting the cohorts carries the risk of introducing an ascertainment bias to the experiment as these criteria could cause the sample to not accurately represent the intended population. For example, it is unclear what happened to the patients that had only one screening, why they had only one screening and what effect their inclusion would have had if they had had subsequent screenings.

There are other limitations in the creation and interpretation of these cohorts. The numbers in the data extracted for this experiment are not intended to represent the actual number receiving school screenings in the areas and should not be interpreted as such. The selection criteria exclude patients on several criteria, including DQ criteria and accordingly, the numbers are less than the actual numbers seen. It is also unclear whether the population in the EHR is representative of the entire population. It is unknown whether the patients in the EHR received treatment outside the public health system i.e. in private practice. It is possible that additional patients received their first screenings at an age other than 7, 8, or 9. Patients whose DMFT was not 0 at the time of screening were also excluded and analysed separately to remove some confounding factors.

7.4.5 Conclusions

This question aims first to examine whether the oral health outcomes and treatment processes vary between the cohorts that received their first screenings at varying ages (7, 8, or 9). The experiment has clearly shown that cohorts can be created and oral health outcomes for the cohorts can be calculated. It has been shown that PM can discover treatment pathways followed by these cohorts fulfilling the success criteria for the experiment.

Though analysis and process models have been shown capable of delivering insights on the significance of the age at which school screenings are first delivered, the limitations of the data and the potential confounding factors dictate that insights from this data be treated with caution.

Adjusting the query to include other ages or any number of screenings is trivial, as is toggling between starting DMFT=0, starting DMFT>0 or indeed any value for this or an alternative oral health outcome.

7.5 Rejected Validating Question

One of the validation experiments originally proposed by the author was to investigate the effect of fissure sealants (FSs) on oral health outcomes and subsequent treatment process. The question would have investigated whether application of FSs to 6's (first permanent molars) leads to a better oral health outcome at age 12/13. This experiment was to use data mining, PM and visualisations to demonstrate the impact of FSs on the 6's and to compare the outcomes and treatment processes of two cohorts - one receiving a school screening and FS in 2007, the other receiving no FS.

Following discussion with domain experts some issues with the experiment design were raised. The existence of the cohort not receiving FS raised the following question: Given that there was a blanket FS policy in place at the time, why did they not receive the intervention? Possibilities raised were: teeth were not erupted or partially erupted, children were free of dentine caries, had good oral hygiene, lived in a fluoridated area and were considered to be at low risk of caries. Was there another clinical reason? Were FSs already in place – perhaps placed under the private system? Perhaps fewer FSs were placed in specific clinics which were under resource pressure or perhaps the teeth had already been restored.

The discussion with the experts clarified that the cohort without FSs consisted of patients who never had FSs completed in the HSE system. However, there remained the possibility that the patients had FSs completed elsewhere, most likely in the private dental healthcare system. To identify patients who might have received FSs elsewhere it was not sufficient to look at the treatments performed by the HSE. It was also necessary to look in a separate part of the BridgesPM1 extract containing a clinical description called 'Conditions'. This is a description of the condition of the patient when examined and charted and contained information on pre-existing FSs. When this was checked it was found that many of individuals in Cohort 2 already had FSs in place when they were screened. These were most likely placed in private practice though it is possible that they were placed within the HSE and graphically charted but not entered in the 'Treatment Items' list. There is no way to verify this one way or the other. In any case, the number of patients remaining in the cohort without FS was quite small and it was decided to abandon this experiment and focus on the frequency of screenings and times between screening experiments. This highlighted the necessary to have domain expertise at all stages in the research process as it provided insight and expertise not available from the data in isolation.

7.6 What Data is Needed in an EHR for Effective PM? (RQ5)

7.6.1 Introduction

This question aims to identify data that would be needed in an EHR if applying the new PM approach to discover dental care pathways and facilitate the evaluation of policy implementation i.e. RQs 1 through 4. The objective of this exercise is to enhance the initial BridgesPM1 data model, which describes the research dataset, with additional desirable entities and attributes identified from existing standards in the literature. The experience gained in this research has provided further information on additional desirable entities and attributes. The primary approach taken is to create a Dental Data Reference Model based the initial BridgesPM1 entity relationship, existing standards and also, on the experience of this research.

The method used studies the existing standards applicable to dental PM in addition to using the experience of this research to enhance the BridgesPM1 data model.

7.6.2 What is a data reference model and why do we need one?

According to the Organization for the Advancement of Structured Information Standards (<https://www.oasis-open.org/committees/soa-rm/faq.php>), a reference model is an abstract framework for understanding significant relationships among the entities of some environment i.e. in a domain-specific ontology, and for the development of consistent standards or specifications supporting that environment. A key element of a reference model is that it should be based on a small number of unifying concepts and be fit for use as a basis for education and explaining standards to a non-specialist. A reference model is useful by defining how these concepts relate to one another using a particular data management technology, e.g. an entity relationship diagram. OASIS also maintain that a reference model must not be directly tied to any standards or technologies, but seeks to provide common semantics, used unambiguously across and between different implementations. This frame of reference should then be capable of being used to communicate ideas clearly among members of the same community. Simply put, it is a model to improve communication between people. It addresses the question: "Is this what you want?" In this research a reference model was necessary to define and document the BridgesPM1 data extract and to address RQ5.

7.6.3 Dental Data Reference Model Development Method

Similar to the approach taken by Mans et al (2015, p. 28), this research used the entity relationship (ER) underlying the Bridges EHR data extract as the starting point for the

dental data reference model (See Figure 4-3). Creating a data reference model in the form of an entity relationship diagram for Emergency Room data is also a preparatory step in the question driven PM methodology (Rojas, et al., 2017). This model exists at the logical data model level as it is not technology specific but contains full entity and attribute lists. The model, in conjunction with the data dictionary in Appendix 10.3 provided detailed information on the entities (tables) and attributes (columns) available in the data.

Two standards relevant to PM of dental EHR data were identified in the literature: the PM healthcare reference model (Mans, et al., 2015) and the ANSI EHR standard (American National Standard/American Dental Association, 2013). These standards provided the basis for first, assessing the ‘completeness’ of the available BridgesPM1 dataset and second, for making recommendations for an ‘ideal’ dataset.

The research data was compared to the standards and a gap analysis was completed. This step positioned the dataset within the proposed standards and produced a generalisable benefit in informing future dental EHR designers wishing to accommodate process and data mining. It also proved useful as a framework to record and manage missing entities and attributes as they arose during the research.

7.6.4 Results

7.6.5 The Healthcare Reference Model

The Healthcare Reference Model (Mans, et al., 2015) (HRM) was created to address the complexity of Hospital Information Systems’ (HIS). This complexity often makes them difficult to understand and difficult to locate within them, the data required for PM data science. The model consists of 122 classes in the form of a UML class diagram showing entities, their attributes and the relationships between them and is intended to assist locating the available PM data in these systems. This is a highly detailed schematic with input from three hospitals and their HIS professionals. For this research, it suffices for us to use the higher-level categorisation of these classes (Mans, et al., 2015, p. 29).

The classes were roughly categorised into 9 groups as in Figure 7-27: general patient and case data, process steps, medication, patient transport, radiology, document data, organisations and buildings, nursing plans and pathways. Their relationships can be approximated as in Figure 7-27 below, with patients and their illnesses at the centre of the design and having a 1 to n relationship to the other entities.

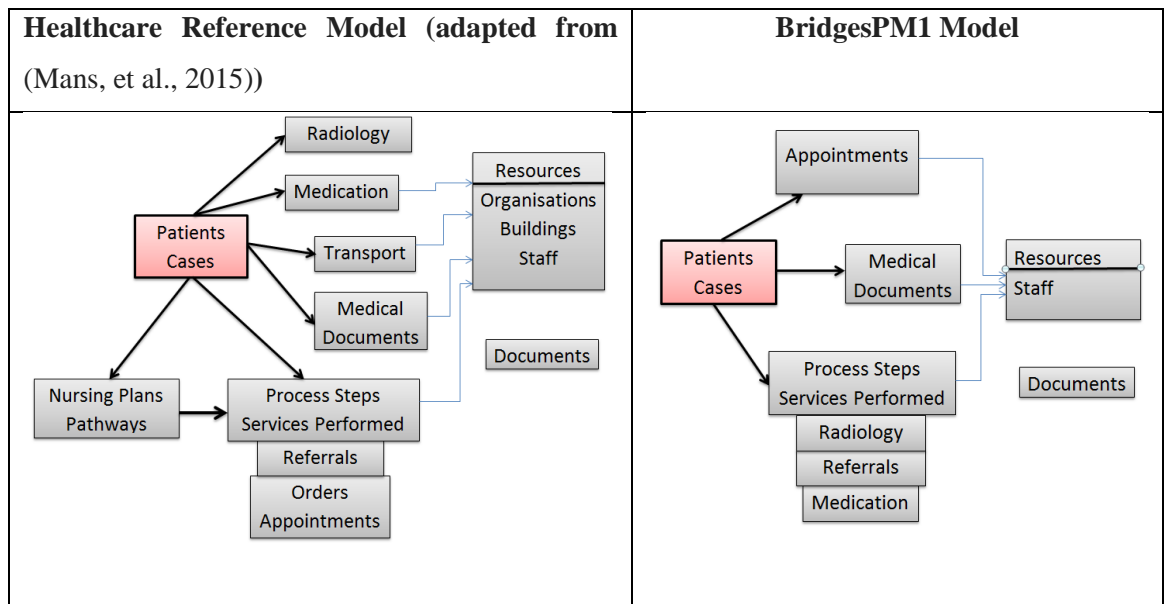


Figure 7-27: HRM comparison to BridgesPM1

The Patient and their Cases (illnesses) are central to the model and during their treatment, services are performed, process steps are executed, resources are utilised, and documents are created in many of these steps. Transport, Organisations, Buildings, Orders, Nursing Plans, and Pathways do not exist in the Bridges Dental EHR. In the Bridges EHR, Medication, Referrals and Radiology are not separate classes, rather, they are recorded as Process Steps (treatments) in Figure 7-27. All the elements use staff resources. Documents are created by many of the classes.

In the HRM, Nursing Plans and Pathways are built into the patient care process and are used as guides and checklists for the treatment of the patient. These do not in the Bridges EHR but would be valuable as evaluation of PM's uses for comparing the actual treatment process with the clinical guidelines is one of this research's objectives.

To identify data that would enhance the value of dental EHR data for PM, the HRM groups are now compared to their corresponding BridgesPM1 group.

7.6.5.1 Mapping the BridgePM1 classes to HRM

The BridgesPM1 classes are often analogous to the HRM classes. Also, both models contain classes which are not represented in the other, unsurprising given that Bridges is a dedicated dental EHR and such dental information would not normally be collected in general hospital settings. On the other hand, the HRM is based on much broader hospital information systems incorporating systems and data not normally found in dental settings. Table 7-9 below indicates the mapping of the classes in the HRM General Patient and Case group to the Bridges-PM1 classes.

In the HRM, it is possible to assign VIP status to a patient who may have multiple health problems and multiple risk factors. They must have at least one case (illness) leading to diagnoses, and possible complications. The case can be assigned to a person or another case or further 'classified', but it is unclear from the documentation what this means.

Table 7-8: HRM Classes mapped to BridgesPM1 Classes

HRM Classes: General Patient and Case Data	BridgesPM1 Classes
Patient	Is similar to PMClient
Case	Is similar to PMTreatment Course
Diagnoses	Is similar to PMTreatments, PMConditions
Complications	None. Free text notes (in original EHR) not in Bridges-PM1
Health Problems	Is similar to PMQuestion, PMQuestionnaire, PMQuestionAnswers
Assignment of a Case to a person	Is a part of PMTreatments
Assignment of a Case to another case	None
Patient Risk Factors	None. Caries Risk Assessment functionality (in original EHR) not implemented
Logging for VIP	Similar to red/green flag alerts for allergy etc
Case classification	None
None	PMChart
None	PMTooth & PMToothpart

Each of the cases above can have several steps. The HRM classifies three groups of classes forming process steps at various degrees of granularity as: referral, diagnosis, & treatment, and orders and appointments. HRM's process-step classes and the corresponding Bridges-PM1 classes are shown in Table 7-9 below.

Table 7-9: HRM Process Steps Classes mapped to BridgesPM1 Classes

HRM Classes: Process Steps	Bridges-PM1 Classes
Case	Is similar to PMTreatment Course
Reference referral data	Is a part of PMTreatment
Referrals	Is a part of PMTreatment
Surgery Diagnosis	None
Surgery Complications	None
Service Catalogue	None (in original EHR)
Movements for Case	None
Services Performed	Is similar to PMTreatments
Involved staff Members	Is a part of PMTreatment
Organisational Units	None
Building Units	None
Occurred Events	None
Surgery, Radiology, Cardiology, Medical Service, Non-Medical Service, Context of Service	None
Patient	Is similar to PMClient
Clinical Order, Item of Clinical Order	None
Appointments	Is similar to PMAppointment, PMAAttendances

The HRM includes a detailed group and structure for the prescribing and administration of medication. This is a much simpler process in the Bridges EHR, with a prescription being a self-contained item within a treatment course as shown in Table 7-10 below.

Table 7-10: HRM Medication Classes mapped to BridgesPM1 Classes

HRM Classes: Medication	Bridges-PM1 Classes
Drug Order	Is a part of PMTreatment
Case	Is similar to PMTreatment Course
Patient	Is similar to PMClient
Multiple detail tables	None

Similarly, minimal detail is recorded for radiology in the BridgesPM1. X-rays are managed as a treatment item. Also, there is no transport element in the recording of dental service in BridgesPM1. While various documents have been created and stored in the original Bridges EHR, for the purposes of maintaining the anonymity of the patients, these have not been included in the BridgesPM1 extract. Organisations and buildings are not managed in the Bridges EHR. Likewise, Nursing Plans and Pathways are not managed.

7.6.5.2 The ANSI/ADA Standard No. 1067-2013

The Electronic Dental Record System Standard Functional Requirements is published by the American National Standard/American Dental Association Standard No 1067, known as ANSI/ADA 1067-2013 (American National Standard/American Dental Association, 2013). Informing “...*those concerned with secondary use of EHR data and national infrastructure what functions can be expected in an EHR System.*” (American National Standard/American Dental Association, 2013, p. 12) is identified as a typical use of these functional requirements.

The standard provides guidelines and recommendations for functions to be performed by dental computer systems to document dental health services in a care environment, described in a conceptual hierarchy, employing the concept of functional granularity as presented in the HL7 Functional Model, an international standard that presents an organized list of functions associated with an EHR system. The requirements do not specify how the EHR system is to perform these functions, merely whether the function is mandatory (SHALL), recommended (SHOULD) or optional (MAY).

The standard defines many functional requirements fundamental to facilitating PM research e.g. recording the events and steps in a treatment plan.

Events e.g. (10.1) The electronic dental system SHOULD have a capability to manually enter the order in which a care recipient is to receive diagnostic services, (11.3 & 11.4).

Timestamps e.g. The electronic dental system SHALL provide a capability to identify and persist the date and time of the health care event, (15.27) The electronic dental

system SHALL provide the ability to capture dates associated with medications such as start date, fill date, and end date, (32.1) The electronic dental system SHALL have the capability to track the completion status of individual steps or tasks in a care plan, (36.1) The electronic dental system SHALL have the capability to track the completion status of individual steps or tasks in the delivery of indirect healthcare services.

Resource Usage e.g. The standard also provides requirements for tracking the resource usage associated with care options. This is a valuable function for PM. The standard recommends that the electronic dental system record an estimate of resources required to deliver healthcare services; (11.1). The electronic dental system SHALL provide a capability to identify the location at which health care services were delivered to the care recipient. (26.1) The electronic dental system SHALL have the capability to associate a specific type of dental equipment needed for a care option, (26.2) The electronic dental system SHOULD have a capability for the care provider to associate specific items of dental instruments or equipment with care options, (12.1) The electronic dental system SHALL have a capability to record the routing of the care recipient to receive services, (39.2) The electronic dental system SHALL have the capability to capture the details of all components that are used in the preparation of materials and devices used in care support, (29.28) The electronic dental system SHOULD provide the ability to display a list of care plans and instructions indexed by provider, problem, and date.

Outcomes e.g. The standard provides guidance for the EHR's ability to support the health care provider's decision process; (19.1) The electronic dental system SHOULD have the capability to provide decision support for the providers' clinical decision processes, (27.1) The electronic dental system SHALL have the capability determine an expected outcome for each care option, (27.5) The electronic dental system SHALL have the capability to present a list of outcomes with expected probabilities to the clinician or care recipient, (27.6) The electronic dental system SHOULD have the capability to analyse the outcomes achieved in population of care recipients treated by the clinician to identify outcomes, (29.30)

Care Pathways/Standards e.g. The electronic dental system SHALL provide the ability to present health standards and practices appropriate to the user's scope of practice Interestingly, the standard provides some guidance on general functionality required for research; (45.1) The electronic dental system SHOULD manage information about the inclusion or exclusion of a subject in a research study, (45.2) The electronic dental system SHALL have the capability to de-identify data associated with a research study, (45.3) The electronic dental system SHALL have the capability to communicate research study

data to the responsible organization, (49.5) The electronic dental system MAY provide an EHR data mining and analysis capability, (49.19) The electronic dental system SHOULD provide a capability to summarize information based on date or date range, chronology, patient characteristics, clinical fact, diagnosis, problem, etc.

The ANSI/ADA standard is vague on role a dental EHR in supporting clinical decision support, data mining and research. The three are intimately linked and EHR users and designers would benefit from a consistent approach to them.

7.6.6 Proposals for a Data Reference Model for Dental Research

The entity relationship for the BridgesPM1 research dataset presented in Figure 4-3 above provides a strong starting point for reference data model to facilitate PM in dentistry. While data in the model provides rich information for creating patient cohorts, calculating associated oral health outcomes and generating process models, there is potential to enhance the functionality by developing the data model in number of directions using the direction from the above standards and the specific experience of this research.

7.6.6.1 Care Pathway Functionality

Nursing Plans and Pathways are built into the HRM model. This facilitates introduction of SOPs and clinical guidelines to the treatment process. Similar functionality would be a valuable addition to the BridgesPM1 model. The underlying dental EHR should integrate clinical guidelines and recommended treatment pathways such as those proposed by the NHS and the (National Institute for Health and Care Excellence (NHS England, 2009; NHS, 2012; National Institute for Health and Care Excellence, 2018). These should then be included as a new class in the data extract facilitating rapid creation of reference care pathways for conformance and compliance checking. The data model would be expanded to include the new pathway entities and relationships in Figure 7-28.

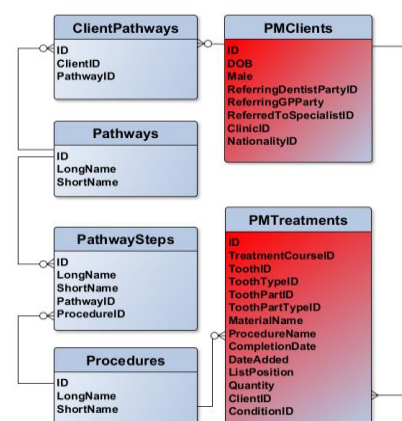


Figure 7-28: Data Model extension to cater for Care Pathways

Once a care pathway is chosen for a patient, a list of proposed pathway steps would be created in the PMTreatments table. It is proposed that this could be altered by the user to suit the individual circumstances.

7.6.6.2 Diagnosis-treatment Pairs

The HRM makes a clear distinction between diagnoses and treatment steps. This distinction is valuable and should facilitate incorporation of diagnosis-treatment pairs as facilitated in SNOMED-CT if required. While BridgesPM1 does facilitate automatic generation of treatment plans based on a dental charting (diagnoses & pre-existing treatments), it does not distinguish between diagnoses and treatments in such a clear-cut manner. The author believes that diagnoses-treatment pairs can help introduce evidence-based guidelines and that this would be a valuable addition to the model. This will involve an enhancement to the existing model as in Figure 7-29.

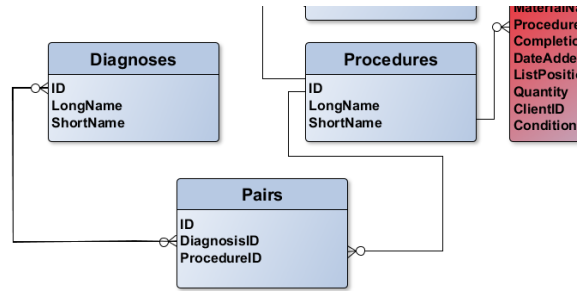


Figure 7-29: Data Model extension to cater for Diagnosis-Treatment pairs

7.6.6.3 Oral Health Outcomes Extension

The data extract would benefit from a broader approach to measuring oral health outcomes. BridgesPM1 currently only uses DMFT as an outcome. Other measures such as Quality of Life measures, ICDAS, Community Periodontal Index of Treatment Needs (CPITN), would add further depth to the research. A general data structure facilitating the creation of user-defined oral health measures would add to the value of the data extract by maximising its usefulness. These oral health measures could be implemented as extensions to treatments such as ‘Initial Exam’ or any other event considered appropriate. The feasibility of using a specific oral health outcome measure is dependent on the appropriate data existing in the EHR in the first place e.g. full ICDAS coding records six levels of decay. This level of detail is not recorded in the Bridges EHR.

7.6.6.4 Free-text Notes

The data extract would be enhanced by inclusion of free-text notes which often contain valuable additional clinical information. X-rays, images and other documents similarly add value. The drawback, and the reason these data elements were excluded from the BridgesPM1 data extract, lies in the increased danger of re-identification of the individuals given the unique nature of some of these data elements. Inclusion of this data with the extract would require pre-processing to remove any data with the potential to

identify an individual. From a data perspective, this enhancement would simply involve additional attributes e.g. 'Notes', at the patient level and at the treatment level.

7.6.6.5 Periodontal Data/Other Specialisations

Specific elements relating to periodontal health would add value to the dataset and present additional opportunities for investigating links to other health problems. A summary periodontal status could be implemented using the CPITN system where the mouth is divided into sextants and a single value recorded for each. This could be implemented by expanding the charting entity as in Figure 7-30.

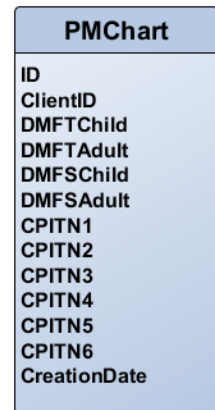


Figure 7-30: Data Model extension to cater for CPITN

A comprehensive periodontal charting could also be considered. This would consist of recording six pocket-depth readings per tooth and additional tooth properties such as furcation, mobility, suppuration & bleeding. Data required for other specialisations such as orthodontics/endodontics would also be desirable.

7.6.6.6 Fluoridation Status

Fluoridation status has been shown to play a significant role influencing oral health outcomes. Inclusion of this data adds value especially when evaluating the data quality against gold standards. One implementation of this would require recording of the individual's water-fluoridation status to be stored with the other demographic information. An alternative implementation would match the individual's address to a water-fluoridation knowledge-base. This could be done using postcodes or small-area-codes and removes the burden of collecting the data from dental service providers and solves the problem of a patient not knowing if their water-supply is fluoridated.

Both approaches suffer from the shortcoming that a person's water-supply fluoridation status may change over time. A more complex solution is required to address this, and this would entail recording a history of a patient's fluoridation status. It is unknown if this is practical or if it would be of any value.

7.6.6.7 Procedure/Event Mapping

Mapping of events to standards such as SNOMED/SNODENT and mapping to high abstraction levels e.g. 'Restoration' or 'Prevention' would help to address the inherent complexity of healthcare processes and potentially improve the quality of the process

models. If done at the EHR level, this would ensure that domain expertise is used to create and verify the mappings. There are two options for executing this. Attributes can be added to the event table, in this case the treatment items table, or, a separate normalised table containing the mappings could be implemented, the latter shown in Figure 7-31.

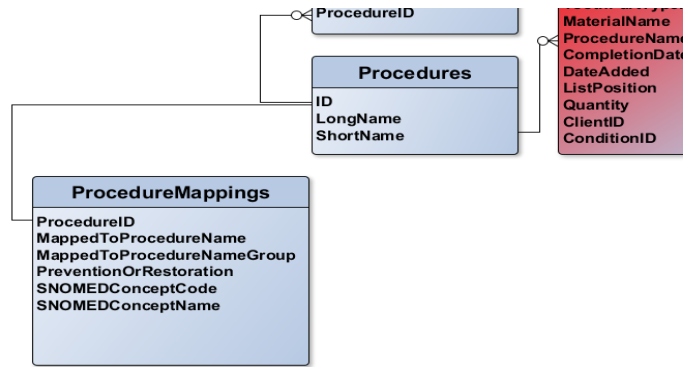


Figure 7-31: Data Model extension with Procedure Mappings

7.6.6.8 Proposed Dental Data Reference Model

Combining these proposals with the existing data model is presented in Figure 7-32.

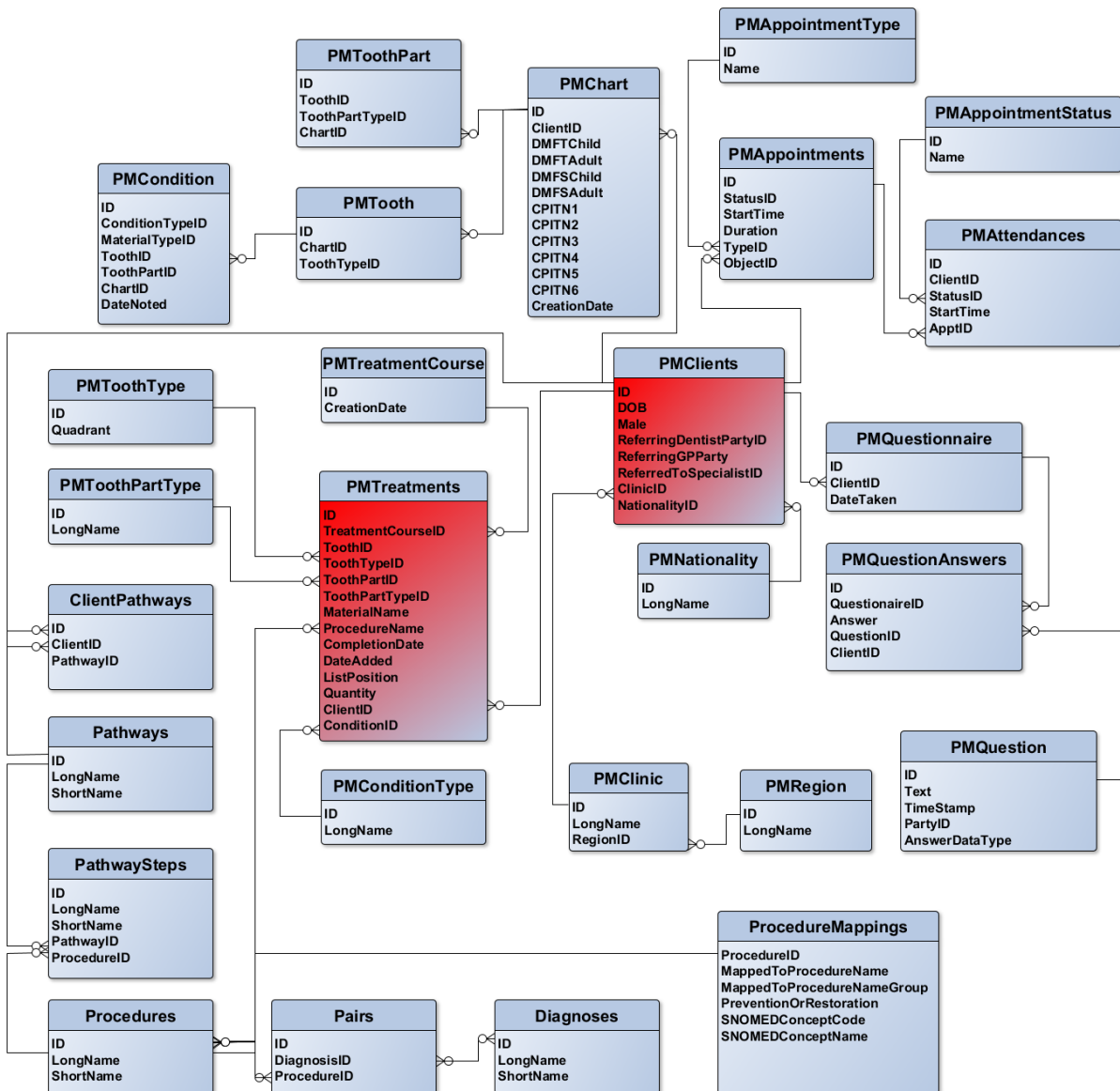


Figure 7-32: Proposed Dental Data Reference Model

7.6.7 Discussion / Limitations

The development of this model was not a perfect process. The HRM is generated from the information systems of several hospitals without any specified dental service and many components were not relevant to the dental service on which this research is based. The ANSI model is a functional description of what an EHR should be capable of rather than a logical data model format which would facilitate direct comparison. However, together they provided an external point of reference and helped identify gaps in the starting data model. It is notable that the starting model was from an operational EHR was the result of intensive work over several years with many dental professionals in the HSE, the EHR's host organisation. However, the validity of the model would have been enhanced if it could have been compared with additional EHRs in other organisations. While useful, it is also limited by the 'unknowability' of other potential RQs. There are many areas of specialism in dentistry not considered (endodontics, orthodontics etc.) each having their own specialist data requirements to assess process and outcomes. Only when the RQs are finalised could an ideal dataset be described.

7.6.8 Conclusions

This question aimed to identify data that would be needed in an EHR if applying the new PM approach to discover dental care pathways and facilitate the evaluation of policy implementation and took the approach of enhance the initial BridgesPM1 data model, with additional desirable entities and attributes identified from existing standards and the experience gained in this research

8 Discussion

8.1 Introduction and Overview

This research commenced with a literature review of the existing peer-reviewed dentistry related PM publications and publications relating to primary care and public health in the area of process-oriented data science. The review showed that while data mining has been carried out in many areas of dentistry, all existing PM research was focused on the steps involved in the delivery of individual high-end treatments, crowns and implants. Within that work, dentistry was merely a case-study in two of the three dental PM publications. The data used in those publications came from private practices and their associated laboratory. There was no prior research on PM of large EHR datasets to establish the processes of delivery of dental care as a whole, and no research using ‘big data’ in a healthcare context, data often originating from the delivery of public health or from large insurance databases. This motivated the author to address this gap.

8.2 Reflections on the Approach

While synthesising the broader literature on PM, variability in the use of common terms became apparent. This is understandable as PM is a relatively new area of research and it is being approached from several perspectives by researchers with informatics, computer science, statistics, machine learning, and other backgrounds. Previous authors often approached the topic from their unique perspective, bringing their own vocabulary and domain language to their publications. The author decided that it would be worthwhile to standardise the terms for this research and proposed a basic, concise vocabulary to describe the main data components and vocabulary in common use in healthcare PM and how they are used in this thesis. Some of the ambiguities in the use of PM terms were identified in this thesis and the process of creating a consistent vocabulary in the area was initiated. This is a useful starting point for the PM in healthcare community to develop a comprehensive, consistently used vocabulary or ontology of concepts, hierarchies and relationships and the idea has been well received and recognised as necessary at international conferences. The author does not expect his interpretation or suggested resolution of any ambiguities to be universally agreed, rather that they start a conversation with others and ultimately lead to an accepted terminology and agreed usage. This will ease the task of communicating with domain experts and other stakeholders. Future work to develop a comprehensive, universally agreed, vocabulary of the emerging discipline of PM in healthcare would be a valuable contribution.

The author received ethical approval and data controller permissions to access an anonymised extract of the Bridges/HSE EHR from the dental public health service in Ireland in order to apply data analysis techniques on, with an emphasis on process-oriented data science. An additional aim was to document this research and analysis in a detailed methodology that will provide useful guidance for executing future analysis of such large dental datasets.

This appears to be the first time that these technologies have been applied to a large clinical dental EHR dataset and accordingly it required a fresh and detailed approach in the research methodology. Existing PM methods were assessed and a methodology, known for convenience as PM4D, appropriate to for this dental EHR research, was documented. PM4D added additional steps to the existing methods in order to facilitate PM of dental EHR data in a research environment. The experience of this research is described in Section 6.4. PM4D identifies in detail the inputs and outputs for each step, the artefacts created and where they are to be found in this thesis. PM4D addressed the acquisition of the data; the process of obtaining ethical approval and the data-owner's permissions. Within this, some of the unique requirements of using healthcare data were addressed such as the issues of anonymizing, transfer and securing of the data and the research environment. PM4D provides a structured approach and may act as a checklist for future research in this area. During this research, a secondary methodological question emerged – how to address policy and strategy questions using EHR data. This necessitated some steps additional to PM4D and provides a way of thinking about using large datasets to evaluate policy or strategic initiatives. This approach was applied consistently in the experiments in RQ4.

8.3 Managing the Data Environment, the Data Quality, and the Data Analysis

8.3.1 Data Environment

Creating a stable technology environment for this research was a key step. The research was carried out in the Windows 10 environment. The primary research database was created in SQL Server 2017 using scripts executed with the SQL Server Management Studio and incorporating functionality from SQL Server Integration Services and SQL Server Analysis Services. However, most of the data analysis was programmed in Python, utilizing the Spyder and Jupyter Notebooks modules within the Anaconda Integrated Development Environment. The database creation, data transforms, data profiling and

analysis and the creation of the event logs for PM are all scripted and reproducible. The automated scripts can be easily edited and rerun if necessary. This flexibility proved invaluable as it was frequently necessary to rerun elements of this research, due to its novel nature and the iterative nature of defining the cohorts, defining outcomes and deciding on analysis methods.

What is the ideal dataset for PM research using dental EHR data? To answer this question, the author looked at the available standards and models and proposed an enhanced dental data reference model for consideration in future research. PM's healthcare reference model (HRM) provides a comprehensive, if aspirational, ideal dataset capable of addressing questions and delivering insights within the many types of PM. It was clear from the outset that such comprehensive data was not available to this research. The dental EHR consisted of data from a single organization, primarily dedicated to the delivery of school dental screening services and dental care to under-16s and special-needs adults. Comparison with the HRM and the ANSI standard generated an important output of the research, namely, proposals for data model to facilitate PM from a dental EHR. These proposals for the data model are of benefit both to those wishing to undertake PM on dental EHR data and to those designing EHRs, ensuring that process-oriented data analysis is facilitated at the design stage i.e. that the EHR is 'process aware'. While acknowledging that each research project has its own data requirements, the model presented can be used as a valuable and timesaving starting point for other researchers seeking access to EHR data. It provides a framework for discussion of what data is needed for the research, what data is available and assessing the impact of this gap and these ideas have been presented, discussed and well received at international conferences.

The data-use agreement for this research required anonymisation of the individual-level dataset and many attributes of the patients were not available due to this. Access to a more comprehensive dataset with full details would expand the range of data mining techniques applicable and linking the dental record to the patient's general health record would also open new avenues for research. These areas require careful crafting of data agreements and adherence to data protection requirements to ensure that EHR data's use in research is developed in a sustainable and secure manner.

8.3.2 Strategies for Data Quality

One of the key issues arising from the advent of the use of EHR data for research is data quality (DQ). DQ issues are dealt with theoretically in many publications and various

dimensions for classification of quality issues have been proposed. An initial assessment of the dataset revealed many potential DQ issues arising from differing sources – from the developers of the application, the users, the data extraction process, and potentially from the research itself. Previous work using similar data from the dental EHR highlighted some quality issues, for example, inconsistencies in recording fluoridation status, trauma status and gender. This research also benefitted from the author's intimate familiarity with the EHR, its design and its day-to-day usage, allowing a birds-eye overview of possible sources of DQ issues and their impact. The author co-designed the underlying data structures and much of the user interface as well as implementing the EHR application in the clinical setting. He defined the research dataset for extraction and executed the technical data transformations within the research. Accordingly, the author was ideally positioned to identify potential DQ issues arising through all the phases of the data's existence. Classifying and managing the numerous issues remained problematic as it became apparent that they arose from various sources e.g. application users; could affect the data at different levels e.g. row or field level; and were identified by various means. Further, the impact of a data issue was dependent on the RQ or experiment e.g. date-of-birth was essential for some queries and irrelevant for others. The author chose to examine these issues in a structured manner and to document and audit every change or transformation made to the data, whether such a transformation was to address a DQ issue or to enrich the data for analysis purposes.

The complexity of the DQ issues was such that it necessitated a formal framework i.e. the care pathway data quality framework (CP-DQF) for managing and, if possible, mitigating these data issues. The framework facilitated the systematic identification, recording, managing and, in some cases, mitigating of the quality issues. It also facilitated reporting of the issues and their scale. A database of potential DQ issues was established, both from the author's own experiences with the application development and with the data itself and from the existing published literature on DQ and forms another output from this phase. This proved to be a valuable and productive undertaking and demonstrated that formal DQ assessment is an essential step in research using EHR data. The framework developed has the potential to be generalised to other research using EHR data and the author believes that the framework and the list of discovered DQ issues can assist other researchers to discover, manage and mitigate the DQ issues in their own work. It provides a valuable, timesaving, pragmatic starting point for other researchers undertaking research using EHR data.

Future work should develop the DQ framework in several ways as detailed in Section 10.19.13.11. First, for ease of use, a graphical user interface would be very useful, rather than the scripting environment used in this research. The framework requires further validation through application to scenarios such as multiple heterogeneous data sources, differing data models and database technologies. This area also requires further DQ metrics in addition to the simple metrics in use in this research i.e. percentage-defects. These developments could deliver a valuable easy-to-apply module for researchers to assess the quality of their data and report on the quality issues in a consistent fashion. The addition of these features would make application of the framework to additional datasets easy and increase the generalisability of the work.

8.3.3 The Data Analysis

When the quality of the research data was in-hand it was then possible to enrich it with aggregations and other calculated values such as oral health outcomes and other markings and to answer the question: What is in the dataset? As there were many data tables with complex interrelationships and containing tens of millions of data-rows, answering this question was not trivial and demanded the use of efficient data querying methods and, in many instances, innovative visualisation techniques. Here, histograms and distributions, heat-maps, bar-charts etc. were used to convert the large dataset, key entities, and attributes such as procedures, patient ages at treatment, and their oral health status into comprehensible formats and facilitated the communication of complex information. Visualisations of the distribution of DMFT over the individual teeth were created as well as the trends over the timespan of the dataset. The potential for other visualisations such as geo mapping and heat maps to enhance understanding of the data were also introduced. Gaining an intuitive understanding through this data profiling is an essential step, but an often-ignored research step, especially in published articles. In this research, the data profile defined the environment and created the context within which additional unique experiments could be carried out i.e. the PM and validation experiments. The technologies were very flexible and agile once implemented and highlighted a key benefit of using EHR data in research where an iterative approach to answering the RQs can be employed at low cost.

The code for all transformations was fully documented and retained. It is the author's view that this is an essential step in using EHR data for research and if possible, publications should routinely incorporate this information. This enhances the reproducibility of the research, adds rigour to the methodology and is a valuable step in

gaining confidence in the results generated by EHR data research. The motivation for almost all the data transformations in this research was to simplify the substantive research queries such as those creating cohorts and outcome measures – both from a coding perspective and for better computational performance.

While the methodology described provides a structure to execute process-oriented data science on dental EHR data it must ultimately deliver clear benefits to dental policy makers and strategists. To help do this, validating experiments were carried out using the data. First, established care pathways and clinical guidelines were used to see if PM could assist in assessing compliance with these. PM of a small data extract provided an automatic breakdown of the process flow into two clear pathways – one being routine and one demonstrating the common emergency treatment pathways, very similar to that proposed in the Steele Report (NHS England, 2009). This was initially surprising as the proposed pathway is a recent innovation and clearly not implemented at the data's source but immediately demonstrated that the EHR data was recorded at the appropriate level of detail for comparison with such *de jure* models. Future work would develop these comparisons to identify important model variations and deviations at various levels – clinic deviations, deviations by individual practitioners, and system-wide deviations.

The research then investigated an undesirable outcome, extraction under general anaesthetic, to establish if PM technologies could offer insights on the pathways leading to this. The results showed that most teeth extracted under general anaesthetic had had no prior treatment in the service. The average waiting time for those patients who received a tooth dressing before the extraction was approximately six months and indicated a service under resource pressures. The experiment demonstrated the value of PM for addressing specific clinical question. Other clinical questions could also be addressed with similar techniques and this area has significant potential for further research.

The research investigated if PM could produce insights around the effects of policy and strategy decisions. It examined PM's ability to assess effects of age at first screening on DMFT outcome at 12 & 13 and examined its ability to assess effects of frequency of screening on DMFT outcome at 12 & 13. The results showed that we could find cohorts in the dataset representing the different sides of policy or strategy decisions. We could calculate their oral health outcomes and generate the treatment process models for the cohorts. While the cohorts showed some differences in health outcomes at first glance, this aspect was not developed with statistical proofs as the aim of the research was to show the potential of applying PM to EHR data and not to draw conclusions from the

data itself. The treatment process models also could demonstrate the differences between the pathways followed by the cohorts and shows how PM of large clinical EHRs can be used to assess strategy and policy decisions.

That said, it was not the aim of this research to wade into the debates on any of the above questions, rather to demonstrate the applicability of the technologies in this research to addressing such questions and to show that data mining and PM technologies can be successfully applied to a large dental dataset. The oral health outcome used, DMFT, has well documented strengths and weaknesses and is subject to many confounding factors, few of which were directly addressed in this research. The key deliverable of this research is the methodical application of these emerging technologies to large dental datasets and the assessment of these technologies' usefulness in assessing the impacts of strategy and policies that are visible in the datasets.

8.4 Principal Outputs of this Research

8.4.1 Data Reference Model Proposals for Dental Process Mining

The literature review identified that there had been no previous research applying process-oriented data science techniques to large dental datasets. This raised questions about what types of data are necessary for such research and what data would be optional but valuable in creating additional insights. The entity relationship (ER) for the research data presented in Figure 4-3 provides a starting point for a data reference model for PM in dentistry. It is enhanced using the Healthcare Reference Model (Mans, et al., 2015, p. 29) and the dental EHR standard (American National Standard/American Dental Association, 2013). Further enhancements arose during the research and were incorporated into a set of proposals for development of the ER used in this dataset. These proposals include integration of clinical guidelines, cross references to external data, additional information on periodontal health status etc. The proposals are presented in Section 7.6.6 and are a valuable resource for both EHR designers and those with access to large dental datasets for PM and similar analyses. They could also result in recommendations for a reduction in the gap between the data necessary to assess clinical guideline compliance and the data available from a pragmatic, operational EHR.

Some of the weaknesses in the model in Figure 4-3 are addressed in the proposals e.g. the proposals noted that increasing the detail in the dataset increases the risk of re-identification of individuals. Additionally, the proposals do not incorporate a mechanism for linking the dental record to the patient's general health record.

Collecting clinical information in EHRs takes time, often the clinician's time. There should be a clear benefit, visible to the clinicians, to make this additional effort worthwhile. Incorporating the information into a decision support mechanism or a learning healthcare system would be an ideal way to motivate accurate recording of pertinent information. Facilitating this is not addressed in the model.

8.4.2 Addressing Data Quality (DQ) of a Public Health Dental EHR

This research developed a framework for managing DQ issues in this research. The decision to design the framework resulted from a realization that a structured approach was required to the management and mitigation of the multiple sources of DQ errors, sources of information about these errors, and differing levels at which they affected data and experiments. The framework facilitated the organised identification, classification and management of many DQ issues. No user interface was developed in the implementation of the framework and this would be a valuable aid to both demonstrating its usefulness and encouraging its further usage. Other areas for potential further development were identified in Section 10.19.13.11.

The second output from the DQ assessment is the list of DQ issues discovered in this research's dataset. This has value to others who may use this or similar datasets in the future and provides a strong starting point for their assessment of their data's quality.

8.4.3 Architecture and Environment for Process Mining Dental EHR Data

The architecture in use in this research as shown in Figure 4-19 has been enhanced to reflect the potential benefits to be gained by having consistent domain expertise input in ontological and clinical matters. This, while desirable, is often impractical due to resource constraints. Linking to external datasets opens many possibilities for exploratory analysis but was not achieved in this research. A user interface allowing public health decision makers directly access the data through the technologies used in this research is also a desirable enhancement. This enhanced architecture is represented in Figure 8-1 below.

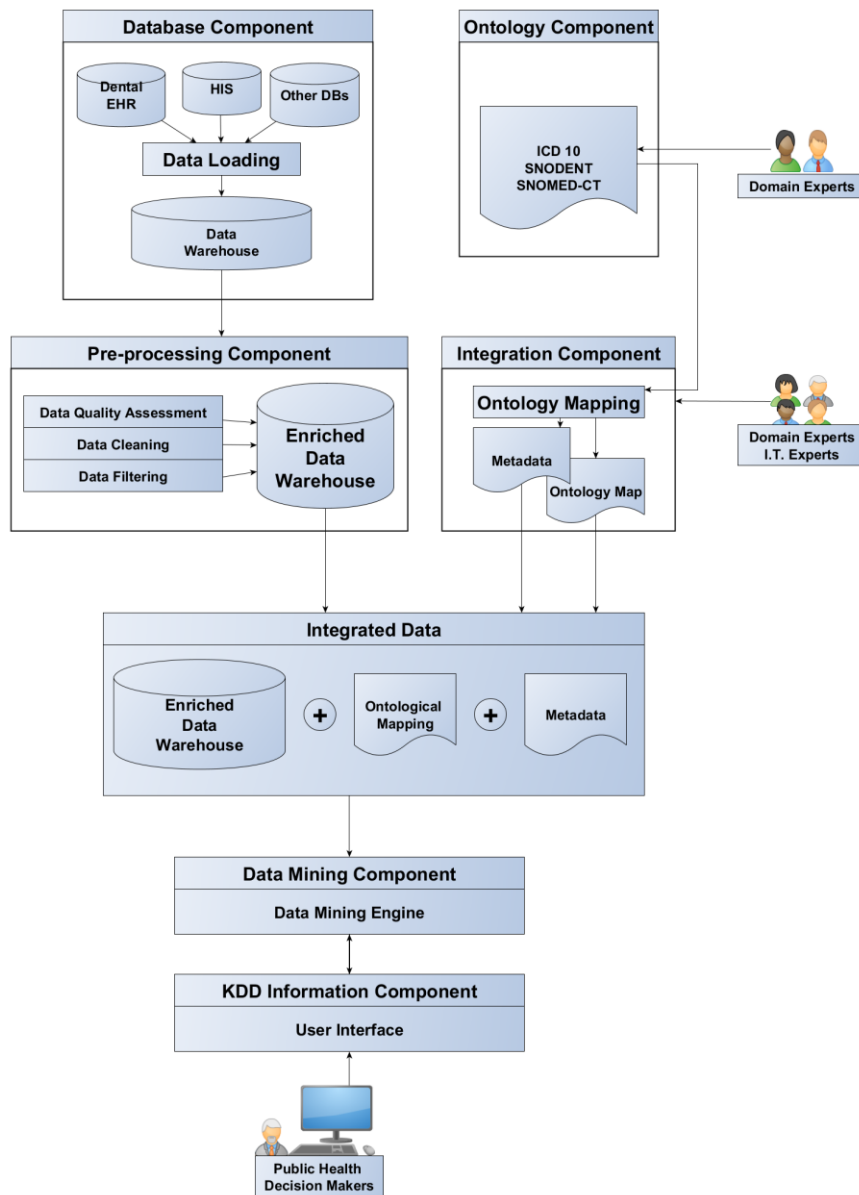


Figure 8-1: Proposed System Architecture adapted from Santos et al. (2013, p. 275)

8.4.4 Visualisation and Profile of a Dental Public Health EHR

Getting an overview of the contents of a large dataset is achieved by using data visualisations. Many of these visualisations are well known to us in our daily lives and are very effective at communicating information, examples being, bar & line charts, pie charts etc. Other visualisations such as stacked charts, histograms, heat maps, bubble charts, polar graphs, regression charts and combinations of these facilitate communication of more complex information, relationships between variables, trends etc. In addition to these, interactive and multidimensional visualisations offer even more scope to communicate the essence of vast and complex datasets and interrelationships to diverse audiences. Initial profiling and visualising our EHR's data were achieved with a geo-map showing concentrations of patients, bar charts & histograms of patients, ages etc. while

DMFT was addressed with stacked bar-charts showing values distributed over the individual teeth. At a geographical level heat maps were used. At the analysis stage combinations of bar charts, histograms and scatter plots were used to depict multiple data dimensions simultaneously such as age, DMFT, and numbers of patients.

These tools were indispensable for describing the data and its analysis, however, they are also limited, and the author regularly encountered difficulty explaining elements of the research whether at poster presentations or oral presentations. The limitations imposed by presenting complex data in static two-dimensional formats are significant and to fully exploit the benefits of the available technologies, interactive use of the data and animation are essential. Temporally driven animations would also be useful for showing development of data values over time. Also, linkages with external datasets such as income-levels and other environmental factors would benefit from such animations.

8.4.5 Initiated Development of a Vocabulary for PM

This research has initiated development of a consistent vocabulary for PM and propositions for resolutions to some conflicts in the literature. It was clear from the literature that there were multiple areas that would benefit from definitions of the terms and vocabulary in use. The author has started this process and the main outputs are the proposed clarifications in Chapter 2.6. Also, the graphical organisation of the terms as proposed in Figure 2-6 summarise the work to date. The main shortcoming of this work is that it is not complete and requires additional effort to identify further areas of ambiguity. Second, assessment of the quality of a process model is a complex and unsettled mathematical area and the author's knowledge of this area is limited and needs additional input from domain experts.

8.4.6 Documenting a Methodology for this Research's PM of Dental EHR data

This methodology integrates and builds on several PM research approaches. PM4D, the overarching methodology for applying PM to this dentistry research is summarised in Section 6.4 Additional steps for investigating the effects of policy or strategy decisions was also developed and summarised in Table 6-3. PM's ability to compare *de facto* process with established care-pathways was demonstrated as well as its ability to answer specific clinical questions. The value of PM to assessing the impact of policy changes on service delivery and oral health outcomes was examined. Whether situations viewable as policy strategy changes affected the oral health outcomes and the process of oral healthcare experienced by patients was investigated. Specifically, it looked at whether

varying scenarios from a school dental screening programme affected outcomes and treatment processes. This was, first-of-all, a process discovery exercise and compared ‘before and after’ scenarios as well as groups subjected to the intervention and groups not subjected to the intervention. It looked at the case perspective where cases can be characterised by corresponding or associated data elements such as timing, frequency of events and outcomes.

8.4.7 Demonstration of the Flexibility of Using EHR data in Research

The technologies and techniques used, while not trivial to operationalize, were very flexible once implemented. For example, while the data-profile and the validation experiments focused on data from 2005 and the following 5 years, it would be a trivial adjustment to carry out the same analysis for 2006 or indeed for any other year or combination of patient ages, number of screenings, etc. Other visualisations such as the distribution of DMF values over the individual teeth lends itself to displaying the distribution of diagnoses, treatments, or other indices such as ICDAS in a similar fashion. This highlighted a key benefit of using EHR data in research where an iterative approach can be employed to parameterize, fine-tune, and re-run data queries and subsequent analysis at very low cost.

8.5 Limitations of the Study

Using EHR data for research is a relatively new area, seeking to make novel findings from the large datasets now being generated as a result of the increasing use of computers for health records and insurance claims databases. Applying PM to these datasets offers a unique perspective otherwise unavailable with standard data mining techniques. The sequence of events and the temporal relationship between events is one of the unique outputs and combined with thorough data profiling can provide data-owners and researchers with novel insights into the care pathways being experienced by patients. The data in epidemiological studies is gathered using focussed methods and protocols, tried and tested over many years, and this is not always the case with EHR data which exists as a by-product of administration and recording of a patient’s clinical conditions and treatments. This requires that research using EHR data be approached cautiously to ensure that the findings are valid, reliable, and reproducible. EHR data does not have the same provenance as data from epidemiological studies so steps must be taken to establish confidence in it. DQ assessment is a key step. Faithful recording of data transforms is another. Applying strong, auditable methods to answering RQs using EHR data will

increase confidence in the findings. This research has addressed these issues but remains aware of the DQ challenges of using EHR data for research.

Another limitation of this research is the exclusive use of DMFT as the outcome measure. Its limitations have been well documented in this research and elsewhere however, it was the only one of the commonly used measures clearly available from the dataset. There are several other measures that could be considered in ideal circumstances such as ICDAS and quality of life measures.

In this research the focus was on the methodology and its validation rather than providing definitive answers to the validating questions. However, there is clearly room for the application of statistical analysis to the results of some of the experiments in this research, in particular those investigating the relationship between DMFT and the number of screenings and the age at first screening. Vital supporting information on fluoridation and socio-economic-status were missing from the dataset making any such analyses more difficult. On another track it can also be argued that the data is very close to the full population of school children in the area and hence, statistical testing may not be necessary. Any statistical testing would need to address the assumption of normality often associated with DMFT analyses and the appropriateness of various tests both parametric and non-parametric in use in the literature.

Also, it is most likely that less than 100% of the relevant patients were identified in the experiment cohorts, rather, it identifies those fulfilling the criteria in the experiments.

Although there are many PM techniques and algorithms and an increasing number of commercially available products are incorporating these developments, this research found that very few of these algorithms could produce process models comprehensible to dental domain experts. Healthcare processes' ad-hoc, complex, dynamic characteristics lead to spaghetti-type models and the shortcomings of existing algorithms were clear. While it is likely that valuable insights are hidden in these models, the limitations of presenting them on paper or small screen formats are clear. Metrics for dataset size/algorithm combinations would provide a valuable starting point for researchers. For example, it is difficult to interpret more than 30 different event types (nodes) with 60 connections (arcs) on an A4 sheet.

8.6 What unique insights does PM bring to analysing healthcare processes?

It could legitimately be asked what PM can do that cannot be done with traditional data querying approaches. For example, as addressed in RQ4, the mean time between tooth

dressings and GAx could have been established with a structured query of the data. Likewise, the total number of GA extractions and the proportions of teeth having no treatments prior to GAx could also be so established - without the help of PM techniques. What does PM add in comparison to direct data queries?

PM, when used as an exploratory data analysis tool, delivers a rapid, time-ordered overview of the data. This has the potential to present opportunities to develop interesting hypotheses for further research. Ad-hoc questions can be quickly tried out at very low cost. While many such questions could theoretically be answered using traditional queries, these queries are often very complex, prone to error, and accordingly, require a high competence and skill level from the researcher whereas in PM tools the coding complexity is mostly hidden in the algorithms.

Also, process models often provide more information than would be present in the results of a structured query, giving a richness and context not necessarily obvious in query results e.g. alternative or less-travelled care pathways, associations between pathways and oral health outcomes and temporal aspects of the pathways that would not be immediately visible from query results. PM also brings the benefits associated with data visualisation, reducing large datasets to comprehensible pictures, and providing an accessible and valuable tool for discussions between PM and oral health domain experts.

PM's types and perspectives as detailed in Section 2.6.3 offered many other ways in which PM can deliver insights from perspectives previously unavailable and traditional querying would, in this author's opinion, be prohibitively complex and prone to error. The established algorithms are in constant use and are being iteratively improved and many have their own inbuilt quality measures such as fitness and precision.

It is this author's view that PM goes far beyond hypothesis generation and offers much more than suggesting that X is associated with Y. However, it is only one tool in the data analysts armoury along with the traditional data mining, visualisation, and machine learning tools.

It is also worth noting that many of the limitations identified in this research would be mitigated in an operational environment not curtailed by research constraints such as the requirement for anonymised data.

8.7 Meanings and Implications for Clinicians and Policymakers

This research introduces the ability to monitor patient's care pathways and compare them to established rules and standards. This facilitates better oversight of the delivery of dental services and the identification of exceptions, outliers and unusual cases. It introduces

ways to monitor and evaluate the effects of policy and strategy changes and provide some much-needed feedback on the success or otherwise of such decisions.

The research shows how PM can add to the traditional ways of looking at policies and help identify how and why policies and parts of policies work or don't work. It facilitates asking questions in different ways and showed rich potential for exploring the process of health and disease over time in a novel way.

The research proposes a basic vocabulary for PM relevant to dentistry along with the proposed data reference model. These can be used as tools for effective communication with which policy makers can have productive discussions with information technology providers to ensure that systems are process-aware and have the functionality to provide them with the data necessary for effective policy and management decisions.

The research documents the detailed methodology applying these technologies to a large dental EHR extract and highlights the necessity for the data to be of good quality. Many of the DQ problems cannot be undone or repaired by the researchers, policymakers, or other secondary users and this research provides a structure that is available to all to check for DQ issues and then to manage and mitigate them if possible. This should provide focus for everyone to improve DQ and ultimately give the policymakers and clinicians confidence that they are maximising the utility of the EHR data and confidence in the decisions they are making using this data.

9 Conclusions

9.1 Review of Research Questions

This work has applied PM to a large dental EHR extract for the first time using the PM4D methodology documented in this research. This section reviews each of the research questions from Chapter 3 and assesses whether they were successfully addressed and answered.

Research Question 1: *Can PM discover care pathways, from a dental EHR?*

PM of the EHR data satisfied both of the success criteria. First, the research data proved suitable for creating PM event logs and producing models recognisable and comprehensible to our PM and dental domain experts. The models produced were comparable with the established care pathways and clinical guidelines allowing insight as to the degree of compliance of the *de-facto* models.

Research Question 2: *Can PM help assess compliance of real-world processes with recommended care pathways and clinical guidelines?*

It can be seen that the technologies used, when appropriately tailored to the research data, produced models that were comprehensible and legible. The pathways in the models were comparable with the recommended pathway from the Steele Report. Although the Steele Report was not intended for implementation and was not implemented in the HSE it serves as a template of a ‘typical’ public policy guideline and was useful for the purposes of exploring the data requirements to assess implementation of the guideline. The limitations of using dental EHR data for assessing compliance with highly granular, detailed, clinical, standard operating procedures was also demonstrated.

Research Question 3: *Can PM discover dental care pathways associated with a specific outcome – e.g. extraction under general anaesthetic?*

Again, the research data was shown to be suitable for creating PM ELs and producing models recognisable and comprehensible to our PM and dental domain experts. The models were of significant interest to our dental experts showing the proportions of patients who received no treatment prior to the GA extraction and average waiting times of 6 months between tooth dressing and GA extraction. These were interesting insights and could be applied to other clinical questions in a similar fashion.

Research Question 4: *Is PM and PM4D capable of assessing the impact of policy changes on service delivery and oral health outcomes, from the dental EHR.*

Though significantly more difficult to answer than the previous questions, requiring complex analysis and computer coding, the research data was shown to be suitable for creating PM ELs and producing models recognisable and comprehensible to our PM and

dental domain experts. The data analysis and process models were shown capable of delivering insights on the significance of the policy changes by establishing the most commonly occurring pathways, analysis of the oral health outcomes and temporal features of the pathways though it was unable to draw strong conclusions about the data itself due to the presence of confounding factors and the lack of availability of some data such as fluoridation status and socio-economic status.

Research Question 5: What Data is Needed in an EHR for Effective PM?

Answering this RQ was achieved by enhancing the initial BridgesPM1 data model with additional desirable entities and attributes identified from existing standards in the literature and through experience gained in the course of this research. The value of this enhanced model is difficult to be definitive about at this point, however, at the very least it provides a starting point for future oral health PM researchers to identify their research data requirements and to use as a communication tool with other stakeholders.

Research Question 6: When applying PM to routine dentistry data, what challenges does one encounter and how can these be overcome?

Answering this question in Chapter 5 identified some of the challenges encountered when applying PM to routine dentistry data. In particular, data access and the use of the Anonymisation Decision Framework to detail and mitigate the risks associated with using healthcare data for research was outlined. Data quality emerged as a key issue in this research and was dealt with in a structured manner with the Data Quality Framework. The issues of model complexity and spaghetti-models were examined, and the techniques employed by this research to reduce the effect were detailed.

The PM experiments do not stand on their own and were positioned in a stable environment, using data of known quality and provenance. Further intuition and unique insights to the dental EHR dataset and to the process of delivery of dental public health services were gained by profiling the data using advanced data and PM visualisations. Profiling the research dataset in this way gave an intuitive feel for the data and this research emphasises this as an essential step when using EHR data for research, although it is often ignored.

Addressing these public health questions in a way that delivers defensible results required meticulous preparation, data management, and auditing. Much of the existing PM literature fails to adequately address and document all the necessary steps and this thesis adopted a structured and detailed approach documented under the methodological title

PM4D. This methodology resulted in several key artefacts: The proposed architecture and environment for PM of dental EHR data incorporates the experience gained in this research and should expedite operationalising future research in this area. Data reference model proposals for dental PM are of significant value to both future researchers seeking research data and to EHR designers wishing to make their systems ‘process-aware’. Development of the care pathway data quality framework and identification of many potential data quality issues and sources provides an advanced starting point for assessing and managing data quality in future research and is an additional contribution.

In summary, the research demonstrated how process mining, data mining, and effective visualisations can provide much-needed insights to the process of delivery of dental public health services. The methodology followed in this research, applied to data from the Irish public dental health system, should extend to U.K. and international large datasets from public health and insurance claims for the purposes of managing and assessing care pathways. The further development of such methods should be a priority for both dental and medical healthcare providers.

9.2 Future Research Opportunities

Several publications relating to this thesis are planned. The CP-DQF will be further developed as detailed in Section 10.19.13.11 and the results published. There may also be an opportunity to publish the list of DQ issues found in this research as this would be a useful asset for future researcher’s undertaking similar research. The DQ research in this thesis was an important contribution to a recent ADVOCATE international data conference which aimed to develop a standard operating procedure for requesting and managing EU-wide large dental datasets and imminent publications are anticipated from that conference. There may also be opportunities to develop and publish the dental data reference model as presented in Section 7.6.6. Further publications may be possible in the dental domain looking at process mining’s applicability to the Steele Report as shown in Section 7.1.4 and its utility to address questions around specific outcomes such as extractions under general anaesthetic as shown in Section 7.2.

A further study using data from the Salud dental EHR in the University of Leeds, School of Dentistry is underway, and an abstract submitted to the BSODR meeting in Leeds, September 2019. It helps assess and monitor the care pathways of patients subject to a novel facial-pain intervention. This helps illustrates the generalisability of the methodological approach taken in this research. Consideration of the patients’ value

judgements as to whether proposed changes to the pathway are beneficial or not would help develop the links with patient-oriented technologies such as value-stream-mapping and could be incorporated into this planned study. A further abstract introducing dental EHR's potential to show links between oral health and general health has also been submitted to the BSODR meeting.

There is also potential to investigate other areas of debate e.g. comparing the outcomes and processes for treatments such as fissure sealants and topical fluoride applications.

Additional applications of PM4D in scenarios with multiple heterogeneous data sources will enrich many of the steps and add significant depth to the data quality framework, the proposed reference data model, the proposed architecture, and the data pipelines. Further experience using the methodology would also deepen our understanding of the range of DQ issues in different scenarios such as heterogeneous data sources and issues arising specific RQs and help to make these steps more generalisable for further use. This research's proposals on DQ and the proposed vocabulary for process mining in healthcare (Section 2.6) have already been well received at the PODS4H International workshop and it is anticipated that the author will contribute to the conversations in these areas on completion of these studies.

Production of generic code to evaluate the probability prediction of next event from a given decision point would be useful and could be combined with a simulation application such as NETIMIS to simulate the impacts of choosing the options available at decision points in the process model. This could also be potentially developed to optimise the care pathway automatically leading to better outcomes i.e. suggesting modifications to care pathways using predictive models and simulation. Future work using predictive modelling based on the characteristics of the patients could yield valuable results e.g. using medical questionnaire results and previous caries experience to create clusters of high-risk patients, facilitating focussed targeting for prevention and early detection of disease.

Policy makers should have a strong voice in the development of these technologies and should act to ensure that information systems are capable of answering the questions that will guide their decision making. A clear articulation of the types of information required by policy makers and their expectations from information systems would be a valuable further development.

References

- Abanto, J. et al., 2014. Effectiveness of a preventive program based on caries risk assessment and recall intervals on the incidence and regression of initial caries lesions in children. *International Journal of Paediatric Dentistry*, Volume 25, pp. 291-299.
- Agrawal, R., Gunopolous, D. & Leymann, F., 1998. *Mining Process Models from Workflow Logs in: Proceedings of the 6th International Conference on Extending Database Technology, Lecture Notes in Computer Science, Volume 1377 PP469-483*. s.l., Springer.
- American Dental Association, 2018. *Dental Informatics*. [Online] Available at: <https://www.ada.org/en/member-center/member-benefits/practice-resources/dental-informatics> [Accessed 1 Nov 2018].
- American Dental Association, 2018. *Glossary of Dental Clinical and Administrative Terms*. [Online] Available at: <https://www.ada.org/en/publications/cdt/glossary-of-dental-clinical-and-administrative-term> [Accessed 5 Jan 2018].
- American National Standard/American Dental Association, 2013. *The Electronic Dental Record System Standard Functional Requirements*. s.l.: is published by the American National Standard/American Dental Association Standard No 1067.
- Anker, J. et al., 2011. *Root Causes Underlying Challenges to Secondary Use of Data*. s.l., AMAI Symposium Proceedings. BDA, 2018. *Clinical Implications of Digital dentistry*. [Online] Available at: <https://bda.org/news-centre/blog/clinical-implications-of-digital-dentistry> [Accessed 15 October 2018].
- Beirne, P., Clarkson, J. & Worthington, H., 2007. Recall intervals for oral health in primary care patients (review). *Cochrane Database of Systematic Reviews*, Oct(4).
- Beirne, P., Forgie, A., Clarkson, J. & Worthington, H., 2005. Recall intervals for oral health in primary care patients. *Cochrane Database of Systematic Reviews*, Apr(18).
- Bhardwaj, A. et al., 2016. Measuring up. Implementing a dental quality measure in the electronic health record context. *JADA*, 147(1), pp. 35-40.
- BioMed Central Ltd, 2018. *INTERVAL Dental Recalls Trial*. [Online] Available at: <http://www.isrctn.com/ISRCTN95933794> [Accessed 1 Nov 2018].
- Bloemen, V. et al., 2018. *Maximizing Synchronization for Aligning Observed and Modelled Behaviour*. Sydney, 16th International Conference BPM.
- Bokhari, S., Basharat, I. & Ahmed, B., 2015. *A framework for clustering dental patients' records using unsupervised learning techniques*. London, Science and Information Conference.
- Bose, J., Mans, R. & van der Aalst, W., 2013. *Wanna Improve Process Mining results? It's high time we consider data quality issues seriously*. [Online] Available at: <http://bpmcenter.org/wp-content/uploads/reports/2013/BPM-13-02.pdf> [Accessed 20 December 2016].
- Botsis, T., Hartvigsen, G. C. F. & Weng, C., 2010. *Secondary Use of EHR: Data Quality Issues and Informatics Opportunities*. s.l., AMIA Joint Summits on Translational Science Proceedings..
- Bozkaya, M., Gabriels, J. & van der Werf, J., 2009. *Process Diagnostics: A Method based on Process Mining*. s.l., IEEE, International Conference on Information, Process, and Knowledge Management.
- Brathall, D., 2000. Introducing the Significant Caries Index together with a proposal for a new global oral health goal for 12-year-olds. *International Dental Journal*, Volume 50, pp. 378-384.

- Broadbent, J. & Thomson, W., 2005. For debate: problems with the DMF index pertinent to dental caries data analysis. *Community Dentistry and Oral Epidemiology*, 33(6), pp. 400-409.
- Brody, S., 2016. *A retrospective cohort study investigating the effects of dental recall visit intervals on the oral health of Irish school children and recommendations on prioritisation of a restricted service.*, Cork: Dissertation for Masters in Dental Public Health, UCC.
- Burattin, A. et al., 2018. *Online Conformance Checking Using Behavioural Patterns*. Sydney, 16th International Conference BPM.
- Campbell, H., Hotchkiss, R., Bradshaw, N. & Porteous, M., 1998. Integrated care pathways. *BMJ*, Volume 316, pp. 133-137.
- Central Statistics Office (Ireland), 2012. *Population Density, Town & Country*. [Online] Available at: https://www.cso.ie/en/media/csoie/census/documents/census2011vol1andprofile1/Profile1_Town_and_Country_Entire_doc.pdf [Accessed 24 October 2018].
- Chenail, R. J., 2011. Interviewing the Investigator: Strategies for Addressing Instrumentation and Researcher Bias Concerns in Qualitative Research. *The Qualitative Report*, 16(1), pp. 255-262.
- Chen, P., 1975. The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems (TODS) - Special issue: papers from the international conference on very large data bases*, 1(1), pp. 9-36.
- Choudhary, K. & Bajaj, P., 2015. Automated Prediction of RCT(Root Canal Treatment) using Data Mining Techniques: ICT in Health Care. *Procedia Computer Science*, Volume 46, pp. 682-688.
- Cook, J. & Wolf, A., 1998. Discovering models of software processes from event-based data. *ACM Transactions on Software Engineering and Methodology*, 7(3), pp. 215-249.
- Cosgun, E., Durukan Koese, S. & Koese, T., 2015. Clustering the Oral and Dental Health Centres in Turkey with Data Mining Methods according to the service they offer. *International Journal of Business and Social Science*, 6(8(1)), pp. 46-56.
- Daly, B., Batchelor, P., Treasure, E. & Watt, R., 2013. *Essential Dental Public Health*. 2nd ed. Oxford: Oxford University Press.
- DAMA UK Working Group, 2013. *The six primary dimensions for data quality assessment*. [Online] Available at: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf [Accessed 30 1 2018].
- Danciu, I. et al., 2014. Secondary use of clinical data: The Vanderbilt approach. *Journal of Biomedical Informatics*, pp. 28-35.
- Datta, A., 1998. Automating the discovery of as-is business process models - probabilistic and algorithmic approaches. *Information Systems Research*, Volume 9, pp. 175-301.
- Davenport, C. et al., 2003. The effectiveness of routine dental checks: a systematic review of the evidence base. *British Dental Journal*, 195(2), pp. 87-98.
- De Weerd, J., De Backer, M., Vanthienen, J. & Baesens, B., 2012. A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs. *Information Systems*, Volume 37, pp. 654-676.
- deLeoni, M., van der Aalst, W. & Dees, M., 2016. A general process mining framework for correlating predicting and clustering dynamic behaviour based on event logs. *Information Systems*, Volume 56, pp. 235-257.

Department of Health (Ireland), 1988. *Report of working group appointed to review the delivery of dental services (The Leyden Report)*, Dublin: Department of Health (Ireland).

Department of Health (Ireland), 1994. *Shaping a healthier future: a strategy for effective healthcare in the 1990's*, Dublin: Department of Health (Ireland).

Dhar, V., 2013. Data Science and Prediction. *Communications of the ACM*, 56(12).

Dragon1, 2018. *Data Mining Definition*. [Online] Available at: <https://www.dragon1.com/terms/data-mining-definition> [Accessed 4 Nov 2018].

Elliot, M., Mackey, E., O'Hara, K. & Tudor, C., 2016. *The Anonymisation Decision-Making Framework*. Manchester: UKAN Publications.

Erdogan, T. & Tarhan, A., 2016. *Process Mining for Healthcare Process Analytics*. s.l., (IEEE Explore) 2016 Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement.

Erdogan, T. & Tarhan, A., 2018. Systematic Mapping of Process Mining Studies in Healthcare. *IEEE Access*, Volume 6, pp. 24543-24567.

European Commission, 2014. *Research and Innovation Performance in the EU*, Luxembourg: Publications Office of the European Union.

Fernandez-LLatas, C., Martinez-Milanna, A., Martinez-Romero, A. B. J. & Traver, V., 2015. *Diabetes care related process modelling using Process Mining techniques*. s.l., Conf Proc IEEE Eng Med Biol Soc. 2015;2015:2127-30. doi: 10.1109/EMBC.2015.7318809.

Few, S., 2004. *Eenie, Meenie, Minie, Moe: Selecting the Right Graph for Your Message*. [Online] Available at: http://www.perceptualedge.com/articles/ie/the_right_graph.pdf [Accessed 4 Jan 2016].

Fluxicon, 2017. <https://fluxicon.com/blog/2010/10/prom-tips-mining-algorithm/>. [Online] Available at: <https://fluxicon.com/blog/2010/10/prom-tips-mining-algorithm/> [Accessed 9 5 2018].

Gansky, S., 2003. Dental Data Mining: Potential Pitfalls and Practical Issues. *Advances in Dental Research*, Volume 17, pp. 109-114.

Gehrke, N. & Werner, M., 2013. Process Mining. *WISU, die Zeitschrift fuer den Wirtschaftsstudenten*, Volume 7/13.

Goedertier, S., Martens, D., Vanthienen, J. & Baesens, B., 2009. Robust Process Discovery with Artificial Negative Events. *Journal of Machine Learning Research*, Volume 10, pp. 1305-1340.

Government of Ireland, 1953. *Health Act*. Dublin: Office of the Attorney General.

Government of Ireland, 2000. *Health(Dental Services for Children) Regulations, 2000 (S.I. No 248 of 2000)*. Dublin: Office of the Attorney General.

Harris, R. & Bridgman, C., 2010. Introducing care pathway commissioning to primary dental care: the concept.. *British Dental Journal*, Volume 209, pp. 233-239.

Harzing.com, 2018. <https://harzing.com/resources/publish-or-perish>. [Online] Available at: <https://harzing.com/resources/publish-or-perish> [Accessed 9 Oct 2018].

Health Research Authority, 2018. *GDPR Guidelines*. [Online] Available at: <https://www.hra.nhs.uk/planning-and-improving-research/policies-standards-legislation/data-protection-and-information-governance/gdpr-guidance/definitions/> [Accessed 14 6 2018].

Hebbal, M. & Nagarajappa, R., 2004. Does School-Based Dental Screening for Children Increase Follow-Up Treatment at Dental School Clinics. *Journal of Dental Education*, 69(3), pp. 382-386.

Hey, T., Tansley, S. & Tolle, K., 2009. *The Fourth Paradigm. Data Intensive Scientific Discovery*. [Online] Available at: https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf [Accessed 1 Jan 2018].

Hoffmann-Axthelm, W., 1981. *History of Dentistry*. Berlin: Quintessence Publishing.

Hripcsak, G. & Albers, D., 2013. Next-generation phenotyping of electronic health records.. *Journal of the American Medical Informatics Association*, 20(1), p. 117–121.

Hsin-Fang, L., 2013. *Data Mining and Pattern Discovery using exploratory and visualisation methods for large multidimensional datasets*. [Online] Available at: https://uknowledge.uky.edu/epb_etds/4 [Accessed 2 1 2018].

IEEE CIS Task Force on Process Mining, 2010. *Minutes of the meeting of the Task Force at BPM 2010*. [Online] Available at: http://www.win.tue.nl/ieeetfpm/doku.php?id=shared:minutes_of_the_meeting_of_the_task_force_at_bpm_2010 [Accessed 16 5 2018].

IEEE, 2011. Process Mining Manifesto. In: *Lecture Notes in Business Information Processing*. Berlin, Heidelberg: Springer.

Information Commissioner's Office, 2018. *What is personal data?*. [Online] Available at: <https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-personal-data/> [Accessed 14 6 2018].

Information Standards Board for Health and Social Care, 2013. *Anonymisation standard for publishing health and social care data specification*, s.l.: Crown Copyright.

International Team for Implantology, 2007. *Glossary of Oral and Maxillofacial Implants*. [Online] available at: <https://www.iti.org/GOMI>. [Accessed 10 Dec 2017].

Irish Dental Association, 2015. *10,000 children under 15 are being hospitalised every year for dental extractions under general anaesthetic*. [Online] Available at: <https://www.dentist.ie/latest-news/10000-children-under-15-are-being-hospitalised-every-year-for-dental-extractions-under-general-anaesthetic.6764.html> [Accessed 5 Jan 2017].

Irish Oral Health Services Guideline Initiative, 2010. *Pit and Fissure Sealants: Evidence-based guidance on the use of sealants for the prevention and management of pit and fissure caries*, Cork: [Online] Available at: <http://ohsrc.ucc.ie/html/guidelines.html>. [Accessed 5 May 2017]

Irish Oral Health Services Guideline Initiative, 2012. *Oral Health Assessment: Best practice guidance for providing an oral health assessment for school-aged children in Ireland*, Cork: [Online] Available at: <http://ohsrc.ucc.ie/html/guidelines.html>. [Accessed 5 May 2017]

Johnson, A. et al., 2016. Data Descriptor: MIMIC-III, a freely accessible critical care database. *Scientific Data*, pp. 1-9.

Johnson, O. et al., 2018. *The ClearPath Method for Care Pathway Process Mining and Simulation: International Workshop on Process-Oriented Data Science for Healthcare*, Sydney. Sydney, IEEE.

Kahneman, D., 2012. *Thinking, Fast and Slow*. s.l.:Penguin, Randon House, UK.

Kahn, M. et al., 2016. A harmonised data quality assessment terminology and framework for the secondary use of Electronic Health Record Data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 4(1).

- Kakilehto, T., Salo, S. & Larmas, M., 2009. Data mining of clinical oral health documents for analysis of the longevity of different restorative materials in Finland. *International Journal of Medical Informatics*, Volume 78, pp. 68-74.
- Kaymak, U., Mans, R., van de Steeg, T. & Dierks, M., 2012. *On Process Mining in Healthcare*. Seoul, 2012 IEEE International Conference on Systems, Man, and Cybernetics.
- Kennebeck, S., Timm, N., Farrell, M. & Spooner, S., 2012. Impact of electronic health record implementation on patient flow metrics in a pediatric emergency department. *J Am Med Inform Assoc*, 19(3), pp. 443-447.
- Knapp, R., Z, M. & Rodd, H., 2017. Treatment of dental caries under general anaesthetic in children. *British Dental Journal*, Issue 17116.
- Knowlton, J. et al., 2017. A Framework for Aligning Data from Multiple Institutions to Conduct Meaningful Analytics. *eGEMs (Generating Evidence & Methods to improve patient outcomes*, 5(5), pp. 1-7.
- Kurniati, A., Johnson, O., Hogg, D. & Hall, G., 2016. *Process Mining in Oncology: a Literature Review*. Hatfield, UK, Proceedings of the 6th International Conference on Information Communication and Management (ICICM 2016).
- Kurniati, A., Rojas, E., Hogg, D. & Johnson, O., 2018. The Assessment of Data Quality Issues for Process Mining in Healthcare Using Mimic-III, a Publicly Available e-Health Record Database. *Health Informatics Journal (Submitted for Publication)*.
- Ledley, R. & Lusted, K., 1959. Reasoning Foundations of Medical Diagnosis. *SCIENCE*, Volume 130, pp. 9-12.
- Lewsey, J. & Thomson, W., 2004. The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. *Community Dental Oral Epidemiology*, Volume 32, pp. 183-189.
- Lillrank, O. & Liukko, M., 2004. Standard, Routine and Non-Routine Processes in Health Care. *International Journal of Health Care Quality Assurance*, 17(1), pp. 39-46.
- Liu, C., Wang, F., Hu, J. & Xiong, H., 2015. *Temporal Phenotyping from Longitudinal Electronic Health Records: A Graph Based Framework*. Sydney, NSW,, KDD.
- Mans, R., 2011. *processmining.org*. [Online] Available at: <http://is.tm.tue.nl/staff/rmans/thesisRonnyMansFinal.pdf> [Accessed 8 6 2016].
- Mans, R., Reijers, H., Wismeijer, D. & van Genuchten, M., 2012. *Mining Processes in Dentistry*. Miami, Florida, IHI '12.
- Mans, R., Reijers, H., Wismeijer, D. & van Genuchten, M., 2013. A process-oriented methodology for evaluating the impact of IT: A proposal and an application in healthcare. *Information Systems*, 38(8), pp. 1097-1115.
- Mans, R., Reijers, H., Wismeijer, D. & van Genuchten, M., 2013. *bpmcentre.org*. [Online] Available at: <http://bpmcenter.org/wp-content/uploads/reports/2013/BPM-13-08.pdf> [Accessed 27 May 2017].
- Mans, R. et al., 2008. Process Mining Techniques: an Application to stroke care. *Studies in Health Technology and Informatics*, Volume 136, pp. 573-578.
- Mans, R. et al., 2008. Application of process mining in healthcare - a case study in a Dutch hospital. *Biomedical Engineering Systems and Technologies*, Volume 25, pp. 425-438.

- Mans, R. S., P., v. d. A. W. M. & Vanwersch, R. J., 2015. *Process Mining in Healthcare. Evaluating and exploiting operational healthcare processes*. s.l.:Springer-Verlag.
- Mans, R., van der Aalst, W. & Vanwersch, R., 2013. *Process Mining in Healthcare: Opportunities Beyond the Ordinary*. [Online] Available at: <http://bpmcenter.org/wp-content/uploads/reports/2013/BPM-13-26.pdf> [Accessed 10 5 2018].
- Mans, R., van der Aalst, W., Vanwersch, R. & Moleman, A., 2013. *Process Mining in Healthcare: Data Challenges when answering frequently posed questions*. s.l., In: Lenz R., Miksch S., Peleg M., Reichert M., Riaño D., ten Teije A. (eds) *Process Support and Knowledge Representation in Health Care*. Lecture Notes in Computer Science, vol 7738. Springer, Berlin, Heidelberg.
- Mathworks, 2018. *Machine Learning Types*. [Online] Available at: <https://uk.mathworks.com/help/stats/machine-learning-in-matlab.html> [Accessed 4 Nov 2018].
- McKinsey Global Institute, 2011. *Big Data: The next frontier for innovation, competition and productivity*. [Online] Available at: <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation> [Accessed 19 July 2017].
- Microsoft, 2012. *SQL Server Data Quality Dimensions*. [Online] Available at: <https://social.technet.microsoft.com/wiki/contents/articles/3919.data-quality-services-dqs-faq.aspx> [Accessed 30 Jan 2018].
- Milsom, K. et al., 2006. School dental screening does not increase dental attendance rates or reduce disease levels. *Journal of Dental Research*, pp. 924-928.
- Milsom, K., Tickel, M. & Blinkhorn, A., 2008. Is School Dental Screening a Political or a Scientific Intervention?. *Journal of Dental Research*, 87(10), pp. 896-899.
- Moen, R., 2010. *Foundation and history of the PDSA cycle*. [Online] Available at: https://s3.amazonaws.com/wedi/www/FileManager/PDSA_History_Ron_Moen.pdf
- Murphy, E., 2011. *Can BRIDGES Dental Informatics System play a role in the development of an Evidence-based HSE Public Dental Service (MDPH Dissertation)*, Cork: Department of Oral Health and Development, National University of Ireland.
- National Institute for Health and Care Excellence, 2004. *Dental checks: intervals between oral health reviews*. [Online] Available at: <https://www.nice.org.uk/guidance/cg19/resources/dental-checks-intervals-between-oral-health-reviews-pdf-975274023877> [Accessed 25 June 2018].
- National Institute for Health and Care Excellence, 2018. <https://pathways.nice.org.uk/pathways/oral-and-dental-health>. [Online] Available at: <https://pathways.nice.org.uk/pathways/oral-and-dental-health> [Accessed 17 July 2018].
- National Institute for Healthcare Excellence, 2017. *Oral and dental health Overview*. [Online] Available at: <https://pathways.nice.org.uk/pathways/oral-and-dental-health> [Accessed 20 June 2017].
- NHS Digital, 2017. *Anonymisation standard for publishing health and social care data*. [Online] Available at: <https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb1523-anonymisation-standard-for-publishing-health-and-social-care> [Accessed 14 12 2017].
- NHS England, 2009. *NHS dental services in England: An independent review led by Professor Jimmy Steele*, s.l.: NHS.
- NHS, 2012. *NHS dental Contract Pilots – Care Pathway Review*, London: Department of Health.
- Nomura, Y. et al., 2004. A survey on the risk factors for the prevalence of dental caries among preschool children in Japan. *Pediatric Dental Journal*, 14(1), pp. 79-85.

November, J., 2011. Early Biomedical Computing and the roots of Evidence-Based Medicine. *IEEE Annals of the History of Computing*, Volume April-June, pp. 9-17.

O'Mullane, D. et al., 2016. Fluoride and Oral Health. *Community Dental Health*, Volume 33, pp. 69-99.

O'Neil, C. & Schutt, R., 2014. *Doing Data Science*. 3 ed. Sepastapol: O'Reilly.

Pannucci, C. & Wilkins, E., 2010. Identifying and Avoiding Bias in Research. *Plast Reconstr Surg*, 126(2), pp. 619-625.

Pedrinaci, C. & Dominique, J., 2007. Towards an Ontology for Process Monitoring and Mining. *Semantic Business Process and Product Lifecycle Management*, Volume 251, pp. 76-87.

Petersen, P., 2008. World Health Organisation global policy for improvement of oral health - World Health Assembly 2007. *International Dental Journal*, Volume 58, pp. 115-121.

Phantumvanit, P. et al., 2018. WHO Global Consultation of Public Health Intervention against Early Childhood Caries. *Community Dentistry and Oral Epidemiology*, Volume 46, pp. 280-287.

Raedel, M., Hartmann, A., Bohn, S. & Walter, M., 2015. Three-year outcomes of apicectomy: Mining an insurance database. *Journal of Dentistry*, Volume 1218-1222, p. 43.

Rebuge, A. & Ferreira, D., 2012. Business Process Analysis in Healthcare Environments: A methodology based on process mining. *Information Systems*, 37(2), pp. 99-116.

Rehse, J. & Fettke, P., 2018. *Process Mining Crimes – A Threat to the Validity of Process Discovery Evaluations*. Sydney, 16th International Conference BPM.

Riley, P., Worthington, H., Clarkson, J. & Beirne, P., 2013. Recall Intervals for oral health in primary care patients. *Cochrane Database Systematic Review*, Dec(12).

Ring, M., 1985. *An Illustrated History of Dentistry*. s.l.:Abradale Press Harry N Abrams Inc.

Rojas, E., Arias, M. & Sepulveda, M., 2015. *Clinical Processes and Its Data, What can we do with them?*. Lisbon, Proceedings of the international conference on Health Informatics.

Rojas, E., Munoz-Gama, J. S. M. & Capurro, D., 2016. Process Mining in Healthcare: A Literature Review. *Journal of Medical Bioinformatics*, Volume 61, pp. 224-236.

Rojas, E. et al., 2017. Question-Driven Methodology for Analyzing Emergency Room Processes Using Process Mining. *Applied Sciences*, Volume 7.

Rovani, m., Maggi, F., de Leoni, M. & van der Aalst, w., 2015. Declarative process mining in healthcare. *Expert Systems with Applications*, Volume 42, pp. 9236-9251.

Rozinat, A. & van der Aalst, W., 2008. Conformance checking of Processes based on monitoring real behaviour. *Information Systems* 33, pp. 64-95.

RTE, 2015. *Varadkar queries dentist group's extractions claim*. [Online] Available at: <https://www.rte.ie/news/2015/1015/734926-teeth-children/> [Accessed 5 Jan 2017].

Santos, R., Malheiros, S., Cavalheiro, S. & de Oliveira, J., 2013. A data mining system for providing analytical information on brain tumors to public health decision makers.. *Computer Methods and Programs in Biomedicine*, Volume 109, pp. 269-282.

Saunders, M., Lewis, P. & Thornhill, A., 2012. *Research Methods for Business Students*. 6 ed. s.l.:Pearson Education MUA.

- Schimm, G., 2003. *Mining Most Specific Workflow Models from Event-Based Data from : BPM International Conference*. Eindhoven, Springer.
- Schleyer, T., 2003. Dental Informatics - An emerging Biomedical informatics area. *Advanced Dental Research*, 17(December), pp. 4-8.
- Schmier, J., Kane, D. & Halpern, M., 2005. Practical applications of usability theory to electronic data collection for clinical trials. *Compemporary Clinical Trials*, Volume 6, pp. 376-385.
- Schrijvers, G., van Hoorn, A. & Huiskes, N., 2012. The care pathway: concepts and theories: an introduction. *International Journal of Integrated Care*, Volume 12.
- Selwitz, R., Ismail, A. & Pitts, N., 2007. Dental Caries. *The Lancet*, 369(9555), pp. 51-59.
- Song, M., Kaihong, L., Abromitis, R. & Schleyer, T., 2013. Reusing Electronic Patient Data for Dental Clinical Research: A Review of Current Status. *Journal of Dentistry*, 41(12), pp. 1148-1163.
- Taleb, N., 2010. *The Black Swan*. 2 ed. s.l.:Penguin.
- Tamaki, Y. et al., 2009. Construction of a dental caries prediction model by data mining. *Journal of Oral Science*, 51(1), pp. 61-68.
- Tax, N., Sidorova, N., Haakma, R. & van der Aalst, W., 2016. *Event Abstraction for Process Mining Using Supervised Learning Techniques*. London, SAI Intelligent Systems Conference 2016.
- The Parliamentary Office of Science and Technology, 2017. *Big Data and Public Health (POSTNOTE 474)*. [Online] Available at: <http://researchbriefings.parliament.uk/ResearchBriefing/Summary/POST-PN-474> [Accessed 10 May 2017].
- Thiede, M. & Fuerstenau, D., 2016. *The technological maturity of Process Mining* [Online] Available at https://www.researchgate.net/publication/293593416_The_Technological_Maturity_of_Process_Mining_An_Exploration_of_the_Status_Quo_in_Top_IS_Journals [Accessed 31 May 2017].
- UCC/HRB, 2005/6. *Results of school dental screening situation analysis*, Personal Contact with Oral Health Research Service Centre, UCC, Cork
- University of Derby, 2018. *Research Onion Diagram*. [Online] Available at: <https://onion.derby.ac.uk/> [Accessed 31 July 2018].
- Van der Aalst, W., 1998. The application of Petri nets to workflow management. *J. circuits, Syst. Comput*, 8(01), pp. 21-66.
- van der Aalst, W., 2011. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Berlin: Springer-Verlag.
- van der Aalst, W., 2013. Extracting Event Data from Databases to unleash process mining.
- Van der Aalst, W., 2015. <https://www.coursera.org/lecture/process-mining/4-3-introduction-to-conformance-checking-pi2L2>. [Online] Available at: <https://www.coursera.org/lecture/process-mining/4-3-introduction-to-conformance-checking-pi2L2> [Accessed 18 Sept 2018].
- van der Aalst, W., 2016. *Process Mining: Data Science in Action*. 2 ed. Heidelberg: Springer.
- van Eck, M., Lu, X., Lemans, S. & van der Aalst, W., 2015. *PM2: a Process Mining Project Methodology; Advanced Information Systems Engineering : 27th International Conference, CAiSE 2015, Stockholm, Sweden, June 8-12, 2015. Proceedings / Ed. J. Zdravkovic, M. Kirikova, P. Johannesson. - Berlin : Springer*.

- van Genuchten, M., Mans, R., Reijers, H. & Wismeijer, D., 2014. Is Your Upgrade Worth it? Process Mining can tell. *IEEE Software*, 31(5).
- vanden Broucke, S., De Weert, J., Vanthienen, J. & Baesens, B., 2013. *A Comprehensive Benchmarking Framework (CoBeFra) for Conformance Analysis between Procedural Process Models and Event Logs in ProM*. s.l., IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 04/2013.
- Verbeek, H. & van der Aalst, W., 2015. *Decomposed Process Mining: The ILP Case In: Fournier F., Mendling J. (eds) Business Process Management Workshops. BPM 2014. Lecture Notes in Business Information Processing, vol 202..* Cham, Springer.
- Wang, J., Wong, R., Ding, J. G. Q. & Wen, L., 2013. Efficient Selection of Process Mining Algorithms. *IEEE TRANSACTIONS ON SERVICES COMPUTING*, 6(4), pp. 484-496.
- Ward, M. et al., 2014. The Effect of Electronic Health Record Implementation on Emergency Department Operational Measures of Performance. *Annals of Emergency Medicine*, 63(6), p. 723–730.
- Weber, P., Bordbar, B. & Tiño, P., 2012. A Framework for the Analysis of Process Mining Algorithms. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS*, 43(2), pp. 303-317.
- Weijters, A., Aalst, W. M. P. & Medeiros, A., 2006. *Process Mining with the Heuristics Miner-algorithm.*, s.l.: TU Eindhoven.
- Weiskopf, N. & Weng, C., 2013. Methods and Dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, Volume 20, pp. 144-151.
- Whelton, H. et al., 2006. *North South survey of children's oral health in Ireland 2002*. [Online] Available at: <http://www.lenus.ie/hse/bitstream/10147/119028/1/OralHealthReport.pdf> [Accessed 01 Feb 2018].
- Whelton, H. et al., 2017. *Effectiveness of Water Fluoridation at 0.7ppm*. [Online] Available at: <https://iadr.abstractarchives.com/abstract/17iags-2638086/effectiveness-of-water-fluoridation-at-07ppm> [Accessed 6 Jan 2019].
- Williams, R., Rojas, E., Peek, N. & Johnson, O., 2018. Process Mining in Primary Care: A Literature Review. *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*, pp. 376-380.
- Wilson, G. et al., 2014. Best Practices for Scientific Computing. *PLOS Biology*, 12(1), pp. 1-6.
- Wilson, M. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, pp. 1-9.
- World Dental Federation (FDI), 2016. *FDI's definition of oral health*. [Online] Available at: <https://www.fdiworlddental.org/oral-health/fdi-definition-of-oral-health> [Accessed 5 Jan 108].
- World Dental Federation (FDI), 2017. *Mouth Smart. Your guide to oral health policies*. [Online] Available at: http://www.worldoralhealthday.org/sites/default/files/assets/2017_WOHD-toolkit-advocacy.pdf [Accessed 5 Jan 2018].
- World Health Organisation, 2003. *Oral Health Promotion: An Essential Element of a Health-Promoting School. WHO INFORMATION SERIES ON SCHOOL HEALTH DOCUMENT ELEVEN*. [Online] Available at: http://apps.who.int/iris/bitstream/handle/10665/70207/WHO_NMH_NPH_ORH_School_03.3_eng.p

[df/iris/handle/10665/70207;jsessionid=F4261621AA6510BF70ED9F81B3E2F129?sequence=1](https://iris.handle/10665/70207;jsessionid=F4261621AA6510BF70ED9F81B3E2F129?sequence=1)[Accessed 25 June 2018].

Yang, W. & Su, Q., 2014. *Process Mining for Clinical Pathway- Literature Review and future directions*. Beijing, International Conference on Service Systems and Service Management (ICSSSM).

Zhou, W. & Piramuthu, S., 2010. Framework, strategy and evaluation of health care processes with RFID. *Decision Support Systems*, Volume 50, pp. 222-233.

10 Appendices

10.1 Data Management Plan

Research Data Leeds <http://researchdata.leeds.ac.uk/> email:
researchdataenquiries@leeds.ac.uk

Adapted from University of Cambridge (<http://www.lib.cam.ac.uk/preservation/datatrain/documents.html>) &
University of Edinburgh MANTRA
http://datalib.edina.ac.uk/mantra/Data_management_plan_template_MANTRA.docx

Basic Data Management Plan Template

(download at <http://bit.ly/2htlnrO>)

Project title and brief description:

Applying the emerging technology of process mining to dentistry

1. What data will be produced?

An extract of a dental EHR (Bridges) will be created, including demographics, treatment items, appointments. Code to create the extract will be written. The extract is from a SQL Server database and is in CSV format. It will be imported into a research SQL Server database and is known as Bridges-PM1. The entity relationship diagram is attached. Code to anonymise the extract will be written. A general profile of the data will be created. The treatment data will be mapped to SNODENT and a cross reference will be maintained. The extract will be mapped to the Process Mining Healthcare Reference Model and a cross reference will be maintained. Experiments will be performed on the extract and visual representations of treatment processes and care pathways will be created and maintained. Statistical analyses of the experiments will be executed, and the results will be preserved

2. How will data be documented and described?

Metadata regarding the original EHR and the extract will be created. A description of the data source, provenance and collection method will be included in the research write-up. The results of the experiments and analyses will be written up and maintained. The code to extract and anonymise the data will be commented and described. The CSV files (20 in total) contain header information naming the column. The data extract will be stored in an SQL Server database for the duration of this research project. Github will be used to store the Experiment Documentation and any software written for this research. It is unclear whether this data will ultimately be re-usable, however, we are proceeding on the assumption that it will be both useful and reusable.

3. How will data be structured and stored?

The raw data is ~ 20GB. Including subsets for experiments, this will increase to ~500Gb. In the University of Leeds, The data will be stored in the Leeds Institute for Data Analytics (LIDA) and subject to normal LIDA backup procedures. It is intended to use SQL Server as the data storage tool. All tables and columns will have human readable names prefixed with 'PM' and dictionary type tables will be prefixed with an additional 'D'. Directory structures will be designed as appropriate when requirements crystallise. Github will be used for version control. There is no specific retention schedule at this point.

4. Are there any 'special' requirements for your data?

The use of the data is subject to a data use agreement with the Health Service Executive(HSE) of Ireland. Access to the data is currently restricted to the research team though it is hoped that on

conclusion of the research the data will be made more generally available. The data is anonymised, individual level data and accordingly is to be treated with care. It is currently encrypted and password protected. The data controller (HSE) has agreed to provide the data for this research with provisos and requirements detailed in governing documents and other communications which we are treating as amounting to a Data-Use Agreement.

5. What are the plans for data sharing and access?

Access to the data is currently restricted to the research team though it is hoped that on conclusion of the research the data will be made more generally available. Consent was waived as the data has been appropriately de-identified and anonymised. It is currently unknown as to whether the data will be stored in a repository following completion of the research.

6. What are your main data challenges? Who can help?

My main data challenges currently (March 1st 2017) are: Creation of a stable, virtual research environment within LIDA. Enabling secure, remote access to that environment. Selection of appropriate datasets to execute research experiments. Managing select of datasets and results in compliance with the research governing documents.

7. Who is responsible for managing the data? What resources will you need?

Currently, me, the lead researcher. I currently need assistance for the creation of a stable, virtual research environment within LIDA with remote access for me enabled.

Basic Data Management Plan Template: Prompt Sheet

1. What data will be produced?

- What physical data will you study? (e.g. artefacts, samples, paper archives, etc.)
- What digital data will you generate? (e.g. field-notes, images, spreadsheets, audio interviews, survey data, annotated bibliography, etc.)
- What file formats and software will you use?

2. How will data be documented and described?

- Will others understand your data? Write documentation. Make sure table and spreadsheet values are clearly labelled.
- What information about data collection methodology will be recorded?
- Is it important for the research to be reproducible? Why/why not? If so, what additional documentation or pointers will be required?
- Will you write software? Where will this be documented and stored for future use?

3. How will data be structured and stored?

- Estimate how much data you will produce over time – do you have enough storage?
- Are you making full use of University provided, fully backed-up storage? How will data generated in the field be saved to safe University storage?
- Do you have a logical file naming convention and directory structure?
- How will you use versioning so you can identify the current version of documents / data?

4. Are there any 'special' requirements for your data?

- Is your data sensitive? Is it stored and encrypted appropriately? (*For a definition of 'sensitive personal data' please see: <https://goo.gl/4xRFQu>. Guidance on the classification of data can be found in the University Information Protection Policy - <https://goo.gl/c7gXOC>).*
- Will you anonymise your data?

- Does your research funder have specific data management and sharing requirements?
- Should some data be destroyed? When and how?

5. What are the plans for data sharing and access in the short and long term?

- Have you discussed data sharing with your research collaborators/ supervisor?
- If your research involves people, have you obtained appropriate consent for data sharing?
- Can your data be released immediately, or should you embargo (delay access to) the data?
- What data will you keep? Who decides?
- Will data be openly available to everyone or will there be access restrictions?
- How long will / should data be available for?
- Will you use a data repository? Which one?

6. What are your main data challenges? Who can help?

- Do you need training or support? What is available?
- What University policies are relevant to your project? Have you read and understood them?

7. Who is responsible for managing the data? What resources will you need?

- Who is responsible for data at different stages in its lifecycle?
- Are sufficient resources (skills, people, storage, technology) available to deliver your plan?

10.2 Data Mappings For Standardisation and SNOMED

Variation mapping: Here, the ProcedureName was mapped onto a ProcedureNameGroup. An example of this is the mapping of “Amalgam Filling–1 Surface”, “Amalgam Filling–2 Surface” and other variations to “Amalgam Filling”.

SNOMED mapping: The ProcedureNameGroup was mapped to the corresponding SNOMED concept name. A matching SNOMED concept was available for almost all the ProcedureNameGroup entries.

Prevention or Restoration mapping: This is an additional higher level of abstraction and maps each ProcedureNameGroup to be either ‘Restorative’ or ‘Prevention’, if feasible. To maintain the structure of a course of treatment, some procedures were not mapped. Specifically, procedures that typically start and conclude a course of treatment were left as-is. This proved useful in reducing the complexity of the process models.

BridgesProcedureName	Mapped (for simplification)	SNOMEDConceptName	Prevention Or Restoration
Amalgam Filling - 2 Surface	Amalgam Filling	Insertion of amalgam restoration into tooth (procedure)	Restorative
Amalgam Filling - 4 Surface	Amalgam Filling	Insertion of amalgam restoration into tooth (procedure)	Restorative
Amalgam Filling - 5 Surface	Amalgam Filling	Insertion of amalgam restoration into tooth (procedure)	Restorative
Amalgam Filling - 1 Surface	Amalgam Filling	Insertion of amalgam restoration into tooth (procedure)	Restorative
Amalgam Filling - 3 Surface	Amalgam Filling	Insertion of amalgam restoration into tooth (procedure)	Restorative
Cancelled	Cancelled Appointment	Appointment canceled by patient	DNA

Cancelled Appointment	Cancelled Appointment	Appointment canceled by patient	DNA
casual	Casual Attendance	None Found	Restorative
Casual Attendance	Casual Attendance	None Found	Restorative
Completed Case	Completed Case	Previously initiated dental therapy completed	Completed Case
Compomer Filling - 1 Surface	Compomer Filling	Restoration- resin (procedure)	Restorative
Compomer Filling - 3 Surfaces	Compomer Filling	Restoration- resin (procedure)	Restorative
Compomer Filling - 2 Surfaces	Compomer Filling	Restoration- resin (procedure)	Restorative
Composite Filling - 2 Surfaces	Composite Filling	Restoration- resin (procedure)	Restorative
Composite Filling - 5 Surfaces	Composite Filling	Restoration- resin (procedure)	Restorative
Composite Filling - 3 Surfaces	Composite Filling	Restoration- resin (procedure)	Restorative
Composite Filling - 4 Surfaces	Composite Filling	Restoration- resin (procedure)	Restorative
Composite Filling - 1 Surface	Composite Filling	Restoration- resin (procedure)	Restorative
consent	Consent	Restoration- resin (procedure)	NULL
Crown Completed	Crown Completed	Fitting of dental crown to tooth (procedure)	Restorative
Crown Fracture	Crown Fracture	Tooth crown fracture (disorder)	Restorative
Crown Preparatory Work	Crown Preparatory Work	Crown preparation of tooth (procedure)	Restorative
Deciduous Pulp Treat Prep Work	Deciduous Pulp Treat Prep Work	Endodontic procedure (procedure)	Restorative
Deciduous Pulp Treatment	Deciduous Pulp Treatment	Endodontic procedure (procedure)	Restorative
Denture Bite	Denture Bite	Adjust denture (procedure)	Restorative
Denture Ease	Denture Ease	Adjust denture (procedure)	Restorative
Denture Impression	Denture Impression	Take impression for denture (procedure)	Restorative
Denture Preparatory Work	Denture Preparatory Work	Take impression for denture (procedure)	Restorative
Denture Reline	Denture Reline	Reline denture (procedure)	Restorative
Denture Repair	Denture Repair	Repair to denture (procedure)	Restorative
Denture Try-in	Denture Try-in	Try-in of denture (procedure)	Restorative
Diet Analysis	Diet Analysis	Nutritional counseling for control of dental disease (procedure)	Prevention
Diseased Extraction	Diseased Extraction	Tooth extraction (procedure)	Restorative
dna 6th class insp	DNA	Did not attend (finding)	DNA
dna fill appt	DNA	Did not attend (finding)	DNA
Did Not Attend	DNA	Did not attend (finding)	DNA
DNA	DNA	Did not attend (finding)	DNA
DNA HYG APPT	DNA	Did not attend (finding)	DNA
DNA HYGIENE APPT	DNA	Did not attend (finding)	DNA

Failed Appointment	DNA	Did not attend (finding)	DNA
Emergency appointment	Emergency appointment	None Found	Restorative
Exposing crown of tooth	Exposing crown of tooth	None Found	Restorative
Familiarisation Visit	Familiarisation Visit	None Found	Prevention
Filling - 3 Surfaces (no material)	Filling (No Material)	Restoration of part of tooth using filling (procedure)	Restorative
Filling - 1 Surface (no material)	Filling (No Material)	Restoration of part of tooth using filling (procedure)	Restorative
Filling - 2 Surfaces (no material)	Filling (No Material)	Restoration of part of tooth using filling (procedure)	Restorative
Fissure Seal - 2 Surfaces	Fissure Seal	Fissure seal tooth (procedure)	Prevention
Fissure Seal	Fissure Seal	Fissure seal tooth (procedure)	Prevention
Fissure Seal - 1 Surface	Fissure Seal	Fissure seal tooth (procedure)	Prevention
Full Upper or Full Lower Fit	Full Upper or Full Lower Fit	Complete upper denture (procedure)	Restorative
Full/Full Fit	Full/Full Fit	Fit complete upper and lower dentures (procedure)	Restorative
GA Extraction	GA Extraction	General anesthesia (procedure)	Restorative
Glass Ionomer Filling - 3 Surfaces	Glass Ionomer Filling	Insertion of glass-ionomer restoration into tooth (procedure)	Restorative
Glass Ionomer Filling - 1 Surface	Glass Ionomer Filling	Insertion of glass-ionomer restoration into tooth (procedure)	Restorative
Glass Ionomer Filling - 2 Surfaces	Glass Ionomer Filling	Insertion of glass-ionomer restoration into tooth (procedure)	Restorative
Glass Ionomer Filling - 4 Surfaces	Glass Ionomer Filling	Insertion of glass-ionomer restoration into tooth (procedure)	Restorative
Impression for Appliance	Impression for Appliance	Take oral or dental impression (procedure)	Restorative
Initial Exam	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 2nd. class 07/08.	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 2nd. class 08/09.	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 6th. class 07/08.	Initial Exam	Initial oral examination (procedure)	Initial Exam
6th class insp	Initial Exam	Initial oral examination (procedure)	Initial Exam
DNA 2ND CLASS INSP	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 2nd. class 06/07.	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 2nd. class 10/11.	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 2nd. class 11/12.	Initial Exam	Initial oral examination (procedure)	Initial Exam

Insp. 2nd. class 13/14.	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 4th. class 10/11.	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 6th. class 08/09.	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 6th. class 10/11.	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 6th. class 12/13.	Initial Exam	Initial oral examination (procedure)	Initial Exam
2nd class	Initial Exam	Initial oral examination (procedure)	Initial Exam
2nd class insp	Initial Exam	Initial oral examination (procedure)	Initial Exam
6TH CLASS 2013/14	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 4th. class 08/09.	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 6th. class 13/14.	Initial Exam	Initial oral examination (procedure)	Initial Exam
2nd class 2013/14	Initial Exam	Initial oral examination (procedure)	Initial Exam
6th class	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 3rd. class 13/14.	Initial Exam	Initial oral examination (procedure)	Initial Exam
Insp. 6th. class 11/12.	Initial Exam	Initial oral examination (procedure)	Initial Exam
L.A. Extraction	L.A. Extraction	Tooth extraction (procedure)	Restorative
Lance abscess	Lance abscess	Drainage of oral abscess (procedure)	Restorative
Luxation	Luxation	Dislocation - complete (morphologic abnormality)	Restorative
mum rang pain	Mum rang Pain	None Found	Restorative
No Treatment Required	No Treatment Required	treatment required for (contextual qualifier) (qualifier value)	No Treatment Required
NWG ortho	NWG ortho	None Found	Ortho
OHI	OHI	Oral health education (procedure)	Prevention
Ortho Appliance Fitted	Ortho Appliance Fitted	Insertion of complete orthodontic appliance (procedure)	Ortho
Ortho treatment completed	Ortho treatment completed	None Found	Ortho
Ortho Treatment On-Going	Ortho Treatment On-Going	None Found	Ortho
Ortho XGA	Ortho XGA	None Found	Restorative
Ortho XLA	Ortho XLA	None Found	Ortho
Orthodontic adjustment	Orthodontic adjustment	Adjust orthodontic appliance (procedure)	Ortho
Orthodontic Check	Orthodontic Check	None Found	Ortho
Orthodontic Extraction	Orthodontic Extraction	None Found	Ortho
Partial Upper or Lower Fit	Partial Upper or Lower Fit	Fit partial denture (procedure)	Restorative
POIG	POIG	None Found	NULL

Polished FGS	Polished FGS	None Found	Restorative
Post-XL Suture	Post-XL Suture	None Found	Restorative
Prescription	Prescription	Prescription (procedure)	Restorative
Prescription - Antibiotic	Prescription	Prescription (procedure)	Restorative
Prescription - Other	Prescription	Prescription (procedure)	Restorative
Preventative Restoration - 2 Surfaces	Preventative Restoration	Insertion of preventive resin tooth restoration (procedure)	Prevention
Preventative Restoration - 1 Surface	Preventative Restoration	Insertion of preventive resin tooth restoration (procedure)	Prevention
Re-Appointment	Re-Appointment		Re-Appointment
Recall appointment	Recall appointment	Recall arranged (finding)	Prevention
Recall	Recall appointment	Recall arranged (finding)	Prevention
Recement Crown	Recement Crown	Recement crown (procedure)	Restorative
Refer for general anaesthetic	Refer for general anaesthetic	None Found	Restorative
Refer for Oral Surgery	Refer for Oral Surgery	None Found	Restorative
Refer to hygienist	Refer to hygienist	None Found	Prevention
Referral for OPG	Referral for OPG	None Found	Ortho
Referral for Paediatric Secondary Care	Referral for Paediatric Secondary Care	None Found	Restorative
Referral to Ortho. Unit	Referral to Ortho. Unit	None Found	Ortho
Relative Analgesia	Relative Analgesia	None Found	Restorative
Review Appointment	Review	None Found	Prevention
Review	Review	None Found	Prevention
Review appointment	Review	None Found	Prevention
Root Treatment Preparatory Work	Root Treatment Preparatory Work	None Found	Restorative
Root Treatment Work Completed	Root Treatment Work Completed	None Found	Restorative
Scale & Polish	Scale & Polish	Scale and polish teeth (procedure)	Prevention
prophy	Scale & Polish	Scale and polish teeth (procedure)	Prevention
Special Tray Impression	Special Tray Impression	Take impression for dental or oral tray (procedure)	Restorative
Splint	Splint	Fit bite raising appliance (procedure)	Restorative
Stainless Steel Crown Completed	Stainless Steel Crown Completed	Prefabricated stainless steel crown- primary tooth (procedure)	Restorative
Subluxation	Subluxation	Subluxation of tooth (disorder)	Restorative
Surgical Extraction	Surgical Extraction	Surgical extraction (procedure)	Restorative
Temporary Crown	Temporary Crown	Construct temporary dental crown (procedure)	Restorative
Tip Replacement	Tip Replacement	Insertion of composite tip tooth restoration (procedure)	Restorative

Tip Restoration	Tip Restoration	Insertion of composite tip tooth restoration (procedure)	Restorative
Tooth Dressing	Tooth Dressing	Dress tooth (procedure)	Restorative
Topical Fluoride Application	Topical Fluoride Application	Topical application of fluoride - tooth (procedure)	Prevention
Urgent Ortho Referral	Urgent Ortho Referral	None Found	Ortho
Bite Wing	X-Ray	Radiography of teeth (procedure)	X-Ray
Bitewing - Single	X-Ray	Radiography of teeth (procedure)	X-Ray
OPG	X-Ray	Radiography of teeth (procedure)	X-Ray
Anterior Occlusal X-ray	X-Ray	Radiography of teeth (procedure)	X-Ray
Bitewing - Pairs	X-Ray	Radiography of teeth (procedure)	X-Ray
Intra-Oral X-Ray	X-Ray	Radiography of teeth (procedure)	X-Ray
Periapical	X-Ray	Radiography of teeth (procedure)	X-Ray
X-ray (no type specified)	X-Ray	Radiography of teeth (procedure)	X-Ray

10.3 BridgesPM1 Data attributes

Class of Data		Description	Number of Rows	Size
PMClients		Client Demographics	231,760	37 Mb
Validated	Attributes		Used or Not	
Y (5)	DOB	Date of birth	Yes	
N	ReferringDentistPartyID	Dentist referring for care if applicable	No	
N	ReferringGPPartyID	GP referring into dental service if applicable	No	
N	ReferredToSpecialistPartyID	Specialist to whom client has been referred	No	
Y (6)	ClinicID	Clinic Attended	Yes	
N	NationalityID	Nationality	No	
Y	PMClientID	Unique Identifier of Client	Yes	
PMTreatments		Treatment Event Description	3,169,864	1.44 Gb
Validated/Code	Attributes		Used or Not	
Y (10)	TreatmentCourseID	Course to which this item belongs	No	
N	ToothID	Tooth on which treatment was executed	No	
N	ToothTypeID	Tooth Type as above	No	
N	ToothPartID	Tooth Part as above	No	
N	ToothPartTypeID	Tooth Part Type as above	No	
N	MaterialName	Material used in the treatment	No	
N	ProcedureName	Name of Procedure	Yes	

Y (11)	CompletionDate	Date Procedure marked as completed	Yes	
N	CarriedOutByPartyID	Who carried out the treatment	No	
Y (12)	ClinicID	Where was the treatment carried out	Yes	
N	ConditionID	Condition name associated with treatment	No	
N	DateAdded	Date the treatment entry was created	No	
Y (13)	ListPosition	Order within the treatment plan	Yes	
N (14)	Quantity	Quantity of procedure carried out	No	
Y (1)	ClientID	Unique Identifier of Client	Yes	
N	PMTreatmentID	Unique Identifier of Treatment	Yes	
Y (15)	ClientAge	Calculated from DOB	Yes	
Y (16)	AssociatedChartID	Best guess based on dates	Yes	
N	DMFTAdult	DMFT Adult	Yes	
N	DMFSAdult	DMFS Adult	Yes	
N	DMFTChild	dmft child	Yes	
N	DMFSChild	dmfs child	Yes	
N	ChartCreationDate	Date Chart was created	Yes	
N	MonthsToDMF	Calculated – months between Initial Exam Date and associated chart DMF value	Yes	
N	NoOfInitialExams	How many screenings for this patient	Yes	
N	FirstExam	Is this the first screening?	Yes	
Y (9)	MappedToProcedureNameGroup	Mapped Name	Yes	
PMTreatmentCourses		Treatment Course Identifiers	285,518	27 Mb
	Attributes	Table not used	Used or Not	
N	PMTreatmentID	Unique Identifier of Treatment		
N	CreationDate	Date TreatmentCourse was created		
PMCharts		Chart Identifier and DMF measure	1,016,197	145 Mb
	Attributes	Table Used	Used or Not	
Y (19)	CreationDate	Date Charting was completed	Yes	
Y (20)	DMFTChild	dmft	No	
Y (21)	DMFTAdult	DMFT	Yes	
Y (22)	DMFSChild	Dmfs	No	
Y (23)	DMFSAdult	DMFS	No	
Y (1)	ClientID	Unique Identifier of Client	Yes	
Y	PMChartID	Unique Identifier of Chart	Yes	
PMTooth		Tooth Description	32,219,452	3.7 Gb
	Attributes	Table not used	Used or Not	
N	ChartID	Linking tooth to a PMChart Entry		
N	ToothTypeID	Link to ToothType		
N	PMToothID	Unique Identifier of Tooth		

PMToothPart		Tooth part Description	16,649,791	4.2 Gb
	Attributes	Table not used	Used or Not	
N	ToothID	Unique Identifier of Tooth		
N	ChartID	Unique Identifier of Chart		
N	ToothPartTypeID	Link to Tooth Part		
N	PMToothPartID	Unique Identifier of ToothPart		
PMCondition		Tooth Condition Description	32,291,681	8 Gb
	Attributes	Table not used	Used or Not	
N	ConditionTypeID	What Type of Condition		
N	MaterialTypeID	What material is used		
N	ToothID	Linking to Tooth		
N	ToothPartID	Linking to ToothPart		
N	ChartID	Linking Condition to a PMChart Entry		
N	DateNoted	Date Condition was charted		
N	PMConditionID	Unique Identifier of Condition		
PMAppointments		Appointment time and duration	1,760,923	376 Mb
	Attributes	Table not used	Used or Not	
Y (1)	ClientID	Link to Client		
N	StartTime	Start time of appointment		
N	Duration	Duration of appointment		
N	StatusID	Link to Appointment Status		
N	TypeID	Appointment Type		
N	PMAppointmentID	Unique Identifier of Appointment		
PMAttendances		Attendance history	5,516,738	1.2 Gb
	Attributes	Table not used	Used or Not	
Y (1)	ClientID	Link to Client		
N	StatusID	Link to Appointment Status		
N	StartTime	Start time of appointment		
Y (4)	ApptID	Link to Appointment		
N	PMAttendancesID	Unique Identifier of Attendance		
PMQuestionnaire		Medical Questionnaire Identifier	332,600	43 Mb
	Attributes	Table not used	Used or Not	
Y (1)	ClientID	Link to Client		
N	DateTaken	Date of Questionnaire		
N	PMQuestionnaireID	Unique Identifier of questionnaire		
PMQuestionAnswers		Medical Questionnaire Answers	9,754,820	2 Gb
	Attributes	Table not used	Used or Not	
N	QuestionID	Link to Question		
N	Answer	Answer to question		
Y (1)	ClientID	Link to Client		
Y (3)	QuestionnaireID	Link to Questionnaire		
N	TimeStamp	Time of answer		
N	PMQuestionAnswersID	Unique Identifier of answer		

PMQuestions		Medical Questionnair. Questions	16,912	37 Mb
	Attributes	Table not used	Used or Not	
N	Text	Answer Test		
N	TimeStamp	Time question created (?)		
N	PartyID	Who Created the Question		
N	AnswerDataType	Freetext/Int etc		
N	PMQuestionID	Unique Identifier of question		
(D)PMToothType		Dictionary of tooth parts		5 Kb
	Attributes	Table not used	Used or Not	
N	ID	Unique Identifier of tooth type		
N	FDI	FDI Notation		
N	QuadrantID	Quadrant ID Notation		
N	US	US Notation		
N	GenericTypeID	Generic Notation		
N	Quadrant	Quadrant Notation		
N	ImageKey	Application control		
N	QuadrantShortName	Application Control		
(D)PMToothPartType		Dictionary of tooth part types		8 Kb
	Attributes	Table not used	Used or Not	
N	ID	Unique Identifier of tooth part type		
N	SuperTypeID	Application control		
N	ChartTypeID	Application control		
N	LongName	E.G. Mesia/Distal		
N	ShortName	E.G. M/D		
N	ToothViewTypeID	Application control		
(D)PMConditionType		Dictionary of condition types		8 Kb
	Attributes	Table not used	Used or Not	
N	ID	Unique Identifier of Condition type		
N	LongName	E.G. Missing		
N	ShortName	E.G. M		
N	ImageKey	Application Control		
N	ListPosition	Application Control		
N	ToothViewTypeID	Application control		
N	Standalone	Application control		
N	RequiresMaterial	Application control		
N	RequiresSurface	Application control		
N	RequiresTreatment	Application control		
N	ToothUnavailable	?		
(D)PMNationality		Dictionary of nationalities	25	2 Kb
	Attributes	Table not used	Used or Not	
	ID	Unique Identifier of Nationality		
	LongName	Usual Name		
	ShortName	Abbreviation		
	ListPosition	Position on dropdown		

(D)PMClinic		Dictionary of clinic names		7 Kb
	Attributes	Table not used	Used or Not	
	ID	Unique Identifier of Clinic		
	Name	Clinic name		
	BlobServerLocation	Location of scanned documents		
	RegionID	Associated Region		
	RefPrefix	Clinic prefix for associated clients		
	LastClientRef	Integrity measure		
	InstanceID	Integrity measure		
	APBookRefreshRateSeconds	Refresh rate of appointment books		
(D)PMRegion		Dictionary of region names		1 Kb
	Attributes	Table not used	Used or Not	
	Name	Region Name		
	ID	Unique Identifier of Region		
(D)PMAppointmentStatus		Dictionary of appointment statuses		1 Kb
	Attributes	Table not used	Used or Not	
N	ID	Unique Identifier of appt status		
N	Name	Emergency/casual etc		
N	ImageKey	Application Control		
N	ListPosition	Application Control		
N	IsHistory	Application Control		
(D)PMAppointmentType		Dictionary of appointment types		2 Kb
	Attributes	Table not used	Used or Not	
N	ID	Unique Identifier of appt type		
N	Name	Emergency/casual etc		
N	SuperTypeID	?		
N	RequiresClient	Application Control		
N	ShowStatus	Application Control		
N	BackColour	Application Control		
N	ForeColour	Application Control		
N	ImageKey	Application Control		
N	ListPosition	Application Control		

10.4 Ethical Approval & Data-Owner Permission



COISTE EITICE UM THAIGHDE CLINIÚIL
Clinical Research Ethics Committee
Lancaster Hall,
6 Little Hanover Street,
Cork,
Ireland.

Our Ref ECM 4 (f) 06/09/16

2nd August 2016

Professor Martin Kinirons
Dean of Dental School
Oral Health Services Research Centre
University Dental School & Hospital
Wilton
Cork



Re: OHSRC00516: Applying the emerging technologies to process mining to dentistry.

Dear Professor Kinirons

Full approval is granted to carry out the above study at:

- > Oral Health Services Research Services Centre.

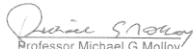
The following documents were approved:

- > Cover Letter dated 21st July 2016
- > Insurance Details
- > Application Form signed 22nd July 2016
- > Data Collection Sheet
- > CV for Chief Investigator.

We note that the co-investigators involved in this study will be:

- > Frank Fox, Bridges Software, Dr Helen Whelton, School of Dentistry, Mr Owen Johnson, School of Computing, Leeds, Dr Michael Thornton, HSE Dental Clinic and Dr Mairead Harding, OHSRC.

Yours sincerely


Professor Michael G Molloy
Chairman
Clinical Research Ethics Committee
of the Cork Teaching Hospitals

The Clinical Research Ethics Committee of the Cork Teaching Hospitals, UCC, is a recognised Ethics Committee under Regulation 7 of the European Communities (Clinical Trials on Medicinal Products for Human Use) Regulations 2004, and is authorised by the Department of Health and Children to carry out the ethical review of clinical trials of investigational medicinal products. The Committee is fully compliant with the Regulations as they relate to Ethics Committees and the conditions and principles of Good Clinical Practice.

Óráilte na hOllscoile Corcaigh - HSE - 06/09/16 - 06/09/16

----- Forwarded message ----- From: **Keane, Shirley - National PCT Programme** <shirley.keane@hse.ie> Date: Wed, Jan 18, 2017 at 6:15 PM Subject: Frank Fox - Applying the emerging technologies of process mining to dentistry To: frnkfx@gmail.com Cc: Emma Benton <emma.benton@hse.ie>, "Pye, Virginia" <Virginia.Pye@hse.ie>, "Keane, Shirley - National PCT Programme" <shirley.keane@hse.ie>, "Kavanagh, Dympna" <dympna.kavanagh@hse.ie>, "Murphy, Brian (Head of Planning, Performance & Programme Management)" <Brian.Murphy@hse.ie>
Dear Frank,

I wish to advise that the Primary Care Research Committee considered the documentation you forwarded to provide clarity on the issues raised by the PCRC members at their last meeting, in particular I refer to the documentation issued by the Data Commissioner in relation to your project.

I wish to advise that the following reflects the discussion and decision of the Primary Care Research Committee:
Frank Fox - Applying the emerging technologies of process mining to dentistry – Arising from last PCRC meeting (Nov 2016):

Application deemed within scope.

The Oral Health lead is aware of this application, the PCRC had two queries to be clarified:

1. Does Frank have access to the data before it is anonymised?
2. If answer to Q1 is yes the PCRC need to see the previously signed data confidentiality agreement to ensure that this covers research activity.
3. If this agreement does not cover research this application will need to be referred to the HSE Data Protection Lead for advice.

Decision: Clarity on above before decision is made.

Documentation received from Frank Fox on 08/01/2017 for consideration and decision at PCRC meeting of 17/01/2017. The response of the Data Commissioner office was noted.

Decision: Approved with the condition that Frank Fox as Researcher is not involved in any data anonymisation process in order for the data to be processed for research purposes.

I note that you have confirmed in the attached documentation that you are in agreement with the above approach (if the PCRC members proposed this condition).

In relation to the approval decision you will note that the PCRC protocol requires that the Primary Care Research Committee will have sight of the final draft report prior to publication and that their

opinion will be considered in relation to the publication, in particular items that may have a bearing on the HSE's reputation, a copy of the protocol is available if required.

I would like to take this opportunity to wish you well with your research.

Kind regards,
Shirley Keane.

On behalf of Chair, Primary Care Research Committee.

Shirley Keane,
Business Planning and Development Manager,
Office of Head of Planning, Performance and Programme Management,
Primary Care Division.
Tel: 091 775922
Mobile: 087 7975674
Email: shirley.keane@hse.ie
Eircode: H91 N973

From: "Michael A. Thornton (Principal Dental Surgeon)" <MichaelA.Thornton@hse.ie>
Date: Thursday, 18 May 2017 at 17:29 **To:** Helen Whelton <H.Whelton@leeds.ac.uk> **Cc:** "Teresa O'Donovan (Head of Primary Care)" <Teresa.ODonovan2@hse.ie>, "Denis Hickey (Project Manager)" <Denis.Hickey@hse.ie> **Subject:** RE: Approval for data use from the PCRC

Dear Helen,

In relation to request to use BRIDGES data for this Research project.

As the dataset has been anonymised, your assurance that the data will be used solely and exclusively for the purposes of Mr. Fox's PhD study and that approval for project has been received from PCRC then I am in agreement that the data can be used.

I wish Mr. Fox every success in his research.

Best regards

Mike

M. Thornton

Principal Dental Surgeon

10.5 Other Governing Documents



10.6 Anonymisation Standard Planning Record

Anonymisation standard planning record			
Source dataset:	Bridges-PM1 DB		
Completed by:	Frank Fox		
Date:			
	Assessed as	Reasoning	help text
Assess threat level associated with data and its release	Normal	Motivation is the major determinant of the threat level associated with the data and its release. This is	In column B, record threat as "high" or "normal". In column C, record your reasoning.
Assess risk of extra information being used to try to reveal identity	Normal	There is no skewed distribution in the data (e.g. Sickle cell anaemia)There is minimal special knowledge, the subjects being previously anonymised before data was released to the researcher. There is no known availability of especially relevant information.	In column B, record "high" or "normal". Note that if threat is "high", then this must be "high". In column C, record your reasoning.
Select anonymisation plan	1	Where cells to be published relate to population > 1,000 people, derive aggregate data without statistical disclosure control. Risk is normal, aggregated data	For column B, choose one of the following: 1 derive aggregate data without statistical disclosure control (normal risk) 2 derive aggregate data with statistical disclosure control (normal risk) 3 derive individual-level data to "weak" k-anonymity (normal risk) 4 derive aggregate data without statistical disclosure control (high risk) 5 derive aggregate data with statistical disclosure control (high risk) 6 Derive individual-level data to "strong" k-anonymity (high risk) Record your reasoning in column C.
Refine anonymisation plan and specify anonymisation			If you decide on any changes to the standard anonymisation plan chosen, record these in column B, and also record any decisions you make on data items to withhold. Record your reasoning in column C
Other comments			Record any additional comments in column C.

10.7 Sample Bridges EHR Application screen

The screenshot shows the 'Client Properties for Test' window with ID 16497. The interface includes several tabs: General, Insurers, Referring, Appointments, Attendance History, Documents, Medical History, Images, and Categories. Below these are sub-tabs for Notepad, Claims, and Finance. The main area is divided into sections for Contact Information, Reference Information, Guardian Information, School Information, Recalls, and Account Information. The patient's name is 'Test', born on 07/July/1977, with a mobile number 0871234567. The account shows a balance of €0.00 and insurance of €0.00. The system administrator's name is visible at the bottom.

10.8 Medical Questionnaire Questions (Alphabetically)

Are there any other aspects concerning your health that you think the dentist should know about?

Allergic to any pills* drugs*medicines* foods or materials?

Do you carry a warning card?

Have you taken steroids in the last two years?

Do you have high blood pressure?

Do you suffer from epilepsy?

Receiving treatment from a doctor?

Do you have a pacemaker* or have you had any form of heart surgery?

Have you had rheumatic fever chorea (St. Vitus Dance)?

Have you had angina or any other heart problem?

Have you ever had difficulty with past dental treatment?

Are you taking any pills*drugs or medicines from your doctor?

Do you have fainting attacks* giddiness or blackouts?

Do you bruise easily or bleed for a long time?

Have you ever had a bad reaction to a general or local anaesthetic?

Do you or anyone in your family suffer from diabetes?

Do you suffer from hay fever* eczema or any other allergy?

Do you suffer from bronchitis* asthma or other chest condition?

Have you had jaundice* liver* kidney disease or hepatitis* HIV?

Have you ever had a joint replacement?

Do you ever get cold sores?

Do you know of any bleeding problem in the family?

Do you have a history of abnormal bleeding after extractions?

Are you suffering from any illness?

Have you had a heart attack?

Have you had a heart murmur or a history of one?

Are you an expectant mother?

Have you ever had your blood refused by the Blood Transfusion Service?

10.9 How DMFT is calculated in this research.

Each patient can have multiple charts – created on different dates. Each charting has the indices calculated for it at the time it was created. DMFT is calculated using all the available permanent teeth. dmft is calculated using all the available deciduous teeth. D, M, & F are separately calculated and totalled to give a DMF value for the chart. We retain all the values.

DMFT at the time of an initial examination is often used as a criterion when selecting cohorts in this thesis. As DMFT is calculated from the charting, there is sometimes no charting on the same day i.e. a time-gap exists between the initial exam and the charting. However 98% of initial exams had a charting within two months.

From the BridgesPM1 Condition list, the following are considered as D

Cavity	Counted as Decayed
Replacement Filling	Counted as Decayed
Root Remaining	Counted as decayed (all surfaces corresponding to no of surfaces of the particular tooth)
Preventative Restoration	This is a tricky one, we will need to keep it in a separate Decayed category, so that the dmft and DMFT can be reported both with and without it included.
Preventative Restoration Required	This is a tricky one, we will need to keep it in a separate Decayed category, so that the dmft and DMFT can be reported both with and without it included.
Let Deciduous Fall	Count as decayed

From the BridgesPM1 Condition list, the following are considered as M

Missing	Must ensure that it was extracted for decay. It should not include extracted for orthodontic purposes
For Extraction	Must be for dental decay
Denture	Must be for dental decay
Denture Required	Must be for dental decay
Recent Extraction	Counted if due to dental decay
Bridge	Must be for dental decay
Bridge Required	Must be for dental decay
Root Remaining	Counted as decayed (all surfaces corresponding to no of surfaces of the particular tooth)
Crown	Must be for dental decay
Crown Required	Must be for dental decay
Space Closed Up	Must be for dental decay

From the BridgesPM1 Condition list, the following are considered as F


Filling	Counted as Filled - but must not be a filling placed because a tooth was fractured in an accident - Typically these are recorded as Incisal tip. Mesial and Distal cavities on Incisors are typically due to decay.
----------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

10.10 Posters and Oral Presentations

School of Dentistry
FACULTY OF MEDICINE AND HEALTH UNIVERSITY OF LEEDS

Data-Mining, Process-Mining, and Visualising a Dental Public Health Dataset

Frank Fox¹, Owen Johnson², Helen Whelton³, Vishal Aggarwal¹
¹School of Dentistry, ²School of Computing and Leeds Institute for Data Analytics, University of Leeds, U.K.
³College of Medicine and Health, University College Cork, Ireland



Background

Activities taking place in dental clinics leave tracks in the dental Electronic Health Record (EHR). Dental Process Mining extracts information about treatment processes from this routine data. Process Discovery and Mining provide a method to map day to day processes as an alternative to traditional back line observations and questionnaires. This remains under-researched in dentistry.

What is Electronic Health Record Data?

- data normally stored on a computer often in a database
- data arises as a result of normal operations e.g. treating patients

What is "Data Mining"?

- Looking for patterns in these large datasets

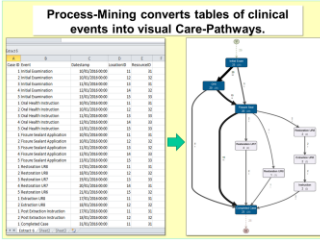
What is "Process Mining"?

- Looking for process patterns - often called Care-Pathways in large datasets

What is "Visualisation"?


- Converting large datasets into comprehensible formats
- Facilitates communication of complex data

Process-Mining converts tables of clinical events into visual Care-Pathways.



Why is this important?

Comparing Process-Mining outputs with optimised Care-Pathways facilitates assessment of dental public health service delivery



Research Data Description

An anonymised extract of 200,000 dental Electronic Health Record from the local public health service

- The data results from the school screenings of children between 2000 and 2014

The "BridgesPM1" database is suitable for the following reasons:

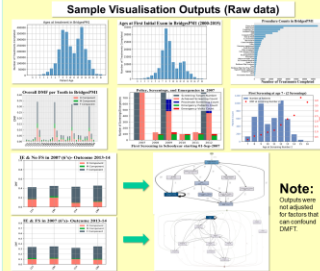
- dataset covers 10 years of annual coverage and treatments
- dataset contains clinical, administrative, and health data

Class of Data	Description	Number of Rows
Patients	Patient Demographics	n = 253,246
Treatments	Treatment Event Descriptions (Events)	n = 3,129,854
Treatment Courses	Treatment Course Descriptions (Events)	n = 285,090
Diagnoses	Chief Complaint and associated ICD10	n = 1,478,182
Health Conditions	Health Condition Descriptions	n = 32,293,486

Using Care-Pathways is well established in dental public health.

- Clinical Guidelines¹
- Steele Report²
- NICE Guidelines³
- Dental Contract Reform Pilot⁴

Sample Visualisation Outputs (Raw data)



Biggest Challenges?

- EHR Data Quality⁵
- Complex Process Models
- Process Model quality

Conclusions

- Technologies were readily applied
- They were reusable, agile and flexible
- Development of these technologies should be a priority for investigating large datasets for use in dental public-health policy decision making

Aim & Objectives

To provide previously unavailable information to dental policy makers and stakeholders by converting routinely collected dental data into visual processes and actionable insights.

Objectives:

- Develop a process mining methodology for dental research
- Show how process mining can evaluate strategic initiatives and conform with Standards and Guidelines
- Develop recommendations for guidelines and EHR design

General method for using Process-Mining of EHR data to evaluate a policy/strategy or decision.

- Identify a situation that represents a policy or strategy change or decision of interest.
- Assemble evidence that this policy actually happened in the real world.
- Is the policy/strategy visible in the EHR?
- What are the appropriate outcomes to measure the effects of a policy/strategy?
- Which of these outcomes are present on the EHR?
- With the objective of ensuring cohorts are from a level playing field, identify potential exposures, outcomes, confounders and mediators and mitigate if possible.
- Develop specific Research Questions around the policy/strategy, answerable with the EHR data.
- Identify cohorts on all sides of the policy/strategy.
- Establish outcomes for these cohorts.
- Are the outcomes different for the cohorts?
- Establish the treatment process maps experienced by the cohorts.
- Are the treatment process maps of adequate quality?
- Are the treatment processes different?
- Analyse and Discuss.

References

1. The National Institute for Health and Care Excellence (NICE) Clinical Guidelines. <https://www.nice.org.uk/guidance>
2. Steele Report. <https://www.dentalcontractreform.co.uk/>
3. National Institute for Health and Care Excellence (NICE). <https://www.nice.org.uk>
4. Dental Contract Reform Pilot. <https://www.dentalcontractreform.co.uk/>
5. EHR Data Quality. <https://www.ehrdataquality.com/>

Poster Presentation, University of Leeds, School of Dentistry Research Day,

School of Dentistry
FACULTY OF MEDICINE AND HEALTH

UNIVERSITY OF LEEDS

Process Mining in Dentistry

Frank Fox¹, Owen Johnson², Helen Whelton¹, Vishal Aggarwal¹
¹School of Dentistry, ²School of Computing and Leeds Institute for Data Analytics
l.fox@leeds.ac.uk



Background

- Process Mining is a branch of Data Mining.
- Activities taking place in dental clinics leave tracks in the dental Electronic Health Record (EHR).
- Dental Process Mining** extracts information about dental treatment processes from this routinely collected data.
- Process Discovery and Mining** provide a method to map out day to day processes without resorting to the traditional tools of observation and questionnaires.
- This remains under-researched in dentistry

Aim & Objectives

AIM:
To provide previously unavailable information to dental policy makers and other stakeholders by converting routinely collected dental data into visual processes and actionable insights.

OBJECTIVES:

- Develop a process mining methodology for dental research.
- Identify the data needs of dental Policy Makers
- Show how process mining can evaluate strategic initiatives and conformance with Standards and Guidelines
- Develop recommendations for future Guidelines and EHR designs

Process Mining in Dentistry converts tables of clinical events into visual process maps.

Why?

- Discover and monitor care pathways and variants.
- Compare actual treatment processes with Standard Operating Procedures or Clinical Guidelines.
- Enhance discovered processes for improvements and prediction capability.

Summary PM² Process Mining Methodology⁴

Phases: Research Planning → Data Extraction → Data Processing → Mining & Analysis → Evaluation → Process Improvement and support

Inputs/Outputs: Research Planning (Inputs: Research Objectives, Data Sources; Outputs: Data Requirements) → Data Extraction (Inputs: Data Sources; Outputs: Data) → Data Processing (Inputs: Data; Outputs: Cleaned Data) → Mining & Analysis (Inputs: Cleaned Data; Outputs: Process Maps) → Evaluation (Inputs: Process Maps; Outputs: Process Performance) → Process Improvement and support (Inputs: Process Performance; Outputs: Recommendations)

Healthcare research requires additional efforts to secure the data e.g.:

- Ethics clearance
- Data controller permissions
- Software supplier cooperation
- Data anonymisation protocols & artefacts supporting these steps.

Methods - Research Data Description

- An anonymised extract of BRIDGES dental Electronic Health Record from the Irish public health service
- The data results from the school screenings of children between 1999 and 2014

The BridgesPM1 database is notable for the following reasons:

- The dataset spans 15 years of dental school screenings and resultant treatments
- The dataset contains clinical, administrative, oral health and KPI data

Class of Data	Description	Number of Rows
Patients	Patient Demographics	230 k
Treatments	Treatment Event Description (events)	3.2 million
Treatment Courses	Treatment Course Identifiers (cases)	285 k
Charts	Chart Identifier and associated DMF measure	1 million
Tooth Condition	Tooth Condition Description	32.3 million
Attendances	Attendance history	5.5 million
Answers	Medical Questionnaire Answers	9.7 million

Preliminary Results

- A 15-day subset of the BridgesPM1 database was selected
- Data pre-processing steps were executed
- Initial results using Disco™ yielded the expected spaghetti type process map shown below in Figure 1.3
- Filtering the data led to more readable process maps as shown below in Figure 2.



Preliminary Conclusions

- Process mining facilitates the creation of useful visualisations from operational dental data
- Pre-processing and filtering of the data is required to balance accuracy and comprehensibility
- Careful experiment design and verification of the results with domain experts will be essential
- Further work is now being undertaken to verify and validate this methodology using data from the School of Dentistry's dental Electronic Health Record

References

1. [Using Guidelines Process for Oral Health Assessment of 5 year olds. Adapted from Irish Oral Health Services Guidelines Initiative 2012, pp. 6-7](#)
2. [Helen R. Whelton, Owen Johnson, Vishal Aggarwal, Frank Fox, 'Mining Process Mining in Dentistry', 1st January 2018, Miami Florida](#)
3. [Helen R. Whelton, Owen Johnson, Vishal Aggarwal, Frank Fox, 'Mining Process Mining in Dentistry', 1st January 2018, Miami Florida](#)
4. [PM2: A Process Mining Methodology for Healthcare Systems Engineering. The International Conference on Health Informatics, 2018, Proceedings Ed. J. Zaslavski, M. Gellera, P. Janssens, Berlin: Springer, 2018.](#)

Oral Presentations

- 1) University of Leeds, School of Dentistry Research Day 2018
- 2) University of Leeds, Faculty of Medicine & Health Postgraduate Research Day 2018

Abstract Submission

Oral Presentation Abstract: Frank Fox

Supervisors: Dr Vishal Aggarwal, Mr Owen Johnson, Prof Helen Whelton

Title: Data-mining, process-mining and visualising an electronic clinical record database

Aims:

- Apply data-mining, process-mining and data-visualisation to an electronic clinical record database in a public dental service.
- Demonstrate the flexibility and agility of the technologies and their potential for assessing the impacts of policy and strategy on oral health outcomes.

Background: Data and process-mining tools were applied to an Irish Health Service Executive dental public-health database to profile and visualise the data.

Methods: Methods were developed facilitating flexible and rapid selection of patient cohorts based on criteria such as age, treatments received and oral health outcome (DMFT). Process-Mining methods were applied to visualise the treatment processes experienced by these patients. Method and technologies were validated by examining effect of fissure-sealants on DMFT. Cohorts, one receiving school dental screenings and fissure-sealants in 2007 at age 8, and one receiving no fissure-sealants were compared on their DMFT at age 13/14.

Results/Findings: Technologies were readily applied, were repeatable, agile and flexible. Visualisations generated were easy to understand and interpret, including tooth specific data. Validation, by examining the effect of fissure-sealants on DMFT showed, as expected, that application of fissure-sealants at age 8 was associated with a lower DMFT at age 13/14. Treatment processes were readily demonstrated using process-mining techniques.

Conclusions or recommendations: For the first time, data-mining and process-mining technologies were applied to visualise and interpret a dental public health clinical dataset. The validating example, although producing the expected outcome, was not adjusted for factors that can confound DMFT. Development of these technologies should be a priority for investigating large dental datasets for use in dental public-health policy decision making.

3) New York, IEEE International Conference of Health Informatics, June 2018

A Data Quality Framework for Process Mining of Electronic Health Record Data

Frank Fox School of Dentistry University of Leeds Leeds, U.K. dnfgf@leeds.ac.uk

Vishal. R. Aggarwal School of Dentistry University of Leeds Leeds, U.K. V.R.K.Aggarwal@leeds.ac.uk

Helen Whelton College of Medicine and Health University College Cork Cork, Ireland h.whelton@ucc.ie

Owen Johnson School of Computing University of Leeds Leeds, U.K. O.A.Johnson@leeds.ac.uk

Abstract:- *Reliable research demands data of known quality. This can be very challenging for electronic health record (EHR) based research where data quality issues can be complex and often unknown. Emerging technologies such as process mining can reveal insights into how to improve care pathways but only if technological advances are matched by strategies and methods to improve data quality. The aim of this work was to develop a care pathway data quality framework (CP-DQF) to identify, manage and mitigate EHR data quality in the context of process mining, using dental EHRs as an example.*

Objectives: *To: 1) Design a framework implementable within our e-health record research environments; 2) Scale it to further dimensions and sources; 3) Run code to mark the data; 4) Mitigate issues and provide an audit trail.*

Methods: *We reviewed the existing literature covering data quality frameworks for process mining and for data mining of EHRs and constructed a unified data quality framework that met the requirements of both. We applied the framework to a practical case study mining primary care dental pathways from an EHR covering 41 dental clinics and 231,760 patients in the Republic of Ireland.*

Results: *Applying the framework helped identify many potential data quality issues and mark-up every data point affected. This enabled systematic assessment of the data quality issues relevant to mining care pathways.*

Conclusion:

The complexity of data quality in an EHR-data research environment was addressed through a re-usable and comprehensible framework that met the needs of our case study. This structured approach saved time and brought rigor to the management and mitigation of data quality issues. The resulting metadata is being used within cohort selection, experiment and process mining software so that our research with this data is

based on data of known quality. Our framework is a useful starting point for process mining researchers to address EHR data quality concerns.

4) Sydney, Process-Oriented Data Science for Health (PODS4H), September 2018

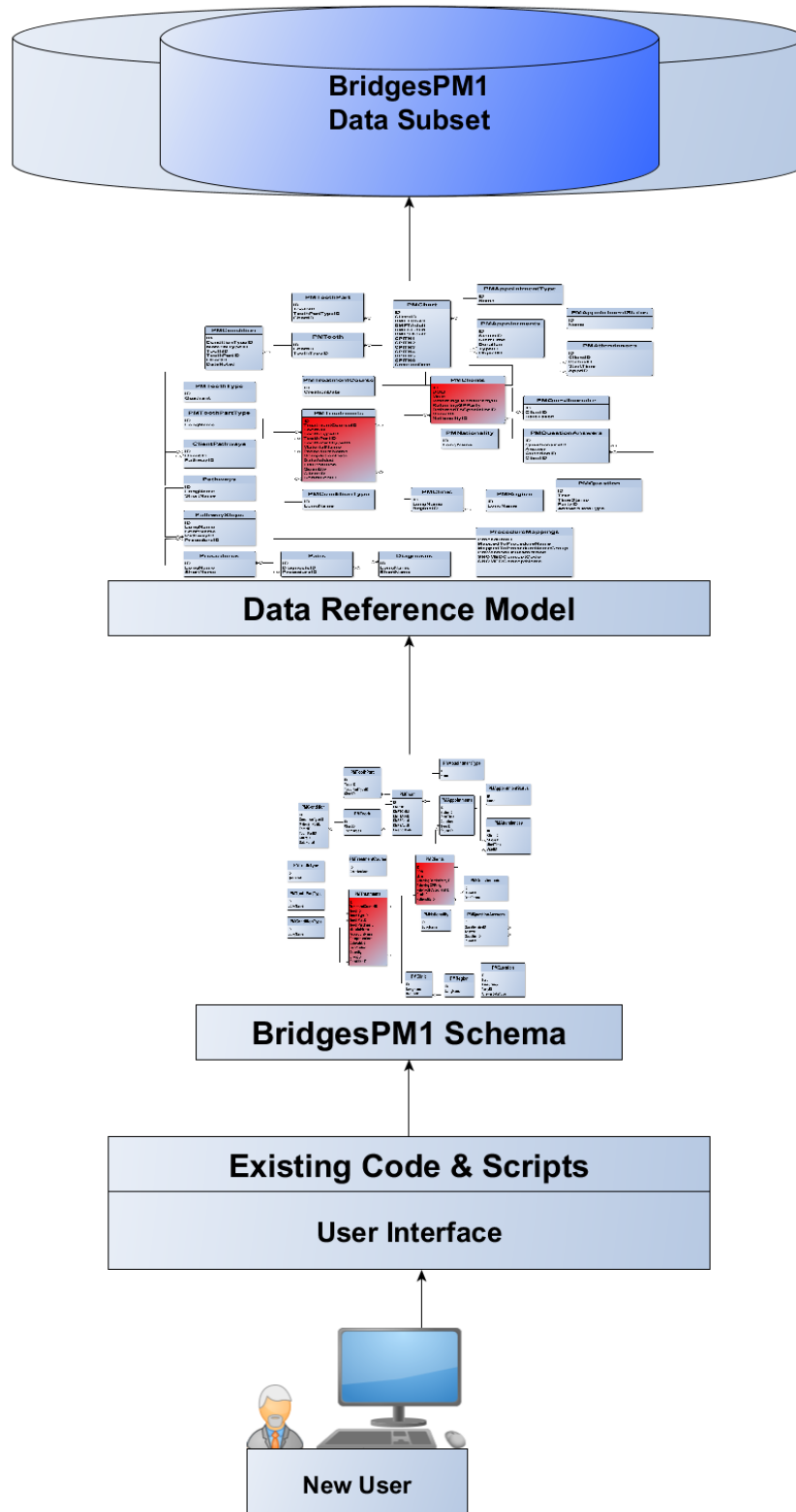
Research using data from electronic health records (EHRs) is in its infancy when compared to research using traditional methods for clinical trials and epidemiological studies. Researchers must develop a detailed understanding of the EHR data's provenance, quality, and suitability before they can trust the data enough to answer the RQs being asked. A key step to achieving that trust is to face the challenges of EHR data quality (DQ) head on: Find and document the issues, manage them, assess their impact and relevance to the research, mitigate their effects where possible, and report clearly on these steps. We have developed a data quality framework to achieve these aims and applied this to a Dental EHR-based process-mining research project in the University of Leeds. The framework is based on existing EHR and process mining data quality literature, and is implementable as an automated, software solution.

5) Leeds, ADVOCATE International Data Conference November 2018, "Quality and utility of data"

- An overview of 'what is quality data' and why it is important utilising the DQF.
- How can we know what is quality data for our research question?
- How can the quality of data collected for healthcare use be improved?
- What level of quality is required?
- How can data-owners who have an interest in quality improvement initiatives ensure they collect the right data in the right form so that it can be easily processed further by other data users? (e.g. researchers).
- What steps to do we need to take to improve the quality and availability of data in the long term? What is realistic?

10.11 Code Reuse Guide

The Python/Jupyter Notebook code could be reused with new data in certain circumstances. The notebook is designed for use with SQL Server and any new data would have to fit into the existing BridgesPM1 database schema as described in Section 4.1.5.2. The code is restricted to the original schema and does not accommodate the new schema as proposed for the data reference model. The code will ‘look through’ the enhanced schema (the data reference model) and any data stored in the enhanced entities and attributes. This can be approximately represented as in the figure below.



10.12 RQ1 SQL cohort selection code (as sample)

This code snippet is the SQL code to create the cohort for the RQ in Section 7.1.1. This cohort is used in establishing the feasibility of using our EHR data for comparison with established care pathways.

```
IF OBJECT_ID ('dbo.CohortCarePathway') IS NOT NULL
```

```
Drop Table CohortCarePathway
```

```
go
```

```
Create Table CohortCarePathway
```

```
(ClientID uniqueidentifier,
```

```
CompletionDate DateTime)
```

```
go
```

/* This statement creates a list of patients who received an 'Initial Exam' or attended for an emergency appointment between Sept 1st and Sept 5th 2007 inclusive. It excludes patients who had had a previous exam or emergency appointment. It excludes data of bad quality */

```
Insert into CohortCarePathway
```

```
Select distinct(Tr.ClientID), Tr.CompletionDate as visitdate from PMTreatments Tr
where Tr.MappedToProcedureNameGroup in ('Initial Exam', 'Emergency
appointment')
```

```
and Tr.CompletionDate between '01-sep-2007' and '5-sep-2007'
```

```
and Tr.BadRow is Null
```

```
and Tr.clientID not in
```

```
(select distinct ClientID from PMTreatments where MappedToProcedureNameGroup
not in ('Initial Exam', 'Emergency appointment')
```

```
and CompletionDate < Tr.CompletionDate
```

```
)
```

```
IF OBJECT_ID ('dbo.MinProcedureCount') IS NOT NULL
```

```
Drop table MinProcedureCount go
```

/*This statement creates a list of procedures occurring more than 20 times for these patients. This reduces the spaghetti affect */

```
select MappedToProcedureNameGroup into MinProcedureCount from PMTreatments
where ClientID in (Select ClientID from CohortCarePathway)
```

```
and CompletionDate>='01-Sep-2007'
```

```
group by MappedToProcedureNameGroup
```

```
having count(MappedToProcedureNameGroup) >20
```

/* This statement creates the Event Log for this cohort – selecting all treatment events after the initial exam or emergency appointment where the occurrence frequency was >20 */

```
select distinct ClientID, MappedToProcedureNameGroup, CompletionDate from
PMTreatments where ClientID in (Select ClientID from CohortCarePathway)
```

```
and CompletionDate>='01-Sep-2007'
```

```
and MappedToProcedureNameGroup in (
```

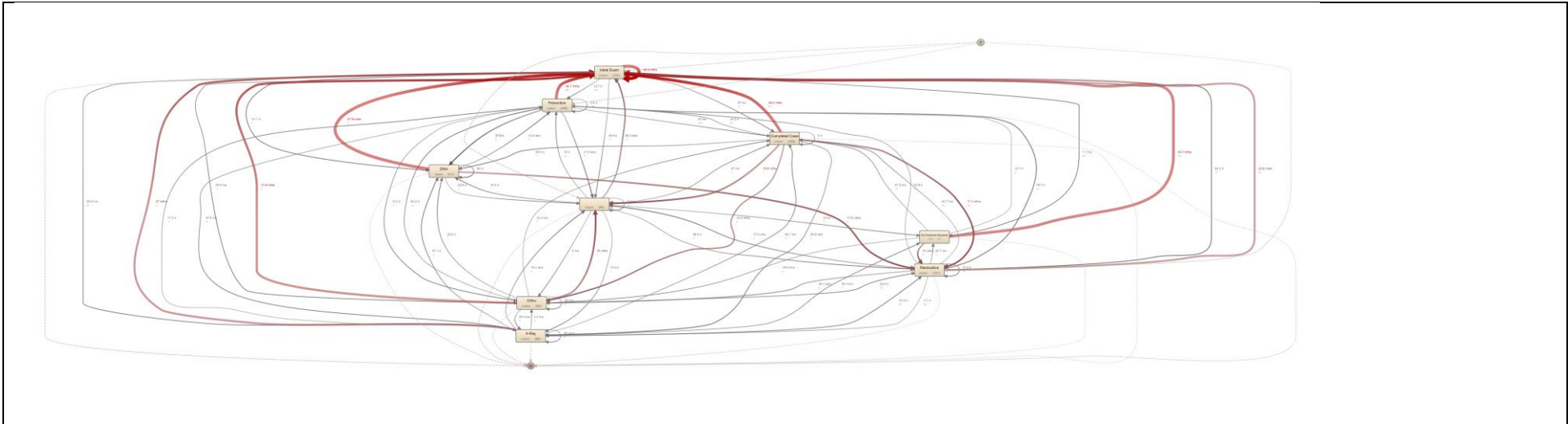
```
select MappedToProcedureNameGroup from MinProcedureCount)
```

```
order by CompletionDate
```

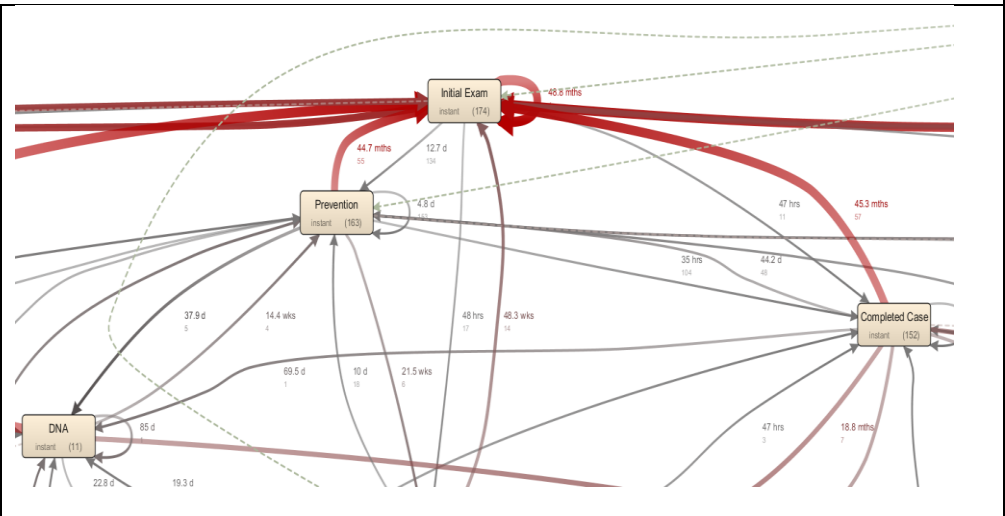
10.13 Frequency of Screening Details

Cohorts		Case/Event Histogram
2 Screenings (Kerry)		
No. of Cases	174	
Variants	174	
No. of Events	2,501	
Events per case	14.37	
3 Screenings (North Lee)		
No. of Cases	63	
Variants	63	
No. of Events	1,678	
Events per case	26.63	
3 Screenings (South Lee)		
No. of Cases	43	
Variants	43	
No. of Events	1,017	
Events per case	23.65	
3 Screenings (West Cork)		
No. of Cases	221	
Variants	220	
No. of Events	4,935	
Events per case	22.33	
4 Screenings (North Cork)		
No. of Cases	2	
Variants	2	
No. of Events	75	
Events per case	37.5	

Figure 10-1: Event Log Characteristics (Frequency of school screening)



Better readability and comprehension of the models can be achieved with the pan and zoom functionality available in the Disco PM product. On the right, the performance detail available in models containing all the data can be seen. Knowing that the model is showing all the executed paths and activities also engenders a higher degree of trust in the results, however, increasing complexity eventually makes the model more difficult or impossible to interpret.



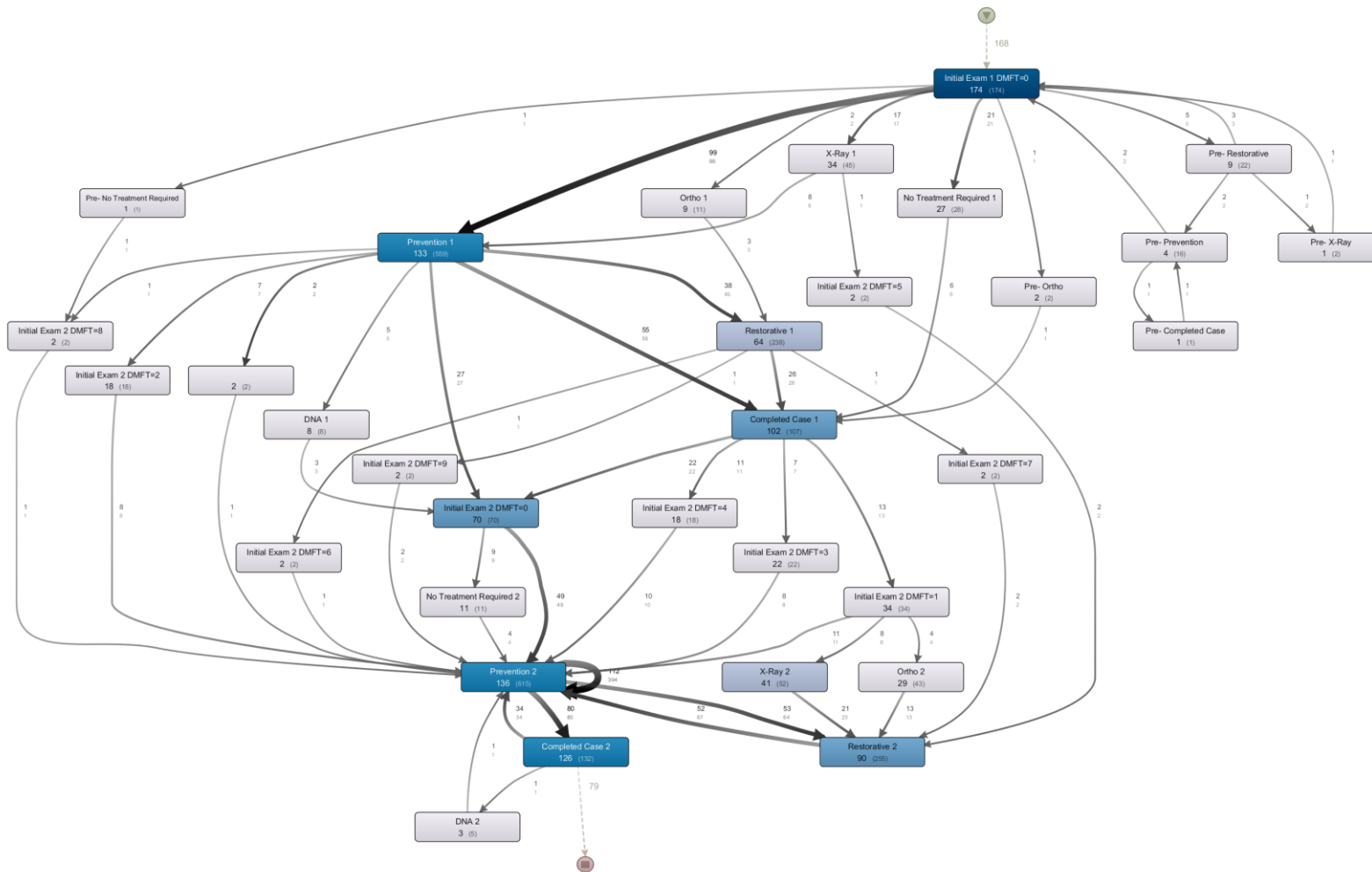


Figure 10-2: Frequency model enhanced with ‘rank & DMFT’ for 2 Screenings (Kerry). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.

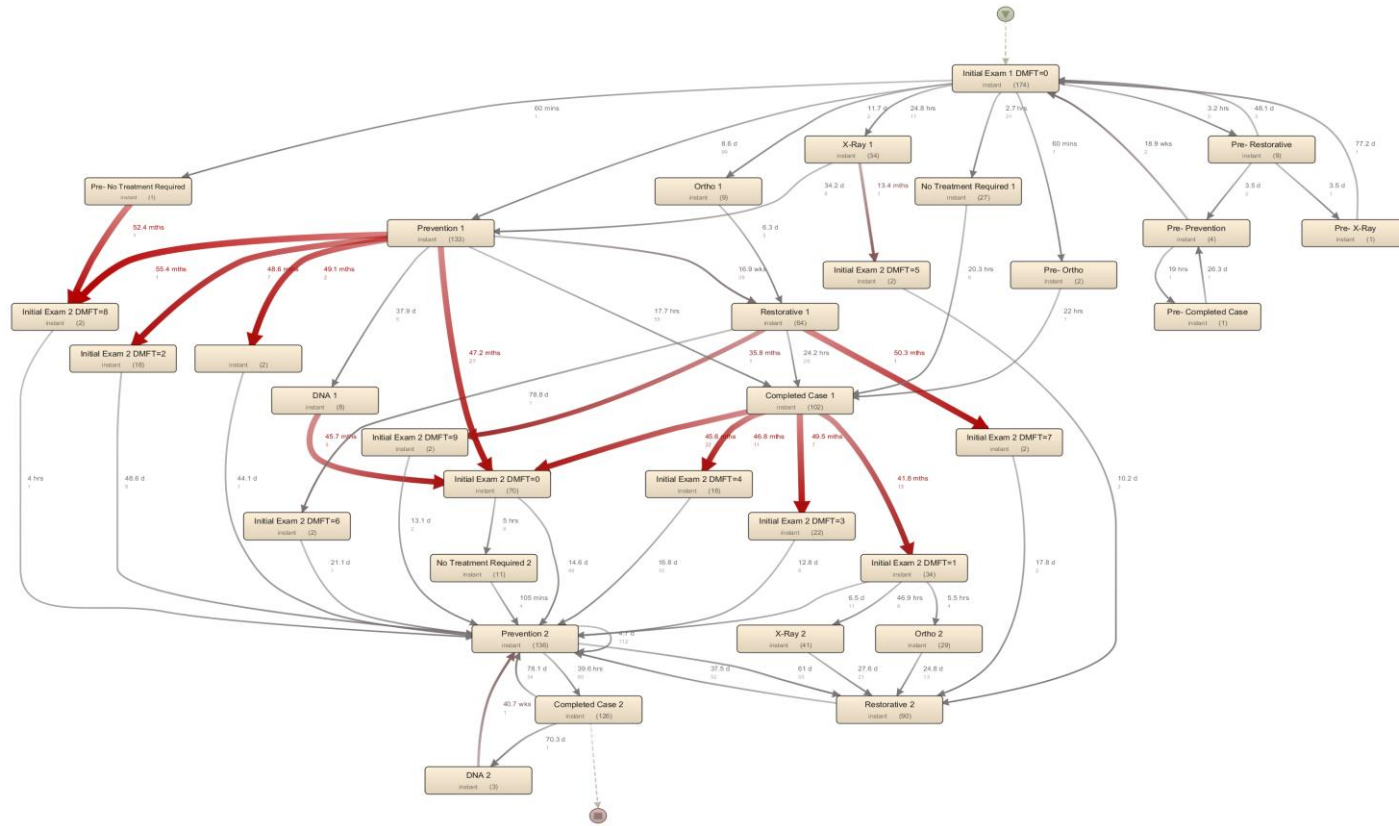


Figure 10-3: Performance model enhanced with ‘rank & DMFT’ for 2 Screenings (Kerry). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.

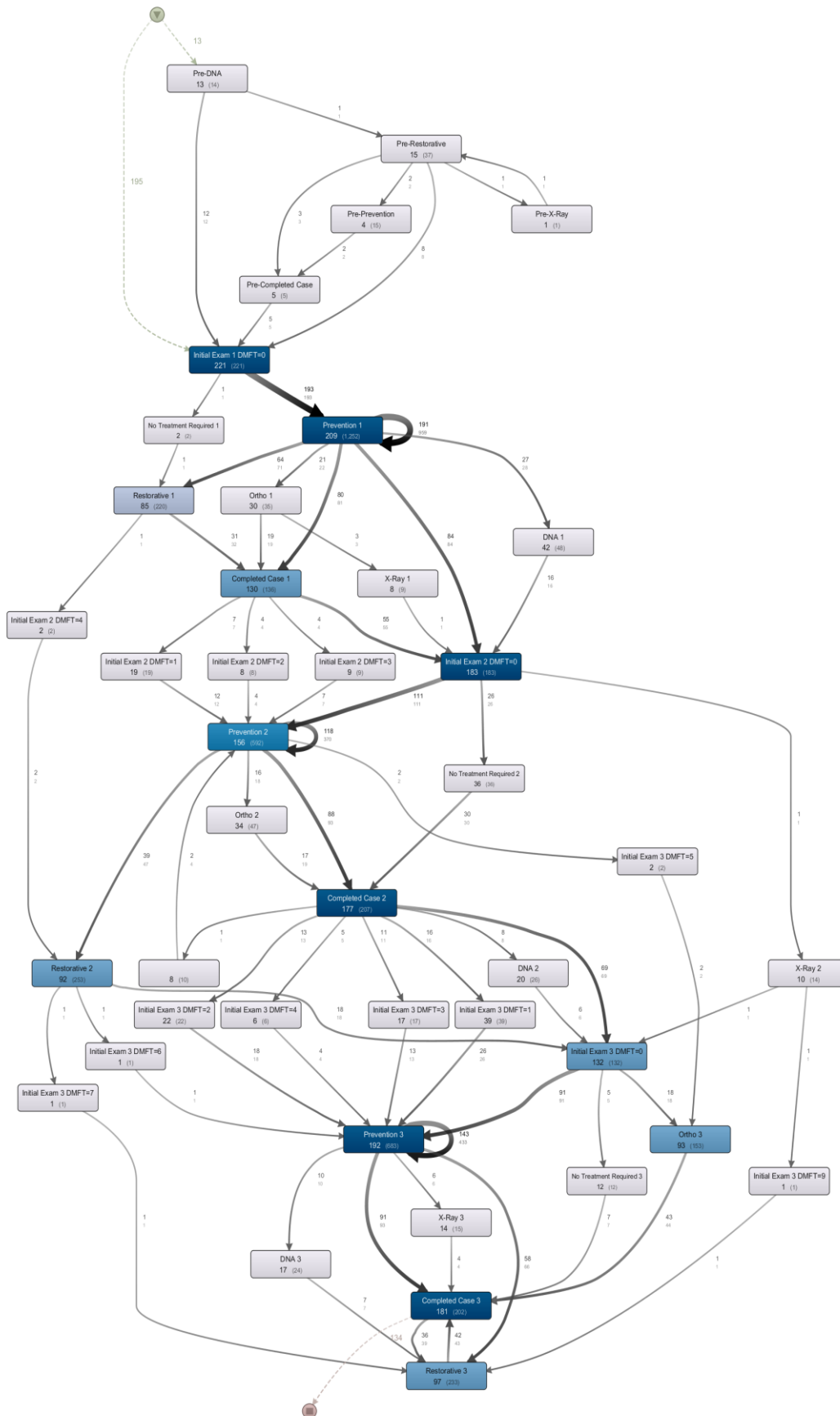


Figure 10-4: Frequency model enhanced with ‘rank & DMFT’ for 3 Screenings (West Cork). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.

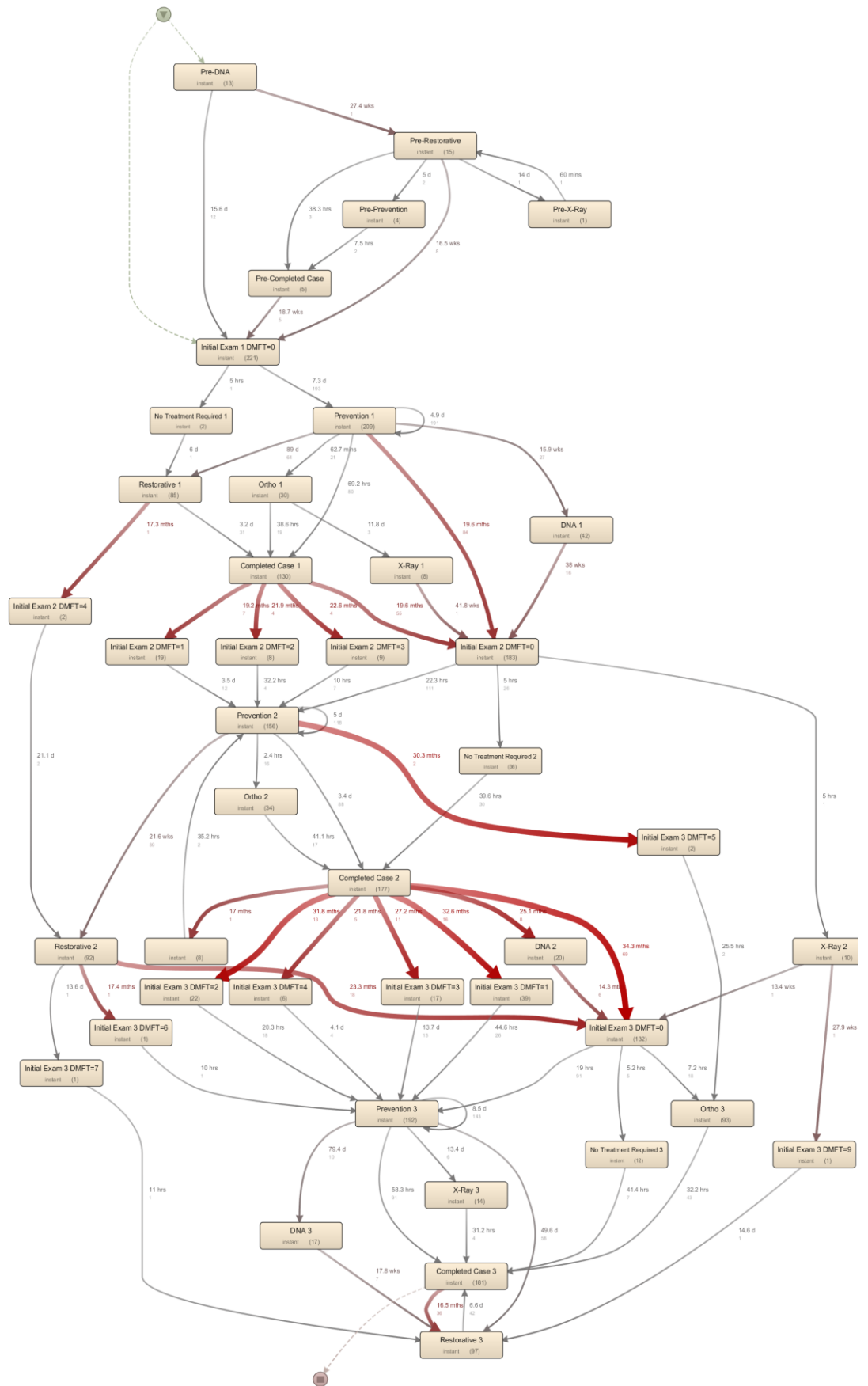


Figure 10-5: Performance model enhanced with ‘rank & DMFT’ for 3 Screenings (West Cork). Temporal sequence for children receiving their first initial exam in academic year 2005/2006 and all subsequent treatment up to 2015.

Note: How to read the Screening Profiles data below.

Each of the five Community Care Areas (Local Health Area) for which we have data has a profile below. The profile consists of three groups of two graphs (Figures a, b, & c) containing increasingly more detailed information. The left column is the cohorts with starting DMFT=0 and the right column has all starting DMFT values

Colour Codes:

Grey Bar – Screening Policy target (based on number of patients compliant with policy & surviving data quality and other restrictions)

Salmon Bar – Screening Target achieved (in first year this is the same as the target)

Navy Bar – Proximate screenings carried out

Green & Black Bar – Emergency Patient and visit numbers – (of the targeted patients)

Figure (a) shows only the Screening Policy Target and the numbers screened that adhered strictly to the policy.

Figure (b) shows the targets and achieved screenings as per figure (a) and also shows, as an addition to the strictly achieved screenings, screenings carried out in the school year adjacent to the strict target year i.e. if a child was targeted and seen in year one, and was targeted but not seen in year 4, but was seen in year 3 or 5, this is seen as a ‘proximate’ screening and shown as the blue section in figures (b) & (c) This varies slightly between areas as they have different policies. The adjacent screened for Kerry 6th class counts those seen in the year prior and the year after. The adjacent screened for North Cork 4th class counts those seen in third class and adjacent screened in 6th class counts those seen in 5th class. For North Lee, South Lee and West Cork the adjacent screened for 3rd class counts 2nd and 4th class, and the adjacent screened for 6th class counts 5th class and the year after.

Figure (c) shows the number of targeted patients also presenting for emergency treatment in each year (green bar). Multiple visits for these patients are represented by the black bar.

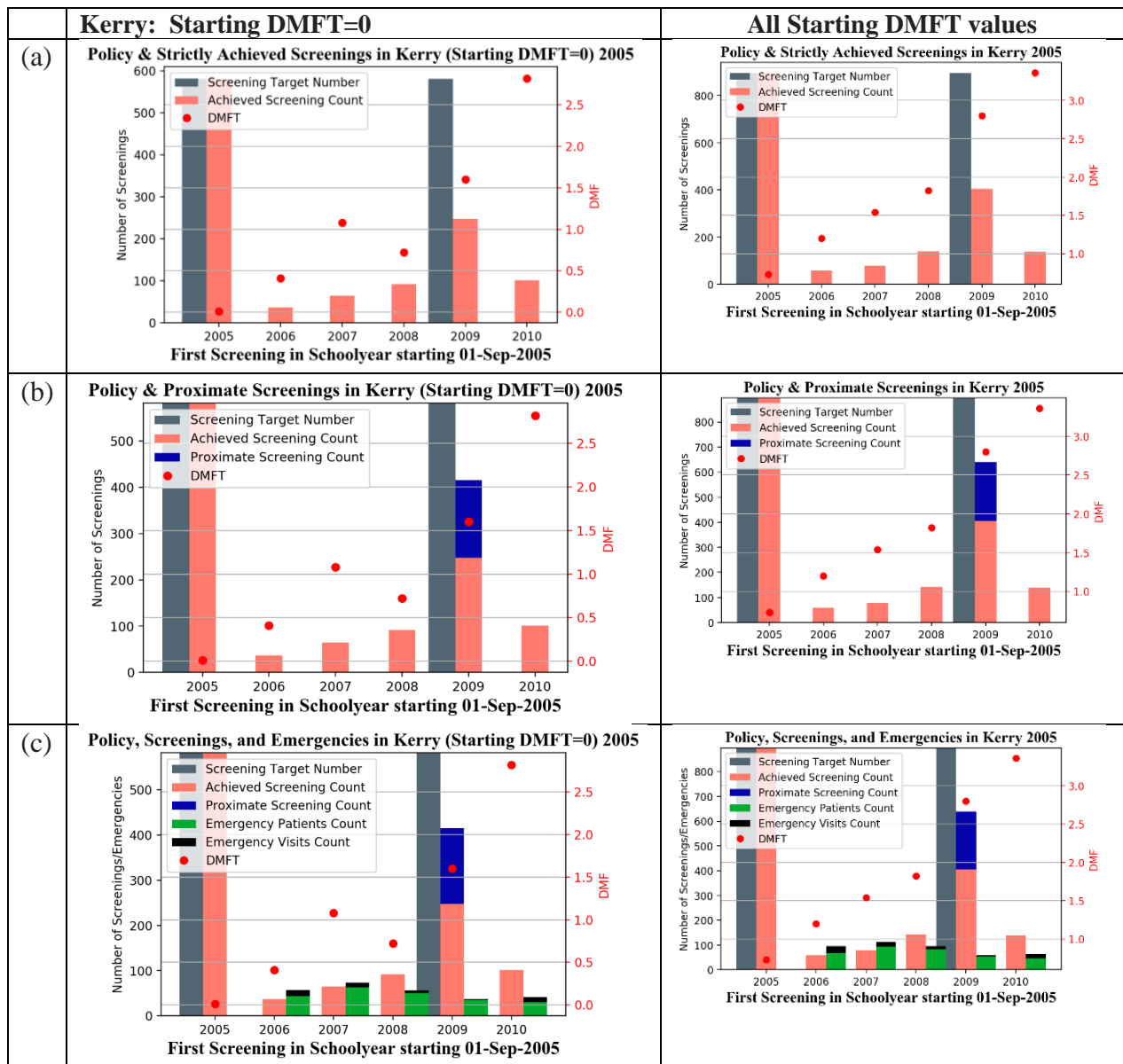


Figure 10-6: Policy & Screening Profile for Area (Kerry)

What are these charts telling us?

- Chart (a), Starting DMFT=0, shows that a cohort of 581 patients complying with the requirements could be identified with Screening (Initial Exam) carried out between September 1st of target year (2005) and 31st August of following year (2006). This was the first screening for those patients. The patients were aged 7,8 or 9 at the time of the screening. The data quality was acceptable. Starting DMFT was approaching 0.
- Chart (a), showing policy and strictly adhered to numbers, shows adherence of 248 patients in the only other policy year, 2009.
- However, Chart (b), where patients seen in adjacent years are stacked on top, the total of the original cohort of 581 seen for their 2nd screening is over 415.
- Chart (c) shows that approximately 10% of the original cohort are seen for emergency appointments annually.
- DMFT (the red dots) generally increases with time between screenings.
- Charts for all starting DMFT values show similar trends.

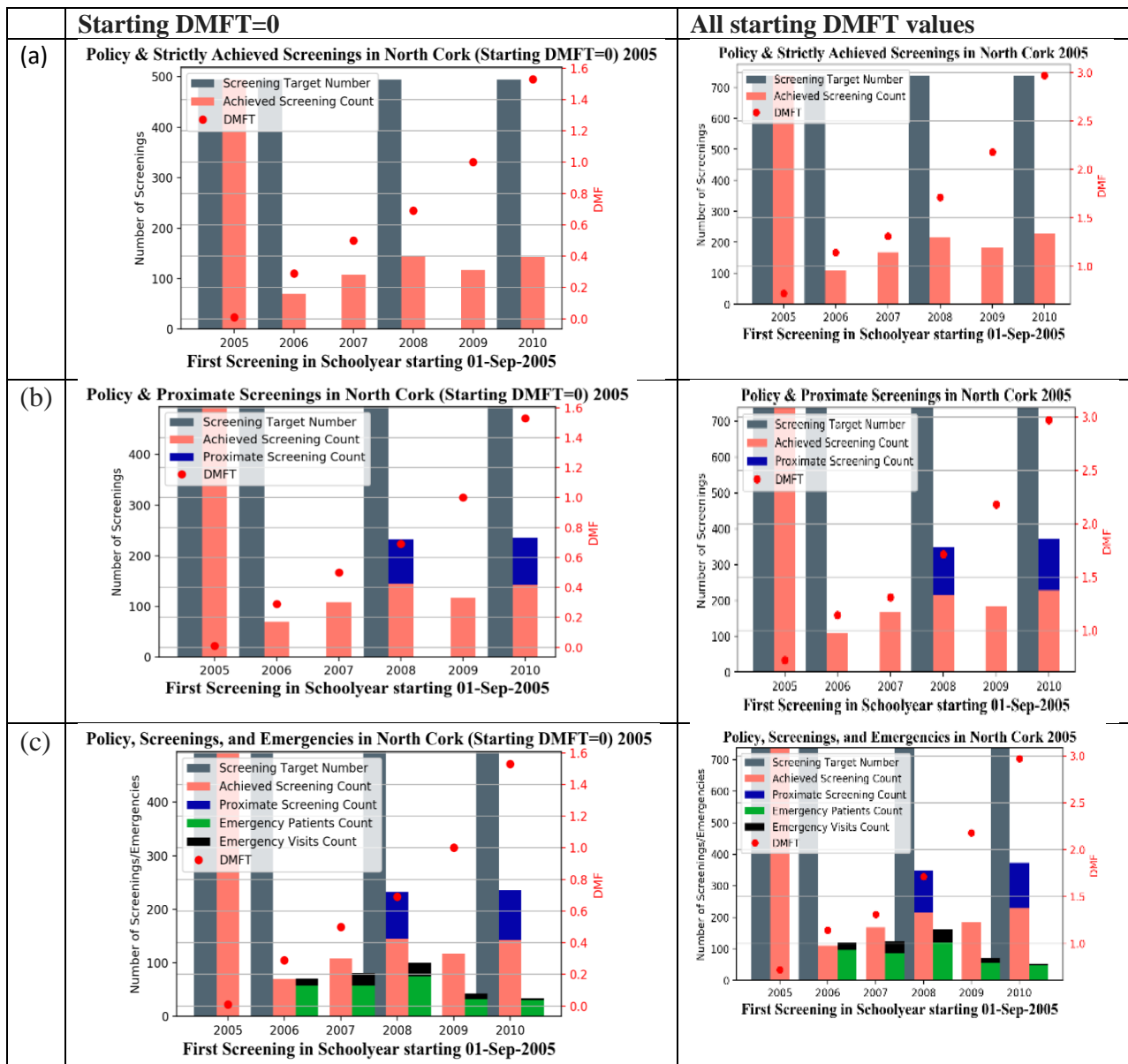


Figure 10-7: Policy & Screening Profile for Area (North Cork)

What are these charts telling us?

- Chart (a), Starting DMFT=0, shows that cohort of 494 patients complying with the requirements could be identified with Screening (Initial Exam) carried out between September 1st of target year (2005) and 31st August of following year (2006). This was the first screening for those patients. The patients were aged 7,8 or 9 at the time of the screening. The data quality was acceptable. Starting DMFT was approaching 0.
- Chart (a), showing policy and strictly adhered to numbers, shows adherence of 70 patients in year 2, 145 in year 4, and 143 in year 6.
- However, Chart (b), where patients seen in adjacent years are stacked on top, the total seen in year 4 is 233 and year 6 is 236.
- Chart (c) shows that approximately 10% of the original cohort are seen for emergency appointments annually.
- DMFT (the red dots) increases with time between screenings.
- Charts for all starting DMFT values show similar trends.

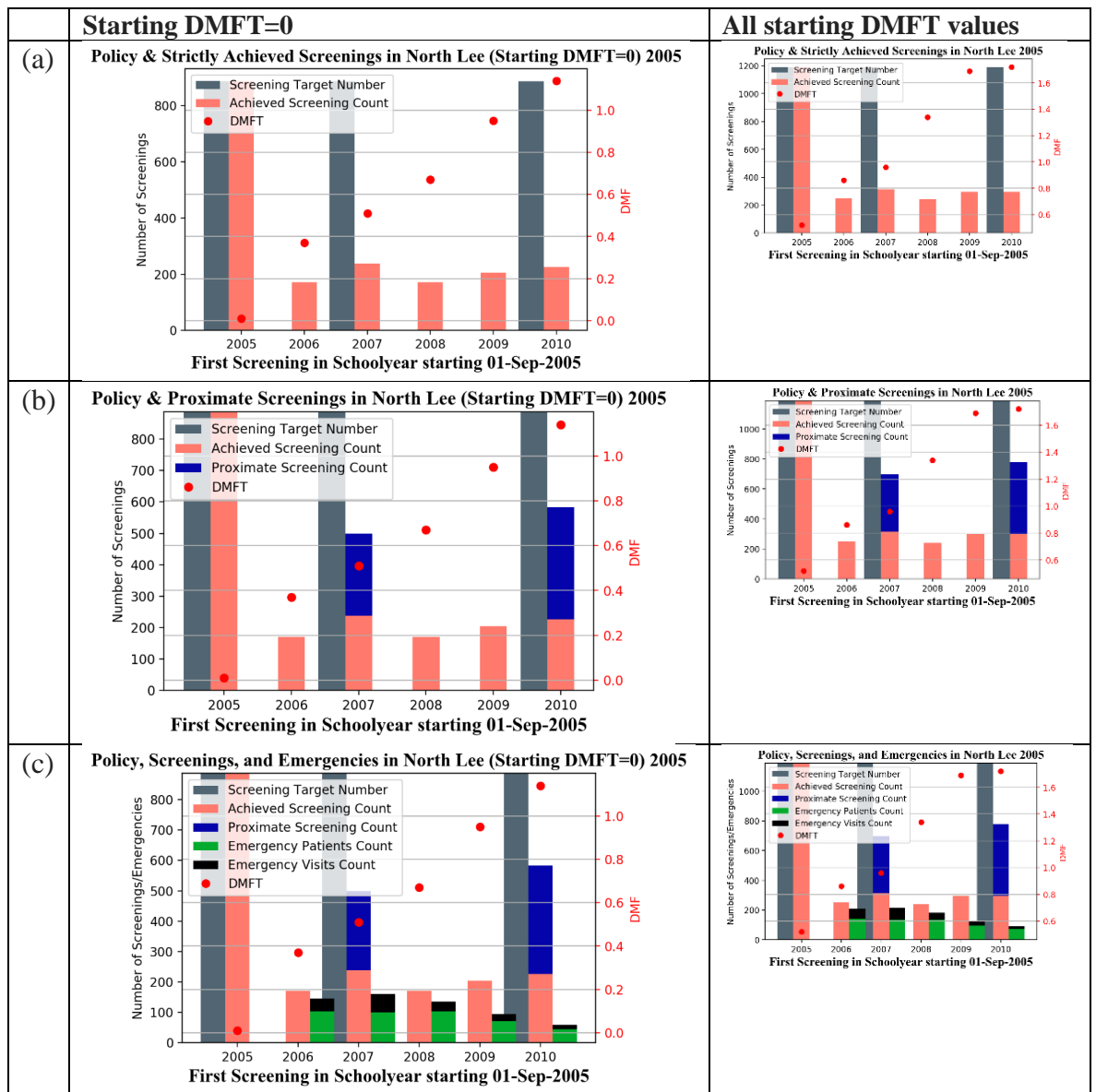


Figure 10-8: Policy & Screening Profile for Area (North Lee)

What are these charts telling us?

- Chart (a), Starting DMFT=0, shows that a cohort of 886 patients complying with the requirements could be identified with Screening (Initial Exam) carried out between September 1st of target year (2005) and 31st August of following year (2006). This was the first screening for those patients. The patients were aged 7,8 or 9 at the time of the screening. The data quality was acceptable. Starting DMFT was approaching 0.
- Chart (a), showing policy and strictly adhered to numbers, shows adherence of 238 patients in year 4, and 226 in year 6.
- However, Chart (b), where patients seen in adjacent years are stacked on top, the total seen in year 4 is 500 and year 6 is 583.
- Chart (c) shows that approximately 10% of the original cohort are seen for emergency appointments annually.
- DMFT (the red dots) increases with time between screenings.
- Charts for all starting DMFT values show similar trends.



Figure 10-9: Policy & Screening Profile for Area (South Lee)

What are these charts telling us?

- Chart (a), Starting DMFT=0, shows that a cohort of 1090 patients complying with the requirements could be identified with Screening (Initial Exam) carried out between September 1st of target year (2005) and 31st August of following year (2006). This was the first screening for those patients. The patients were aged 7,8 or 9 at the time of the screening. The data quality was acceptable. Starting DMFT was approaching 0.
- Chart (a), showing policy and strictly adhered to numbers, shows adherence of 448 patients in year 4, and 225 in year 6.
- However, Chart (b), where patients seen in adjacent years are stacked on top, the total seen in year 4 is 780 and year 6 is 736.
- Chart (c) shows that approximately 10% of the original cohort are seen for emergency appointments annually.
- DMFT (the red dots) increases with time between screenings.
- Charts for all starting DMFT values show similar trends.

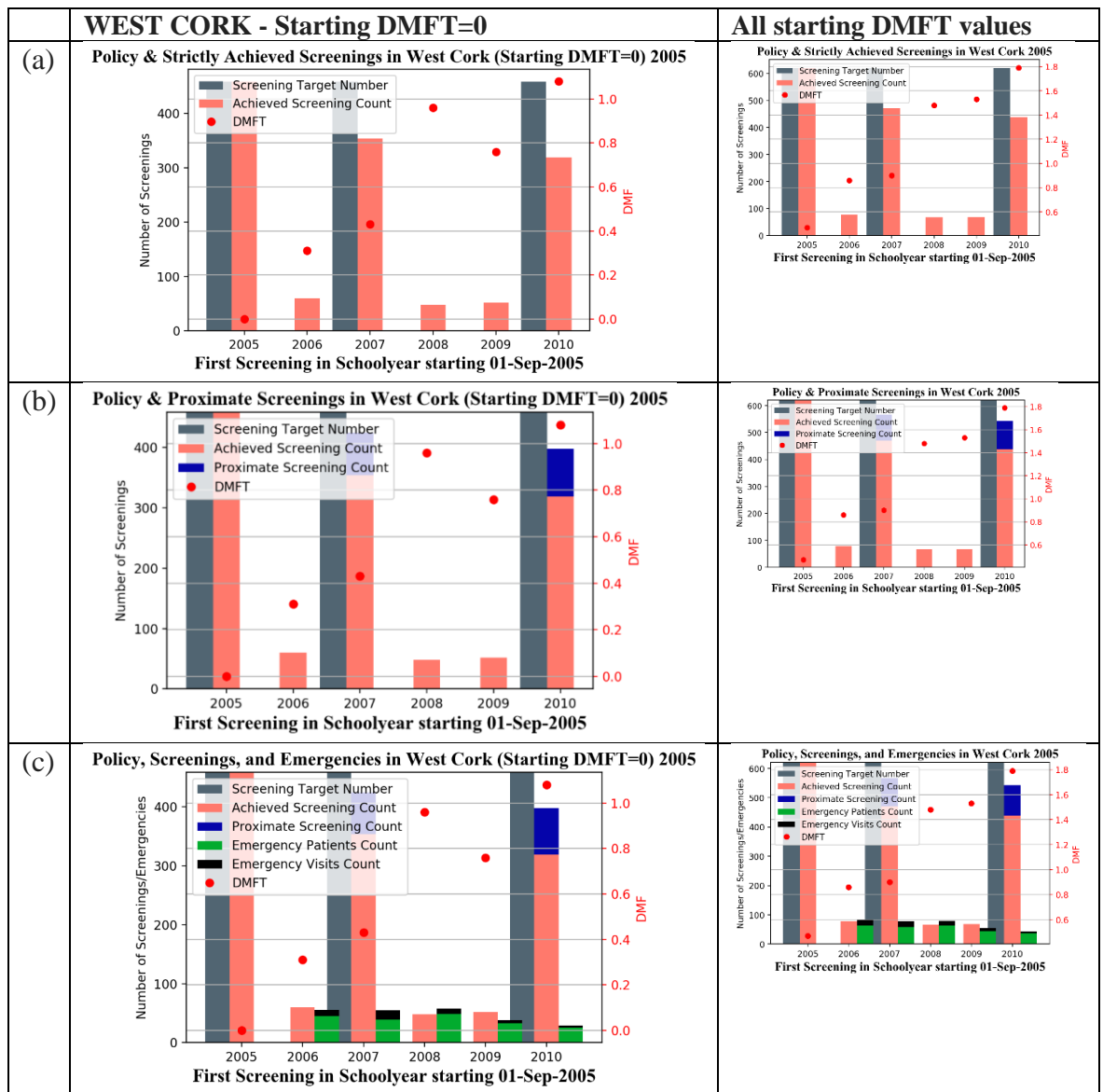


Figure 10-10: Policy & Screening Profile for Area (West Cork)

What are these charts telling us?

- Chart (a), Starting DMFT=0, shows that a cohort of 458 patients complying with the requirements could be identified with Screening (Initial Exam) carried out between September 1st of target year (2005) and 31st August of following year (2006). This was the first screening for those patients. The patients were aged 7,8 or 9 at the time of the screening. The data quality was acceptable. Starting DMFT was approaching 0.
- Chart (a), showing policy and strictly adhered to numbers, shows adherence of 354 patients in year 3, and 319 in year 6.
- However, Chart (b), where patients seen in adjacent years are stacked on top, the total seen in year 3 is 424 and year 6 is 398.
- Chart (c) shows that approximately 10% of the original cohort are seen for emergency appointments annually.
- DMFT (the red dots) increases with time between screenings.
- Charts for all starting DMFT values show similar trends.

Table 10-1: Base frequency of screening data with patients having initial DMFT=0

Area	Kerry	North Cork	North Lee	South Lee	West Cork
NoOfScreenings	2	4	3	3	3
StartDate	2005-09-01	2005-09-01	2005-09-01	2005-09-01	2005-09-01
EndDate	2006-08-31	2006-08-31	2006-08-31	2006-08-31	2006-08-31
Year1Targeted	581	494	886	1090	458
AllPolicyYearsScreened	174	2	63	43	221
Year2Policy	0	494	0	0	0
Year3Policy	0	0	886	1090	458
Year4Policy	0	494	0	0	0
Year5Policy	581	0	0	0	0
Year6Policy	0	494	886	1090	458
Year2Screened	37	70	171	276	60
Year3Screened	65	108	238	448	354
Year4Screened	92	145	171	216	48
Year5Screened	248	117	205	504	52
Year6Screened	101	143	226	225	319
Year2EmergencyPatients	44	58	103	81	45
Year3EmergencyPatients	63	58	100	105	39
Year4EmergencyPatients	50	75	104	82	49
Year5EmergencyPatients	35	32	71	70	33
Year6EmergencyPatients	30	31	44	51	25
Year2EmergencyVisits	57	71	146	104	56
Year3EmergencyVisits	74	81	161	126	55
Year4EmergencyVisits	56	100	136	101	58
Year5EmergencyVisits	37	43	95	88	38
Year6EmergencyVisits	42	34	59	59	29
Year2AdjacentScreened	0	0	0	0	0
Year3AdjacentScreened	0	0	262	332	70
Year4AdjacentScreened	0	88	0	0	0
Year5AdjacentScreened	167	0	0	0	0
Year6AdjacentScreened	0	93	357	511	79
Year1DMFT	0.01	0.01	0.01	0	0
Year2DMFT	0.41	0.29	0.37	0.24	0.31
Year3DMFT	1.08	0.5	0.51	0.36	0.43
Year4DMFT	0.72	0.69	0.67	0.58	0.96
Year5DMFT	1.6	1	0.95	0.82	0.76
Year6DMFT	2.82	1.53	1.14	0.88	1.08
AllPolicyYearsScreenedDMFT	1.6	2.33	1.12	1.05	1.04
Year1DMFTSTDEV	0.15	0.1	0.11	0.09	0
Year2DMFTSTDEV	0.83	0.74	0.79	0.71	0.61
Year3DMFTSTDEV	1.65	0.91	0.96	0.86	0.89
Year4DMFTSTDEV	1.63	1.2	1.17	1.05	1.43
Year5DMFTSTDEV	2.1	1.52	1.43	1.43	0.98
Year6DMFTSTDEV	3.6	2.04	1.62	1.35	1.79
AllPolicyYearsScreenedDMFTSTDEV	2.1	2.08	1.63	1.65	1.77

Table 10-2: Base frequency of screening having initial DMFT >0 values

Area	Kerry	North Cork	North Lee	South Lee	West Cork
NoOfScreenings	2	4	3	3	3
StartDate	2005-09-01	2005-09-01	2005-09-01	2005-09-01	2005-09-01
EndDate	2006-08-31	2006-08-31	2006-08-31	2006-08-31	2006-08-31
Year1Targeted	895	737	1188	1392	620
AllPolicyYearsScreened	405	8	118	103	396
Year2Policy	0	737	0	0	0
Year3Policy	0	0	1188	1392	620
Year4Policy	0	737	0	0	0
Year5Policy	895	0	0	0	0
Year6Policy	0	737	1188	1392	620
Year2Screened	58	110	250	345	78
Year3Screened	78	168	315	568	471
Year4Screened	141	214	241	273	67
Year5Screened	405	184	297	647	68
Year6Screened	138	229	297	277	439
Year2EmergencyPatients	68	96	141	106	63
Year3EmergencyPatients	93	86	137	131	59
Year4EmergencyPatients	84	120	137	108	64
Year5EmergencyPatients	54	56	95	92	45
Year6EmergencyPatients	46	47	72	68	36
Year2EmergencyVisits	95	121	211	133	83
Year3EmergencyVisits	113	122	213	153	77
Year4EmergencyVisits	96	163	181	134	79
Year5EmergencyVisits	59	71	126	111	55
Year6EmergencyVisits	64	53	94	78	43
Year2AdjacentScreened	0	0	0	0	0
Year3AdjacentScreened	0	0	382	420	95
Year4AdjacentScreened	0	134	0	0	0
Year5AdjacentScreened	234	0	0	0	0
Year6AdjacentScreened	0	144	482	653	103
Year1DMFT	0.73	0.72	0.52	0.39	0.47
Year2DMFT	1.2	1.14	0.86	0.58	0.86
Year3DMFT	1.54	1.31	0.96	0.76	0.9
Year4DMFT	1.82	1.71	1.34	1.01	1.48
Year5DMFT	2.8	2.18	1.69	1.28	1.53
Year6DMFT	3.36	2.97	1.72	3.36	1.79
AllPolicyYearsScreenedDMFT	2.8	3.8	1.63	1.26	1.72
Year1DMFTSTDEV	1.31	1.24	1.12	0.92	0.95
Year2DMFTSTDEV	1.4	1.58	1.23	1.16	1.33
Year3DMFTSTDEV	1.9	1.61	1.46	1.32	1.47
Year4DMFTSTDEV	2.57	2.6	1.89	1.52	1.78
Year5DMFTSTDEV	2.99	3.14	2.12	1.8	1.95
Year6DMFTSTDEV	3.87	3.42	1.99	3.87	2.51
AllPolicyYearsScreenedDMFTSTDEV	2.99	2.04	1.97	1.73	2.48

10.14 Age at first screening details

Cohorts		Case/Event Histogram
Age 6, 3 screenings, Starting DMFT=0		<p>Events per Case - Cohort Age 6, 3 Screenings, 2005 DMFT0</p>
No. of Cases	790	
Variants	788	
No. of Events	20,684	
Events per case	26.1	
Age 7, 3 screenings, Starting DMFT=0		<p>Events per Case - Cohort Age 7, 3 Screenings, 2005 DMFT0</p>
No. of Cases	2,081	
Variants	2,050	
No. of Events	47,735	
Events per case	22.9	
Age 8, 3 screenings, Starting DMFT=0		<p>Events per Case - Cohort Age 8, 3 Screenings, 2005 DMFT0</p>
No. of Cases	3,322	
Variants	3,247	
No. of Events	73,494	
Events per case	22.1	
Age 9, 3 screenings, Starting DMFT=0		<p>Events per Case - Cohort Age 9, 3 Screenings, 2005 DMFT0</p>
No. of Cases	1,671	
Variants	1,624	
No. of Events	35,285	
Events per case	21.1	

Figure 10-11: Event Log Characteristics (Age at 1st school screening)

As this experiment is primarily to see if the age at first screening had an impact on oral health outcomes or treatment processes at age 12/13, the pertinent data was extracted for cohorts having DMFT=0 and is shown in Table 10-3 and Table 10-4 with the data cohorts having starting DMFT>0 in Table 10-5 and Table 10-6.

The complete dataset was extracted for 2 groups of cohorts. The first group of cohorts had a starting DMFT=0 at the time of their first examination. The second group of cohorts had a starting DMFT>0 at the time of their first examination. Within each of these groups, four separate groups were identified, those receiving their first school screening at 6, 7, 8 or 9. Within each of these age-groups, two subgroups were identified: those receiving 2 screenings and those receiving 3 screenings. i.e. 16 cohorts in total.

The visualisation of the cohort with DMFT=0 at 1st screening is represented in Figure 10-12, from Table 10-3 and Table 10-4. The cohort with DMFT > 0 at 1st screening is represented in Figure 10-13, from Table 10-5 and Table 10-6.

Note: How to read the Age at Screening Profile data below.

Chart title indicates 3 things:

- Age when cohort received their first screening e.g. ‘First Screening at age 6...’
- Number of screenings for this cohort e.g. ‘... (2 Screenings) ...’
- Initial DMFT at first screening e.g. ‘...Starting DMFT=0’...’

Blue Bars indicate the numbers of patients from this cohort receiving their final screening at each of the other ages, i.e. of those patients in the cohort described in the chart title, how many received their 2nd screening at age 6,7,8,9, etc.

Red Dots indicate the patients’ DMFT at the time of the final screening.

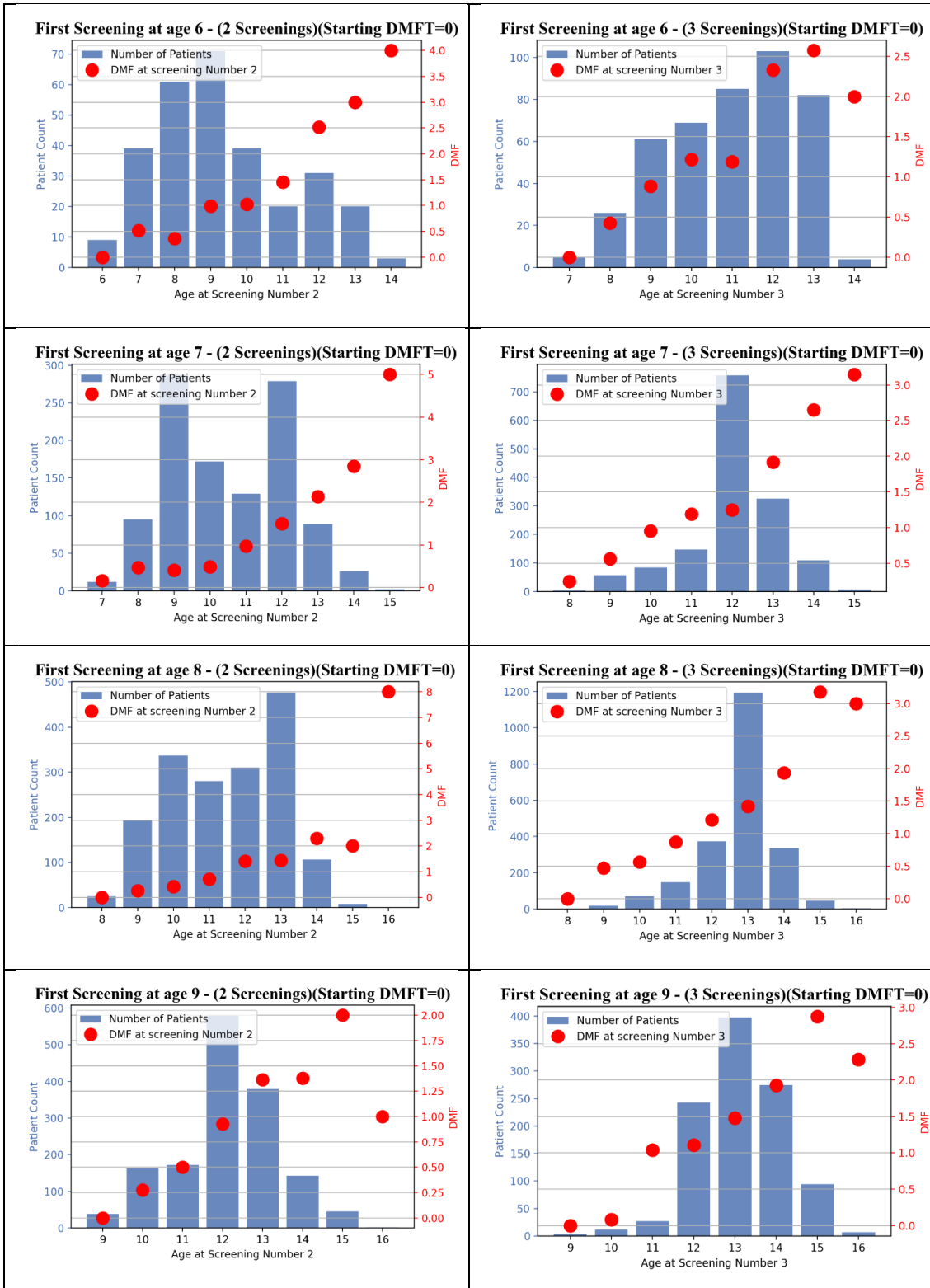


Figure 10-12: Results for 2 & 3 Screenings, Initial DMFT=0

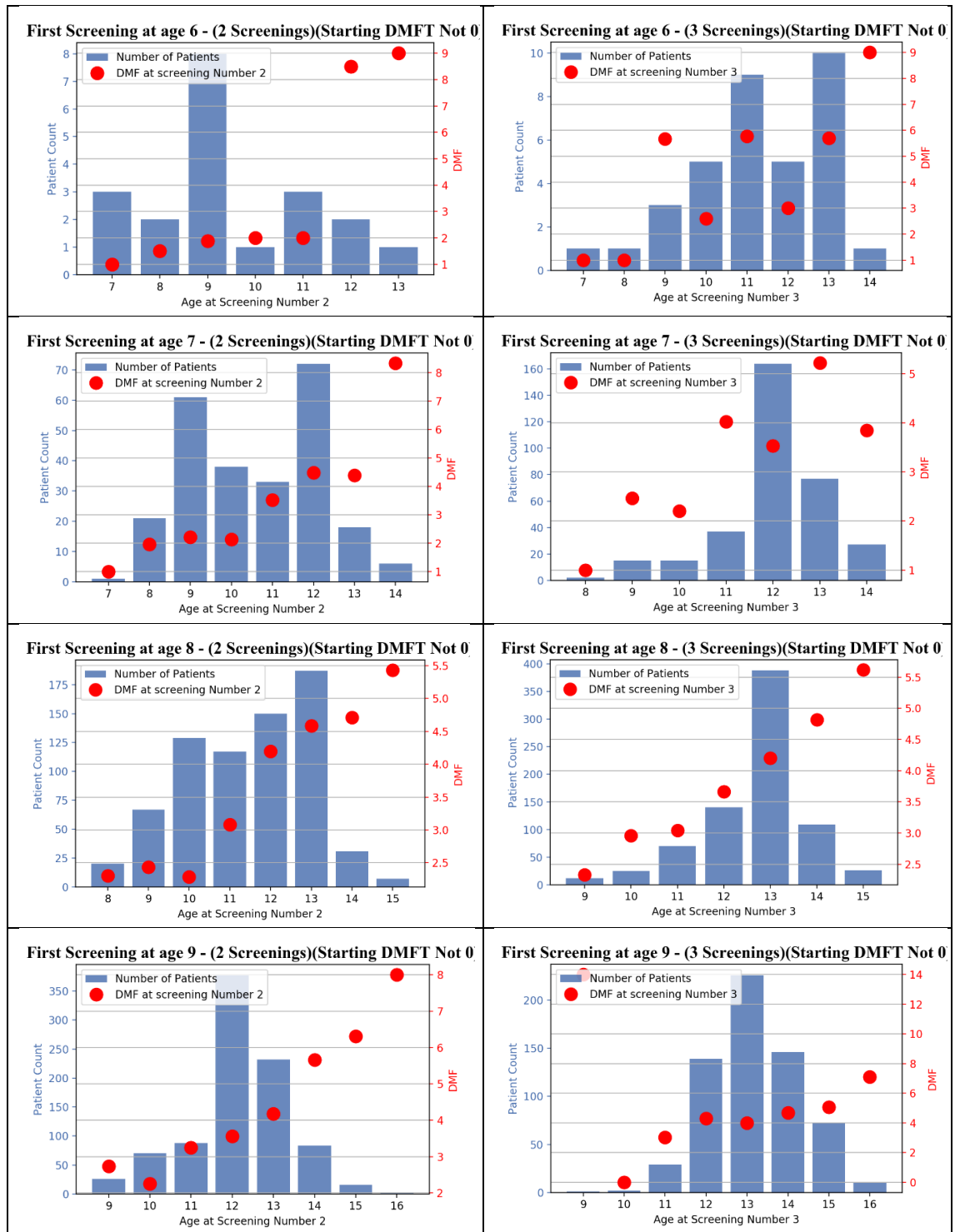


Figure 10-13: Results for 2 & 3 Screenings, Initial DMFT > 0

Table 10-3: Starting DMFT=0, Number of school screenings =2

First Screening At Age	Number Of Screenings	Screening Rank	Age At Rank Screening	Client Count	Average DMFT	STDEV DMFT
6	2	1	6	468	0.00	0.00
6	2	2	12	31	2.52	3.17
6	2	2	13	20	3.00	3.31
7	2	1	7	1362	0.00	0.00
7	2	2	12	279	1.53	2.04
7	2	2	13	89	2.13	2.33
8	2	1	8	2665	0.00	0.00

8	2	2	12	310	1.43	2.07
8	2	2	13	477	1.44	2.15
9	2	1	9	2106	0.00	0.00
9	2	2	12	579	0.94	1.52
9	2	2	13	379	1.38	1.92

Table 10-4: Starting DMFT=0, Number of school screenings =3

First Screening At Age	Number Of Screenings	Screening Rank	Age At Rank Screening	Client Count	Average DMFT	STDEV DMFT
6	3	1	6	790	0.00	0.00
6	3	3	12	103	2.33	2.36
6	3	3	13	82	2.67	2.40
7	3	1	7	2081	0.00	0.00
7	3	3	12	758	1.29	1.73
7	3	3	13	325	1.96	2.43
8	3	1	8	3323	0.00	0.00
8	3	3	12	374	1.25	1.73
8	3	3	13	1194	1.45	1.97
9	3	1	9	1671	0.00	0.00
9	3	3	12	243	1.13	1.87
9	3	3	13	398	1.49	1.96

Table 10-5: Starting DMFT>0, Number of school screenings =2

First Screening At Age	Number Of Screenings	Screening Rank	Age At Rank Screening	Client Count	Average DMFT	STDEV DMFT
6	2	1	6	33	1.33	0.54
6	2	2	12	2	8.50	4.95
6	2	2	13	1	9.00	NaN
7	2	1	7	297	1.82	1.01
7	2	2	12	72	4.49	2.69
7	2	2	13	18	4.39	3.31
8	2	1	8	983	2.04	1.17
8	2	2	12	150	4.19	2.64
8	2	2	13	187	4.58	2.98
9	2	1	9	1248	2.20	1.25
9	2	2	12	377	3.62	2.77
9	2	2	13	232	4.30	2.62

Table 10-6: Starting DMFT>0, Number of school screenings =3

First Screening At Age	Number Of Screenings	Screening Rank	Age At Rank Screening	Client Count	Average DMFT	STDEV DMFT
6	3	1	6	59	1.68	0.82
6	3	3	12	5	3.00	2.00
6	3	3	13	10	5.70	3.68
7	3	1	7	468	1.75	0.97
7	3	3	12	164	3.57	2.28
7	3	3	13	77	5.58	3.41
8	3	1	8	1117	1.94	1.23
8	3	3	12	140	3.69	2.83
8	3	3	13	388	4.27	3.05
9	3	1	9	810	2.12	1.21
9	3	3	12	139	4.32	3.17
9	3	3	13	226	4.10	2.59

10.15 Screening Base Data

Table 10-7: Base data with patients having initial DMFT=0

First Screening At Age	Number Of Screenings	Screening Rank	Age At Rank Screening	Client Count	Average DMFT	STDEV DMFT
6	2	1	6	468	0.00	0.00
6	2	2	6	9	0.00	0.00
6	2	2	7	39	0.51	1.12
6	2	2	8	61	0.37	0.69
6	2	2	9	71	0.99	1.36
6	2	2	10	39	1.08	1.16
6	2	2	11	20	1.45	1.61
6	2	2	12	31	2.52	3.17
6	2	2	13	20	3.00	3.31
6	2	2	14	3	4.00	3.61
6	3	1	6	790	0.00	0.00
6	3	2	6	21	0.00	0.00
6	3	2	7	105	0.20	0.61
6	3	2	8	291	0.54	1.06
6	3	2	9	328	0.85	1.25
6	3	2	10	101	0.92	1.29
6	3	2	11	65	1.59	2.00
6	3	2	12	12	1.83	2.12
6	3	2	13	2	2.00	0.00
6	3	3	7	5	0.00	0.00
6	3	3	8	26	0.44	0.96
6	3	3	9	61	0.89	1.14
6	3	3	10	69	1.22	1.44
6	3	3	11	85	1.22	1.66
6	3	3	12	103	2.33	2.36
6	3	3	13	82	2.67	2.40
6	3	3	14	4	2.00	1.83
7	2	1	7	1362	0.00	0.00
7	2	2	7	12	0.17	0.39
7	2	2	8	95	0.47	1.02
7	2	2	9	287	0.41	0.83
7	2	2	10	172	0.49	0.92
7	2	2	11	129	0.97	1.54
7	2	2	12	279	1.53	2.04
7	2	2	13	89	2.13	2.33
7	2	2	14	26	2.85	2.72
7	2	2	15	2	5.00	5.66
7	3	1	7	2081	0.00	0.00
7	3	2	7	39	0.13	0.47
7	3	2	8	364	0.28	0.80
7	3	2	9	1030	0.44	0.89
7	3	2	10	557	0.50	1.00
7	3	2	11	215	1.03	1.72
7	3	2	12	89	1.38	1.73
7	3	2	13	17	2.69	2.09
7	3	2	14	1	6.00	NaN
7	3	3	8	4	0.25	0.50
7	3	3	9	57	0.57	0.89
7	3	3	10	84	0.96	1.37
7	3	3	11	147	1.21	1.60

First Screening At Age	Number Of Screenings	Screening Rank	Age At Rank Screening	Client Count	Average DMFT	STDEV DMFT
7	3	3	12	758	1.29	1.73
7	3	3	13	325	1.96	2.43
7	3	3	14	109	2.65	2.62
7	3	3	15	7	3.14	1.86
8	2	1	8	2665	0.00	0.00
8	2	2	8	25	0.00	0.00
8	2	2	9	192	0.25	0.65
8	2	2	10	337	0.43	0.87
8	2	2	11	280	0.72	1.28
8	2	2	12	310	1.43	2.07
8	2	2	13	477	1.44	2.15
8	2	2	14	106	2.34	2.66
8	2	2	15	8	2.29	1.50
8	2	2	16	1	8.00	NaN
8	3	1	8	3323	0.00	0.00
8	3	2	8	111	0.15	0.57
8	3	2	9	790	0.28	0.67
8	3	2	10	1416	0.43	0.90
8	3	2	11	640	0.67	1.17
8	3	2	12	174	1.22	1.81
8	3	2	13	90	1.79	1.91
8	3	2	14	4	1.00	1.41
8	3	3	8	1	0.00	NaN
8	3	3	9	19	0.47	1.31
8	3	3	10	71	0.57	0.89
8	3	3	11	148	0.91	1.55
8	3	3	12	374	1.25	1.73
8	3	3	13	1194	1.45	1.97
8	3	3	14	336	1.97	2.51
8	3	3	15	46	3.56	2.92
8	3	3	16	4	3.00	1.63
9	2	1	9	2106	0.00	0.00
9	2	2	9	38	0.00	0.00
9	2	2	10	163	0.28	0.66
9	2	2	11	172	0.51	0.94
9	2	2	12	579	0.94	1.52
9	2	2	13	379	1.38	1.92
9	2	2	14	143	1.41	1.74
9	2	2	15	45	2.09	2.00
9	2	2	16	2	1.00	1.41
9	3	1	9	1671	0.00	0.00
9	3	2	9	109	0.21	0.59
9	3	2	10	554	0.25	0.69
9	3	2	11	445	0.46	1.10
9	3	2	12	284	1.03	1.71
9	3	2	13	125	1.54	2.39
9	3	2	14	18	1.61	1.79
9	3	3	9	4	0.00	0.00
9	3	3	10	12	0.08	0.29
9	3	3	11	27	1.08	1.72
9	3	3	12	243	1.13	1.87
9	3	3	13	398	1.49	1.96
9	3	3	14	275	1.97	2.57

Table 10-8: Base data with patients having initial DMFT > 0

First Screening At Age	Number Of Screenings	Screening Rank	Age At Rank Screening	Client Count	Avg DMFT	ST DEV	First Screening At Age	Number Of Screenings	Screening Rank	Age At Rank Screening	Client Count	Avg DMFT	ST DEV
	2	1	6	33	1.33	0.54	7	3	2	11	49	2.92	2.05
6	2	2	7	3	1.00	1.00	7	3	2	12	17	4.53	3.64
6	2	2	8	2	1.50	0.71	7	3	2	13	4	4.75	1.71
6	2	2	9	8	1.88	1.55	7	3	3	8	2	1.00	0.00
6	2	2	10	1	2.00	NaN	7	3	3	9	15	2.47	1.25
6	2	2	11	3	2.00	2.00	7	3	3	10	15	2.75	1.54
6	2	2	12	2	8.50	4.95	7	3	3	11	37	4.14	3.78
6	2	2	13	1	9.00	NaN	7	3	3	12	164	3.57	2.28
6	3	1	6	59	1.68	0.82	7	3	3	13	77	5.58	3.41
6	3	2	7	7	4.43	6.55	7	3	3	14	27	4.00	3.17
6	3	2	8	18	2.50	1.62	8	2	1	8	983	2.04	1.17
6	3	2	9	17	2.71	2.39	8	2	2	8	20	2.30	1.30
6	3	2	10	4	3.00	0.82	8	2	2	9	67	2.43	1.64
6	3	2	11	5	4.60	4.93	8	2	2	10	129	2.29	1.52
6	3	2	12	1	3.00	NaN	8	2	2	11	117	3.10	2.43
6	3	2	13	1	9.00	NaN	8	2	2	12	150	4.19	2.64
6	3	3	7	1	1.00	NaN	8	2	2	13	187	4.58	2.98
6	3	3	8	1	1.00	NaN	8	2	2	14	31	4.71	2.69
6	3	3	9	3	5.67	2.08	8	2	2	15	7	5.43	5.22
6	3	3	10	5	2.60	1.82	8	3	1	8	1117	1.94	1.23
6	3	3	11	9	5.78	3.80	8	3	2	8	29	2.03	1.09
6	3	3	12	5	3.00	2.00	8	3	2	9	280	2.08	1.45
6	3	3	13	10	5.70	3.68	8	3	2	10	428	2.45	1.73
6	3	3	14	1	9.00	NaN	8	3	2	11	245	3.25	2.14
7	2	1	7	297	1.82	1.01	8	3	2	12	95	3.96	2.60
7	2	2	7	1	1.00	NaN	8	3	2	13	45	5.47	3.41
7	2	2	8	21	1.95	1.20	8	3	2	14	9	6.89	4.76
7	2	2	9	61	2.21	1.51	8	3	3	9	12	2.33	1.50
7	2	2	10	38	2.19	1.49	8	3	3	10	25	2.96	1.24
7	2	2	11	33	3.63	2.49	8	3	3	11	70	3.13	2.15
7	2	2	12	72	4.49	2.69	8	3	3	12	140	3.69	2.83
7	2	2	13	18	4.39	3.31	8	3	3	13	388	4.27	3.05
7	2	2	14	6	8.33	5.79	8	3	3	14	109	5.05	3.30
7	3	1	7	468	1.75	0.97	8	3	3	15	26	5.84	3.70
7	3	2	7	8	2.00	1.51	9	2	1	9	1248	2.20	1.25
7	3	2	8	70	2.41	1.44	9	2	2	9	26	2.73	1.46
7	3	2	9	210	2.09	1.45	9	2	2	10	70	2.29	1.54
7	3	2	10	142	2.70	1.96	9	2	2	11	88	3.45	2.51

9	2	2	12	377	3.62	2.77
9	2	2	13	232	4.30	2.62
9	2	2	14	83	5.66	3.47
9	2	2	15	16	6.73	4.93
9	2	2	16	2	8.00	2.83
9	3	1	9	810	2.12	1.21
9	3	2	9	48	2.42	1.67
9	3	2	10	227	2.41	1.77
9	3	2	11	225	2.74	1.84
9	3	2	12	207	3.89	2.67
9	3	2	13	77	4.16	2.55
9	3	2	14	11	5.60	2.76

9	3	2	15	1	8.00	NaN
9	3	3	9	1	14.00	NaN
9	3	3	10	2	0.00	0.00
9	3	3	11	29	3.03	1.66
9	3	3	12	139	4.32	3.17
9	3	3	13	226	4.10	2.59
9	3	3	14	146	4.84	3.22
9	3	3	15	72	5.69	3.85
9	3	3	16	11	7.80	5.33

10.16 Age at First Screening Process Mining Output

The default output from Disco© for the 6-year-old cohort is shown in Figure 10-14 and shows the 466 patients getting their first screening (Initial Exam) at age 6. The model gives an initial overview of the treatment process and the most common paths. Of the initial 446 presenting for screening, the model shows 335 of those proceeding directly to 'Prevention' and 409 were marked as 'Completed Case'. The darker coloured boxes (events) indicate higher frequency of execution of these procedures and the heavier arrows indicate the most travelled pathways. The larger font number within the box indicates the number of patients (cases) receiving the treatment and the smaller font number within the box indicating the number of times the treatment was executed reflecting that a patient may receive a treatment on multiple occasions.

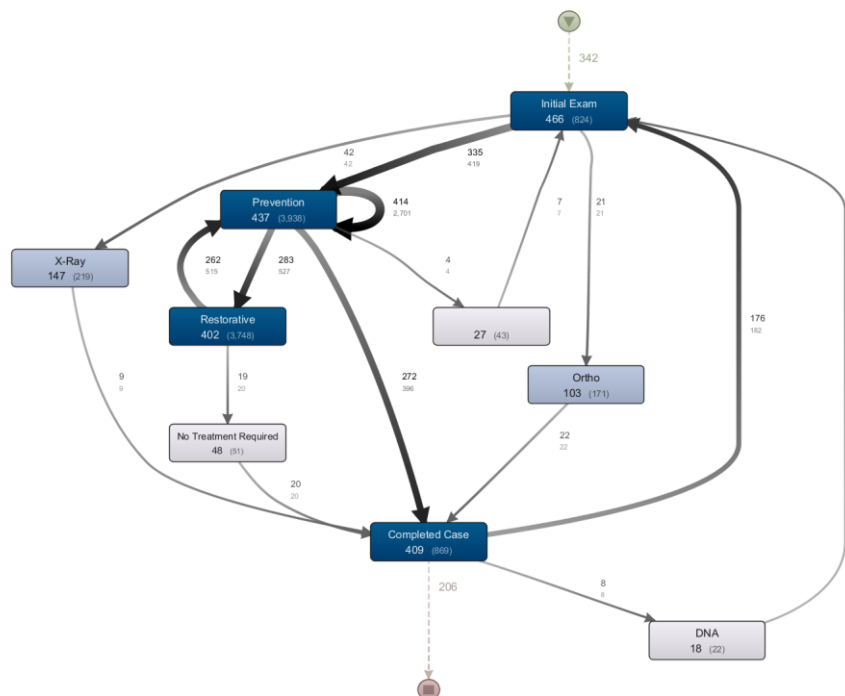


Figure 10-14: Default output from Disco for 6-year olds. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008

The process model with an emphasis on performance offers an enhanced view. In particular, this variation shows the mean time between the events. Again, BridgesPM1 has no record of the time required to complete an individual step e.g. a screening, hence the value 'instant' is recorded in the event box. This performance model is shown in Figure 10-15 below.

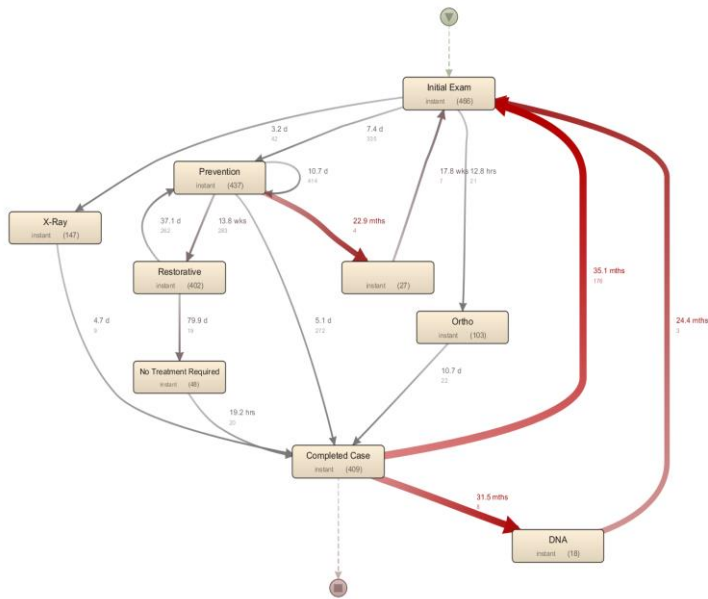


Figure 10-15: Performance output from Disco for 6-year olds. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008

The default settings for this PM tool aims to present the main features of the dataset: the most frequent activities and the most frequent paths. It omits less frequent events and pathways and these simplifications can be misleading. These process models require careful examination to ensure that they represent the data and the real-world process correctly.

Disco© allows adjustment of the amount of detail presented in the models. Adjusting to show 100% of the paths and activities results in the process model presented in Figure 10-16 below. It can be clearly seen how the complexity of the process model increases and its readability and comprehensibility are reduced accordingly. It is also important to note that most of the possible dental treatments have already been simplified to 'Preventive' and 'Restoration'.

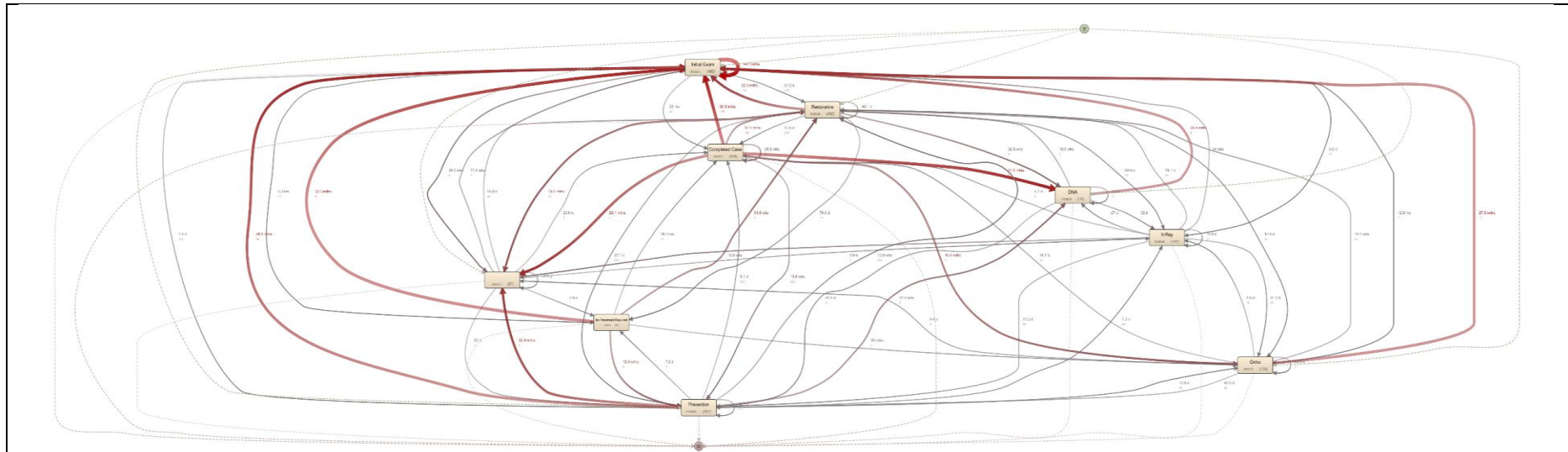
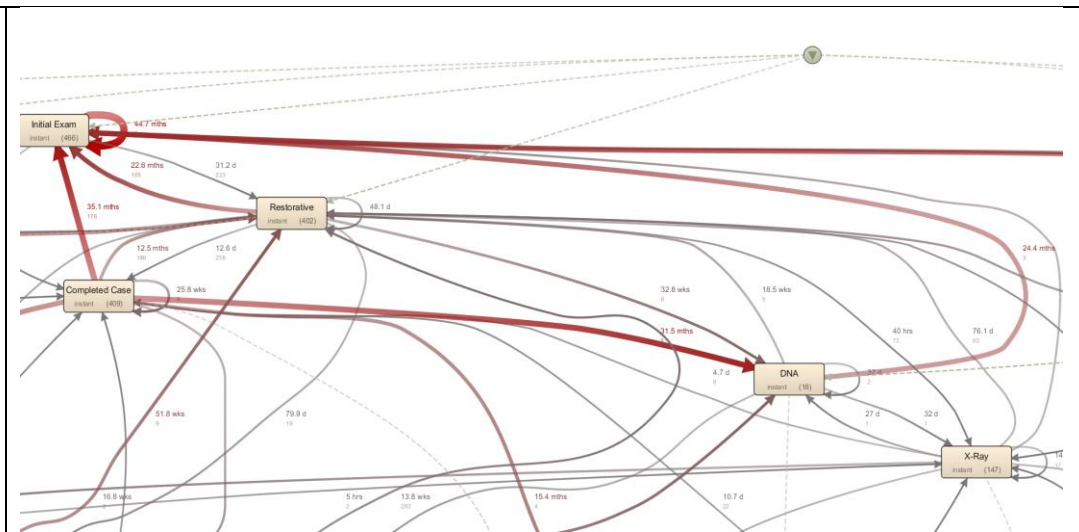


Figure 10-16: Disco output showing 100% detail, excerpt below right.

Better readability and comprehension of the models can be achieved with pan and zoom functionality. On the right the performance detail available in models containing all of the data can be seen. Knowing that the model is showing all of the available paths and activities also engenders a higher degree of trust in the results.



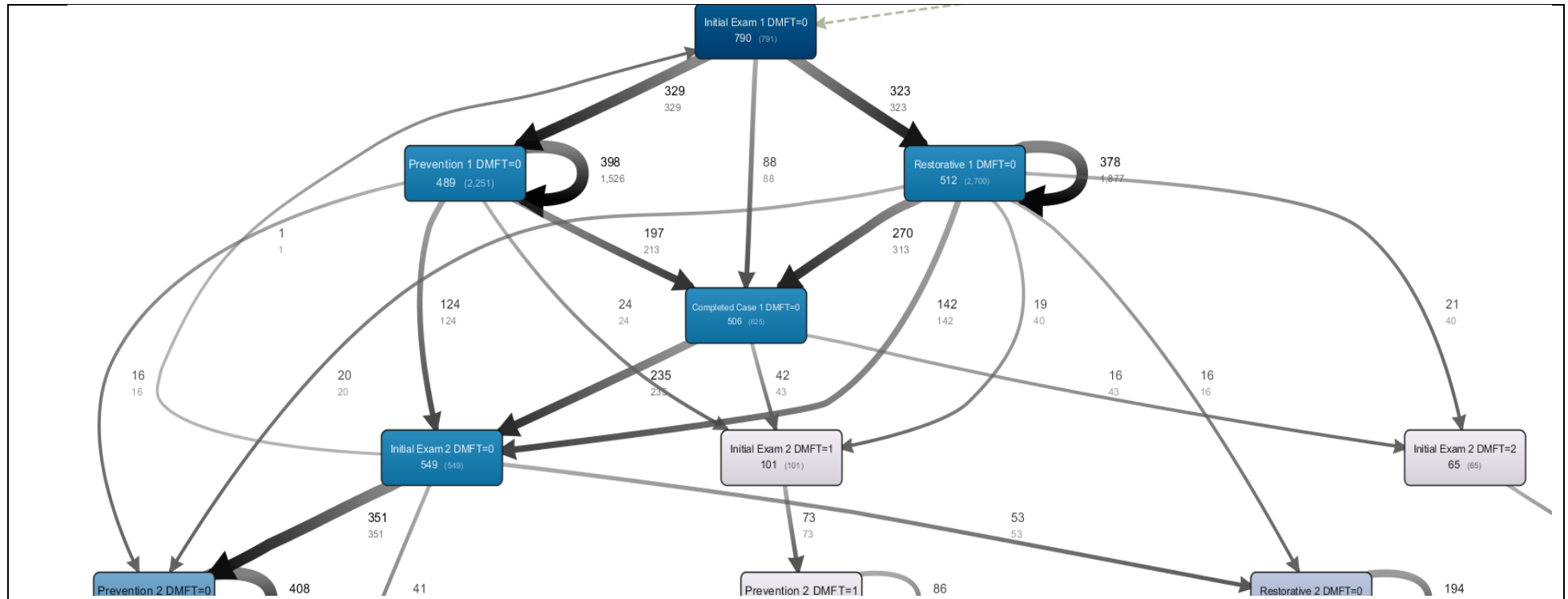


Figure 10-17: Process model detail for age at first screening = 6. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008.

Age	Starting Year	Starting DMFT	Number of Screenings	Number of Rows (Events)	Number of Cases	Variants	Number of unique Events	Activities %	Paths %	Frequency & Performance	Legibility on A4 Print. Landscape/Portrait	Appendix No.
6	2005	0	3	20,684	790	788	74	49.1	9.8	Y	With Difficulty (L)	I
Age	Prevention after 1 st Screening	Restorative after 1 st Screening	DMFT 0 after 2 nd Screening	DMFT 1 after 2 nd Screening	DMFT 2 after 2 nd Screening	DMFT 3 after 2 nd Screening	DMFT 4 after 2 nd Screening	DMFT 5 after 2 nd Screening	DMFT > 5 after 2 nd Screening			
6	329	323	549	101	65	NV/39	NV/23	NV/5	NV/6			

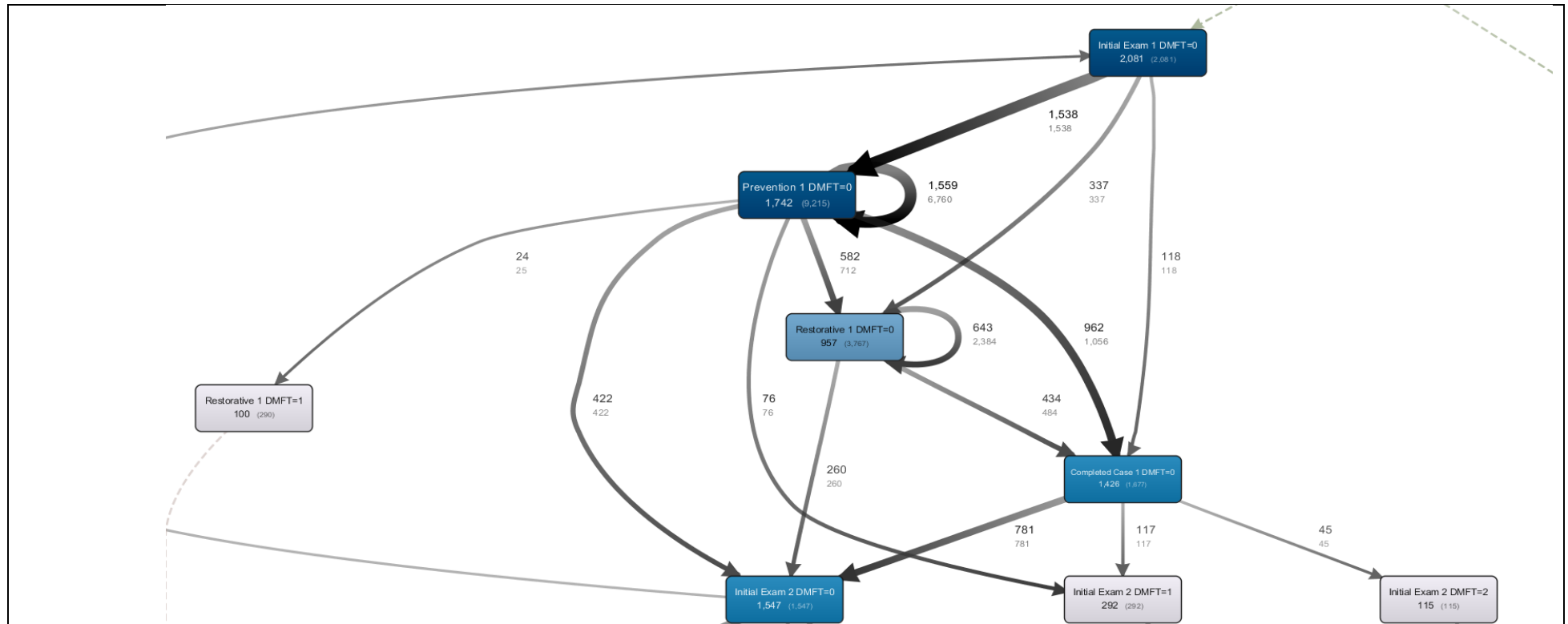


Figure 10-18: Process model detail for age at first screening = 7. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008.

Age	Starting Year	Starting DMFT	Number of Screenings	Number of Rows (Events)	Number of Cases	Variants	Number of unique Events	Activities %	Paths %	Frequency & Performance	Legibility on A4 Print. Landscape/Portrait	Appendix No.
7	2005	0	3	47,735	2081	2050	76	49.1	4.6	Y	Not Legible	J
Age	Prevention after 1 st Screening	Restorative after 1 st Screening	DMFT 0 after 2 nd Screening	DMFT 1 after 2 nd Screening	DMFT 2 after 2 nd Screening	DMFT 3 after 2 nd Screening	DMFT 4 after 2 nd Screening	DMFT 5 after 2 nd Screening	DMFT > 5 after 2 nd Screening			
7	1538	337	1547	292	115	NV/68	NV/32	NV	NV/9			

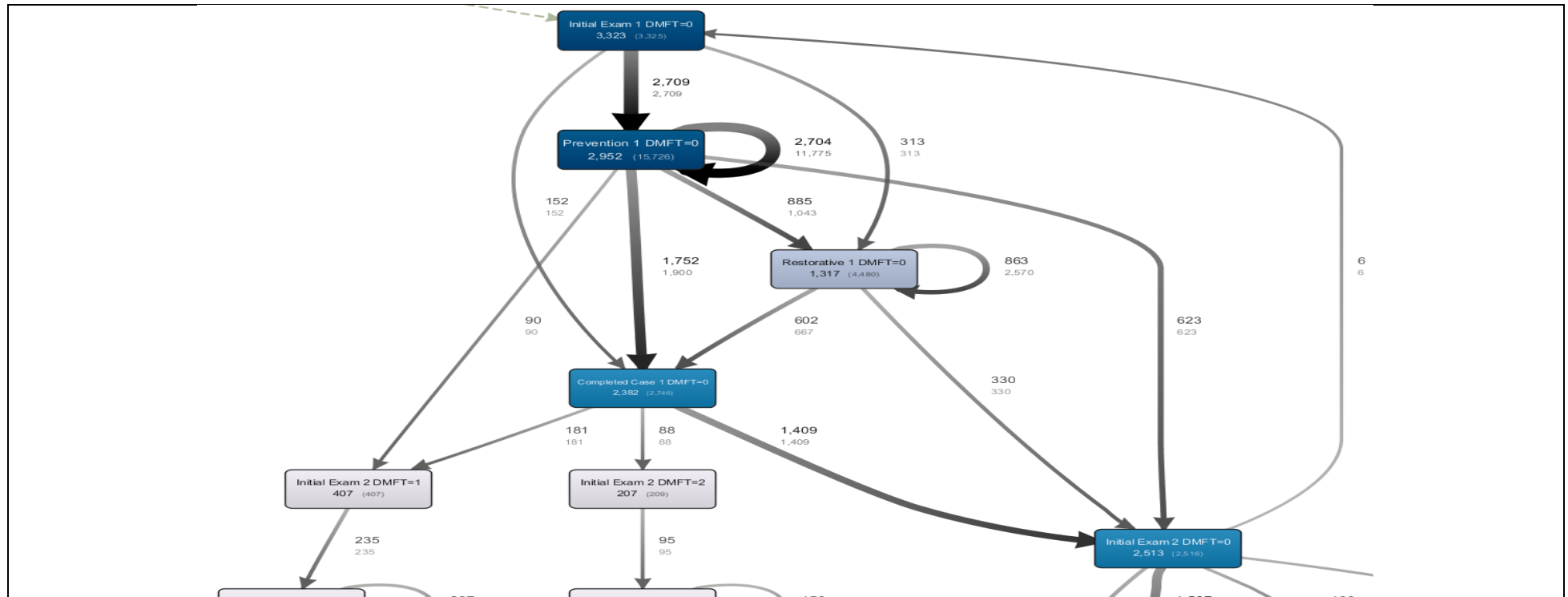


Figure 10-19: Process model detail for age at first screening = 8. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008

Age	Starting Year	Starting DMFT	Number of Screenings	Number of Rows (Events)	Number of Cases	Variants	Number of unique Events	Activities %	Paths %	Frequency & Performance	Legibility on A4 Print. Landscape/Portrait	Appendix No.
8	2005	0	3	73,494	3322	3247	77	49.1	4.6	Y	With Difficulty (P)	K
Age	Prevention after 1 st Screening	Restorative after 1 st Screening	DMFT 0 after 2 nd Screening	DMFT 1 after 2 nd Screening	DMFT 2 after 2 nd Screening	DMFT 3 after 2 nd Screening	DMFT 4 after 2 nd Screening	DMFT 5 after 2 nd Screening	DMFT > 5 after 2 nd Screening			
8	2952	1317	2513	407	207	NV/94	NV/44	NV/9	NV/16			

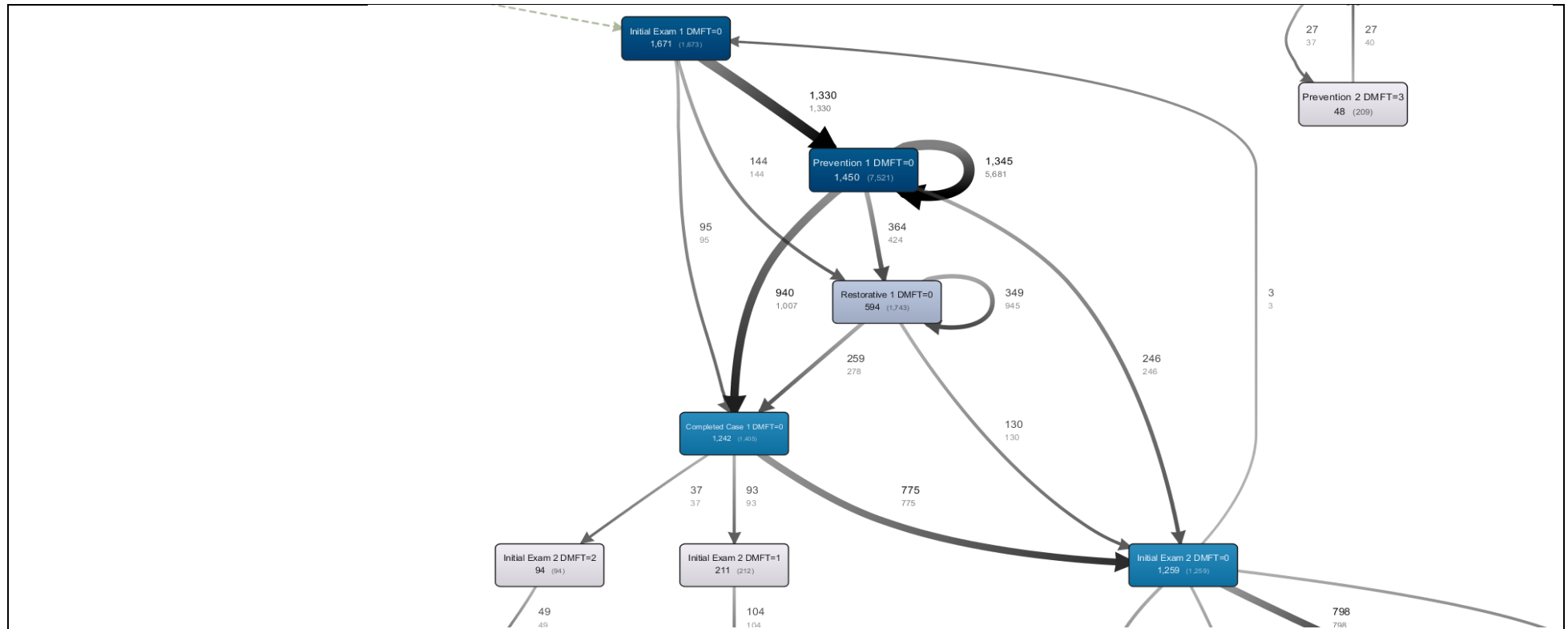


Figure 10-20: Process model detail for age at first screening = 9. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008.

Age	Starting Year	Starting DMFT	Number of Screenings	Number of Rows (Events)	Number of Cases	Variants	Number of unique Events	Activities %	Paths %	Frequency & Performance	Legibility on A4 Print. Landscape/Portrait	Appendix No.
9	2005	0	3	35,285	1671	1624	74	49.1	4.6	Y	With Difficulty (P)	L
Age	Prevention after 1 st Screening	Restorative after 1 st Screening	DMFT 0 after 2 nd Screening	DMFT 1 after 2 nd Screening	DMFT 2 after 2 nd Screening	DMFT 3 after 2 nd Screening	DMFT 4 after 2 nd Screening	DMFT 5 after 2 nd Screening	DMFT > 5 after 2 nd Screening			
9	1330	144	1259	211	94	NV/42	NV/25	NV/8	NV/15			

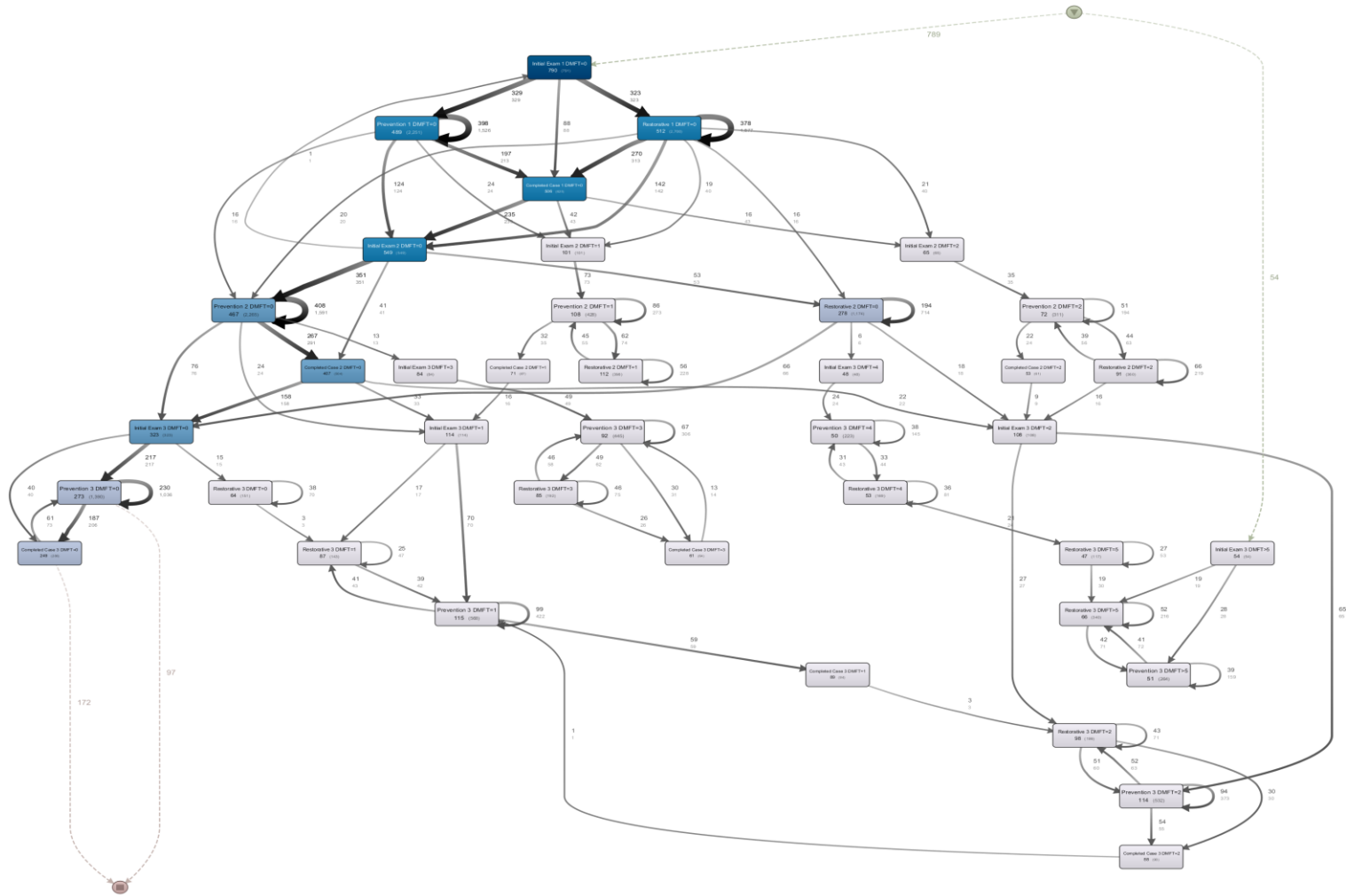


Figure 10-21: First Screening Age 6 – Frequency. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008

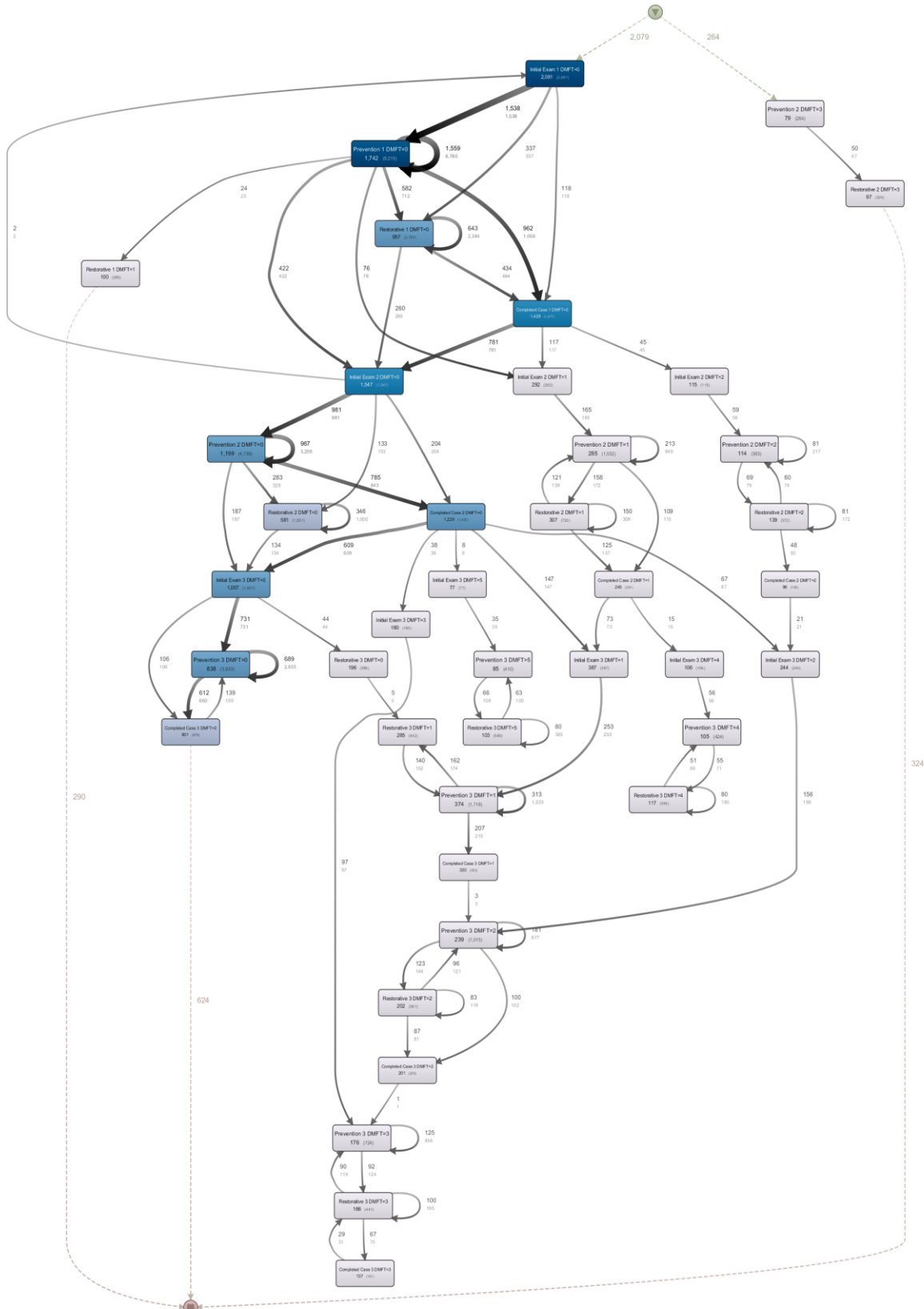


Figure 10-23: First Screening Age 7 – Frequency. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008.

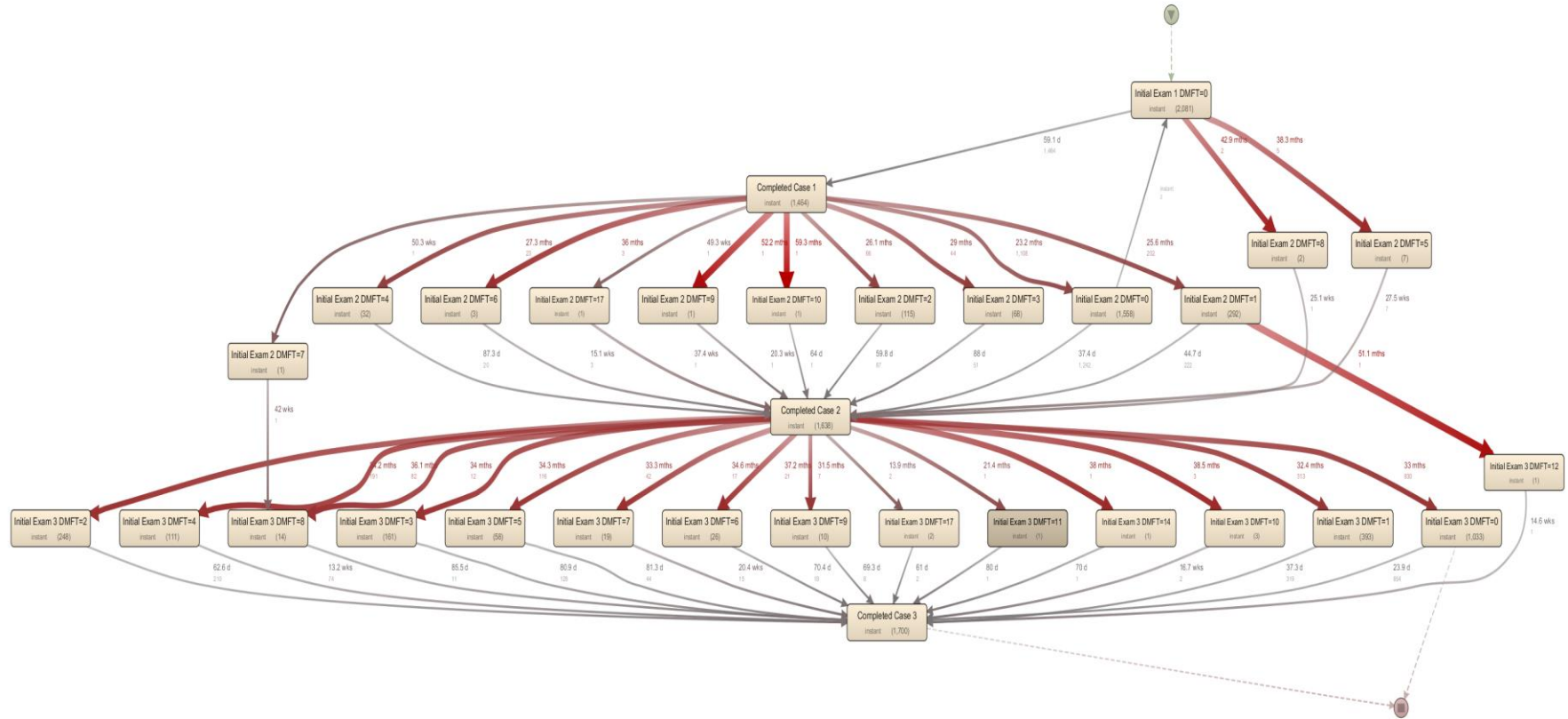


Figure 10-24: First Screening at age 7 – Performance. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008.

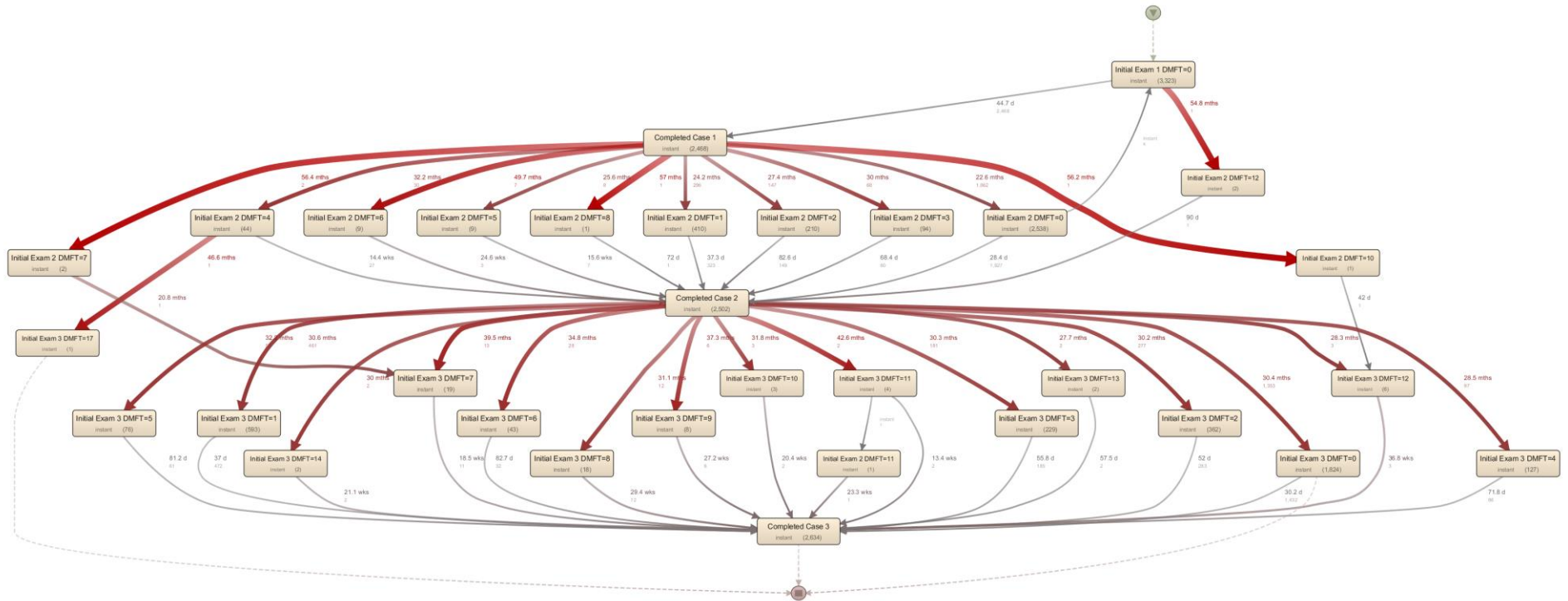


Figure 10-26: First Screening at age 8 – Performance. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008.

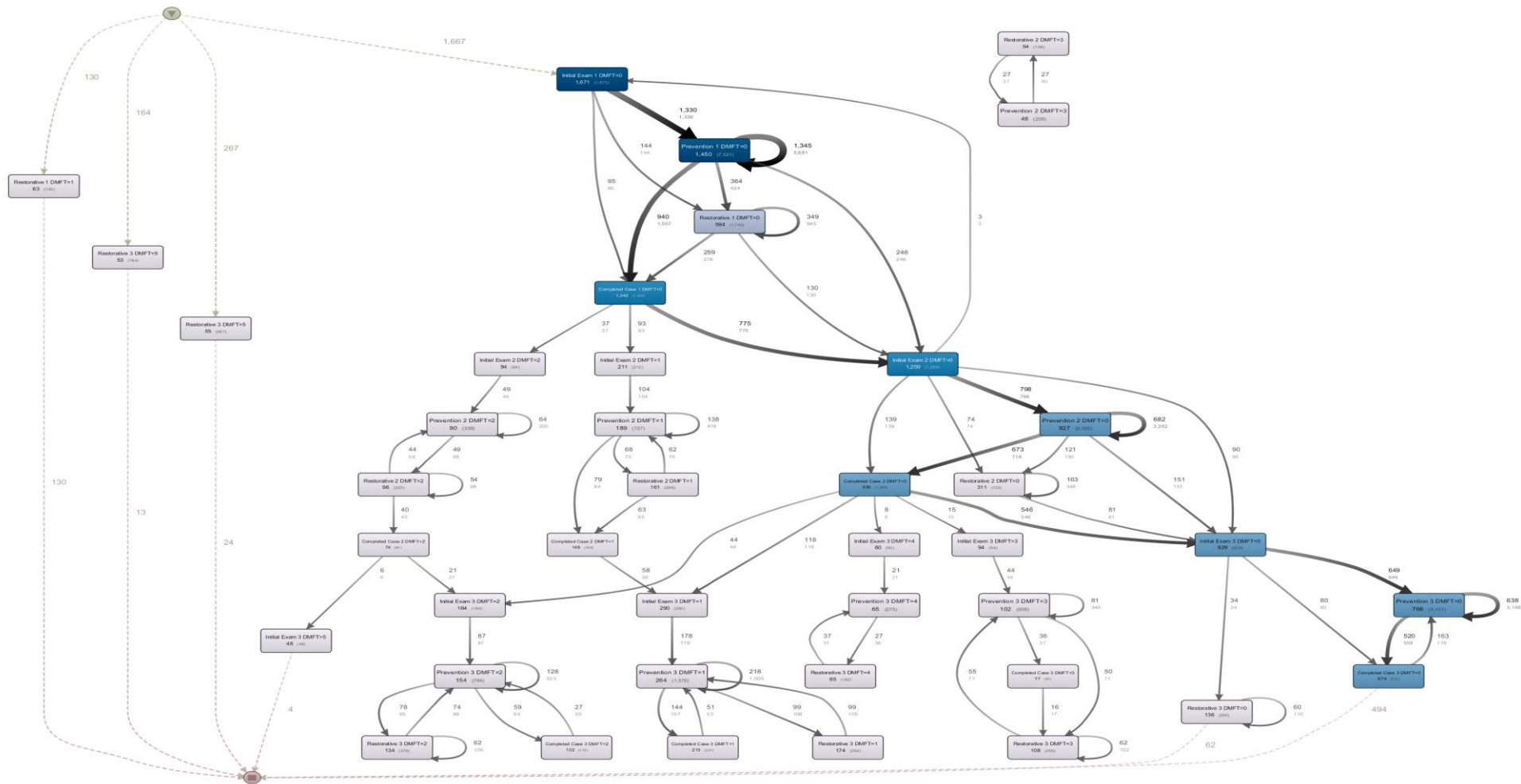


Figure 10-27: First Screening at age 9 – Frequency. Temporal sequence for patients receiving first screening between January 1st, 2004 and December 31st, 2008.

10.17 Data Quality Issues

DataIssueName	ShortName	Level	Source	Dimension	NoOfRows	% Defect
All entries in PMAppointments must have a corresponding Client in PMClients	No such client exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	240449	13.65
All entries in PMAttendances must have a corresponding Client in PMClients	No such client exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	679728	12.32
All Clients in the PMClient must have a ClientAge between 0 and 100	Invalid Age	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	22444	9.68
All Treatments in the PMTreatments must have a CompletionDate > =1990-01-01 00:19:02.000	Invalid Treatment Completion Date	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	197352	6.23
All entries in the following tables must have a corresponding Appointment in PMAppointments: PMAttendances	No such appointment exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	292167	5.30
All entries in PMTreatments must have a corresponding Client in PMClients	No such client exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	48330	1.52
All treatments in the PMTreatments should have a ClinicID in the PMClinics table	Invalid Treatment Clinic	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	30940	0.98
All Treatments in the PMTreatments must have a MappedToProcedureNameGroup in the PMProcedureCountGreaterThan100 table. The purpose is to reduce noise from rarely occurring procedures (<100 times)	No MappedToProcedureNameGroup	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	18316	0.58
All entries in PMQuestionnaire must have a corresponding Client in PMClients	No such client exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	1813	0.55
All entries in PMQuestionAnswers must have a corresponding Client in PMClients	No such client exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	50806	0.52
All entries in PMChart must have a corresponding Client in PMClients	No such client exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	3267	0.32
All treatments in the PMTreatments should have been carried out when the patient was aged 0-100	Invalid Age	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	805	0.03
All charts in PMCharts should have a DMFSCild between 0 and 60	Invalid dmfs	Field Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	98	0.01

All charts in PMCharts should have a DMFSAdult between 0 and 96	Invalid DMFS	Field Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	22	0.00
All charts in PMCharts should have a CreationDate >1995	Invalid Chart CreationDate	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	3	0.00
All entries in PMTooth must have a corresponding Client in PMCharts	No such chart exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	60	0.00
All entries in PMToothPart must have a corresponding Client in PMCharts	No such chart exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	16	0.00
All entries in PMCondition must have a corresponding Client in PMCharts	No such chart exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	1	0.00
All entries in PMQuestionAnswers must have a corresponding Client in PMQuestionnaire	No such questionnaire exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	0	0.00
All entries in PMQuestionAnswers must have a corresponding Client in PMQuestionnaire	No such questionnaire exists	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	0	0.00
All Clients in the PMClient must have a ClinicID in the PMClinics table	Invalid Clinic	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	0	0.00
All treatments in the PMTreatments should have a list Position >0	Invalid List Position	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	0	0.00
All treatments in the PMTreatments should have a Quantity >0	Invalid Quantity	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	0	0.00
All charts in PMCharts should have a DMFTAdult between 0 and 32	Invalid DMFT	Field Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	0	0.00
All charts in PMCharts should have a DMFTChild value between 0 and 20	Invalid dmft	Field Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	0	0.00
All treatments in the PMTreatments should have an AssociatedChartID in the PMCharts table	Invalid ChartID	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	0	0.00
All Treatments in the PMTreatments must have a TreatmentCourseID in the PMTreatmentCourses table)	Invalid TreatmentCourseID	Row Level Data Issue	Software Developers & DBA (Bridges)	Incorrect (Mans et al)	0	0.00
Fluoridation Status not consistently recorded	Imprecise Fluoridation Status	Dataset Level Data Issue	Previous research work using this or similar data	Imprecise (Mans et al)	0	0.00
Trauma Status not consistently recorded	Imprecise Trauma Status	Dataset Level Data Issue	Previous research work using this or similar data	Imprecise (Mans et al)	0	0.00

Gender not consistently recorded	Imprecise Gender Status	Dataset Level Data Issue	Previous research work using this or similar data	Imprecise (Mans et al)	0	0.00
'Initial Exam' not consistently recorded	Imprecise 'Initial Exam' recording	Dataset Level Data Issue	Previous research work using this or similar data	Imprecise (Mans et al)	0	0.00
Process Mining Missing Cases (1) – No way to verify that all eligible children were screened	Missing Cases – Process Mining(1)	Dataset Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining Missing Events (2) – Cannot verify that all treatment items (events) were recorded. Dental Professionals are incentivised to record all steps as their activity levels are based on this.	Missing Events – Process Mining(2)	Dataset Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining Missing Case Relationship (3) – Case cannot exist without a valid client. This referential integrity is enforced by Rule 1	Missing case relationships – Process Mining(3)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining Missing Case Attribute (4) – The sole case attribute currently used is 'DOB', Logical DOB checks for age at treatment Issue 21	Missing case Attribute – Process Mining(4)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining Missing Position (5). This is only relevant if we have no timestamps. In this dataset it is enforced at the UI.	Missing Position – Process Mining (5)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining Missing Activity Names (6). In this dataset it is enforced at the UI and at the Process Name Mapping Stage	Missing Event Names – Process Mining (6)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining Missing TimeStamp. In this dataset it is enforced at the time of creation of the entity	Missing TimeStamp – Process Mining(7)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining – Missing Resource information. In this dataset, the resource carrying out the event/activity is enforced at the UI. Dental Professionals are incentivised	Missing Resource Information– Process Mining(8)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00

to record all steps as their activity levels are based on this						
Process Mining – Missing Event Attributes. In this dataset, Event attributes such as DMFT could be missing	Missing Event Attribute– Process Mining(9)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining – Incorrect Cases. In this dataset, An incorrect case could arise if events were incorrectly recorded leading, e.g. incorrect recording of ‘Initial Exam’ Could lead to a misleading case being created – perhaps an outlier as a result	Incorrect Case – Process Mining(10)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Process Mining – Incorrect Events. In this dataset, An incorrect event could arise if events were incorrectly recorded leading, e.g. incorrect recording of ‘Initial Exam’	Incorrect Event– Process Mining(11)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Process Mining – Incorrect Relationship. In this dataset. This referential integrity is enforced by Rule 1	Incorrect relationship– Process Mining(12)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Process Mining – Incorrect Case Attribute. Only DOB is used in this research to calculate age at treatment – no way to know if this is incorrectly recorded	Incorrect Case Attribute – Process Mining(13)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Process Mining Incorrect Position (14). This is only relevant if we have no timestamps. In this dataset it is enforced at the UI. (manipulated with listposition to get a sequence – except ‘Initial Exam’ = 0 , and ‘Completed Case’ = 23	Missing Position – Process Mining(14)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining Incorrect Event/Activity Names (15). In this dataset it is enforced at the UI and at the Process Name Mapping Stage – Users may have incorrectly recorded this – known issue with ‘Initial Exam’	Incorrect Ev/Act Names – P M(15)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining Incorrect TimeStamp. In this dataset it is enforced at the time of ‘completion’ of the entity – this may not	Incorrect TimeStamp – Process Mining(16)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00

necessarily be contemporaneous with the treatment						
Process Mining – Incorrect Resource information. In this dataset, the resource carrying out the event/activity is enforced at the UI. Dental Professionals are incentivised to record all steps as their activity levels are based on this	Incorrect Resource Information– Process Mining(17)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining – Incorrect Event information. In this dataset, many opportunities for incorrect event information	Incorrect Event Information– Process Mining(18)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining – Imprecise Relationship. In this dataset. This appears unlikely to arise in this research	Imprecise relationship– Process Mining(19)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Process Mining – Imprecise Case Attribute. Only DOB is used in this research to calculate age at treatment enforced at UI	Incorrect Case Attribute – Process Mining(20)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Process Mining – Imprecise Position. This appears unlikely to arise in this research	Incorrect Position – Process Mining(21)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Process Mining – Imprecise Activity Name. This appears unlikely to arise in this research	Imprecise Activity Name – Process Mining(22)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Process Mining – Imprecise Timestamp (23). This is what is manipulated with listposition to get a properly ordered sequence – except ‘Initial Exam’ = 0, and ‘Completed Case’ = 23	Imprecise Timestamp – Process Mining(23)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Process Mining – Imprecise Resource information. In this dataset, the resource carrying out the event/activity is enforced at the UI. Dental Professionals are incentivised to record all steps as their activity levels are based on this	Imprecise Resource Information– Process Mining(24)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00
Process Mining – Imprecise Event information(25). In this dataset, many opportunities for incorrect event information	Imprecise Event Information– Process Mining(25)	Row Level Data Issue	Process Mining Literature	Incomplete (Mans et al)	0	0.00

Process Mining – Irrelevant Cases (26). – Problematic because it causes unnecessary complexity in the process maps	Irrelevant Case – Process Mining(26)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Process Mining – Irrelevant Events. Problematic because it causes unnecessary complexity in the process maps	Irrelevant Event– Process Mining(27)	Row Level Data Issue	Process Mining Literature	Incorrect (Mans et al)	0	0.00
Data Entry – Unexplainable difference between charting of Left & Right Quadrants	L-R Charting Bias	Table Level Data Issue	Common Sense	Incorrect (Mans et al)	0	0.00
Data Entry – Invalid Tooth Extraction e.g. tooth extracted twice	Invalid Tooth Extraction	Row Level Data Issue	Common Sense	Incorrect (Mans et al)	0	0.00

10.18 Data Transforms

10.18.1 Mapping the Procedure Names

The table containing the list of treatment events, PMTreatments, has a field called ProcedureName. This normally contains a name such as 'Initial Exam' or 'Casual Attendance'. This is the most important field in this research for defining an 'event' or a discrete step in a treatment process or care pathway. In the Bridges application creating this dataset, users had the option for many years of customising the ProcedureName field. This led to several variations on common procedure names as well as several rarely occurring procedure names. To deal with this the following steps were taken:

- Determine frequency of appearance of each ProcedureName.
- Map procedure name variations to the standard procedure name. (MappedToProcedureNameGroup)
- Count frequency of standard procedure name.
- Ignore rarely occurring procedures.

Of the total entries in PMTreatments (3,169,864), 18,443 (0.6%) were not mapped to a standard procedure name. On examination, most of these were one-off customisations and accordingly, were not used in this research.

The mappings are stored in a table called PMProcedureNameMappings. The standard procedure name was added to the PMTreatments table and the original ProcedureName was retained. The mappings table also has capacity for other abstractions of the data e.g. Preventive or Restorative and also has a column for SNOMED/SNODENT codes. The entries and their mappings are shown in Appendix 10.2. The mappings table was also created here.

10.18.2 Updating the PMTreatments Table with ClientAge

Often it is significant to know the age of the client (patient) at the time they received a treatment. To simplify the queries requiring this information, the client age at the time of treatment was calculated and added to the treatment record.

10.18.3 Updating the PMTreatments CompletionDate with ListPosition

A process-step ordering problem arose when testing the suitability of the data for PM. The CompletionDate is exactly that - the date and has no time component. This means that treatments completed on the same day are inseparable from an 'order' perspective. Fortunately, Bridges stored a 'ListPosition' which often indicates the order in which treatments will be completed and certainly indicates the order they were created in. The value of the ListPosition was added to the 'hours' component of the CompletionDate of a treatment. This allowed the PM algorithms to separate the event times.

ListPosition occasionally goes to 200+ but the hh field keeps incrementing so the order will still be good. Presumably the 'Date' will have incremented over the course of 100 treatments solving that problem. Initially it was hoped to update the seconds or milliseconds field, but the installation of SQL Server did not accurately store datetime to this level and applied a random figure irrespective of the value of ListPosition.

10.18.4 Query Adding BadRowCode Field to BridgesPM1 Table

To allow marking of individual rows with a DQ measure, each table in BridgesPM1 got an additional metadata field BadRowCode which is then updated if the row has any DQ issues. The list of possible issues is stored in DataQualityIssuesRegister

10.18.5 Query Creating DataQualityIssuesRegister table and Inserting Rows

Four new tables were created to manage the PM issues around the research: DataQualityIssuesRegister to store the individual issues along with code to mark data or mitigate the problem, DataQualityIssueLevels to categorise an issue as row-level, field-

level, table-level, or dataset level, DataIssueDimensions to identify the category from which the issue emerged e.g. General Literature and DataIssueSources to identify the exact source of the DQ issue i.e. who was the information source for the specific issue.

10.18.6 Query Adding DMFT and ChartID to Treatments.

DMFT is the oral health outcome used in this research and it is often required to know the DMFT status at the time that treatment is carried out. To simplify queries, the DMFT values were added to the PMTreatments table. Also, as DMFT is measured at the time that a dental charting is completed and is not directly tied to the treatment therefore an additional field called MonthsToDMF is also added. This gives us a number indicating the number of months elapsing between the date that the treatment was completed and the date that the associated chart with the DMFT was created.

10.18.7 Ordering and Ranking Initial Exams (IE) - Count IEs and Identify 1st IE

It is valuable to know whether an 'Initial Exam' in the PMTreatments table is the first initial exam for a patient. Likewise, it is useful to know the number of 'Initial Exams' a patient experienced. To simplify future queries, NoOfInitialExams and FirstExam fields are created and calculated.

An additional table called RankedScreenings is created containing an entry for each IE and ranked according to its completion date. For clarity, the rank of a screening is its place in the sequence of screenings, i.e. the first screening has rank 1, the second screening has rank 2 etc.

10.18.8 Cross Tabulation of Screenings and Patient Age at Screening

Ranked Screenings is used to create ScreeningAgeCrossTab again allowing us to view the data associated with IEs from several perspectives and simplifying the queries.

10.18.9 Adding DMF Tooth Columns and Calculating

This query adds a D, M, & F column for each tooth to the chart table PMCharts and applies conditions in the PMConditions table to them giving a DMF score for each tooth.

10.18.10 Add PreventionOrRestoration to PMTreatments Table

To add a higher level of abstraction to our events (treatments), an additional column was added to the treatments table and the event was characterised as 'Preventive' or 'Restoration' if possible. 'Initial Exam' was left as-is, as were 'x-rays', 'Completed Case' and some others. This was done to simplify the process models into the two paradigms, 'Preventive' and 'Restoration', and helps simplify some process models and address problems caused by the ad-hoc, flexible characteristics, typical of healthcare processes.

10.18.11 Add Emergency marker to PMTreatments Table

An additional column called EmergencyCasual was added to the treatments table. Emergency or casual visits to the dental service seem to account for up to 10% of the activity. It appears that some practitioners were registering an 'Initial Exam' on some emergency appointments and the option to exclude these from some queries was required, as they are not scheduled screenings. To do this, the query would have to reference the appointments table for each treatment item and this would have had a serious performance impact in many places therefore each treatment was marked if it had taken place on the same date as an emergency appointment for that patient. 11.03% of all treatments were carried out on the same day as an emergency appointment.

10.18.12 Create Summary Table for Medical Questions and DMFT Outputs

To give a sense of the frequency of various medical conditions as registered on the patient medical questionnaire and their associated DMFT outcome, a summary table of commonly positively answered questions was created called DMFTDistributions. This provided a preliminary overview of these data as presented in Section 4.1.6.4.

10.19 Data Quality Framework

As the data in this research is an extract from an operational dental EHR, assessing the DQ is an essential step. To achieve this, the following are now addressed:

- What are the DQ dimensions relevant to the use of EHR data for research?
This will be achieved referencing the recent literature in the area of EHR data quality.
- Which of these are relevant for this research?
This will be achieved by reference to the PM DQ literature and the EHR research literature above.
- What are the DQ information sources in this research?
This will be achieved by reference to existing literature and on the author's own experience with the generation of the research data through the EHR and
- Present the Data Quality Framework.
- What DQ issues were discovered? An appendix of data quality issues will be provided

10.19.1 Dimensions of Data Quality

10.19.1.1 Dimensions of EHR data quality

Dimensions of DQ allow us to identify data features that can be measured. Weiskopf & Weng (2013) reviewed the literature on dimensions of EHR DQ and methods of DQ assessment identifying Completeness, correctness, concordance, plausibility and currency as the dimensions. Seven broad categories of assessment methods were also identified. They further suggested that concordance and plausibility could be handled within the 'correctness' dimension. Many other dimensions were identified in this review which they rationalized to the five named above. This framework has been successfully applied to MIMIC-III, a publicly available e-health record database (Kurniati, et al., 2018). Incomplete or missing data, inconsistent and inaccurate data are confirmed as major issues (Song, et al., 2013; Danciu, et al., 2014; Anker, et al., 2011; Botsis, et al., 2010). A variation of these dimensions is also proposed by the (DAMA UK Working Group (2013) and by Microsoft (2012): completeness, conformity, consistency, accuracy, validity and duplication. Kahn et al. (2016) produced a harmonized DQ assessment terminology and framework for the secondary use of EHR data incorporating several existing EHR DQ frameworks. Their output consisted of harmonized DQ terms and an organizing framework. They further rationalized DQ dimensions into 3 categories; 'conformance' with subcategories value, relational and computational, 'completeness', and 'plausibility' with subcategories uniqueness, atemporal and temporal. These categories can be applied in two assessment contexts; 'verification' (internal to the data) and 'validation' (referencing external benchmarks). Intrinsic data features were included in the scope of the study with extrinsic features including fitness for a specific analysis excluded. DQ issues caused by deficiencies in the data representation or the data model and 'relevancy' were also excluded.

10.19.1.2 Dimensions of PM DQ

Mans et al. (2015) and Bose et al. (2013) identified four broad DQ issues that could exist in event logs: missing data, incorrect data, imprecise data and irrelevant data. This further dimension, 'irrelevant,' is important in PM because superfluous information increases the complexity of process models and can detract from their comprehensibility. These dimensions were further detailed in 27 types of quality issues relating to the case, event and attribute levels of the data in an event log.

The Process Mining Manifesto (IEEE, 2011), proposes a rating system indicating data quality ranging from 1-star to 5-star. 3-star systems typically automatically record events. The PM event log is deemed as trustworthy though not necessarily complete. Examples of 3-star event logs are tables in ERP systems, event logs of CRM systems etc. Event

logs resulting from traditional Business Process Management workflow systems might be considered for 4-star status whereas the 5-star status is reserved for logs that are trustworthy and complete, events are automatically recorded, well defined, systematic and have clear semantics (IEEE, 2011, p. 7). This last point resonates with this research as the event names are being mapped to the Standard Nomenclature for Dentistry (SNODENT) with the intention of encouraging reproducible research and allowing further research to build on this research. With this rating system in mind, subjectively assessing the BridgesPM1 dataset would suggest it has a quality rating between 3 and 4-star.

10.19.1.3 Which data quality dimensions are relevant for this research?

Specific to PM, four broad DQ issues that could exist in PM ELs were identified by Mans et al. (2015) and Bose et al. (2013): missing, incorrect, imprecise and data that is irrelevant or superfluous to the investigation. This further dimension, ‘irrelevant’, is very interesting to process miners because superfluous information increases the complexity of process models and reduces their comprehensibility. These dimensions were further detailed in 27 types of quality issues relating to the case, event and attribute levels of the data in an event log. The widely cited Process Mining Manifesto proposes a rating system for DQ ranging from 1-star to 5-star (IEEE, 2011) as detailed in the previous section. The proposed framework allows us to include and tailor those dimensions and categories appropriate for the specific research and to include extrinsic data features as DQ issues.

10.19.2 What are the Data Quality Information Sources in this Research?

Assessing DQ is further complicated by the many potential information sources. Having established the dimensions of DQ for this research, where can information on potential issues be found? Looking at the lifecycle of the data and the potential of many parties and stakeholders to influence the DQ and to provide information on the DQ issues showed that there are many stakeholders capable of valuable commentary on the quality of the BridgesPM1 dataset. The developers and database administrators of the software application can comment on data integrity issues and applied business rules through their knowledge of the database structure and its entities and attributes. Users of the original application can comment on the custom and practice of the system’s use on the ground and the protocols for data recording. Dental professionals can provide domain knowledge and comment on the plausibility of data values. Previous research using the data for earlier work also contains information on DQ. As discussed earlier, the general literature on EHR data quality (Weiskopf & Weng, 2013), the literature on the secondary use of routinely collected health data for research (Danciu, et al., 2014; Anker, et al., 2011; Botsis, et al., 2010) and the literature specific to DQ in data mining and PM (Bose, et al., 2013; Mans, et al., 2015) and ANSI & SNODENT standards all serve to provide us with tools to assess the DQ. In brief, DQ information sources for this research are:

- Software Developers and Database administrators.
- EHR Application Users.
- Domain experts (dental professionals).
- Previous research work using this or similar data.
- General Literature.
- PM Literature.
- Comparison to Standards (SNODENT, ANSI).
- Patients.

10.19.3 Introducing the Care Pathway Data Quality Framework (CP-DQF)

Given the many sources for potential DQ issues, a structured approach to the management of DQ issues has been taken. The DQ management strategy adopted in this research

centres on a registry of DQ issues. The research data is assessed using this registry and any research data affected by these issues is marked. The scale of the issue is recorded and mitigated through code if possible. This is achieved by a Data Quality Framework (DQF). (*A User Interface is not yet developed as of 7/1/2019*). *The registry is currently managed with manual (scripted) inserts and updates.*)

Applying Deming's Plan-Do-Study-Act (Moen, 2010) our approach for the CP-DQF is:

- Plan – Frame the quality questions for the research.
- Do - Identify DQ dimensions. Identify potential sources of information on DQ. List potential DQ issues. Relate the issues to the experiments. Mark the data. Mitigate the DQ issue if possible.
- Study – Analyse the results of the 'Do' phase.
- Act – Take steps to improve future DQ.

The aim here is that data of unacceptable quality is marked as 'bad' i.e. unusable. Imperfect but acceptable data is marked as 'compromised' i.e. it can be used in some situations or experiments. The remaining data is unmarked or 'good' and is available for all purposes.

The framework can incorporate fitness-for-use DQ issues, i.e. DQ issues affecting specific experiments. Involvement of the researchers or principle investigators at this juncture will strengthen the exercise and help eliminate confounders and invalid assumptions. The CP-DQF maintains a registry of DQ issues. Code is written to mark individual data elements (usually rows) affected by the DQ issue. The code is stored with the DQ issue in the registry. In the case-study below, this code consisted of Structured Query Language (SQL) update commands. The research data is assessed against the DQ registry and data records affected by these issues are marked. The scale of each issue is recorded and mitigated through code if possible. The registry is currently managed with manual inserts and updates and building a user interface is in progress. The principle components of the data structure supporting the CP-DQF are shown in the entity relationship in Figure 10-29 below.

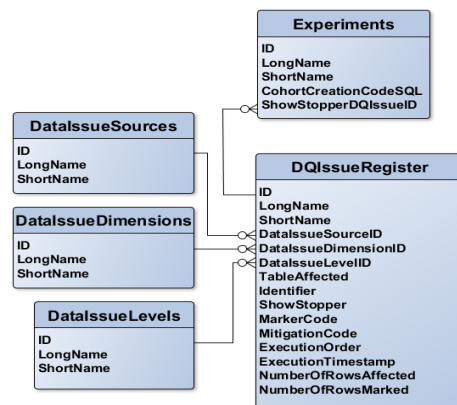


Figure 10-29: CP-DQF Entity Relationship Diagram

How does the CP-DQF help with this complex problem?

The CP-DQF framework helps:

- Identify DQ issues.
- Record DQ issues.
- Mark-up research datasets with DQ metadata.
- Mitigate effects of DQ issues on research by easing exclusion of data.
- Mitigate effects of DQ issues by, for example, imputation of values.
- Report on the extent and impact of DQ issues.

Using this CP-DQF has three principal steps. First, establish the DQ issues register for the research. Some previously known issues may be prepopulated in the register and this will be supplemented with additional issues specific to the research or discovered by the

researchers. Second, push the research data through the CP-DQF. Third, report on what happened.

10.19.3.1 Step 1: Establish the DQ issues register for this research

This phase establishes the DQ issues for the research and links them to the specific research experiments to be carried out. First, it established the DataQualityIssuesRegister for the research by pre-populating the register with known issues and supplementing this with additional issues specific to the RQs.

Phase 1: Establishing the DQ issues register for the research.

- Add general DQ issues to the register.
- Create entries in the experiments table.
- Add any experiment-specific issues to the register.
- Link experiments to entries in the DataQualityIssuesRegister. This will disqualify the data from use in that experiment if marked as a showstopper.

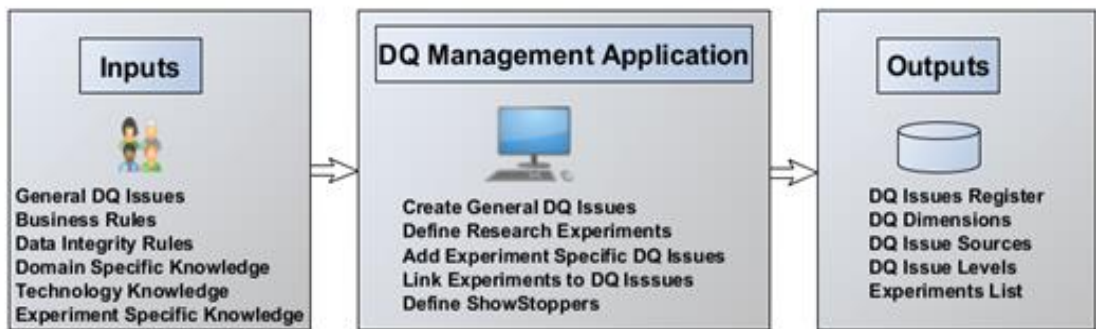


Figure 10-30: CP-DQF (Step 1)

What is in the DataQualityIssueRegister? All the potential DQ issues that the researcher discovered either from the literature, from domain specific knowledge, prior knowledge about the data e.g. EHR users or previous researchers. It contains details of the scope of the issue and perhaps code to mitigate the problem or to mark the data as compromised. What is in the DataIssueLevels? This identifies the level at which the DQ issue exists – Table level, Row Level, Field Level.

What is in the DataIssueDimensions? The major dimensions or categories of DQ issues established from the literature and a rationalisation of these, e.g. ‘Incorrect Data’ is a DQ dimension. ‘Invalid Data’ is also proposed as a dimension in the literature. I believe this can be rationalised and subsumed into ‘Incorrect Data’. Although other dimensions have been proposed in the literature, other dimensions beyond those proposed by (Mans, et al., 2015) do not seem necessary, i.e. Incomplete, Incorrect, Imprecise, and Irrelevant.

What is in DataIssueSources? The specific source of our data quality issues i.e. Who told us about it? e.g. Weiskopf & Weng

Step 2: Applying the CP-DQF to the research data.

10.19.4Eight steps are taken in applying the CP-DQF to the research data.

10.19.5Add Metadata to the research data.

Mark-up fields are added to the research data allowing us to store DQ information with the data element (usually a row). This information can be used to exclude the data from the dataset as it is extracted for a specific experiment. Suggested fields are: a Boolean called BadRow and a vector string called BadRowCodes. The vector string can hold multiple error codes simultaneously.

10.19.6Pre-processing or discussion section?

Decide where the DQ issue is to be dealt with, in pre-processing or by way of discussion. This will determine whether the data can be marked with this issue. If not, this will be addressed in the research discussion.

10.19.7 Does an issue disqualify the data from the experiment?

If the DQ issue is serious for any specific experiment, the experiment should be marked, and the data excluded from use there.

10.19.8 Evaluate the effect of these data disqualifications.

Does it require re-execution of marking or mitigation code? Does it skew results? e.g. Removal of data may violate previously satisfied data integrity constraints.

10.19.9 Write/Run the Marking Code from the CP-DQF against the data.

Executing the code stored with the DQ issue in the register will mark the research data's metadata with information about its DQ.

e.g. Mark orphaned treatments (no client exists) as 'bad'

```
Update PMTreatments
set BadRow =1,
BadRowCodes= Concat(BadRowCodes,' 7')
where ClientID not in (select PMClientID from PMClients);
```

10.19.10 Write/Run the Mitigation Code against the data.

Executing the mitigation code (if exists) will update the research data to improve its quality.

10.19.11 Update the DQ issues register with the results.

Record the scope of the issue and the scope of the mitigation efforts, primarily for reporting purposes.

10.19.12 Write/Run the CohortSelection Code.

Cohort/Dataset selection code can now be written incorporating the metadata as a criterion for exclusion/inclusion in the dataset. In the implementation below, treatment events are only selected if the metadata, BadRow is NULL.

```
e.g. Select * from PMTreatments
where ClientAge = 8
and BadRow is NULL
```

These steps are summarized in Figure 10-31

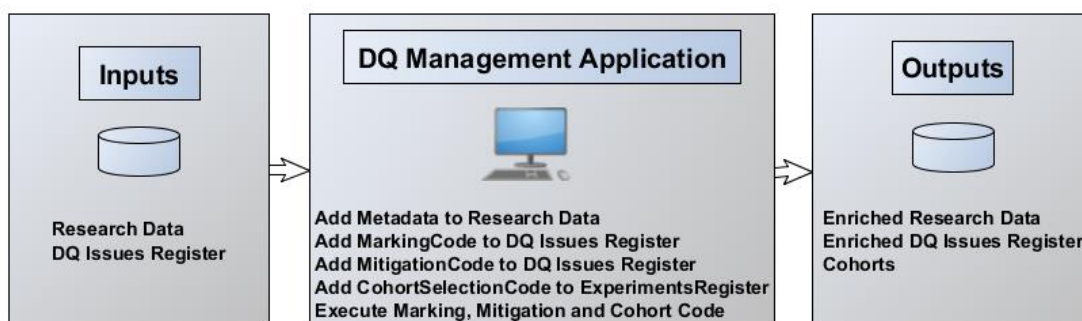


Figure 10-31: CP-DQF (Step 2)

The question arose as to whether all pre-processing of data is a DQ step. e.g. When mapping 'Amalgam Filling 1 surface' and Amalgam Filling 2 surfaces' to a simplified event 'Amalgam Filling', is this a DQ improvement step and, should that DQ issue be present in the DataQualityIssuesRegister? Removing the unnecessary complexity, albeit reversibly, will lead to a less complex, more comprehensible, and better-quality event log. These boundaries remain unclear.

10.19.12.1 Phase 3: Report on Phases 1 & 2

- Report of the data issues, their scope, how much data was affected etc.

- Evaluate the effect of data disqualifications in phase 2 above, e.g. Does it skew results? These steps are summarized in Figure 10-32

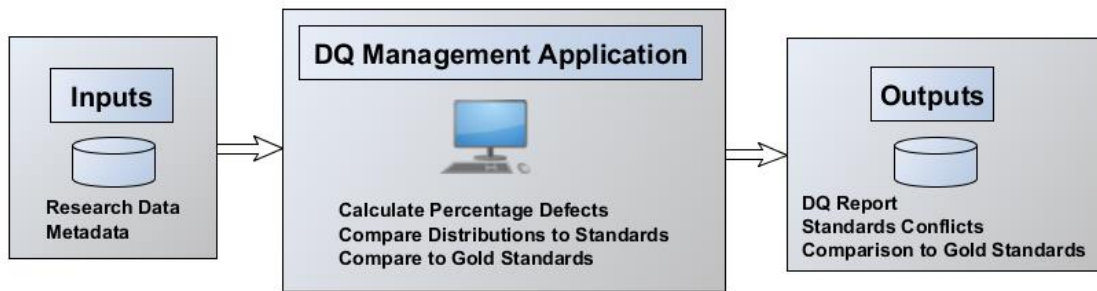


Figure 10-32: CP-DQF (Step 3)

10.19.13 Applying the CP-DQF to this research

To identify the potential data DQ issues the following steps were taken:

10.19.13.1 Identify potential data quality information dimensions

While cognisant of the quality dimensions proposed by Weiskopf & Weng (2013), DAMA UK Working Group (2013) and Microsoft (2012), those proposed by Mans, et al (2015), designed specifically for PM, were used.

- Incomplete (e.g. missing date-of-birth).
- Incorrect (e.g. incorrectly logged timestamp).
- Imprecise (e.g. lacking precision or too coarse).
- Irrelevant (e.g. increasing complexity of process model without contributing value).

Some of the other proposed dimensions were mapped to the above: ‘concordance’ and ‘plausibility’ to ‘incorrect’ (Weiskopf & Weng, 2013); ‘accuracy’ to ‘imprecise’ (DAMA UK Working Group, 2013) and ‘conformity, consistency, validity and duplication’ (DAMA UK Working Group, 2013) to ‘incorrect’, ‘conformance’, and ‘plausibility’ (Kahn, et al., 2016) to ‘incorrect’.

10.19.13.2 Identify potential DQ information sources within these dimensions

Identify potential sources of information on DQ. Each of these sources was assessed and a list of potential DQ issues arising from them was developed. Each identified potential DQ issue received a unique code and an entry in the DataQualityIssuesRegister. The data was assessed to establish if the DQ issue applied in this instance. Queries were developed to mitigate the DQ issue if possible. Rows affected by issues were marked as such and a decision whether the issue disqualifies the row from use in the research was made.

The data issues entity was stored in DataQualityIssuesRegister with the main attributes describing the issue, the source of the data issue, the level of the issue as well as the number of rows affected and the code to mitigate or mark the issue. There are several levels at which a DQ issue could have affected our data. It could affect a field, a row or rows or even a table or the complete dataset. These levels were stored in the DataIssueLevels table. Categories of DQ issue sources were identified in this research and stored in the DataIssueDimensions table. An example of a dimension is ‘Missing Data’. A specific source belonging to that dimension is e.g. (Weiskopf & Weng, 2013).

10.19.13.2.1 From the software developers/database administrator

The software developer/database administrator understood the logic of the database architecture and the entity relationships. This included knowledge of the table structures, their fields and the relationships between these tables. This resulted in data integrity rules and an initial DQ assessment phase enforcing the logic of the database. It primarily marked data in the BridgesPM1 extract that should never have been there in the first place. In the main, this related to orphaned data. An example of this is, treatment items that have no corresponding patient, i.e. entries in the PMTreatments table without a corresponding entry in PMClients. Orphaned records might have existed for a number of reasons. The

original Bridges EHR database, for the purposes of audit and integrity of records, marked records as deleted as opposed to physically deleting them. Other functionality in the original database such as merging of duplicate records could also have contributed to the existence of orphaned records. It was not possible to enforce all such application logic in the original extract from the HSE/Bridges database due to resource constraints. It was a simple step to enforce the integrity constraints before use of the data in this research.

In the BridgesPM1 dataset, 23 such integrity and business rules were identified. Each rule had an entry in the DQ registry and code was written to update each affected record.

10.19.13.2.2 From the Bridges application users

Previous research provided us with valuable details on how the data was created and associated protocols (Murphy, 2011). Users of the original application commented on the practical use of the system, with knowledge of data input on a day-to-day basis, information on the functionality of the system, shortcuts, and weaknesses in the system. Murphy (2011) carried out analysis and verification of data entry and involved all staff in the North Lee LHO Area. Areas that would benefit from improved data entry protocols were identified and gold standard definitions of data entry practices were developed and circulated to all staff, e.g. Initial Exam should be ticked for a screening appointment only. This had a 50% compliance indicating that practitioners were also ticking Initial Exam in circumstances other than school screening appointments. Murphy indicated that this over-counting would inhibit accurate comparisons of service activity levels between clinics or LHO areas. This issue was of concern to this research as ‘Initial Exam’ was used as a starting point for some of the RQs. The issue was mitigated as much as possible in the cohort selection phase. Gender, fluoridation status, and recording of dental trauma were also identified as invalid due primarily to incorrect data entry protocols and have been excluded from this research. These issues are entered in the DQ issues register.

10.19.13.2.3 From Dental Domain Expertise

This was not carried out in the DQ assessment phase of this research and may be more appropriate to the discussion section. No issues from this source were entered in the registry at this time. This area offers strong potential for calculated metrics such as mean, median, and value distributions. Validated oral health benchmark measures such as DMFT (Decayed, Missing & Filled teeth) could be registered here and the research data values compared to this to give an indication of external data validity. Other work, specific to the implementation of dental quality measures in dental EHRs should also provide indications of external data validity (Bhardwaj, et al., 2016). There are no issues from this source entered in the data quality register.

10.19.13.2.4 From earlier research using Bridges data

Several research projects and master’s theses have been carried out with Bridges as a data source. The validity of Bridges Database Query System was validated by Murphy (2011). Further validation of results that did not fit with evidence was also carried out e.g. gender distribution and recording of trauma. Discrepancies in data input protocols were identified as responsible. The query system itself was found to be valid.

The database facilitated access to dental health status through DMF measures and KPIs and has supported research projects including Fluoride and Caring for Children’s Teeth (FACCT) (CARG/2012/34) and Mapping the Divide (MTD) (HRA_HSR/2012/25).

As outlined above, gender, fluoridation status, and dental trauma status were not used, and these DQ issues were ignored in this research. There were no additional issues from this source entered in the DataQualityIssuesRegister.

10.19.13.2.5 General data mining literature

The general data mining literature suggests common issues are representational bias, clinician-related biases regarding missing data and outcomes, non-standardization of data entry, data redundancy, inaccuracy, restriction to retrospective study, and difficulties extracting data (Song, et al., 2013). Root causes for some DQ issues in the secondary use

of data were identified by Danciu et al. (2014) and Anker et al. (2011) (See Section 5.3.1). The authors also proposed some solutions involving formal information exchange mechanisms, clinical registries and personal health records as well as the sharing of effective strategies for secondary use of healthcare data (Anker, et al., 2011).

10.19.13.2.6 From PM DQ Literature

As detailed earlier, Mans et al. (2015) and Bose et al. (2013) identified four broad DQ issues that could exist in PM ELs: missing data, incorrect data, imprecise data, and irrelevant data. These were further detailed in 27 types of quality issues.

The Bridges-PM1 dataset was evaluated for each of these, identifying whether it is likely that the problem exists, how it may have arisen and what its effect is likely to be. Further, steps to mitigate the problem were considered and whether their effect merits the investment. Using the method proposed by Mans et al. (2015) these were tabulated as possible sources of DQ issues. The potential issues have been numbered as in the original research with ‘N’ indicating that the issue does not exist, ‘L’ indicates a low likelihood of the issue being present and ‘H’ indicating a high likelihood. A brief justification of the NLH classification logic has been documented. “(??)” in the table indicates that this needs to be confirmed.

Table 10-9: 27 Data Quality Issues (adapted from (Mans, et al., 2015))

	Missing Data	Incorrect Data	Imprecise Data	Irrelevant Data
Case	1 (L) All schoolchildren are eligible to be screened	10 (L) duplicate records for a patient may exist.	N/A	26 (L) Superfluous data may have been recorded. Most of the data recorded will not be used in this PM exercise
Event	2 (L) Dental Professionals are incentivised to record all steps as their activity levels are based on this	11 (L) Dental Professionals are incentivised to record all steps. It is possible that screenings are over-reported.	N/A The creation of an event is controlled by the user interface.	27 (L) Superfluous data may have been recorded. Some details of the individual events will not be utilised.
Relationship (Belongs to)	3 (N) Primary Key Integrity enforced	12 (N) Primary Key Integrity enforced	19 (N) Primary Key Integrity enforced	N/A
C_attribute	4 (N) Only case attribute is ‘DateCreated’	13 (N) Only case attribute is ‘DateCreated’	20 (N) Only case attribute is ‘DateCreated’	N/A
Position	5 (L) Enforced by user interface/ application	14 (H) No strict protocols exist	21 N/A	N/A
Activity Name	6 (N) Enforced in EHR GUI	15 (L) Dental Professionals are incentivised to record all steps accurately	22 (L) Dental Professionals are incentivised to record all steps accurately	N/A
Timestamp	7 (N) Enforced in EHR Business Rule	16 (N) Enforced in EHR Business Rule	23 (N) Enforced in EHR Business Rule – Completion Date for treatments is ‘date’ only. This may not be detailed enough. See	N/A

			Timestamps note below	
Resource	8 (H) Dental Professionals are incentivised to record all steps as their activity levels are based on this	17 (L) Sometimes procedure marked as completed by Surgery Assistant.	24 (L) Dental Professionals are incentivised to record all steps accurately	N/A
E_attribute	9 (H) Not compulsory (??)	18 (L) Dental Professionals are incentivised to record all steps, however errors are possible.	25 (N)	N/A

Additionally, Bose et al. (2013) proposed categories of process characteristics with the potential to impact the output of PM, summarised in Table 10-10. Many high-tech systems produce logs of very fine granularity leading to spaghetti-like process models. Higher levels of abstraction can be achieved using ontologies (Pedrinaci & Dominique, 2007). Case heterogeneity can also produce spaghetti-like process models. Trace clustering has been shown to be effective i.e. partition event logs into subsets of homogeneous cases. Voluminous data will require ever more efficient and scalable PM algorithms.

Several additional matters affecting the quality of the event data and the resultant models were outlined by van der Aalst (2016). He points out that processes are not necessarily in a steady state. They can be affected by working hours, weekends, contextual factors and concept drift. Processes can alter significantly at shift changeovers, and often show daily, weekly and seasonal patterns (van der Aalst, 2016, p. 318). These issues are not automatically visible in discovered processes and present a significant challenge when trying to use such models for prediction and suggesting improvements. Contextual factors such as case context, process context, social, and external contexts also need to be considered in evaluating data and model quality. He also points out the issue of concept drift, where a process changes as it is being analysed i.e. within an event log. It is then necessary to identify when and what changed. He also introduced a fresh way of categorising data quality issues: missing in log, missing in reality, and concealed in log (van der Aalst, 2016, p. 148). He also identified that DQ issues could themselves have a temporal dimension and hence, be continuous, intermittent or changing. Interestingly, the data quality dimension ‘irrelevant’ does not feature in his analysis.

A general timestamp arose in the BridgesPM1 dataset. A process-step ordering problem arose when testing the suitability of the data for PM. The Completion Date is exactly that - the date, and it has no time component. This problem and its solution has been explained in detail in Section 10.18.3.

Incorrect Timestamps: If cases have an incorrect timestamp, e.g. treatments are not contemporaneously marked as ‘completed’, this is very difficult to establish. There is no reason to suggest that this is a common occurrence. Mixed granular Timestamps are not an issue here.

Table 10-10: Process Characteristics leading to DQ issues adapted from Bose, et al. (2013)

Data Quality Problem	Relevance to BridgesPM1
Voluminous Data	The research data is easily manageable from a volume perspective.
Case Heterogeneity	As in most healthcare processes, spaghetti-type models initially emerged. By streamlining the data and using various other steps in the ETL process, Comprehensible models emerged.

Event Granularity	Excessive granularity often leads to spaghetti-type process models, typical in healthcare environments. In this case, the events are relatively ‘coarse’, mostly involving a significant item of treatment and accordingly this is not an issue here.
Process Flexibility and Concept Drift	Evolutionary Change: This is worth considering and indeed might interfere with the underlying experiments and RQs. Momentary Change: Most PM algorithms have capacity to deal with noise and accordingly should identify this issue as noise or as an outlier.

10.19.13.2.7 From Standards

Care pathways are often highly variable in clinical settings and PM of EHRs often produce logs of high heterogeneity and very fine granularity leading to *spaghetti-like* process models. To untangle the *spaghetti*, abstraction methods using classifiers or ontologies are commonly used, for example, abstractions or standards like SNODENT-CT, (Pedrinaci & Dominique, 2007). Trace clustering has been shown to be effective in identifying patients with similar pathways, which can be then be used to partition event logs into subsets of homogeneous cases.

10.19.13.3 Create & list potential DQ issues from these sources

From the above:

- Identify as many potential DQ issues as possible.
- Code the issues (ID, Fatal/non-Fatal). This may vary between experiments.
- Assess applicability of the issues to our research generally.
- Develop Code/Queries to disqualify/mark the data.
- Develop Code/Queries to mitigate the data issues.

The relevance of each DQ issue to each experiment can then be assessed and mitigation measures applied. When this DQ strategy is applied to the BridgesPM1 database, the result is that some of the data is marked as ‘bad’ i.e. unusable, some as ‘compromised’ i.e. it can be used in some situations and the remaining data is unmarked or ‘good’.

10.19.13.4 Other areas causing DQ issues – not included in the register.

There are many other areas, both general areas and those applicable to this research’s dataset that merit consideration as DQ issues. A brief discussion of some of these follows. No privately funded dental treatment was explicitly included in the research dataset, i.e. no treatment items are present in the PMTreatments table for treatments not carried out within the public health system. However, it is possible that when the patient is examined, and their dental status was charted, some of these externally received treatments will be recorded in the graphical charting. This gives rise to a bias in the research where all that is seen is not all that there is.

Also, it is worth considering if ‘insufficient data’ is a legitimate DQ issue, for example, to use ICDAS as an outcome measure, additional information was required to that required for DMFT calculations. Also, this dataset did not contain socio-economic status (SES) information and although this could perhaps be inferred from the treating clinic, it is inferior to direct evidence. Similarly, the lack of fluoridation information is a shortcoming in the dataset, though again it could often be inferred from the treating clinic.

Anonymisation and de-identification of the data, though clearly necessary, degraded the overall quality of the dataset. Removal of address information denied many research opportunities e.g. both fluoridation and SES could be established to a high degree of accuracy from an accurate home address. Also, free-text often contains valuable additional information supplementing the more structured information in treatment lists. However, this free-text also often contains names of relatives and other identifying information and hence, are mostly removed in the anonymisation process

Datasets sometimes contain seasonal and other temporal DQ issues. This dataset is an extract from a school screening service and potentially contained several such issues. For example, the service is Monday to Friday and the screenings are concentrated during term time and accordingly, higher levels of activity would be expected at these times.

Also, in this dataset, data collection started in 1998 as a pilot, but was not fully operational in all areas of the organisation until 2002 i.e. several years passed while the EHR ramped-up. Time-boxing is a method to ensure that EHR data is optimal for study. Use of data from early use of a database might give rise to unstable data. Identifying a date where use of the system is stable can resolve this e.g. after all the staff have been trained and data entry protocols are in place. These temporal issues need to be considered when defining cohorts and drawing conclusions from data and analyses and in this research, no data from before 2004 was used in the experimental analyses.

Removal of outliers has the potential to create DQ issues. It is possible that the most interesting information is in the noise and the outliers. One strategy to deal with this is to remove the most common 80% of activities/events and examine what's left. In the case of this research, this might involve only looking at the care-pathways and outcomes associated with patients presenting as 'emergency' or 'casual' patients.

The issue of researcher bias is well documented (Pannucci & Wilkins, 2010) and many of these same issue are expanded on by Kahneman (2012). Bias is defined by Pannucci as any tendency which prevents unprejudiced consideration of a question. The degree to which bias exists must also be considered as no research is entirely bias-free. Whereas chance and confounding can be quantified through good study design, not so bias. It is independent of statistical significance and sample size. Pannucci & Wilkins (2010) provide a list of potential biases in clinical-trial research and this can serve as a useful checklist for potential DQ issues cause by biases. These are summarised as pre-trial biases: flawed study design, selection bias, and channelling bias. Bias during trial: interviewer bias, chronology bias, recall bias, transfer bias, exposure misclassification, outcome misclassification, and performance bias. Bias after trial: citation bias and confounding bias. Other similar methodological pitfalls include randomisation errors, information bias i.e. errors in the outcome due to misinterpretation of information or systematic errors in the measurement of research variables. These can be prevented or at least mitigated using hard outcome measures. In this research's case DMFT is used.

It is debatable whether all of these and other factors such as the decision-making flaws identified by Kahneman (2012) and Taleb (2010) are truly DQ issues but there is no doubt that they lead to DQ issues and errors surrounding the interpretation of data. Including them as part of the DQ issues list ensured that they were considered and accordingly created the opportunity to improve the outputs of this research. Some of these biases are considered in the following section.

10.19.13.5 Kahneman's 'Thinking fast and slow' biases

Several flaws or biases in our thinking were identified by Daniel Kahneman (2012) in his book 'Thinking Fast and Slow'. It is based on a model of our decision-making consisting of System 1 thinking which is impulsive and automatic, and System 2 thinking which is thoughtful and conscious. Many of our decision-making errors have their source in our System 1 thinking. Looking at how some of these weaknesses may have affected this research is the final step in this process. The structure is to briefly describe some of the biases and retrospectively identify areas where this might have impacted this research.

- The lazy mind leads to errors. Cognitive ease is when the mind considers everything to be under control and it is more likely to make mistakes than when in a state of cognitive strain. Kahneman used the example of using a small, less-legible font on examination papers leading to fewer 'silly' mistakes. While this is somewhat counterintuitive, it is worth considering whether the more straightforward experiments in Chapter 6 might have

been more susceptible to this error. The experiments in 7.1.1 & 7.2 are significantly simpler than those in 7.3 & 7.4 and the author has reviewed the former experiments in this light.

- Operating on autopilot, also known as ‘priming’, can negatively affect the quality of our thinking, where exposure to a word or a context can lead to the summoning of related words and concepts more easily. Kahneman uses the example of priming with the concept of ‘money’ leading to selfish and individualistic actions. There are many opportunities for this bias to have affected this research. One example is the regular use in the field of data analytics of phrases such as ‘Data is the new oil’. In this author’s opinion, accepting this phrase without utilising the necessary discipline to assess its value with due diligence could lead to assumptions about the value of the data. This research expended significant effort to assess and document the quality of the EHR dataset in use.
- Snap judgments, also known as the halo effect, occurs where the mind oversimplifies a problem when there is insufficient information and fills in the gaps without the necessary justification. This research is clearly open to such biases. The very act of attempting to abstract a treatment process into a process model is already, albeit intentionally, doing this in reverse, and is a key objective of the research. However, it is worth considering that the creation of these models is a clear act of simplifying the ‘data’ and reducing it to a model for the purposes of identifying key features. A second type of snap judgment is known as confirmation bias where one tends to agree with information that supports previously held beliefs or, to believe or accept information suggested to them. This research is based on previous work and literature and could be susceptible to this error. However, the discipline of carrying out literature reviews and critically analysing the prior work as well as critical discussion of this research itself reduces the risk of this bias and its impact.
- A heuristic judgment is where the mind uses shortcuts to make quick decisions. ‘Substitution’ is one type of heuristic judgment, where the mind substitutes one question with another – usually one that is easier to answer. This could have arisen in several phases of the research. Abstraction of the RQs into validating experiments and further abstracting these experiments into algorithms and computer code offers multiple opportunities for this bias to arise. To minimise the likelihood of this happening, clear documentation of all algorithms and enumeration of assumptions should always take place. Another type is the ‘Availability’ heuristic where the mind applies a higher probability to something easy to remember or heard about often. Kahneman points to the example that many people estimate the likelihood of death due to a car accident much higher than that by stroke, whereas the opposite is the case. This is unlikely to have affected this research. Also, the phenomenon of ‘what you see is all that there is’ is an often-invalid assumption and in this research results in the assumption that the dataset contains all of the children in the target area and all of the dental treatments that they have received. Clearly, there may be children not present in the dataset and second, those in the dataset may have received dental treatment outside the public service. This has been acknowledged in the discussion chapter.
- The bias ‘No Head for Numbers’ can manifest itself as base-rate neglect where Bayesian priors are not factored into one’s thinking. An ignorance of the tendency to regress to the mean is another example of this. This research stopped short of applying strict statistical tests to the outputs of the analyses. This reflects the focus of the research being the development of the methodology and its validation.

- Past imperfect is a phrase to examine the possibility of our recollection being incorrect as it often is i.e., we often remember from hindsight rather than from the actual experience. The experiencing self is much more accurate, and the remembering self is subject to two main flaws, peak end and duration neglect. This is not relevant to this research primarily due to it being primarily quantitative research. If quality-of-life measures were used, based on individuals’ recollections then this bias would merit consideration.
- The way probabilities are expressed affects our judgment. Kahneman gives the example of the risk of dying when undergoing a medical procedure e.g. 10% of patients will die versus 10 out of one hundred patients will die. He suggests that the latter is seen as a higher risk. This was considered when reporting the results of the research. Other errors based on utility theory and prospect theory are not obviously applicable to this research.

10.19.13.6 Create entries in the experiments table

In the validation phase, two experiments were set up. First, comparing two cohorts at age 12/13/14 - one having received 2 school dental screenings beginning at age 7/8/9, and one receiving 3 screenings. Second, an experiment was created comparing cohorts at age 12/13/14 - having received their first school dental screenings at age 7, 8, or 9. Cohorts were assessed using the decayed, missing and filled teeth (DMFT) index. The data was not adjusted for factors that can confound DMFT.

10.19.13.7 Add experiment-specific DQ issues to the register.

None arise.

10.19.13.8 Identify DQ ‘Showstoppers’ and mark experiments with them.

If appropriate, the experiment’s showstopper was marked to indicate that there is a showstopper entry in the DataQualityIssuesRegister. This means that data marked with this DQ issue was excluded from the research. The DQ issue was, ‘All entries in PMTreatments must have a corresponding Client in PMClients’ .

10.19.13.9 Validation - Phase 2: Applying the CP-DQF to the research data.

Add Metadata to the research data.

In this implementation, two additional fields were added, a Boolean called BadRow and a vector string called BadRowCodes. The vector string can hold multiple appended error codes if required. This structure is represented in Figure 5 below.

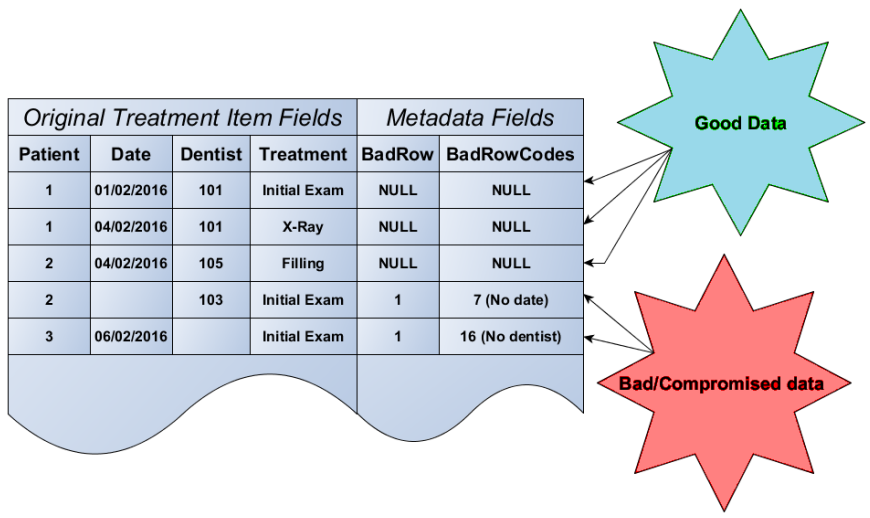


Figure 10-33: Example of Research Data with Metadata added
Pre-processing or discussion section.

The DQ issue here is a data integrity issue and accordingly was dealt with in the DQ pre-processing section.

Decide which of these issues disqualifies the data from use in the experiment.

If the DQ issue was serious for any specific experiment, the experiment was marked, and the data excluded from use. In this case, all treatments should have a valid active client. If there was no client associated with a treatment, vital information was missing e.g. the age of the client. Therefore, this issue disqualified the data from use in this age-dependent experiment.

Evaluate the effect of these data disqualifications.

What was the effect of these disqualifications? Did it require re-execution of previously executed marking or mitigation code? Did it skew results? Depending on the extent of the issue and the underlying causes, this might have caused skewing of the data. No evidence of this was seen in this experiment.

Write/Run the Marking Code from the CP-DQF against the data.

Executing the code stored with the DQ issue in the register marked the research data's metadata with information about its DQ.

Write/Run the Mitigation Code from the CP-DQF against the data.

No mitigation code was applied directly to the data at this point. However, the fact that the data was now annotated with DQ information allowed exclusion of specified data from individual experiments which was intended to have the effect of mitigating the DQ issue. Code to directly mitigate the DQ issue e.g. imputation of missing values is being developed.

Update DQ issues register with the results.

Here, the scope of the DQ issue and the scope of the mitigation efforts were recorded and added to the DataQualityIssuesRegister for reporting purposes.

Write/Run the Cohort Selection Code.

Cohort/Dataset selection code was executed incorporating the metadata as a criterion for exclusion/inclusion in the dataset. In our implementation, treatment events were only selected if the metadata, `BadRow`, is NULL.

10.19.13.10 Step 3: Report on Step 1 & 2.

After executing steps 1 & 2, it is important to know the scope of the DQ issues and a report showing DQ metrics can be run against the DataQualityIssuesRegister to achieve this. The report should list the issues in the register along with frequency and percentage data affected. This may flag issues needing attention and a root cause analysis might be needed leading to improvement steps and better future DQ. The predominant metric used shows a 'percentage' indicating the scale of the DQ issue against the total number of rows. Practically, this only applies to DQ issues at the row or field level. Other metrics, e.g. those comparing calculated values such as mean, median and distributions to expected values are also calculated at this step.

A sample of the data issues in the registry and the number of rows affected is shown in **Table 10.11** below. The complete and detailed list is in Appendix 10.17.

Table 10-11: Sample of Data Quality Registry entries

Data Quality Issue Name	No of rows
All entries in PMTreatments must have a corresponding Client in PMClients	48330
All entries in PMChart must have a corresponding Client in PMClients	3267
All entries in PMAppointments must have a corresponding Client in PMClients	240449

All entries in PMAttendances must have a corresponding Client in PMClients	679728
All entries in PMQuestionnaire must have a corresponding Client in PMClients	1813
All entries in PMQuestionAnswers must have a corresponding Client in PMClients	50806
All entries in the following tables must have an Appointment in PMAppointments: PMAttendances	292167
All Clients in the PMClient must have a ClientAge between 0 and 100	22444
All entries in the PMTreatments must have a MappedToProcedureNameGroup in the PMProcedureGroupNames table - to reduce noise from rarely occurring procedures (<100 times)	18316
All Treatments in the PMTreatments must have a CompletionDate > =1990-01-01 00:19:02.000	197352

10.19.13.11 Limitations & Future Work

- This research describes a scenario where the researcher has direct access to the data through SQL Server Management Studio. This access allowed addition of the metadata fields to the research data, database scripting, and inclusion of additional clauses in the cohort selection process etc. The current framework design incorporates assumptions based on this scenario. Different research scenarios may require alternative approaches, for example, storing the DQ metadata in distinct and separate tables or locations, or database normalization measures.
- The proposed database design fulfils the requirements of the application of the CP-DQF in this research. Other scenarios may require redesign. Simpler case-studies may only require the DataQualityIssuesRegister while more complex scenarios may require further normalization of the database to improve data integrity and reduce data redundancy. It is unknown how this would impact the performance of cohort selection queries.
- This research deals with data from a single, homogeneous EHR source. Consideration needs to be given to additional DQ matters such as ‘Variety’ in scenarios with complex, multi-source, multi-institution research projects using heterogeneous data sources - perhaps as approached by Knowlton et al. (2017).
- This PM research used the DQ dimensions from Mans et al. (2015); Incomplete, Incorrect, Imprecise and Irrelevant. Further work to incorporate the dimensions from Kahn et al. (2016) and others could contribute to a more harmonized and generalizable understanding. The CP-DQF framework is customizable allowing the incorporation of these additional DQ dimension, however, the deeper thinking behind these dimensions must be reconciled with the requirements of PM research work to avoid overlap of dimensions and gaps. In particular, the important extrinsic data features such as ‘fitness-for-use’ and ‘relevancy’, which are central to our PM research, need to be included in the framework.
- The design presented here could be developed to further encompass data management in research using EHR data. This might include logging and auditing other elements of the Extract, Transform, Load process, multiple runs of the same experiment, user management and error handling etc.
- While some of the DQ issues can be identified, marked and perhaps mitigated-against in a pre-processing phase of the research e.g. Missing Date-of-Birth, others are less clear-cut, and might only be adequately dealt with by way of discussion e.g. issues caused by clinician bias, researcher bias, or data model deficiencies. The distinguishing line between these types of issues is undefined and would benefit from further work. It seems likely that many of these types of issues may be difficult or impossible to automatically identify and mitigating these issues may be multi-faceted and require root-cause analysis.

- Future work can include approaches from latent class imputation to mitigate missing data.
- The results presented have focused on a small number of easily quantifiable DQ issues with the easily established metric of ‘% affected’. More complex DQ metrics as detailed above are in development.
- Further metrics could also be added to the data based on the method of DQ assessment employed (Weiskopf & Weng, 2013) e.g. gold-standard assessment methods would give the overall DQ a higher rating.
- Assessing whether exclusion of the quality-affected data impacts the outcomes of specific research experiments would be useful.
- Specific and detailed questions on DQ could be developed and embedded within the live EHR e.g. To the Application Users - *“Is there any possibility that Date of Birth has been incorrectly recorded?”*

Conclusions

The design for the CP-DQF and its application in this research has been presented. It is implementable as a software tool that can be used to manage the DQ issues of research using EHRs. In this thesis the CP-DQF framework has been applied to a large dental EHR and the framework proved useful in providing a structured method to identify and document issues following the DQ dimensions established by the existing literature, notably by Weiskopf & Weng (2013) and Mans et al. (2015). Our example illustrates how code to mark the data to mitigate DQ can be implemented. Intimacy with the data was helpful in identifying many of the information sources and data issues. The case study also showed DQ issues linked to individual experiments in the research and how this can cause affected data to be excluded if appropriate.

The CP-DQF framework has the functionality to be used as an audit trail tool for all data transformations and data cleaning activities. This would satisfy the demands for greater transparency in the pre-processing of EHR-data in preparation for research. By slightly varying the cohort selection criteria, it is also possible to compare research results before and after the exclusion of bad quality data the impact. While the framework was prototyped in the Microsoft SQL Server environment, researchers in other environments could easily replicate this design. The entity design is simple but effective and the dictionaries of sources, dimensions and levels can be tailored to the research.

Use of the CP-DQF may help researchers think about the potential DQ issues in their research, log and manage them in a structured environment, create an audit trail for data transformations, assess and mark their data with quality information, mitigate the issue if possible, exclude data from their experiments if appropriate, compare before and after research outputs and finally, report on DQ metrics.

This will lead to known and more robust EHR DQ, a secure audit trail of DQ transformations, reproducible research steps and more reliable PM results.

Research conclusions can and should be informed by a rigorous assessment of DQ and a structured and auditable approach to marking and mitigating DQ issues. Our framework provides a useful starting point for other PM researchers to address EHR DQ concerns.

10.20 Dental Literature Review Details

Each of the themes in Table 2-1 will now be briefly explained.

10.20.1 Process Types

Process types can be broadly categorised into Medical Treatment Processes and Organisational Processes. Medical Processes can be further categorised into non-elective care and elective care including standard, routine and non-routine processes (Lillrank & Liukko, 2004) (Mans, et al., 2015, p. 13). In (1) the process analysed was that of

diagnosis, placing of implants and the placement of the final restoration. The data was extracted from two information systems, the dental practice's appointment system and from the steps recorded in the dental laboratory producing the crown. As this treatment is voluntary and the process contains elements of both medical treatment and organisational steps, it is an elective treatment/organisational process. In (2), the authors are using PM to analyse the effects of digital technologies on their business processes of 'crown' and 'prosthesis' placement. In (3), the effect of IT upgrades on business processes is analysed.

10.20.2 Data Sources

Using the classification proposed by Mans et al. (2013), data can come from administrative systems e.g. accounting, from clinical support systems e.g. any department specific information system, from healthcare logistics systems such as operational support systems and data from medical devices such as X-Ray machines. Each of these systems, having different objectives and functions, tend to store information at varying degrees on a spectrum of abstraction, accuracy, granularity, directness and correctness. According to Mans, this makes these systems 'more' or 'less' suitable for answering the types of PM questions likely to be posed. For example, administrative systems that are primarily concerned with billing might only record the date on which a procedure was performed. This information will be enough to ensure that the hospital is reimbursed for the service. However, a clinical support system may need more accurate information e.g. what time the last blood pressure test was executed, and hence its information may be more granular. Again, according to Mans, an X-ray machine may automatically collect this information to the millisecond and hence may be both highly granular and accurate. This author believes that there is some overlap between 'data suitability' described above and 'data quality' as described in the work of Mans et al. (2015) and Bose et al. (2013) where four broad data quality issues that could exist in Event Logs were identified: missing data, incorrect data, imprecise data and irrelevant data. The issue of data quality is comprehensively dealt with in Section 5.3.

Publication (1) utilises a combination of administration and clinical support data. While it is likely that X rays were part of the process, there is no specific mention of data from the X-ray machine. The research is primarily explorative however, an analysis of the suitability of the data from the information systems would help highlight limitations of the research at an early stage. (2) follows a similar path while (3) also uses data from the Computer Aided Design (CAD) software. All three articles used data relating to implants, prosthetics and crowns.

There is no discussion of the suitability of the information systems to provide data for the specific questions being asked and this would be useful for future PM studies.

10.20.3 Frequently posed questions

Mans et al. (2015) maintain that PM allows medical process specialists to respond to frequently posed questions about these processes and categorised these questions as follows:

- What happened? i.e. process discovery.
- Why did it happen? e.g. why did this patient deviate from the normal process?
- What will happen? e.g. what is the likely process in the circumstances?
- What is the best that can happen? i.e. how can the process be improved?

More specific to healthcare is the study of questions frequently asked by medical professionals in Mans et al. (2013). Analysing previous PM studies, they established the following questions:

- What are the most followed paths and what exceptional paths are followed?
- Are there differences in care paths followed by different patient groups?

- Are internal and external guidelines being complied with? This is relevant to this research as the potential of PM to assess compliance with guidelines is investigated.
- Where are the bottlenecks in the process?

In the literature, (1) is asking what happened in the process of placing of implants and restorations? Log filtering is applied to exclude unusual events and focus on the paths most likely to be followed. They also examined some aspects of the 2nd category, ‘Why did it happen?’ (2) & (3) also asked what happened and both publications, focussed on the impact of IT on the dentistry process, also deal with categories 3 and 4.

The prospect of these questions being answered is determined by the suitability of the information systems supplying the event logs.

10.20.4 Process Mining Perspectives

The process discovery or control flow perspective, which establishes the order in which activities are executed is used in all 3 of the publications. In (1) it is unclear whether the organisational perspective is distinct from the resource perspective as the terms seem to be used interchangeably. Cross-organisational PM is introduced by including the dental laboratory in the value stream.

The performance and resource perspective are used in (1) (2) & (3) to establish who performs which steps and their duration.

10.20.5 Process Mining Tools

The commonly used PM tools are outlined in Section 2.2 above. All three papers used the ProM tool with no discussion of the alternatives.

10.20.6 Techniques and Algorithms

ProM is a software framework providing many possible analysis techniques or algorithms to produce, optimise and analyse discovered processes. In their analysis of PM in healthcare, Rojas et al. (2016) tabulated the techniques used in their reviewed literature with the Heuristics Miner and the Fuzzy Miner being the most commonly used.

The Heuristics miner and the Social Network Miner were used in Publication (1) and the resulting Petri-nets were analysed with the ‘Performance Analysis with Petri-net’ plugin. Publications (2) & (3) did not specify the algorithms used though Petri-nets were demonstrated. Mans et al. (2013) suggested that appropriate algorithms may not be available for specific requirements and researchers may have to develop these as needed. De Weerd et al. (2012) published a quality assessment of state-of-the-art process discovery algorithms, capable of producing Petri nets, using real-life event logs. They assessed how the various algorithms performed for accuracy and comprehensibility. They described how the Heuristics miner is especially suited for real-life settings supporting the dental research author’s choice. Only assessing algorithms capable of producing Petri nets excluded widely used algorithms such as the Fuzzy, Workflow and Inductive miners.

10.20.7 Methods

While there is no PM method specific to healthcare (Rojas, et al., 2016), there are six main established methods for general PM projects, the L* Life Cycle method (IEEE, 2011), a Process Diagnostics Method (PDM) (Bozkaya, et al., 2009) which addresses some of the complexities of healthcare processes, Business Process Analysis in Healthcare Environments Methodology (Rebuge & Ferreira, 2012) building on the PDM above, PM² (van Eck, et al., 2015) and a Question-Driven Methodology for Analyzing Emergency Room Processes using Process Mining (Rojas, et al., 2017). A further recent approach using discrete event simulation was introduced by Johnson et al. (2018). In the

dental research literature, neither publications (1) nor (3) identified which methodologies were used. In (2) Mans et al. (2013) referred to existing methods but pointed out that none aimed to evaluate changes within the process. They then aimed to develop a method with PM combined with discrete event simulation. All three research efforts could have benefited from applying the discipline inherent in the above methods.

The PM² method offers a structured method with 6 phases: planning, extraction of data, data processing, mining & analysis, evaluation and process improvement and support. Each of the 6 phases has pre-defined inputs and outputs. RQs are derived from project goals, are answered by performance findings and lead to improvement ideas. While acknowledging that the PM² was published after the dental PM literature, these stages will be now looked at in more detail to see how the dental literature addresses them.

10.20.8 Planning

The PM² method proposes three main activities for the planning phase: identifying the RQs from the project goals, selecting the business processes and composing the project team. The following sections look at the extent to which these activities were carried out in the dental PM literature.

10.20.9 Research Questions (RQs)

Publication (1) outlines its purpose to be to demonstrate the usefulness of PM for the domain of dentistry. It is unclear how this will be demonstrated from the outset as no specific RQs have been listed and the research is taking this more abstract, explorative approach. When defining the RQs, it should have been possible to define an ideal dataset, i.e. a dataset that contained all the information required to answer the RQs. This could be developed cognisant of the Healthcare Reference Model (Mans, et al., 2015, pp. 27-52) and Dental EHR Standards (American National Standard/American Dental Association, 2013). This research proposes an additional data reference model in Section 7.6.6 which could be used in future work. This should have resulted in useful artefacts such as data mapping documents, entity relationship diagrams and a gap analysis to facilitate effective PM of future Dental EHR implementations. Further, it would have been useful to describe the data-set in classical data mining terms or schemas such as the star and snowflake schemas (Santos, et al., 2013). The star schema presents the data as a central fact-table linked to several dimension-tables whereas the snowflake schema has additional hierarchical detail with some of the dimensions. The research data in this thesis is presented in the snowflake schema in Figure 4-3 with the PMTreatments table as the fact-table in most scenarios.

10.20.10 Selecting Business Processes

PM² identifies process characteristics and data quality as having a large influence on the achievable results and refers to the work of Bose et al., (2013) where four broad data quality issues that could exist in event logs (ELs) were identified: missing data, incorrect data, imprecise data and irrelevant data. These were further detailed in 27 types of quality issues and the likelihood of its relevance in a specific dataset by Mans et al (2015). Additionally, the Process Mining Manifesto (IEEE, 2011), proposes a rating system indicating data quality. It proposes a quality assessment ranging from 1-star to 5-star. Three-star systems typically automatically record events. The log is deemed as trustworthy though not necessarily complete. Examples are tables in Enterprise Resource Planning (ERP) systems, event logs of Customer Relationship Management (CRM) systems etc. Event logs resulting from traditional Business Process Management (BPM) workflow systems might be considered for 4-star status whereas the 5-star status is

reserved for logs that are trustworthy and complete, events are automatically recorded, well defined, systematic and have clear semantics (IEEE, 2011, p. 7). Addressing this point, researchers could have mapped their event names to the Standard Nomenclature for Dentistry (SNODENT) having the effect of encouraging reproducible research and allowing further work to build on theirs in a predictable fashion. This rating system might have provided a useful overview of the quality of the data used in this research. With the exception of publication (1) detailing a method, resulting in a ProM plug-in, to solve an issue where the same event was being referred to by different names, data quality issues were not addressed in the dental research literature.

10.20.11 Composing the Project Team

The final activity of the planning phase is ensuring that the correct personnel are in place, including business owners, business experts, system experts and process analysts. Clear definition of the roles is helpful and ensures that the correct stakeholders are involved when required. The writers clearly possess expertise in the PM and dentistry domains. Their interaction with the process owners i.e. the dental practice and the dental laboratory is limited to validating the results of the PM. Additional detail would be useful in this area e.g. Was this an iterative process? Mans et al. (2015, p. 3) suggested that traditional methods of gathering the information required for process analyses by observation and interview are flawed due to their subjective nature. It is unclear if the authors of the dental PM literature dealt with potential issues such as bias and subjectivity introduced by the project team as proposed by Chenail (2011) and Pannucci & Wilkins (2010).

10.20.12 Extraction

The PM² method defines the extraction stage as the extraction of event data and optionally process models from the information systems. This includes scoping the data required from a granularity and detail perspective and defining the appropriate time-frame. The authors scope the data required as ‘a group of patients with an implant-borne, single crown restoration’. This is a process known as ‘single crown on implants’ and involved a collaboration with one dental laboratory. Both these organisations provided the researchers with a log and the 55 patients involved were matched up manually. The time period defined was 2008 to 2011 so the criteria for scoping was clearly defined in the paper. PM² describes a final step of ‘Transferring process knowledge’ where tacit information is exchanged between business experts and process analysts. This phase enhances the analysts’ effectiveness in the mining and analysis phases and may provide some *de jure* process models as output. Clearly, the researchers/analysts had contact with the business owners but some documentation of its structure and nature would be useful. This is closely related to the planning phase described above.

10.20.13 Data Processing

The event data as extracted above may not yet be ready for mining and analysis. Subsets of the event data may be required or there may be data issues that will be dealt with at this stage. PM² identifies several steps that may be executed at this phase: creating views, aggregating events, enriching logs and filtering logs. No information regarding specific views of the event data is given by the writers. PM² describes two distinct types of aggregations, ‘is-a’ and ‘part-of’. Aggregating events was carried out as detailed in the data quality activity in the planning phase above. Similar events having differing names were consolidated and this can be seen as an ‘is-a’ aggregation. No ‘part-of’ aggregation was documented. The log in use is a rich log, providing insights on performance and social networks within the organisations as well as the control-flow perspective. No

specific detail is given as to the process of log enrichment. It may either be that the enriched log was presented to the researchers as a *fait accompli* or as a series of related tables that the researchers queried as required. Here again, the planning artefacts such as entity relationship diagrams would be useful as well as additional transparency in the Extract, Transform & Load (ETL) sequence. Filtering is the final phase of data processing proposed by PM² and includes slice-and-dice, variance filtering and compliance-based filtering. Slice-and-dice allows inclusion or exclusion of data based on the values of attributes e.g. time. Variance based filtering groups similar traces with the objective of partitioning an EL to reduce the complexity of resulting models. Compliance based filtering removes traces or events based on rules or their compliance with a given process model. The dental researchers initially encountered spaghetti-type process models and, to arrive at a comprehensible model, applied a strategy where only events that occurred in more than 10% of the process instances were included. This is a type of slice-and-dice filtering. Unfortunately, no discussion was held on the value of the discarded data. Perhaps the deviant processes are also interesting, and it is certainly worth consideration. There is no analysis as to what information was lost in this process, nor its value. It would be essential to assess the omitted information with the help of domain experts.

It is known that many high-tech systems produce logs of very fine granularity leading to spaghetti-like process models. Case heterogeneity can also produce spaghetti-like models. Higher levels of abstraction can be achieved using ontologies (Pedrinaci & Dominique, 2007) e.g. SNODENT. Trace clustering has been shown to be effective i.e. partition event logs into subsets of homogeneous cases (deLeoni, et al., 2016). It would be useful to try these methods before taking the above filtering step.

To enhance the reproducibility of the research, at this stage the authors could have considered the mapping of their events to a standard terminology or ontology such as SNOMED/SNODENT or a treatment specific vocabulary such as the Glossary of Oral and Maxillofacial Implants (International Team for Implantology, 2007).

It would also have been appropriate to consider a data description document such as that available for the critical care database MIMIC III (Johnson, et al., 2016) and as presented for the data in this research in Section 4.1.5, 4.1.6. This would provide useful information on patient characteristics, data classes, description of the data situation and movement, anonymization and de-identification, legal and ethical issues, use cases etc.

10.20.14 Mining and Analysis

PM² identifies four activities that take place at this stage: process discovery, conformance checking, enhancement and process analytics. The dental research is primarily explorative process discovery using the Heuristic Miner and converting the heuristics net to a Petri net. They justify their use of the Heuristics Miner as it can deal with noise and exceptions and allow users focus on the main process flows. While this may be the case, there are many other algorithms such as the Fuzzy and Enhanced WF Miners also offering this functionality (De Weerd, et al., 2012) and it would be interesting to investigate their suitability also.

The quality of a discovered process model can be assessed for 'fitness', measuring how well the discovered model fits the event log (Rozinat & van der Aalst, 2008). The dental researchers found a fitness measure of 0.95 indicating a high model accuracy. There are also several 'Accuracy' quality metrics under the headings of 'Recall' and 'Precision' applicable to discovered models (De Weerd, et al., 2012) e.g. completeness, soundness and behavioural appropriateness. De Weert notes that as there is no 'Generalisability' metric, Accuracy is defined as a function of Recall and Precision exclusively. Additionally, the Process Mining Manifesto suggests adapting data mining techniques such as cross-validation of the model to judge the quality of the output (IEEE, 2011) & (Rovani, et al., 2015). Furthermore, De Weert suggests several metrics measuring the

comprehensibility of a process model. This involves counting the number of transitions, places, joins, splits etc. The dental researchers have not documented their use of such objective measures of model comprehensibility, rather they chose a subjective filtering of events to ‘arrive at a comprehensible model’.

As part of the conformance checking, the discovered model could have been examined *vis-à-vis* clinical guidelines for crown/implant processes (Rovani, et al., 2015).

No justification for the selection of the Social Network Miner for resource analysis was given, although the results were validated by interview with participants. For performance analysis, the ‘Performance Analysis with Petri-net’ was chosen. This projects timing information, averages, and standard deviations without incorporating business rules. This might lead to misleading information as it will not expose rules e.g. ‘3 months must elapse before fitting crown to implant’. This was not addressed in detail in the research.

10.20.15 Evaluation

The evaluation phase takes the discovered process models and the performance and conformance findings as input and aims to find process improvements achieving the projects goals. Apart from validating the process models with the process owners, this phase of PM was not considered by the researchers – the aim of the research being to evaluate the usefulness of PM for analysing dental workflows.

10.20.16 Process Improvement and Support

This phase aims to use the discoveries to improve the studied processes. This phase of PM was not considered by the researchers.

As this is a healthcare setting, other steps above and beyond an industrial setting should be considered in these research papers. The issues of ethics were not mentioned nor was the requirement for patient consent for the use of their personal data. If the research did not require ethical approval, this should have been clarified. If patient consent was not required, then details of the waivers or data anonymization would have clarified this.

10.20.17 Implementation Strategies

Rojas et al. (2016) distinguishes between 'Direct Implementation' and 'Semi-automated Implementation' and 'Integrated suite Implementation' in PM experiments or projects.

These strategies all involve the same steps. They are distinguished by who is responsible for the distinct steps. In direct implementation, the researcher (or end-user) designs and executes the queries to generate the EL and then applies PM techniques and algorithms. In semi-automated implementations a third party gives the completed EL to the researcher who then applies PM techniques and algorithms. Finally, in an 'Integrated Suite' implementation, both the EL creation and the application of the PM techniques and algorithms are 'under-the-hood' and the end user just interprets the results.

It is unclear whether the dental researchers adopted the direct implementation or the semi-automated approach. It would be beneficial to know if the researchers had direct knowledge of and access to the data of the dental practice and the prosthetics laboratory for creation of the EL. This might have also given rise to some formal agreements with the software suppliers, intellectual property owners or data-owners.

10.20.18 Analysis Strategies

Three analysis strategies are identified by Rojas et al. (2016). A basic strategy takes an EL and applies pre-existing techniques and algorithms available in the PM tools. The second strategy involves the development of new techniques or algorithms specific to the field or questions being asked to find novel ways to deal with complex process and data.

The third strategy additionally incorporates knowledge from other domains such as data mining, Online Analytical Processing (OLAP), ontologies and simulation models. Publication (1) used an advanced strategy with new plug-in and some semantic analysis to improve the data quality. Publication (2) involved discrete event simulation while Publication (3) is not specific.

10.20.19 Geographical Analysis

All three publications originated in the Eindhoven University of Technology, Netherlands, ACTA Amsterdam, and data from a local practice and dental laboratory.

10.20.20 Medical Fields

All three publications used the dental implants and prosthetics/crowns process as their case study.

10.21 Application of the ADF

The ADF steps are addressed in turn in the following section.

1. Describe the Data Situation

This research is a dynamic data situation. The data was extracted from a live database and anonymised for sharing with the researchers based in the University of Leeds and University College Cork. The data, after anonymisation, was physically personally transported to the two university locations by the author. The research has resulted in a thesis, conference papers, and potential journal publications where aggregated and tabulated data is presented. It is hoped that permission will be granted for open access to the anonymised data subset. The data situation is summarised in Figure 10-34 below.

A secure server at the OHSRC is used to store electronic data related to the project. The data is encrypted using Windows BitLocker. Access is password protected.

The anonymised data has been encrypted and personally transported to the Leeds Institute of Data Analytics (LIDA) <http://lida.leeds.ac.uk/> in the University of Leeds for further analysis. LIDA is underpinned by an enabling technology platform, the Integrated Research Campus (IRC). The IRC is an advanced computational infrastructure that is highly secure and scalable, to meet the needs of data-intensive research using personal and sensitive data securely.

Following a process of external independent assessment, the IRC has attained accredited certification to the international standard for information security management, ISO/IEC 27001:2013 and will meet U.K.'s NHS Information Governance Toolkit level 3.

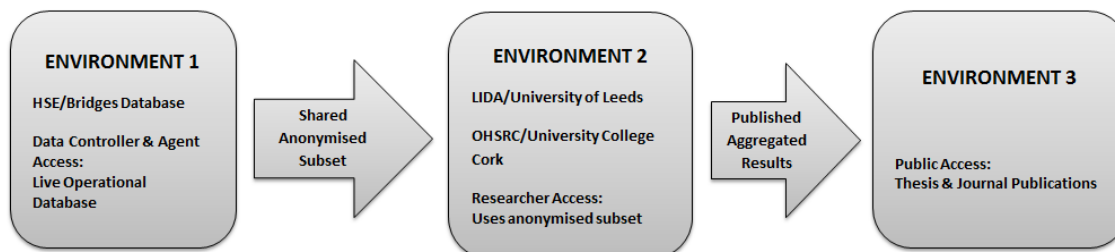


Figure 10-34: Data Flow between multiple environments

2. Understand your legal responsibilities

As the data was transferred from the HSE in Ireland to University College Cork and subsequently to the Leeds Institute of Data Analytics, University of Leeds, the writer has considered the legal and governance issues in both jurisdictions.

Irish Legislation and Guidelines.

The study complied with the requirements of the Data Protection Act 1988 and the Data Protection (Amendment) Act 2003; Data Protection (Access Modification) (Health) Regulations, 1989 (S.I. No.82 of 1989); Data Protection (Access Modification) (Social Work) Regulations, 1989 (S.I. No.83 of 1989); Council Directive on the Protection of Individuals with Regard to Processing of Personal Data (Directive 95/46/EC) (W). The Data Protection Guidelines on Research in the Health Sector (2007) were also carefully considered. The GDPR was also considered. The proposed study received full ethical approval from the Clinical Research Ethics Committee of the Cork University Teaching Hospitals (CREC) on August 2nd, 2016. Following consultation with the Office of the Data Protection Commissioner, the research was approved by the HSE's Primary Care Research Committee (PCRC) at their monthly meeting on 17/01/2017 with conditions that the researcher not be involved in the data anonymisation process.

UK Legislation and Guidelines.

The study complies with the requirements of the U.K.'s Data Protection Act 1998 and the U.K.'s common law Duty of Confidentiality. The Information Governance Framework was adhered to and the IBS Anonymisation Standard also. The transfer of data from Ireland to the UK does not raise any additional issues as both jurisdictions were subject to the European Union Data Protection Directive (95/46/EC) and the GDPR. The international dimension does not currently add to the complexity. The impact of the UK's planned exit from the European Union is currently unclear.

This research is a secondary use of the data. The data was primarily gathered to manage the care of patients. It was acquired during the treatment of the patients. Before release to this researcher, ethical clearance was obtained, the research proposal was cleared by the Office of the Data Protection Commissioner in Ireland and the data was anonymised. As data-owners and controllers, the HSE was satisfied that the data was anonymised.

3. Know Your Data

The ADF suggests a series of questions and provides a data features template to help define the data.

Table 10-12: Data Features from ADF

Feature Type	Question	Answer/Actions
Data Subjects	Who are they?	School-going children, usually under 16 and special-needs adults
	What is their relationship with the data?	Data is a by-product of routine dental treatment and operations
Data type	Microdata, Aggregates or something else	Microdata, individual patient records.
Variable Types	What common indirect identifiers to you have?	Date of Birth, Clinic, Nationality
	What sensitive variables do you have?	Dental Clinical Information Medical Questionnaires
Data Properties	Is the data Accurate?	Data Quality is addressed in the research Section 5.5
	How old is the data?	Data relates to dental treatments administered approximately from 2000-2015
	Is it hierarchical or flat?	Hierarchical
	Is it Longitudinal or Cross-Sectional?	Longitudinal
	Population or Sample? (what fraction)	School screening population of Counties Cork & Kerry. Could be interpreted as a sample of the Irish population.

Anything else of note	Data is anonymised hence, consent is not required (Confirmed by the DPC's office, Ireland)
------------------------------	--------------------------------------------------------------------------------------------

4. Understand the use case

This study looked at the process of delivery of dental healthcare with and extract of data held in the HSE/Bridges dental EHR using the emerging technology of PM. This is a non-interventional study to develop new methods for health services research using historical data from the Bridges (EHR) in the Health Service Executive, Ireland (HSE South, Counties Cork and Kerry). This research examines how PM can deliver worthwhile insights to dental policy makers and develop a roadmap for executing a PM initiative.

PM research is concerned with the extraction of knowledge about a healthcare or business process from its process execution logs. PM aims to discover, monitor and improve real processes by extracting knowledge from event logs readily available in today's information systems.

Process Discovery and Mining provide a method to find out what is happening on the ground without having to employ the traditional tools of observation and questionnaires. The activities taking place in the dental clinics leave markers in the EHR which can be subsequently examined and provide an insight to the care pathways

The research investigated whether PM of the dental EHR can be used to compare execution of treatment processes with ideal and proposed processes. It attempted to measure the impact of specific policy and strategy changes on the process of delivery of dental care experienced by patients. The effect of policy around frequency of school dental screening is examined. The research examines the value of PM to the theoretical monitoring of the impact of policy changes on service delivery and investigates the levels of information-granularity in the proposed guidelines and the data being collected by the EHR to identify differences between the two.

The data required along with an explanation of why the information is required, steps that have been taken to de-identify the datasets, and the anonymization/de-identification process is described in Table 10-13.

Table 10-13: Data Extracted and Anonymisation Steps

Information extracted from the Bridges database	Explanation	Anonymisation / de-identification
BRIDGES reference number	To denote an individual record and facilitate link between several courses of treatment for one person.	Replaced by an unique GUID (ID) within the HSE
Gender	To provide a high-level view of the proportions of each gender.	Unchanged
Age	To allow analysis of process variation by age	Unchanged
Clinic & Region	To allow analysis of process variation by clinic, e.g. some clinics & Regions will not have implemented the PHN Initiative.	Unchanged
Treatment Plan & Chart	To link between the individual treatment items (procedures) and thereby facilitate analysis of the complete treatment plan	Unchanged

Procedure name & details	To position individual treatments within the complete dental process	Initially unchanged. May be permanently mapped to SNODENT or similar standard nomenclature
Appointment start/end times	To crudely estimate time spent on a procedure	Unchanged
Who carried out procedure?	To assist in identifying bottlenecks or resource usage e.g. dentist or dental hygienist	Unchanged (No name data included in dataset)
DMF- Teeth and/or surface conditions. Medical histories.	Outcome and pre-treatment measures. Will be of value in assessing the significance of process variation.	Unchanged
Base name data tables with names of Regions, Clinics, tooth types, tooth condition types, appointment types etc.	Allowing translation of codes stored in treatment & chart tables etc.	Unchanged

5. Meet Your Ethical Obligations

The study proposal received ethical approval from the Clinical Research Ethics Committee of the Cork University Teaching Hospitals (CREC) on August 2nd, 2016. The HSE's Primary Care Research Committee (PCRC) approved data access on 17/01/2017 with conditions that the researcher not be involved in the data anonymisation process. The ethics application documentation is in Appendix 10.4.

6. Identify the processes you will need to assess disclosure risk

The ADF recommends the use of scenario analysis. Given that the objective of anonymisation is to prevent re-identification of individuals from the data, it suggests putting ourselves in the villain's shoes. What resources does the villain need to re-identify individuals? In the current environment, it is difficult so see how any individual could be identified from the data being used in this research. No direct identifiers have been included in the research data subset. Internal client identifiers including GUIDs have been deleted and replaced with new GUIDS. No cross-reference table between the old and new GUIDS exists. From a practical perspective the data has been irreversibly de-identified. Why would a villain attempt to re-identify the data? Spiteful breaches are often the case. There might be a wish to embarrass the data controller. There may be a motivation to steal information on individuals' oral health status for marketing purposes. There may be a motivation to steal demographic information about children.

RISK: If the villain had access to the underlying HSE/Bridges database it is conceivable that, with intimate knowledge of the database structures, entities and attributes, queries could be designed that would identify an individual by virtue of the fact that dental records are distinctive and often unique. This scenario is extremely unlikely. Any such villain would already have access to a much richer and complete, un-anonymised dataset, making such an attack virtually pointless.

The Information Standards Board for Health and Social Care (2013) defines Statistical Disclosure Control as: "*Techniques for obscuring small numbers (e.g. less than "5") that appear in aggregate tables so as to prevent re-identification*".

This means that aggregations at a level that pose a re-identification risk cannot be created. There are four main statistical disclosure attack techniques: identification, attribution, subtraction, and table linkage against which appropriate steps have been taken, primarily by ensuring that any aggregated cells have sufficiently high membership to minimise the risk of re-identification of individuals.

The ADF also suggests using a comparative data situation analysis. If the risk in the new environment is less than in the original environment, then it is probably acceptable to say that it is safe. In this case, all primary identifiers have been removed from the data, access

is restricted to the research team in the secure surroundings of UoL/LIDA and UCC/OHSRC. The complete data was previously open to a significant number of administrative staff and primary providers in several clinics and it is difficult to see how the disclosure risk is higher in the research environment.

As further security measures, one could have considered the use of penetration tests to see if re-identification is possible, perhaps using a crowd sourced hacking challenge. A thermostat approach to releasing data is another option - starting with very cautious risk & developing a slightly more liberal approach as confidence in our methods increases. The use of the data is currently restricted to this specific research and the author believes that further consideration of these methods is not necessary in the scope of this research.

7. Identify the relevant disclosure control processes (The data environment)
In this research the risk is reduced by placing controls on the data. The anonymised data has been encrypted and personally transported to LIDA in the University of Leeds for further analysis. This is further detailed in Section 4.1.4 above.
8. Identify stakeholders and plan communication
To help in the process of building trust and credibility, it is important to establish effective communication with the stakeholders. This requires us to first identify who needs to know about the data share. In this case the stakeholders are the patients (the data subjects), the Data Controller (the HSE), the Universities. It is our intention to keep the data-owner informed regarding the research's publications. Details regarding the anonymisation process will remain unpublished to reduce the hacking risk. Data subject consent is not required as the data has been anonymised and as confirmed by the DPC's office.
9. Plan what happens after sharing or releasing data
Continuing advancements in IT capabilities require us to remain vigilant and to monitor any use of the data prior to release or publication. There are currently no permissions in place for use of the data beyond this research.
10. Plan what you will do if things go wrong
As this work is not at zero risk it is important that a breach policy is in place. To facilitate this, a robust audit trail of all anonymisation activities and a crisis management plan need to be maintained. The anonymisation plan for this research is fully documented. To reduce the hacking risk, this will not be published. The main concerned party is the original Data Controller (HSE) and the Primary Care Research Committee will be informed as soon as practicable in the event of a breach. What are the likely next steps in the event of a breach? The data is classified as 'normal' risk, is relatively mundane dental treatment records without personal identifiers and is unlikely to attract media attention. The key steps would be to identify the source of the breach and the reasons behind it and take steps to ensure that it is not repeated.