DOCTORAL THESIS

Comparison of different statistical methods for the analysis of patient-reported outcomes in randomised controlled trials

Author:

Yirui Qian

Supervisors:

Prof. Stephen J Walters
Dr. Richard M Jacques
Dr. Laura Flight

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

University of Sheffield
Faculty of Medicine, Dentistry and Health
School of Health and Related Research (ScHARR)

30th June 2023

Abstract

Introduction: Patient-reported outcomes (PROs) that aim to measure patients' subjective attitude towards their health or health-related conditions in various fields have been increasingly used in randomised controlled trials (RCTs). The PRO data is likely to be bounded, discrete, and skewed. Although various statistical methods are available for the analysis of PROs in RCT settings, there is no consensus on what statistical methods are the most appropriate for use.

Research Question: What statistical methods are appropriate for the analysis of PROs in RCTs?

Methods: Firstly, two literature reviews were performed to identify what methods have been developed and applied. The identified statistical methods were then filtered and applied to various RCTs and simulated datasets considering a range of scenarios. Finally, recommendations on what statistical methods are the most appropriate were proposed according to their technical details, and their model performances in the empirical analysis and simulation analysis.

Results: The literature reviews found that the majority of publicly funded RCTs (251/303) included PROs as outcomes, and 114 used PROs as primary outcomes. A total of 29 statistical methods were identified in the two reviews, and they were filtered down to 10 statistical methods with justifications. These 10 methods were described and applied to various RCT datasets for empirical analysis. With the results from empirical analysis, the list of 10 methods was further narrowed down to six that were carried forward for simulation analysis. This study found that multiple linear regression (MLR) was associated with little bias of the estimated treatment effect, small mean squared error, and appropriate coverage of the confidence interval under most scenarios. Tobit regression (Tobit) performed similarly to MLR, and it performed slightly better for the analysis of PRO data with a large number of categorical values. Median regression and its extension were relatively inefficient, producing estimates that were multipliers of the difference between two neighbouring categorical values. Beta-binomial regression (BB) failed to converge for outcomes with a small number of categorical values, but it performed well for outcomes with 10 or more categorical values. Beta regression required the compressing process of dimension scores, and its estimates were more scattered than fractional logistic regression (Frac). Ordered logit model tended to generate numerically large bias and small coverage, especially when the true treatment effect was large.

Conclusion: MLR is recommended as the universal statistical method for the analysis of multidimensional PROs in RCT settings if the mean is the targeted summary measure, and Frac is recommended as the universal statistical method if the odds ratio is the targeted population summary measure. Tobit and BB are recommended as alternative methods for PRO dimensions with 10 or more categorical values.

Acknowledgements

I would like to express my deepest gratitude to my supervisors Prof. Stephen Walters, Dr. Richard Jacques, and Dr. Laura Flight for their invaluable guidance, continuous support, and uplifting encouragement throughout the entire research process, both academically and professionally. Their profound expertise and insightful contributions have been instrumental in shaping this work. It is my great pleasure to have worked with such an exceptional supervisory team.

I extend my sincere appreciation to the ScHARR community, especially the Medical Statistics Group, for providing me with the necessary resources, facilities, and a stimulating academic environment to carry out this research. In particular, I would like to thank Prof. Steven Julious for his support in explaining the standardised effect size, Dr. Praveen Thokala for his advice on constructing the multi-criteria decision analysis for the filtration of statistical methods, and Claire Beecroft for her help with developing the search strategy in method review. I would like to thank Prof. Tracey Young and Dr. Rebecca Simpson for their expert views and constructive advice on the work following the confirmation review. I would also like to thank my viva examiners, Prof. Catherine Hewitt and Dr. Ines Rombach, for their insightful comments and thoughtful examination, which has contributed to the refinement and improvement of this work.

My gratitude also goes to my colleagues and fellow researchers who have contributed to my academic journey. Their support, discussions, and exchange of ideas have been immensely beneficial in shaping my research perspectives. Thanks for everything that has brightened up my life during this journey - the casual talks and hangouts with my PhD fellows; the shared dinners and laughter with my flatmates; and the joyful walks and beautiful parks in Sheffield … These experiences compose great memories that worth looking back throughout my lifetime.

I owe a massive thank you to my friends and family, especially my mum and dad, for their unwavering support, unconditional love, and encouragement throughout my academic pursuit. Their support and understanding during the ups and downs of this journey have been invaluable.

Lastly, I would like to thank myself for making the decision to take this journey. I have pushed myself to overcome obstacles, to stay motivated, and to expand my knowledge and skills in research. This demanding process has undoubtedly made great impact on my work and life.

This PhD was made possible by the joint sponsorship from the University of Sheffield and the China Scholarship Council (grant number 201908890049).

In memory of my grandfather who passed away during my PhD study. His integrity, wisdom, resilience, and courage has been making a profound influence on my academic and personal development.

Research achievements

Publications

Qian, Y., Walters, S.J., Jacques, R.M., Flight L. (2021) Comprehensive review of statistical methods for analysing patient-reported outcomes (PROs) used as primary outcomes in randomised controlled trials (RCTs) published by the UK's Health Technology Assessment (HTA) journal (1997-2020), BMJ Open, 11(9), p.e051673.

Oral presentations

Qian, Y., Jacques, R.M., Laura, F., Walters S.J. Comparison of statistical methods for the analysis of patient-reported outcome measures (PROMs) in randomised controlled trials (RCTs): a simulation study. The 7th UK National Patient Reported Outcome Measures Research Conference, Sheffield, United Kingdom, 2023.

Poster presentations

Qian, Y., Jacques, R.M., Laura, F., Walters S.J. Comparison of statistical methods for the analysis of Short-Form 36 (SF-36) in randomised controlled trials (RCTs): an empirical analysis. The 43rd Annual Conference of the International Society for Clinical Biostatistics (ISCB), Newcastle, United Kingdom, 2022.

Qian, Y., Jacques, R.M., Laura, F., Walters S.J. Comparison of statistical methods for the analysis of patient-reported outcomes (PROs) particularly Short-Form 36 (SF-36) in RCTs using standardised effect size: an empirical analysis. The 6th International Clinical Trials and Methodology Conference, Harrogate, United Kingdom, 2022.

Simpson R., Jacques, R.M., **Qian Y.**, Lewis J., Firth N., Bursnall M., Laura, F., Walters S.J. Intra-cluster correlation for patient-reported outcome measures in individually randomised cluster trials. The 6th International Clinical Trials and Methodology Conference, Harrogate, United Kingdom, 2022.

Qian, Y., Walters, S.J., Jacques, R.M., Flight L. A comprehensive review of statistical methods for analysing patient-reported outcome measures (PROMs) as primary outcomes in randomised controlled trials (RCTs) published by the United Kingdom's Health Technology Assessment Journal (1997-2020). The 5th UK National Patient Reported Outcome Measures Research Virtual Conference, Sheffield, United Kingdom, 2021.

List of Contents

List of Figures

List of Tables

List of abbreviations

AIC	Akaike information criterion
ANCOVA	analysis of covariance
ANOVA	analysis of variance
AUC	area-under-the-curve
BB	beta-binomial regression
BCa	bias corrected and accelerated
BDI	Beck Depression Inventory
BLN	binomial-logit-Normal regression
BP	bodily pain
BR	beta regression
EORTC QLQ-C30	European Organization for Research and Treatment of Cancer quality of life questionnaire
CEA	cost-effectiveness analysis
CIs	confidence intervals
CLAD	censored least absolute deviations
CLT	Central Limit Theorem
CONSORT	Consolidated Standards of Reporting Trials Statement
DGM	data-generating mechanism
EQ-5D	EuroQol-5 Dimensions
FDA	Food and Drug Administration
Frac	fractional logistic regression
GEE	generalized estimating equation
GH	general health
GLM	generalized linear model
GLMM	generalized linear mixed models
HADS	Hospital Anxiety and Depression Scale
HRQoL	health-related quality of life
HTA	health technology assessment
HUI	Health Utility index
ICER	incremental cost-effectiveness ratio
ITT	intention-to-treat
LAD	least absolute deviations
MCDA	multi-criteria decision analysis
MCID	minimum clinically important difference
Median	median regression
MH	mental health
MLE	maximum likelihood estimation
MLR	multiple linear regression

MSE	mean squared error
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health and Care Research
OHS	Oxford Hip Score
OKS	Oxford Knee Score
OL	ordered logit model
OLS	ordinary least squares
OP	ordered probit model
OR	odds ratio
OSS	Oxford Shoulder Score
PF	physical functioning
PHQ	Patient Health Questionnaire
PRO	patient-reported outcome
PROM	patient-reported outcome measure
QoL	quality-of-life
RCT	randomised controlled trial
RE	role limitation - emotional
REML	restricted maximum likelihood
RP	role limitation - physical
SD	standard deviation
SE	standard error
SES	standardised effect size
SF	social functioning
SF-36	Short Form-36
SF-6D	Short Form-6 Dimensions
SISAQOL	Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data
SPIRIT	Standard Protocol Items: Recommendations for Interventional Trials
Tobit	Tobit regression
VAS	visual analogue scale
VT	vitality

Chapter 1 Introduction

1.1 Background

Patient-reported outcomes (PROs) are widely used to measure patients' subjective attitude towards their health or health-related conditions in various fields (Fitzpatrick *et al.*, 1998; Weldring and Smith, 2013). PROs enable health researchers to measure, analyse and compare clinical outcomes from the patient perspective, providing clinical effectiveness outcomes and evidence to support decision making in health technology assessment (HTA). For example, the Oxford Hip Score (OHS) and Oxford Knee Score (OKS) are condition-specific PROs that are compulsory to report for hip replacement and knee replacement in England (NHS Digital, 2022). The EuroQol-5 Dimensions (EQ-5D) and the Short Form-36 (SF-36) are generic PROs that allow the comparison of patients' health status under different health conditions. EQ-5D and Short Form-6 Dimensions (SF-6D), a derivation of SF-36, can be converted into preference-based scores, which is recommended for use in economic evaluation by the National Institute for Health and Care Excellence (NICE). The descriptive and scoring system of PROs transform subjective descriptions of an individual's health to numerical scores in a range of dimensions (Brazier *et al.*, 2016). This transformation quantifies health from patient subject perspective, and allows the statistical analysis on patients' health status through PROs.

There is an increasing trend of using PROs to measure the treatment effect in clinical trials (Qian *et al.*, 2021). Randomised controlled trials (RCTs) are regarded as the golden standard for evaluating the effectiveness of interventions (Altman, 1996; Akobeng, 2005). The randomisation process in a well-designed RCT can reduce the selection bias and allocation bias, and inform the causality of the treatment on responses (Moher *et al.*, 2010). These traits of RCTs can simplify the data analysis of PROs.

Despite these traits of RCTs, it still can be complex to analyse PROs in RCT settings considering the multidimensional structure, and the bounded, discrete, skewed features of PROs. First, one PRO can generate multiple outcomes, and these outcomes can be reported in different forms, e.g., using different score types such as subscales or summary score, producing dichotomised scores from continuous or ordinal PROs, generating quality-of-life (QoL) adjusted survivals etc. The various forms can results in multiple endpoints, and potentially increase the complexity of analysing PRO data (J.-F. Hamel *et al.*, 2017; Pe *et al.*, 2018). Second, PRO data are likely to be discrete, skewed, and bounded (i.e. with ceiling effect and floor effect) (Walters, 2009). When analysing PRO data using a general linear model, including *t*-test, analysis of variance (ANOVA), and linear regression, some model assumptions such as the Normality assumption of residuals are likely to be violated (Lumley *et al.*, 2002; Walters and Campbell, 2004). Also, the application of statistical methods may vary depending on various factors such as the distribution of PRO data and the aim of the statistical analysis, but the multidimensional, discrete, skewed and bounded features of PROs may obscure the decision on what statistical methods need to be applied for the data analysis.

An inappropriate statistical analysis of PROs can result in unreliable estimates of clinical effectiveness and accordingly fail to provide accurate and robust results for decision making, with wider confidence intervals (CIs) and larger errors. For example, patients may fail to receive an effective treatment because this treatment is falsely shown not to be clinically effective based on inaccurate estimates; vice-versa, patients can receive a treatment which may potentially harm their health when unreliable evidence supports the use of this treatment. Therefore, applying appropriate statistical methods for the analysis of PROs in trials is crucial to reduce biases of estimates, to accurately evaluate clinical effectiveness and to support healthcare decision-making.

Statistical methods that are alternatives to general linear model such as bootstrapping (Walters and Campbell, 2004), Tobit regression (Austin, Escobar and Kopec, 2000), beta-binomial regression (Arostegui, Núñez-Antón and Quintana, 2007) and quantile regression (D'Silva *et al.*, 2018) have been applied to PRO data to address some abovementioned issues. However, as each method has its own model assumptions and estimation procedures, it is still unknown which method is the most appropriate to analyse PROs, particularly in RCT settings.

Guidelines have been published by government organisations such as the Food and Drug Administration (FDA) and academic groups such as the Standard Protocol Items: Recommendations for Interventional Trials-PRO extension (SPIRIT-PRO) and the Consolidated Standards of Reporting Trials Statement-PRO extension (CONSORT-PRO) to standardise the use of PROs (Bottomley, Jones and Claassens, 2009; FDA, 2009; Calvert *et al.*, 2013, 2018; Kyte *et al.*, 2016). However, these guidelines mainly focused on the reporting of PROs or on the process of PRO development such as what health dimensions to cover, what items to include, and how feasible, valid, and reliable the PROs are. In the guidance for the use of PROs in medical product development for labelling claims by FDA (2009), a series of statistical considerations for the analysis of PRO data were stated, which to some extent can guide future studies on constructing statistical analysis plans. The guidance for inclusion of PROs in clinical trials protocols by the SPIRIT-PRO Extension (Calvert *et al.*, 2018) also stressed the importance of developing a data analysis plan and reporting missing data for PROs. Coens *et al.* (2020) compared different statistical methods for the analysis of PROs in cancer trials using a range of criteria, and made recommendations on the statistical methods that could be used for the analysis of PROs. However, this recommendation focuses on the analysis of PROs in cancer trials and it is purely based on experts' opinions without support from empirical analysis or simulation analysis.

Though various statistical methods have been developed and some recommendations have been made for the analysis of PRO data, there are no consensus or guidelines on what statistical methods should be used for the analysis of PROs in RCT settings.

1.2 Research question

To instruct the statistical analysis of PROs considering the frequent use of PROs in RCTs and demand for sound and feasible statistical methods, this PhD study aims to address the following research question:

> *What statistical methods are appropriate for the analysis of patient-reported outcomes in randomised controlled trials?*

1.3 Research aims

The overall aim of this thesis is to identify, describe, and compare different statistical methods that can be used for the analysis of PROs in RCT settings and make recommendations for appropriate methods of analysis. The research question is split into several specific objectives.

First, two literature reviews on what statistical methods are described in the literature for analysing PROs and what statistical methods are used in practice will be systematically conducted, and the key statistical properties or criteria for the evaluation of statistical methods, with the specific purpose on analysing PROs in RCTs, will be summarised and presented.

Second, a set of desired statistical criteria for the evaluation of statistical methods will be established to filter the identified statistical methods from the two literature reviews.

Third, the technical details of the filtered statistical methods will be described, together with the commands to run in the computational software and the possible interpretation of the estimates from each statistical method. An example will be used to explain the process of applying the different statistical methods to PRO data. The filtered statistical methods will be applied to multiple RCT datasets that used PROs as clinical outcomes, and the list of statistical methods will be narrowed.

Fourth, Monte Carlo simulation methods will be carried out to evaluate the divergence of the estimates produced by each statistical method from the predefined ‘truth’, and these statistical methods will be compared and contrasted according to their model performance.

Finally, recommendations on what statistical methods are the most appropriate for the analysis of PROs in RCT settings will be made according to their technical details and their model performances in the empirical analysis and the simulation analysis.

1.4 Thesis structure

The remainder of this thesis is presented in nine chapters.

In Chapter 2, a literature review on research articles that develop and compare various statistical methods for the analysis of PROs is performed. Literature reviews, guidelines, and standards are also included to present what statistical methods are used in practice and to summarise what statistical properties or criteria can be used to evaluate statistical methods for the purpose of analysing PROs.

In Chapter 3, a review on what statistical methods have been applied in RCTs is systematically conducted, by reviewing reports of RCTs published in the United Kingdom (UK) National Institute for Health and Care Research (NIHR) HTA Journal to identify how frequently PROs are used as primary outcomes and what statistical methods are used for the primary analysis of PROs in RCTs.

Alongside the two reviews, a summary of available statistical methods for the analysis of PROs and of the statistical properties or criteria to consider for the evaluation of statistical methods are collected and summarised in Chapter 4 which also describes the research gap identified from the reviews in Chapter 2 and Chapter 3, and defines the research question and objectives. The summarised statistical methods for PRO analysis are filtered through a set of desired statistical properties that define an appropriate statistical in Chapter 5. An appropriate statistical method for the analysis of PROs in RCTs would be one that can compare two or more treatment arms; can adjust for confounding factors, including baseline PRO scores; can produce an estimate of the treatment effect and associated CIs; can handle a bounded/censored scale; and requires the least amount of recoding to use the statistical method.

The technical details of the filtered statistical methods with example code and interpretation of outputs from computational software are explained in Chapter 6. Then, in Chapter 7, the empirical analysis is conducted by applying the filtered statistical method to a series of RCT datasets that used PROs as their clinical outcomes, and the list of statistical methods are narrowed down according to their performance in estimating the treatment difference of PROs in RCTs.

Chapter 8 presents the simulation protocol that proposes Monte Carlo simulation to compare the model performance of the narrowed list of statistical methods, and Chapter 9 carries out the simulation analysis following the simulation protocol, presents and compares the performance measures of these methods. The performance of the statistical models in the simulations, will be compared using several summary statistics including bias of the estimated treatment effect from the model compared to the true treatment effect; mean square error and empirical standard error to describe the precision of the estimated treatment effect; coverage of the 95% CIs for the treatment effect estimate; Type I error under the null hypothesis of no treatment effect/difference; and power (also known as Type II error) under a variety of alternative hypothesised true (non-zero) treatment effects.

Finally, Chapter 10 will make recommendations on what statistical methods are the most appropriate for the analysis of PROs under different scenarios in RCT settings. The strengths and limitations of these statistical methods will be discussed on the basis of their technical details and model theories, and their model performances in the empirical analysis and the simulation analysis. Comparison of this study to existing evidence will be made, and potential topics for future research will be proposed.

Chapter 2 Review on statistical methods for analysing PROs and statistical properties for method evaluation

2.1 Introduction

Literature review is needed to systematically identify what statistical methods are available for the analysis of PROs. There are various statistical methods that can be used, and a decision on what statistical methods are appropriate to use can result from various reasons. As statisticians, clinicians, policy makers, and other stakeholders conduct analyses from different perspectives, it is possible for them to use different criteria or to weight criteria differently for the decision on what statistical methods to apply for the analysis of PROs. Thus, it is important to establish criteria before making model comparisons and deciding which models are more appropriate than others.

The primary aim of this method review is to identify potential statistical methods available for the analysis of PRO data using evidence from published literature. The secondary aim is to identify potential statistical criteria that could be considered to compare and contrast the statistical methods.

2.2 Methods

The review was systematically performed using three databases (EconLit, Embase and MEDLINE) to identify publications written in English with statistical methods for analysing PROs from 1 January 2000 to 31 August 2021, following the Preferred Reporting Items for Systematic reviews and Meta-Analyses statement (PRISMA) 2020 guideline (Page *et al.*, 2021). Reference tracking was used as a second source to identify potentially useful records by tracking the articles cited in the bibliography of identified records and checking the eligibility for inclusion. This review was conducted systematically to identify studies that developed statistical methods for the analysis of between group differences in PROs; and reviews, guidelines or standards that summarised statistical methods for the analysis of PROs in RCTs.

2.2.1 Definition of patient-reported outcomes

A patient-reported outcome measure (PROM) is defined as a questionnaire that measures health or a health-related outcome as a result of health interventions reported by patients themselves without any interpretation by clinicians or any other proxies. A patient-reported outcome (PRO) is an umbrella term for outcomes used to measure patients’ perceptions of health-related quality of life (HRQoL), broadly QoL, health status, satisfaction with the treatment and health conditions, etc. (Fitzpatrick *et al.*, 1998; Mokkink *et al.*, 2010; Calvert *et al.*, 2013; Brazier *et al.*, 2016).

2.2.2 Search strategy

This review focused on the statistical methods for the analysis of PRO data to compare the treatment difference between groups. The search strategy was developed with the support from an information specialist (CB). The term QoL, a domain of PRO, is more prevalently used than PRO, so in order to increase the sensitivity of this search strategy, the term*'quality of life'* was included in addition to the term *'patient-reported outcome*'*. In order to increase the specificity, the term *'statistic*'* was used to constrain the number of records to those reviews that focused on statistical techniques and to those studies that developed methods in statistics journals. The search terms and strategy are shown below.

1. [(quality of life OR patient-reported outcome*).tw. AND (analys* NOT meta-analys*).tw. AND statistic*.tw.]
2. MEDLINE.tw. OR review.tw. OR meta-analysis.pt.
3. [(quality of life OR patient-reported outcome*).tw. AND (analys* NOT meta-analys*).tw. AND statistic*.jw.]
4. (1 AND 2) OR 3

2.2.3 Inclusion and exclusion criteria

Studies were included if they met one of the following criteria: 1) the studies proposed or extended statistical methods to analyse PRO data; or 2) reviews, guidelines and standards conducted model comparisons or summarised the statistical methods or made recommendations on what statistical methods to use for the analysis of PROs.

Studies focusing on the following topics were excluded: 1) developing, validating or mapping PROs such as confirmatory factor analysis; 2) dealing with missingness, such as methods for imputation; 3) methods for multivariate analyses; 4) methods not able to make between or within group comparisons; and 5) reviews not reporting the statistical methods for analysing PROs. This is because most methods that deal with missing outcome data involve imputation of the missing data and then application of a statistical model to the augmented dataset. This review is interested in the statistical method applied to the augmented dataset and not the statistical method used to impute the missing data. PROs with multiple dimensions tend to analyse each individual dimension separately. In addition, even if a multivariate method is used in the first instance to compare multiple dimensions of a PRO, and if this model produces a statistically significant result, then a series of univariate analyses need to be carried out on each of the individual PRO dimensions separately to determine which dimensions have different outcomes. Studies conducting the data analysis of a single trial; studies looking for correlation and association; protocols; conference papers; and pilot studies were excluded. A detailed table summarising the inclusion and exclusion criteria is available in Appendix A.

2.2.4 Data extraction

As data collected from identified records was in the text format, a narrative synthesis was applied to summarise and analyse the collected information. Results on statistical methods are presented in a table summarising the methods used in trials with evidence from reviews, and in another table listing identified methods for the statistical analysis with evidence from method studies. Results on potential criteria or properties of design analysis and reporting PROs are summarised using identified literature.

2.3 Results

The review retrieved 3,052 records published between 1 January 2000 and 31 August 2021 after duplicates were removed, of which 39 articles met the inclusion criteria. The PRISMAS diagram is presented in Figure 2.1 (Page *et al.*, 2021). Of the included articles (N=39), reviews, standards, and guidelines on statistical methods provided an insight into classical and popular statistical approaches for analysing PROs (N=12); and studies that introduced, developed, or compared different statistical methods shared a mixture of classical and novel methods and various statistical properties for the evaluation of statistical methods for PRO analysis (N=27). The studies that passed primary screening but were excluded after secondary full-text screening are listed in Appendix A.

2.3.1 Reviews on different statistical methods that are applied for the analysis of PROs in trials

A total of 12 reviews, guidelines or standards were identified from the search strategy, including eight reviews on trials and four narrative studies discussing statistical methods for the PRO analysis. Three narrative studies discussed using PRO endpoints in medical research (Fairclough, 2004; Saver, 2011; Shields *et al.*, 2015), and one study that was developed by the Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data (SISAQOL) Consortium made recommendations on statistical methods for the analysis of PROs. They recommended the use of Cox proportional-hazards model for evaluating time-to-event data and linear mixed model for evaluating magnitude of event at a time and a response trajectory over time, and the use of linear regression for evaluating magnitude of event at a specific timepoint (Coens *et al.*, 2020).

Although the SISAQOL Consortium (Coens *et al.*, 2020) developed some criteria with support from experts and made recommendations on specific models for the data analysis, these criteria and recommendations are established especially for cancer studies. In addition, their recommendations are based on expert opinions and are not evidenced by fitting models to cancer datasets and comparing model performances. Therefore, further investigation is needed to extrapolate their recommendations to generic PROs or other disease-specific PROs with focuses on other disease areas such as depression.

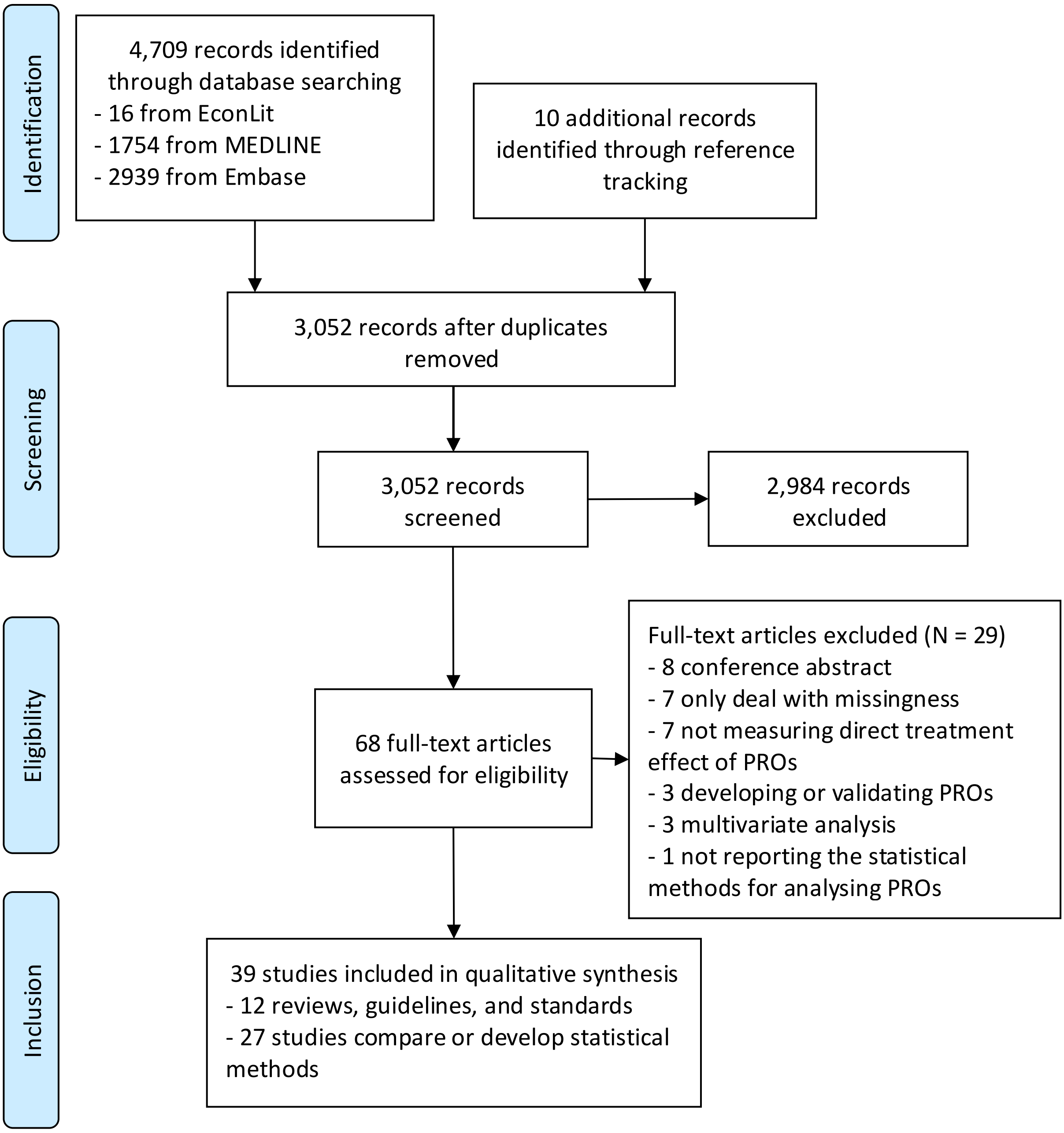

Identification
4,709 records identified through database searching
- 16 from EconLit
- 1754 from MEDLINE
- 2939 from Embase
10 additional records identified through reference tracking
Screening
3,052 records after duplicates removed
3,052 records screened
2,984 records excluded
Eligibility
68 full-text articles assessed for eligibility
Full-text articles excluded (N = 29)
- 8 conference abstract
- 7 only deal with missingness
- 7 not measuring direct treatment effect of PROs
- 3 developing or validating PROs
- 3 multivariate analysis
- 1 not reporting the statistical methods for analysing PROs
Inclusion
39 studies included in qualitative synthesis
- 12 reviews, guidelines, and standards
- 27 studies compare or develop statistical methods

Figure 2.1 PRISMAs flow diagram the inclusion and exclusion of retrieved studies

Table 2.1 summarises the proportion of different statistical methods used in eight identified reviews. It is worth noting that some of the trials reported in the reviews used more than one method for the statistical analysis, and these identified reviews all focused on cancer studies.

In general, the PRO data was analysed as a continuous outcome in both cross-sectional and longitudinal studies. Three studies (Turner-Bowker *et al.*, 2016; Pe, *et al.*, 2018; Nielsen *et al.*, 2019) reported that 45/110 studies used general linear models (e.g. *t*-test, linear regression, ANOVA or ANCOVA) and their non-parametric counterparts (e.g. Mann-Whitney U-test or Wilcoxon signed rank test) for the analysis of PROs. The model extensions such as estimation methods, e.g. generalized estimating equation (GEE), and mixed models were also applied as they relax the model assumptions on data distribution. Linear mixed model, repeated measures ANOVA and estimation using GEE were reported being used in 54/185 studies by six reviews (Fiteni *et al.*, 2016, 2019; Turner-Bowker *et al.*, 2016; Hamel *et al.*, 2017; Pe, *et al.*, 2018; Fiero *et al.*, 2019; Nielsen *et al.*, 2019). In five reviews, time-to-event data, area-under-the-curve (AUC) or survival analysis was conducted in 16/162 studies, and proportion of patients or responder analysis was conducted in 11/162 studies (Fiteni *et al.*, 2016, 2019; Turner-Bowker *et al.*, 2016; Hamel *et al.*, 2017; Pe, *et al.*, 2018). Only Nielsen *et al.* (2019) reported the studies (3/23) with ordinal data analysis using ordinal logistic regression or generalized mixed model for ordinal outcome.

Four literature reviews reported the proportion of studies that did not specify the statistical methods in their statistical summary tables (Turner-Bowker *et al.*, 2016; J.-F. Hamel *et al.*, 2017; Pe *et al.*, 2018; Coomans *et al.*, 2020), and three studies reported the proportion in the text (Fiteni *et al.*, 2016, 2019; Nielsen *et al.*, 2019), with the proportion of 52%, 80% and 100% respectively. One study did not report the proportion of studies that did not specify the statistical methods (Fiero *et al.*, 2019).

Three studies summarised the statistical methods in subgroups. Fiteni *et al.* (2016) split the observation time into three periods and described the trend of different statistical methods used across time. Fiero *et al.* (2019) evaluated the statistical methods in three groups: instruments with pre-specified PRO concepts, instruments with dimension-level analyses and instruments with item-level analyses, as they believe the statistical methods for different instruments in a trial can be different. Coomans *et al.* (2020) summarised statistical methods by research objectives including comparing PRO scores between groups at one timepoint, at multiple timepoints (i.e. cross-sectional), and over time (i.e. longitudinal).

Reviews by Turner-Bowker *et al.* (2016) and Pe *et al.* (2018) provided comparatively detailed statistical classifications. Fiero *et al.* (2019) provided a classification of PRO analysis with specific statistical models in each category, which is identical to the classification proposed by Shields *et al.* (2015), but Fiero *et al.* (2019) did not summarise specific methods for the analysis of PROs. When synthesising evidence from the identified reviews, most classifications of statistical methods in each study were kept as they were reported in the original reviews, but some of them were combined. For example, in the critical review conducted by Fiteni *et al.* (2016), the Fisher’s test, rates of symptom palliation and percentage of patients with at least two points improvement at the beginning of the cycle two were combined into ‘proportion of patients or responder analysis’ category, where the responder analysis includes statistical methods such as Fisher’s exact test and logistic regression for the analysis of proportional data. Three different responder analysis (Fisher's exact test, Chi-square with Bonferroni correction, and Blyth-Still-Casella and Mantel-Haenszel Chi-squared test) used for the analysis of PROs summarised by Turner-Bowker *et al.* (2016) were also grouped into the responder analysis category. The different types of sensitivity analyses listed in the review by Turner-Bowker et al. (2016) were combined into a broad category of sensitivity analysis, as these analyses focused on dealing with missing data and were not of the research interest in this literature review, and other included reviews did not summarise the sensitivity analysis for the PRO analysis.

Table 2.1 A summary of reviews on the statistical methods for the analysis of PROs in trials (N = 8)

Author (year of publication)	Statistical methods	N	%
Coomans *et al.* (2020) (Total = 139)	Descriptive	28	20%
	Mann-Whitney U/Wilcoxon signed rank/Kruskal-Wallis tests	26	19%
	Student's *t*-test	19	14%
	Mixed-effect models	14	10%
	Unknown	13	9%
	ANOVA/ANCOVA	9	6%
	Survival analysis	9	6%
	Responder analysis	8	6%
	Other linear models	5	4%
	Joint model	2	1%
	Regression analysis	2	1%
	Reliable change index	2	1%
	Paired signed-rank test	1	1%
	Kolmogorov-Smirnov test	1	1%
	AUC	1	1%
	Difference-in-difference approach	1	1%
	GEEs	1	1%
Fiteni *et al.* (2016) (Total = 27)	Mean change from baseline	9	33%
	Linear mixed model for repeated measures	6	22%
	Time to HRQoL score deterioration	5	19%
	AUC	2	7%
	Mixed-effects growth-curve model	2	7%
	Proportion of patients or responder analysis	3	10%
	Group comparisons of scale at each time	1	3%
Fiteni *et al.*, (2019) (Total = 15)	Mean change from baseline	8	53%
	Mean score at follow-up timepoints	2	13%
	Linear mixed model for repeated measures	1	7%
	Time to HRQoL score deterioration	1	7%
	Percentage of patient-reported symptoms	1	7%
Hamel *et al.* (2017) (Total = 33)	Cross-sectional comparison	12	36%
	Baseline changes comparison	9	27%
	Mixed model	4	12%
	Repeated measures ANOVA	3	9%
	Individual average scores comparison	3	9%
	AUC	2	6%
	GEEs	1	3%
	Logistic models	1	3%
	Percentage of symptoms over follow-up comparison	1	3%
	Cox proportional hazards model	1	3%
	Not clear	1	3%

Author (year of publication)	**Statistical methods**	**N**	**%**	
Nielsen *et al.* (2019) (Total = 23)	Descriptive analyses	2	9%	
	Non-parametric tests	6	26%	
	Parametric tests			
	t-test, one-way ANOVA	8	35%	
	Linear mixed model of repeated measures, GEEs	11	48%	
	Ordinal logistic regression, generalized mixed model with ordinal outcome	3	13%	
Pe *et al.* (2018) (Total = 66)	Not report or unclear	15	23%	
	Linear mixed models, incl. pattern mixture models	18	27%	
	Wilcoxon rank-sums test or between subjects *t*-test	11	17%	
	ANOVA or linear regression	9	14%	
	Time-to-event	6	9%	
	Repeated measures ANOVA	2	3%	
	Proportion of patients or responder analysis	2	3%	
	Others	3	5%	
Turner-Bowker *et al.* (2016) (Total = 21)	Tests of between and/or within-group difference			
	Unclear method but p-values reported	3	14%	
	t-test	5	24%	
	Repeated measures ANOVA	4	19%	
	Wilcoxon	4	19%	
	General linear mixed model	2	10%	
	Mixed-effects repeated measures model	1	5%	
	ANCOVA	2	10%	
	GEE for general linear model	1	5%	
	Responder analysis	3	14%	
	Sensitivity analysis	5	24%	
	Difference between groups in time to deterioration			
	Not specified	2	10%	
	Time-to-event	3	14%	
Fiero *et al.* (2019) (Total = 25)		Instruments with pre-specified PRO concepts	Instruments with dimension-level analyses	Instruments with item-level analyses
	Time-to-event	23/25	13/25	12/25
	Longitudinal analysis	10/25	21/25	17/25
	Basic inferential test or generalized linear model			
	Continuous	0/25	7/25	6/25
	Responder analysis	7/25	13/25	12/25
	Descriptive summaries	15/25	44/25	43/25

ANCOVA, analysis of covariance; ANOVA, analysis of variance; AUC, area-under-the-curve; GEE, generalized estimating equation.

2.3.2 Studies that developed or compared statistical methods for the analysis of PROs

Retrieved studies that developed statistical methods for the analysis fall into four categories: 1) dealing with data that are bounded, discrete and ordinal, 2) extending existing models to accommodate longitudinal data; 3) dealing with missingness; and 4) multivariate analysis. The first two topics are of interest in this review, and studies that focused on the last two topics were excluded.

Table 2.2 presents a summary of studies that developed, extended or modified, and compared statistical methods for the analysis of PRO data. There is an obvious trend that recent developed methods are more complex than the classical methods that are collected from reviews in Table 2.1. Studies that are summarised in Table 2.2 mainly focused on improving the method capability to deal with the statistical features of PROs that are bounded, skewed, and discrete. Some of the method studies compared the classical methods such as the general linear model (linear regression, *t*-test, ANOVA, ANCOVA, and repeated measures ANOVA) with the improved methods (Austin, 2002; Pullenayegum *et al.*, 2010; Arostegui, Núñez-Antón and Quintana, 2012).

Most classical methods assume that PROs are continuous, but the model assumption such as the Normality of residuals and constant variance can be violated due to the bounded, discrete and skewed properties of PRO data. Facing the violation, the classical methods can still be feasible in some circumstances. Walters and Campbell (2004, 2005) found that the bootstrap technique produced similar results to conventional methods (*t*-test and linear regression) when fitting SF-36 dimensions, and it is explained as the conventional methods are likely to be robust to non-Normality caused by HRQoL data.

To solve the problem caused by bounded and skewed PRO data, several models were proposed. Hutton and Stanghellini (2010) introduced a censored regression model assuming a censored skew-Normal distribution of the PRO data, assuming the skewness can be explained by the clustering at the boundary. Austin (2002) compared three estimation methods, maximum likelihood estimation (MLE), symmetrically trimmed least squares and censored least absolute deviations (CLAD) regression, for Tobit regression, together with ordinary least squares (OLS) and median regression fitting the Health Utility Index (HUI) scores. They found that CLAD and median regression produced similar results, and CLAD was recommended because of its low prediction error and its robustness to heteroscedasticity and non-Normality of errors, whereas Pullenayegum *et al.* (2010) recommended linear regression with robust standard errors (SEs) or nonparametric bootstrap as a simple and valid approach for analysing health utility, and found that both CLAD regression and Tobit regression are not appropriate for the analysis of utility decrement (Pullenayegum *et al.*, 2011). Median regression is a special case of quantile regression which allows the analysis of any specified quantile of the conditional distribution (Austin and

Schull, 2003). Leng and Zhang (2014) proposed a new quantile regression model by combining multiple sets of unbiased estimating equations to accommodate longitudinal data.

Instead of treating the PRO as continuous data, ordinal regression methods have been developed to accommodate the discrete nature of PRO data with few categories. Walters, Campbell and Lall (2001) recommended treating the HRQoL measure with less than seven categories as a discrete scale. Various methods have been developed to analyse the ordinal response such as ordinal regression and beta-binomial regression. Two link functions, i.e. the logit model and probit model, respectively give the ordered logit (proportional-odds) model and the ordered probit model for ordinal regression. Two studies developed statistical methods to estimate the cut-points in the proportional-odds model, and evaluated the developed methods by fitting longitudinal PRO data (Manuguerra and Heller, 2010; Parsons, 2013). Lall *et al.* (2002) introduced partial proportional-odds methods as an extension of proportional-odds model, which are less restrictive on the assumption that each covariate share a constant odds ratio (OR) across the cut-points, i.e. the influence of each covariate on the response variable is independent of the cut-points (Arostegui, Núñez-Antón and Quintana, 2012).

Alternative regression models such as beta regression and beta-binomial regression were also proposed to analyse PRO data. Beta regression is proposed by Ferrari and Cribari-Neto (2004) to analyse responses which are beta distributed. The beta distribution is a continuous probability distribution with two positive parameters alpha and beta (both > 0) defined on the interval [0, 1]. It is later applied to PRO data to deal with the bounded nature (Hunger, Baumert and Holle, 2011; Kharroubi, 2020). Zou, Carlsson and Quinn (2010) proposed a beta-mapping and beta-regression method for the change of ordinal QoL by mapping the change of Likert scale to a beta distribution within [0,1] and using the mapped data to fit the beta regression under the generalized linear mixed model framework. Fractional logistic regression (as known as fractional logit model) has similar features to beta regression, except that it can account for responses at boundaries i.e. 0 and 1 (Meaney and Moineddin, 2014). Arostegui, Núñez-Antón and Quintana (2007) analysed the SF-36 dimensions assuming a beta-binomial distribution where the probability of success in the logit link follows a beta distribution. The other way to estimate the beta-binomial regression is by using ‘the hierarchical generalized linear model as a generic method of performing generalized linear mixed models with non-Normal random effects’ (Najera-Zuloaga, Lee and Arostegui, 2018). The binomial-logit-Normal regression was proposed based on the beta-binomial regression by assuming that the random effects follow a standard Normal distribution, and the probability therefore follows a logit-Normal distribution (Arostegui, Núñez-Antón and Quintana, 2012; Liang *et al.*, 2014). Based on those methods, further statistical methods were proposed to extend the developed methods by introducing new parameter estimation processes to accommodate to longitudinal data (Gheorghe *et al.*, 2017; Najera-Zuloaga, Lee and Arostegui, 2019).

Table 2.2 A summary of statistical methods developed and compared in identified studies (N = 27)

Author (year of publication)	Summary of contents
Methods for continuous data (with bounded or censored feature)	
Austin (2002)	Compared three estimation methods for Tobit model (MLE, symmetrically trimmed least squares and CLAD regression), MLR and median regression.
Walters and Campbell (2004)	Compared bootstrap with standard methods (*t*-tests, linear regression, summary measures, and general linear models fitted with GEE).
Walters and Campbell (2005)	Compared *t*-test and bootstrap (Bca) estimates of CIs of eight dimensions in SF-36, and did not recommend MLR with bootstrapping as the results generated by both methods are similar.
Hutton and Stanghellini (2010)	Developed a censored regression model assuming a skew-Normal distribution, and compared with a censored Normal model and an uncensored skew-Normal model.
Pullenayegum *et al.* (2010)	Compared MLR, Tobit, and CLAD regression for the analysis of EQ-5D utility score in terms of their bias and estimated CIs, and recommended MLR with robust SEs or the non-parametric boostrap as a simple and valid approach.
Pullenayegum *et al.* (2011)	Compared marginal Tobit regression and CLAD regression when facing utility decrement, and concluded that these two methods should not be used under this circumstance.
Leng and Zhang (2014)	Constructed a new quantile regression model by combining multiple sets of unbiased estimating equations to accommodate longitudinal data.
Methods for response data scatter between 0 and 1	
Ferrari and Cribari-Neto (2004)	Proposed the beta regression for analysing response that is beta distributed using a parameterisation of the beta law that is indexed by mean and dispersion parameters.
Zou, Carlsson and Quinn (2010)	Mapped the score change in Likert scale to a beta distribution and conducted beta regression using the mapped data under the generalized linear mixed model, which benefits from using the flexibility of beta distribution.
Hunger, Baumert and Holle (2011)	Compared MLR and beta regression for analysing SF-6D, and suggested that the beta regression, especially with prevision covariates, is a possible supplement to methods currently used in the analysis of health utility data.
Meaney and Moineddin (2014)	Compared MLR, beta regression, and fractional logit regression to estimate covariate effects on (0,1) response data, and found these three methods all performed well.
Gheorghe *et al.* (2017)	Proposed Bayesian mixed beta regression by using Markov chain Monte Carlo methods to model longitudinal HRQoL data.
Kharroubi (2020)	Compared beta regression with MLR for analysing SF-6D, and found that beta regression performed better than MLR in predictive ability using mean prediction error, root mean squared error and deviance information criterion.
Methods for Ordinal data	
Qian *et al.* (2000)	Compared summary measures (worst score, worst score minus pre-treatment score, and AUC) analysed by stratum-adjusted Mann-Whitney test in the form of Cochran-Mantel-Haenszel statistic, and ordinal (or cumulative) logistic regression mixed models for repeated measures using GEE.
Walters, Campbell and Lall (2001)	Compared conventional statistical methods (i.e. *t*-tests and multiple regression), ordinal regression models (proportional-odds, continuation ratio, polytomous, and stereotype) and the bootstrap method.

Author (year of publication)	Summary of contents
Methods for Ordinal data	
Lall *et al.* (2002)	Reviewed ordinal regression models, including proportional-odds model, partial proportional-odds model, and the stereotype model with bootstrap techniques to obtain standard errors.
Arostegui, Núñez-Antón and Quintana (2012)	Compared MLR, with least square and bootstrap estimates, Tobit regression, ordinal logit and probit regressions, beta-binomial regression, binomial-logit-Normal regression, and coarsening.
Lee and Daniels (2008)	Extended the marginalized random effects model to accommodate longitudinal ordinal data, using Quasi-Newton algorithms with Monte Carlo integration of the random effects to calculate the maximum marginal likelihood estimation.
Manuguerra and Heller (2010)	Modelled cut-point parameters using generalized logistic and non-parametric functions in proportional-odds models for continuous ordinal scores derived from visual analogue scales in a Bayesian setting.
Parsons (2013)	Introduced a repeated measures proportional-odds logistic regression model that estimated the cut- point by using an orthogonal polynomial decomposition of the ordered but unstructured cut-points, with model parameters estimated by GEE.
Methods for multinomial data	
Arostegui, Núñez-Antón and Quintana (2008)	Introduced a beta-binomial distribution approach for the analysis of HRQoL data, and compared the model with MLR.
Liang *et al.* (2014)	Proposed the use of binomial-logit-Normal regression as an alternative model for the bounded PRO which can be considered as a candidate model.
Najera-Zuloaga, Lee and Arostegui (2018)	Compared two models for beta-binomial regression: beta-binomial distribution with a logistic link and hierarchical generalized linear models.
Najera-Zuloaga, Lee and Arostegui (2019)	Developed an estimation procedure for the analysis of longitudinal discrete and bounded outcomes using a beta-binomial mixed-effects model, and compared the developed model with generalized additive models for location, scale, and shape.
Semi-parametric methods	
Moerkerke *et al.* (2005)	Adopted a permutation-based approach to evaluate the null distribution of the maximum of many correlated test statistics and used the statistics to build a regression model that explains QoL differences between treatment arms.
Zheng, Qin and Tu (2017)	Developed a semi-parametric generalized partially linear mean-covariance regression, using a Cholesky decomposition for the covariance matrix of the longitudinal responses with parameters estimated by modified GEE.
Wang and Tu (2020)	Developed a three-component mixture model for the proportional data and a density ratio model for the distributions of continuous observations in (0,1), and derived a bootstrap semiparametric homogeneity test for the homogeneity of distributions of multi-group proportional data, and compared it with likelihood ratio tests under parametric distribution assumptions (beta distribution and logit-Normal distribution), rank-based Kruskal-Wallis test, and Wald-type test.

BCa, bias-corrected and accelerated; CLAD, censored least absolute deviations; GEE, generalized estimating equation; HRQoL, health-related quality-of-life; MLR, multiple linear regression; SE, standard error; SF-6D, Short Form-36.

Two semi-parametric models were proposed for the analysis of longitudinal proportional data. One constructed a flexible semiparametric covariance model (Zheng, Qin and Tu, 2017), and the other assumed the distribution of the proportional data follows a semiparametric density ratio model and applied bootstrap to improve the model performance (Wang and Tu, 2020). Moerkerke *et al.* (2005) adopted a permutation-based approach by using a permutation test to construct a non-parametric null distribution for the Wald statistic, conditional on the observed data, and used the statistic to build a regression model that explains QoL differences between treatment arms.

2.3.3 Criteria to assess statistical issues for PRO analysis from the identified reviews

Table 2.3 summarises the assessment criteria to evaluate the PRO analysis listed in three identified reviews (Fiteni *et al.*, 2016; Hamel *et al.*, 2017; Pe *et al.*, 2018). Though the criteria vary in these three reviews, they can be summarised into four categories - data structure, statistical methods of analysis, baseline assessment and missing data management. The report of targeted dimensions and repeated measurements are desired in three reviews. In terms of statistical methods, the clinical relevance and statistical techniques are desired in all three reviews, and other features such as specific hypothesis and multiple comparison management are of interest in some studies. Compared to the other two studies, Hamel *et al.* (2017) did not summarise specific information about baseline assessment. Missing data management is also a highlighted topic in the three reviews on cancer trials.

The criteria presented in these three reviews are in line with the recommended components to report for the statistical analysis of PROs in protocols proposed by the CONSORT-PRO Extension (Calvert *et al.*, 2013) and SPIRIT-PRO Extension (Calvert *et al.*, 2018). However, these abovementioned criteria in the identified reviews did not concentrate on the specific evaluation of statistical methods, but focused on other components in statistical analysis including targeted dimensions, descriptive analysis, compliance rate etc. Recognising the lack of standardised criteria for the methodological issues in PRO analysis, the SISAQOL Consortium (Coens *et al.*, 2020) listed a range of statistical features for PRO analysis and gathered opinions from multiple stakeholders on what features are essential and highly desired. The desired features include whether the method can compare two treatment arms, adjust for baseline score, be clinically relevant, allow for confounding factors, handle missing data, and handle clustered data.

Table 2.3 A summary of criteria to assess statistical issues for PRO analysis

Fiteni *et al.* (2016)	Hamel *et al.* (2017)	Pe *et al.* (2018)
Data structure		

Targeted dimensions;
Number of HRQoL data at baseline and at subsequent timepoints;
	Multidimentionality;
	Longitudinality
		Multiple dimensions;
		Repeated assessments;
		Reporting of descriptive data

Statistical methods of analysis		

HRQoL hypothesis;
Statistical approach for HRQoL analysis;
MCID considered in the statistical analysis;
Multivariate analysis;
Procedure to control the Type I error;
	HRQoL data analysis;
	MCID report;
	Multiple comparisons management;
		Specific hypothesis;
		Primary statistical technique;
		Reporting of clinical relevance;
		Included baseline as a covariate;

Baseline assessment		
HRQoL scores at baseline for each group and each dimension;	NA	Assessed baseline; Compared baseline scores between treatment arms;
Missing data management		
Profile of missing data at baseline; Statistical approaches for dealing with missing data; Study population.	Patient characteristics comparison depending on compliance status; Survival and compliance rate at the end of follow-up; Protocol specified non-attrition.	Strategy to handle missing data; ITT population; Compliance rates.

MCID, minimum clinically important difference; NA, not available; HRQoL, health-related quality of life; ITT, intention-to-treat.

2.4 Discussion

This review provides an insight in what statistical methods have been developed, applied, and extended for the analysis of PROs, and extracts a list of criteria to assess statistical issues for PRO analysis from identified reviews. In general, reviews on statistical methods provided an insight into classical and popular statistical approaches for the analysis of PROs; and studies that introduced, developed or extended statistical methods provided novel and advanced methods for the analysis.

This review has the following findings:

First, there is no consensus on which statistical method is the most appropriate for the analysis of PROs in RCTs. For example, Walters and Campbell (2005) compared the estimations of SF-36 outcome by bootstrapping and *t*-test, and did not recommend bootstrapping over the general linear model as the results generated by both methods are similar. In comparison, Arostegui, Núñez-Antón and Quintana (2012) compared OLS with bootstrapping and OLS of two dimensions from SF-36. Their results showed that bootstrapping is able to detect statistical significance of an estimation whereas OLS cannot, indicating that the OLS estimates might not be reliable when the Normality assumptions are violated and therefore they recommended to use OLS with bootstrapping. This might be because these recommendations were made based on different datasets according to different criteria. As different stakeholders might be interested in different statistical features, they could hold different opinions when selecting the method for the analysis considering their desired criteria.

In addition, the identified reviews showed that the statistical methods for ordered data were not popular for the analysis of PROs in cancer studies (Fiteni *et al.*, 2016, 2019; Turner-Bowker *et al.*, 2016; Hamel *et al.*, 2017; Pe *et al.*, 2018; Fiero *et al.*, 2019; Nielsen *et al.*, 2019; Coomans *et al.*, 2020). There is only one identified review that summarised trials that conducted statistical analysis with ordered logistic regression or generalized mixed model with ordinal outcome (Nielsen *et al.*, 2019). The international standards for the analysis of PROs in cancer trials by the SISAQOL Consortium neither recommended the use of ordinal regression nor compared ordinal regression with other statistical methods (Coens *et al.*, 2020). This might be because the identified reviews focused on the analysis of PROs in cancer trials and typical PROs used in cancer trials are less likely to have the ordinal feature. But meanwhile, ordered logistic regression and other methods for PRO data with ordinal features have been developed and recommended for use (Arostegui, Núñez-Antón and Quintana, 2007, 2012; Zou, Carlsson and Quinn, 2010; Najera-Zuloaga, Lee and Arostegui, 2019). This indicates that there are some disagreements on what statistical methods should be applied for analysing PROs, and these disagreements may be due to different factors such as the nature of the outcome data, the proposal of the analysis, the adherence of the data to the method assumptions, and the criteria set for method evaluation.

Second, it is doubtful when studies applied different methods using a single dataset and drew the conclusion on which method is better than others. Since each dataset has different characteristics (e.g. data distribution, different degree of information loss and unbalanced data), it would be crude to reach a conclusion when fitting models to a single dataset. And even though in a single dataset, various methods might be suitable for the analysis of PROs, especially in the case of multidimensional PROs which may have more than one endpoint of interest. For instance, it could be more meaningful to fit the role functioning - emotional dimension in SF-36 with four possible categorical values using ordinal regression than linear regression, and to fit physical functioning dimension with 21 possible categorical values using linear regression than ordinal regression, although pragmatically one common statistical method to analyse multidimensional PROs such as SF-36 which has eight dimensions may be preferred. Therefore, a specific target dimension of a PRO for analysis is recommended before proposing a statistical method (Fiteni *et al.*, 2016; J.-F. Hamel *et al.*, 2017; Pe *et al.*, 2018).

Furthermore, the way the PROs have been generated is particularly important for the data analysis and assumptions made for the underlying latent variable could lead to different statistical methods (Lall *et al.*, 2002). For example, Tobit regression assumes the response variable is censored, meaning that the latent response variable can exceed the boundaries but cannot be observed (Pullenayegum *et al.*, 2011), but the scores exceeding the boundaries are meaningful.

Finally, the classification of statistical methods may not be appropriate in some identified studies. For example, within and between group differences, such as the mean change from baseline in the classification developed by Fiteni *et al.* (2016, 2019), are the target estimates of the statistical analyses, but not the statistical methods that can be applied for analysing PROs. This classification could be generated under the situation that statistical analyses in some trials were not conducted or were not clearly reported, and thus it was not clear what exact approaches were applied. Instead of summarising the statistical methods with statistical estimates, it would be helpful to report the proportion of studies that clearly reported the statistical approaches, and to summarise the statistical methods and estimates separately. It is understandable that reviewers are interested in different traits of the statistical analysis, and these statistical methods are sometimes hybrid with each other, and accordingly those combined or extended methods own different desired traits. However, a unified classification of statistical methods in reviews that summarise different statistical methods applied in the trials is needed to help researcher have a clearer picture.

This review has the following limitations: first, this study only used '*patient-reported outcome**' or '*quality-of-life*' in the search strategy in order to retrieve a manageable number of results given the time constraints for the research. As this might result in omitting studies summarised or developed methods for bounded and discrete data but not specially designed for PRO or QoL data, reference tracking was used to supplement the database search. Second, this study excluded several conference abstracts, some of which were of research interest. As the full texts of the excluded conference abstracts could not be retrieved, some valuable information could be omitted. Third, the eight identified reviews on the methodology for the analysis of PROs in trials all focused on the cancer studies. This might bias the popular methods that have been used for analysis, as cancer data which usually summarise survival data might have different features from other diseases. Last, only one reviewer (YQ) was involved in the screening and data extraction, but this ensures consistency in the entire process of this review.

This literature review provides an insight in what statistical methods have been developed for the analysis of PROs, and what statistical methods have been applied through identified reviews. However, the identified reviews in this chapter all focused on cancer trials, and they neither share consistent terminology or classification of statistical methods, nor disclose details in the strategy for statistical analysis. Therefore, the next chapter will report a further review investigating what statistical methods are used for the analysis of PROs in publicly funded RCTs to see what methods are used in practice for other disease areas. Additionally, limited information on the criteria for the evaluation of statistical methods was identified from this review. Other criteria and characteristics under this topic will be collected in the next chapter, and will then be presented in Chapter 4.

Chapter 3 Review on statistical methods for the analysis of PROs used as primary outcomes in published RCTs

3.1 Introduction

This chapter aims to review the statistical methods for the analysis of PROs that are used as primary outcomes in RCTs published by the UK's NIHR HTA Journal. This review helps the understanding of statistical analysis of PRO in RCTs, the identification of different statistical methods used for the analysis of PROs in published clinical trials and potential statistical issues in analysing PROs, and the summary of necessary components or criteria for reporting statistical analysis of the PROs in RCTs.

3.2 Methods

This review aims to firstly identify how frequently PROs have been used as primary or secondary clinical outcomes in reports of RCTs published in the UK's NIHR HTA Journal; and secondly, when the PRO is used as the primary outcome for a trial, to summarise what statistical methods have been used to analyse the PRO.

3.2.1 Trial identification

Reports of RCTs published in the UK's NIHR HTA Journal between 1 January 1997 and 31 December 2020 that defined and reported a PRO as clinical endpoints or outcomes for the trial were systematically identified and reviewed. The definition of PRO in this review is adapted from the definition from Chapter 2 Section 2.2.1. The HTA Journal was chosen because the information related to the trial and the PROs are reported in more detail in comparison to major medical journals.

Information related to the use of PROs include the frequency of using PROs as clinical outcomes, whether the PROs were used as primary and/or secondary outcomes, and when the PRO was the primary outcome, the characteristics of the PROs and the statistical methods used for the analysis of PROs. The identification of HTA reports of RCTs used the same search strategy as previous work (Walters *et al.*, 2017). The selection of trials with PROs was conducted by one reviewer (YQ). Three reviewers (SW, RJ, and LF) conducted quality assurance checks on 30% of the included papers after the data extraction was completed, and disagreements were discussed to achieve consensus.

3.2.2 Inclusion and exclusion criteria

The studies included in this review satisfied the following criteria: 1) individually randomised controlled trials; 2) trials with at least one PRO as the primary outcome; and 3) trials with the statistical analysis conducted for the PRO. Studies excluded from this review are cluster RCTs as these have specific statistical issues; influenza trials as these rarely use PROs as clinical outcomes; adaptive or group sequential trials as these have different statistical issues that may influence the choice of analysis; follow-on studies and pilot studies.

PROs identified in this review can be well-established measures from previous studies with feasibility, reliability and validity tested, or self-developed measures by researchers alongside trials. For studies with measures which were not clearly defined as a PRO in the trial, various methods were taken to identify whether the measure was categorised as a PRO, including retrieving the cited paper that developed the measure, identifying signal words such as ‘carers’ and ‘physicians’ for rating or assessing patients’ outcomes in the measure description, and referring to other papers that developed or applied the outcome measure. According to our definition of PROs, trials that only recruited patients, or trials that recruited both patients and proxies when patients were unable to complete PROs were included; trials that only used proxies as informants to complete the PROs were excluded in order to avoid the cases where clinicians respond to health outcome measures on the patients’ behalf.

Trials using the product of PROs, such as a dichotomised outcome and quality-adjusted survivals, were included. Trials that used PROs only as primary cost-effectiveness outcomes, but not as clinical primary outcomes were excluded. Trials that did not actually conduct the statistical analysis were excluded, even if the statistical methods were proposed.

3.2.3 Data extraction

The following information was extracted from the included trials.

1. Characteristics of the trials with PROs as primary outcomes, including the number of participants randomised and analysed, the baseline and post-randomisation assessment, the most frequently used PROs, and special types of PROs (including patient satisfaction, preference-based and proxy-reported).
2. Statistical methods conducted for the primary analysis of the PROs, including the study population, the specific statistical methods, the adjustment for baseline score or other covariates, involvement of random effects, robust SEs and bootstrapping techniques, repeated measures analysis, and strategies for missing data.
3. The quality of reporting PROs, including whether there is a clear definition or justification of the primary outcomes or primary endpoints, statistical methods, and covariates.

3.2.3.1 Characteristics of PROs that are used as primary outcomes

We classified a unidimensional PRO as one that focuses on one dimension, but it can have more than one item used to measure that dimension; and a multidimensional PRO focuses on more than one dimension and should have more than one item. There is at least one item in each dimension.

In this review, we defined four score types of PROs - overall summary score, subscale score, single item score and unidimensional summary score. The former three types can be generated from multidimensional PROs, and unidimensional PROs are able to generate unidimensional summary scores and single item scores. The overall summary score covers more than one dimension, and it is composed of all dimensions covered by the PRO; the subscale score is composed of more than one item and it covers one or more than one dimensions; the single item score is composed of only one item, usually resulting from a global question and in the format of Likert scale or visual analogue scale (VAS).

3.2.3.2 Statistical analysis of PROs that are used as primary outcomes

All statistical methods used for the analysis of PROs that were defined as primary outcomes in the identified trials were extracted. This included different analysing stages (including primary, secondary, longitudinal, and sensitivity analyses) for the analysis of different data types of PROs as primary outcomes (including continuous, ordinal, binary, and others). These data types were defined according to the statistical methods used for the analysis of PROs.

For the purpose of this review, the statistical methods were broadly classified into two categories: univariable methods that do not adjust for any other covariates except the randomised group, e.g. *t*-test, Chi-squared test, and simple linear regression, and multivariable methods that have one or more explanatory variables (e.g. baseline score) in addition to the randomised group, e.g. multiple linear regression (MLR). The multivariable methods were further classified according to the categories of generalized linear model (GLM), including MLR, analysis of covariance (ANCOVA), binary logistic regression, ordinal logistic regression, and their extensions for correlated responses such as models with coefficients estimated by GEEs and mixed effect models with coefficients estimated by MLE or restricted maximum likelihood (REML). Repeated measures analysis for PROs with more than one post-randomisation assessments were classified into four categories: response feature analysis (i.e. using summary measures, such as AUC or post-randomisation mean score), generalized linear mixed models (GLMM) with parameters estimated by MLE or REML, GLM with parameters estimated by GEE, and repeated measures ANOVA (Walters, 2009; Schober and Vetter, 2018).

The information about missing data was also collected in this review, including whether the proportion of missing data was reported and whether a strategy for dealing with missing data was developed.

3.3 Results

In total, 1356 reports were published by the HTA Journal between 1 January 1997 and 31 December 2020, and 928 reports were excluded after screening the titles and abstracts. In the remaining 416 reports, 125 were excluded for various reasons (Figure 3.1). In the 303 published individually randomised controlled trials, 37.6% (114/303) of trials used PROs as primary outcomes and 82.8% (251/303) of trials used PROs as secondary outcomes. Two trials with PROs as primary outcomes were excluded as they were closed without conducting a statistical analysis of the data using the statistical methods that were proposed in the report (Mihaylov *et al.*, 2008; Williams *et al.*, 2017). It should be noted that the first RCT with a PRO as a clinical outcome was published in the HTA Journal in 1999 (Simpson *et al.*, 1999), and the earlier reports published in the HTA Journal were mainly systematic reviews.

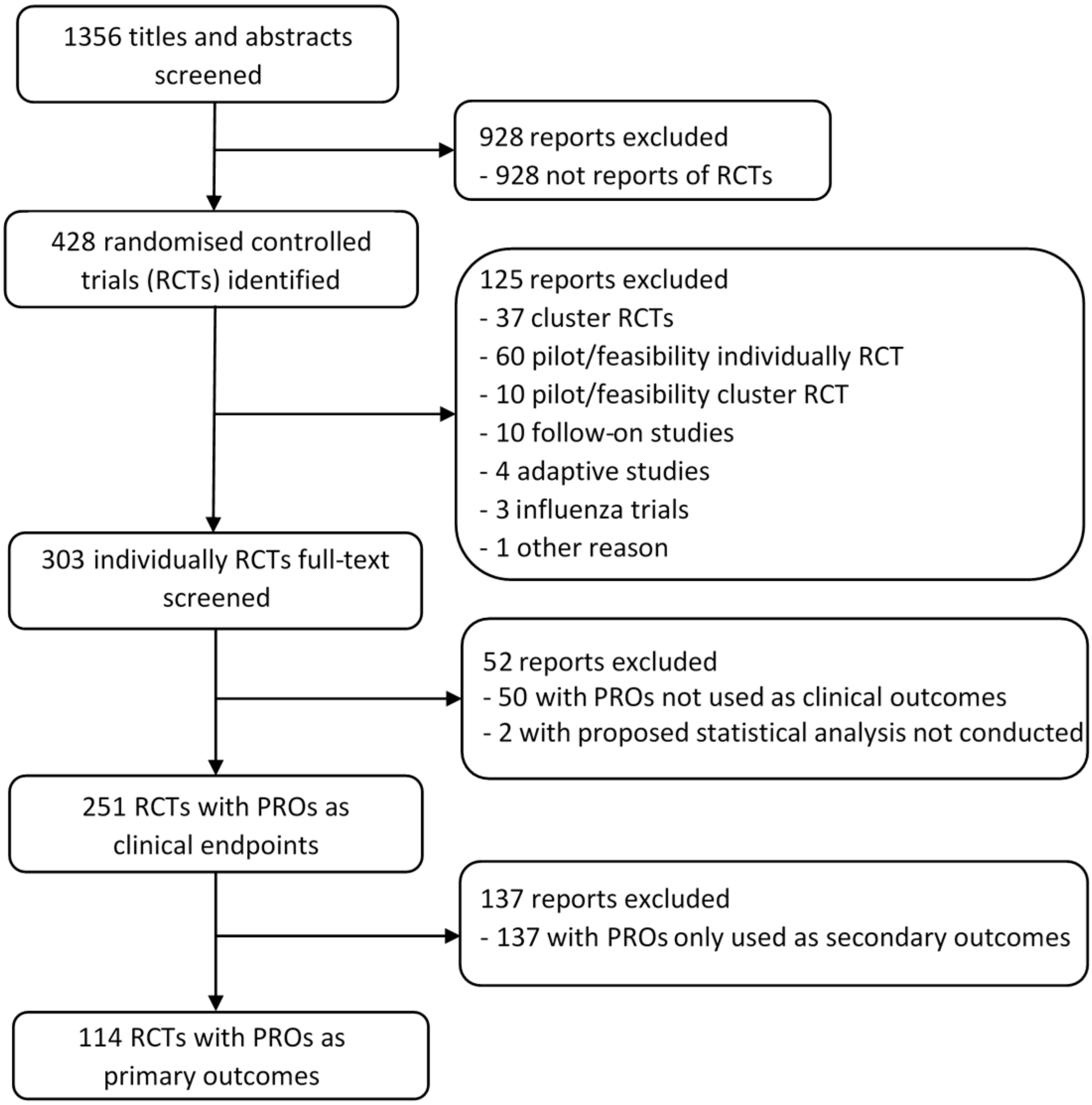

1356 titles and abstracts screened
928 reports excluded
- 928 not reports of RCTs
428 randomised controlled trials (RCTs) identified
125 reports excluded
- 37 cluster RCTs
- 60 pilot/feasibility individually RCT
- 10 pilot/feasibility cluster RCT
- 10 follow-on studies
- 4 adaptive studies
- 3 influenza trials
- 1 other reason
303 individually RCTs full-text screened
52 reports excluded
- 50 with PROs not used as clinical outcomes
- 2 with proposed statistical analysis not conducted
251 RCTs with PROs as clinical endpoints
137 reports excluded
- 137 with PROs only used as secondary outcomes
114 RCTs with PROs as primary outcomes

Figure 3.1 Flow diagram for the inclusion and exclusion of trials published in the UK's NIHR HTA Journal from 1997 to 2020

Table 3.1 shows the number and percentage of trials that used PROs as primary and/or secondary outcomes. The number of HTA trials that used PROs as primary outcomes was around 60% of the number of HTAs that used PROs as secondary outcomes. All identified HTA studies that reported PRO as primary outcomes also employed PROs as secondary outcomes. Of 303 HTA reports, 52 (17.2%) did not use PROs as clinical effectiveness outcomes.

Table 3.1 Number (percentage) of HTAs reporting PROs as primary and/or secondary outcomes

		PROs as primary outcome?		Total
		Yes	No	
PROs as secondary outcome?	**Yes**	114 (37.6%)	137 (45.2%)	251 (82.8%)
	No	0 (0.0%)	52 (17.2%)	52 (17.2%)
	Total	114 (37.6%)	189 (62.4%)	303 (100.0%)

All included trials that used PROs as primary outcomes also used PROs as secondary outcomes. The trend of using PROs as clinical outcomes in trials between 1999 to 2020 is shown in Figure 3.2. Overall, PROs were more frequently used as secondary clinical effectiveness outcomes than as primary outcomes. Except for the earlier years (1999-2003) with a small number of studies, the average proportion of trials with PROs used as secondary outcomes (represented by the red curve) fluctuates around 87%, which is approximately two times higher than the average proportion of the trials with PROs as primary outcomes (represented by the blue curve). Generally, there is an increase in using PROs as clinical outcomes in HTA trials.

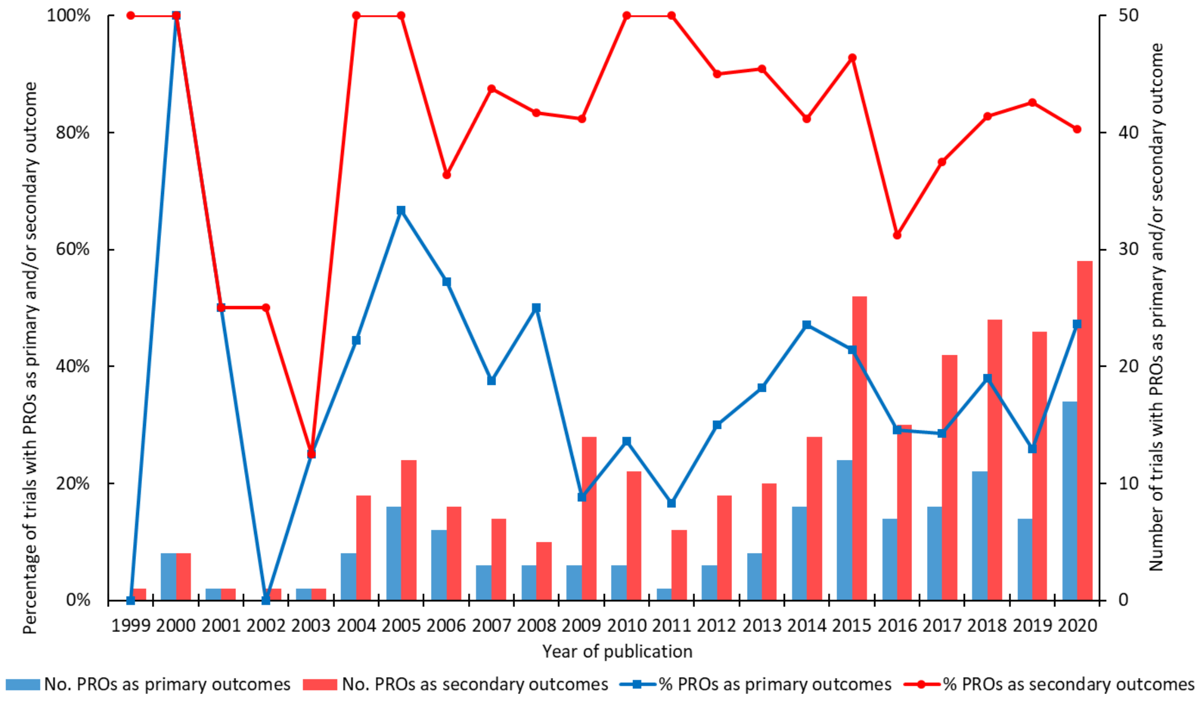


Figure 3.2 Number and proportion of trials using PROs as primary and/or secondary outcomes from 1999 to 2020

3.3.1 Trial characteristics

In total, 83.1% (61,715/74,298) of the participants randomised in the 114 trials were included in the primary analysis (Table 3.2). The characteristics of these trials are summarised in Table 3.3. The most common design was a two-arm parallel group trial. More than half of the trials were in either mental health (30/114) or musculoskeletal conditions (28/114). Most trials collected baseline assessments (101/114) and more than one post-randomisation assessments (107/114). The maximum number of post-randomisation assessment was 24 in a trial on eczema management for children (Santer *et al.*, 2018).

3.3.2 Characteristics of PROs that are used as primary outcomes

Most trials (107/114) clearly defined the primary outcomes. The sample size calculation implied the primary outcomes for the trial was a PRO in six trials that did not explicitly specify the primary outcome, and one trial defined PROs as main outcome measures but used an alternative outcome for the sample size calculation (Kerry *et al.*, 2000). Table 3.4 summarises the PROs used as primary outcomes in four or more included trials. The most popular PROs were mainly generic, i.e. SF-36/SF-6D and EQ-5D, and depression-specific, i.e. Beck Depression Inventory (BDI), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire (PHQ). Eight trials used more than one PRO as the primary outcomes, and 14 trials used non-PRO clinical outcomes as co-primary outcomes.

Preference-based PROs were used as primary outcomes in six trials, including five that used the EQ-5D (Russell *et al.*, 2013; Brittenden *et al.*, 2015; Watson *et al.*, 2017; Sharples *et al.*, 2018; Gazzard *et al.*, 2019) and one that used the SF-6D (Michaels *et al.*, 2006). Seven trials used quality-adjusted survivals, including three that used EQ-5D (Russell *et al.*, 2013; Watson *et al.*, 2017; Sharples *et al.*, 2018) and four that used specific PROs for estimation (Hewison *et al.*, 2006; Bedson *et al.*, 2014; Williams *et al.*, 2016; Pickard *et al.*, 2020). Patient satisfaction was used as primary outcome in two trials (Townsend *et al.*, 2004; Cooper *et al.*, 2019). Proxies were recruited in six trials that primarily aimed to collect PROs, and only recruited proxies for patients who were unable to complete the PROs (Dennis *et al.*, 2006, 2020; Weindling *et al.*, 2007; Banerjee *et al.*, 2013; Francis *et al.*, 2016; Santer *et al.*, 2018). Seven included trials used self-developed PROs as primary or co-primary outcomes. Most of these PROs had one item based on a Likert scale or visual analogue scale (VAS), except for one trial that specially developed a REFLUX questionnaire with 31 items to generate the QoL of patients with gastro-oesophageal reflux disease (Grant *et al.*, 2008). Health outcomes assessed by investigators were not included as primary outcomes, e.g. four trials using Quality of Life Scale or Rankin Scale that were completely assessed by investigators were excluded (Lewis *et al.*, 2006; Langhorne *et al.*, 2017; Bath *et al.*, 2018; Sprigg *et al.*, 2019).

Table 3.2 Recruitment and retention of trial participants included from the 114 reports

Items	Mean	Median	SD	Min	Max	Total
Number of participants randomised	652	480	928	85	8,003	74,298
Number of participants analysed*	541	388	847	65	7,677	61,715

*Number of participants analysed in the primary analysis of PROs; if multiple post-baseline assessments were used for primary outcome, the number of participants analysed at the longest primary endpoint was taken.

Table 3.3 Trial design and assessments characteristics of the 114 trials included in the review

Items		No	%	Total
Trial design				114
	Parallel group	102	87.2	
	Factorial	3	2.6	
	Crossover	0	0.0	
	Other*	9	7.7	
Number of arms				114
	2	83	70.9	
	3	21	17.9	
	4	5	4.3	
	>4	5	4.3	
Clinical area				114
	Mental Health	30	25.6	
	Musculoskeletal	28	23.9	
	Obstetrics and gynaecology	9	7.7	
	Gastrointestinal	7	6.0	
	Respiratory	5	4.3	
	Stroke	5	4.3	
	Primary Care	4	3.4	
	Cardiovascular	4	3.4	
	Dermatology	4	3.4	
	Cancer/Oncology	3	2.6	
	Other^	15	12.8	
Number of trials with a baseline assessment of the PRO				114
		101	86.3	
Timing of primary outcome post-baseline assessments				114
	<1 month	7	6.0	
	1-6 months	28	23.9	
	6-18 months	50	42.7	
	>18 months	27	23.1	
	Missing$	2	1.7	
Number of post-baseline assessments				114
	1	5	4.3	
	2	41	35.0	
	3	29	24.8	
	4	19	16.2	
	>4	18	15.4	
	Missing$	2	1.7	

*patient preference/Zelen's. ^chronic fatigue, minor surgery, multiple sclerosis, neurosurgery, paediatric, sleep disorders, urology, vascular. $Two trials did not specify the timing and number of post-baseline assessments.

Different score types were used to summarise the primary clinical effectiveness measured by PROs. Of the 114 included trials, 99/114 used one score type to report the primary outcome, and 15 of the included trials used more than one score type. Summary scores were the most popular way to measure PROs, and it was widely used in identified trials (83/114), followed by subscales (30/114) and single items (16/114), including Likert scale (11/114) and VAS (5/114). Among 15 trials that used over one score type for summary, 11 studies with one PRO as primary outcome reported subscales and overall summary scores to measure the primary clinical effectiveness, and the remaining four studies had more than one PRO as primary outcomes (Lamb *et al.*, 2010; Brittenden *et al.*, 2015; Orgeta *et al.*, 2015; Glazener *et al.*, 2016).

Except for seven trials with only one post-randomisation follow-up, 107 trials assessed PROs with more than one post-randomisation follow-up (Table 3.5). Of the 114 identified studies, 41 (36%) studies used unidimensional PROs, and 72 (63%) studies used multidimensional PROs as primary outcome. There was one study employing both multi- and unidimensional PROs as more than one primary outcome was used in this study (Lamb *et al.*, 2010).

The PRO scores can be transformed into dichotomised outcomes or quality-adjusted survivals. Five trials used the AUC estimated by overall summary scores from PROs as the primary outcome. Three trials (Russell *et al.*, 2013; Watson *et al.*, 2017; Sharples *et al.*, 2018) used EQ-5D for AUC estimation, and two trials (Hewison *et al.*, 2006; Bedson *et al.*, 2014) used condition-specific PROs for estimation. A number of 10 trials used dichotomised PRO scores as primary outcomes. In 11 trials, the PRO as primary outcomes were also used to generate secondary outcomes using a different score type. Four trials used a dichotomised PRO as primary outcome, and the original PRO as secondary outcome, while six studies employed the original PRO as primary outcome, and the dichotomised PRO as secondary outcome. One study used the AUC derived from PRO as primary outcome, and the original PRO score as secondary outcome (Russell *et al.*, 2013).

3.3.3 Statistical analysis of PROs that are used as primary outcomes

Intention-to-treat (ITT) analysis (i.e. analysis based on the treatment assignment of all participants but not the actual treatment received) (Montedori *et al.*, 2011), including ITT with and without missing data imputation, was used in 111/114 trials. A number of 46 trials used other study populations such as per protocol analysis (i.e. analysis based on the patients who completed the originally treatment assigned), as treated, or complier average causal effect analysis (i.e. analysis of the treatment effect based on the subgroup that completed the originally treatment assigned) for the secondary or sensitivity analysis.

Table 3.4 The most frequently used PROs as primary outcomes in the included trials

PROs	Abbr.	No	%	Reference to the PRO
Short Form-36 Short Form- 6 Dimensions	SF-36 SF-6D	8	7.0	(Ware and Sherbourne, 1992; Brazier, Roberts and Deverill, 2002)
Beck Depression Inventory	BDI	7	6.1	(Beck, Steer and Carbin, 1988)
Hospital Anxiety and Depression Scale	HADS	5	4.4	(Zigmond and Snaith, 1983)
EuroQol-5 Dimensions	EQ-5D	5	4.4	(The EuroQol Group, 1990; Dolan, 1997; Herdman *et al.*, 2011)
Patient Health Questionnaire	PHQ	5	4.4	(Spitzer, Kroenke and Williams, 1999)
Oxford Shoulder Score	OSS	4	3.5	(Dawson, Fitzpatrick and Carr, 1996)
Other*		90	78.9	
Total^		124	108.8	

*Only PROs that were used in four or more trials are listed separately.

^The total number of included trials is 114. Eight trials used more than one PRO as primary outcomes, including two trials that used three PROs, and six trials that used two PROs as primary outcomes.

Table 3.5 Characteristics of PROs that are used as primary outcomes in the 114 included trials

Subjects	N	%	Total
primary outcome clearly reported	107	93.9	114
primary outcome over one PRO	8	7.0	114
co-primary outcome non-PRO	14	12.3	114
primary preference-based PROs	6	5.3	114
primary outcome satisfaction	2	1.8	114
primary outcome proxy-reported	6	5.3	114
primary endpoint			114
single timepoint	71	62.3	
series of timepoints	41	36.0	
missing	2	1.8	
baseline & follow-up timepoints			
baseline collected?	101	88.6	114
over one follow-up?	107	93.9	114
score type (studies with one score type)			99
summary score	68	68.7	
subscale	19	19.2	
single item	12	12.1	
score type (studies with over one score type)			15
summary & subscales	11	73.3	
summary & single	4	26.7	

Note that overall summary score, subscale score, and single item score can be generated from multidimensional PROs, and unidimensional PROs are able to generate unidimensional summary scores and single item scores. The overall summary score covers more than one dimension, and it is composed of all dimensions covered by the PRO; the subscale score is composed of more than one item and it covers one or more than one dimensions; the single item score is composed of only one item, usually resulting from a global question and in the format of Likert scale or visual analogue scale (VAS).

The majority of trials stated the proportion of missing PRO data (109/114), developed strategies to deal with missing data (99/114), and imputed missing data using various methods such as mean imputation (89/114). In 40/114 studies, missing data were imputed as part of a sensitivity analysis to check the robustness of the primary analysis strategy that did not consider missing data.

The statistical methods applied for primary analyses were clearly defined in 79/114 trials, and the use of univariable or multivariable methods for primary analyses were justified in 92/114 trials. Except for two trials that did not specify the timing and number of post-randomisation assessments,(Little *et al.*, 2009, 2014) 72/114 clearly defined the single timepoint used for the primary analysis (e.g. x-months post-baseline), and 40/114 used the repeated post-baseline outcomes for the primary analysis.

The statistical methods for the primary analysis of PROs that were used as primary clinical effectiveness outcomes are shown in Table 3.6. Seven of the 27/114 trials reporting the use of univariable methods for primary analysis employed unadjusted regression methods to estimate treatment effects. These seven trials did not adjust for other covariates in the model besides randomised group, including three trials used linear regression, one used linear mixed model, two used ordinal logistic regression and one used binary logistic regression. The linear mixed model (45/114), linear regression (29/114) and ANCOVA (13/114) were the most popular methods among multivariable methods. Of the 45 trials using linear mixed models for the primary analysis, 23 trials conducted a repeated measures analysis, and the remaining trials did not consider the repeated post-randomisation outcomes in the primary analysis.

Longitudinal data was analysed in 100/114 trials. The repeated measures analysis was conducted as primary analysis in 39/114 trials, including seven used response features analysis with quality-adjusted survivals, and 27 used modelling methods. In seven trials using response feature analysis in primary analysis, four used linear regression or ANCOVA, two used linear mixed models, and one used survival analysis. Repeated measures ANOVA were used in six trials for longitudinal analysis. Around 70% of trials (23/33) conducting repeated measures analysis in primary analysis used linear mixed models. Three trials used GLM with coefficients estimated by GEE for the longitudinal primary analysis, including one as an extension to ordinal logistic regression (Kennedy *et al.*, 2006) and two as an extension to linear regression (Lewis *et al.*, 2006; Molassiotis *et al.*, 2013). Only one trial with quality-adjusted survivals used Cox regression for the primary analysis (Russell *et al.*, 2013).

Bootstrapped CIs were calculated after using a general linear model (including *t*-test, ANCOVA, linear regression) or linear mixed model in six trials due to the skewness of PRO scores (Kerry *et al.*, 2000; Morrell *et al.*, 2000; Symmons *et al.*, 2005; Weindling *et al.*, 2007; Sharples *et al.*, 2018; Shawo *et al.*, 2020). One trial did not conduct any other statistical analysis except for calculating the bootstrapped CIs for the treatment estimate (Wiggins *et al.*, 2004). Robust SEs can be used to estimate CIs and the calculation of test statistics (and associated P-values). Six trials reported the use of robust SEs based on regression methods for the primary analysis (Kennedy *et al.*, 2003, 2006; Weindling *et al.*, 2007; Beard *et al.*, 2020; Francis *et al.*, 2020), and three trials reported the use of robust SEs for the longitudinal analysis in the non-primary analysis (Chalder *et al.*, 2012; Molassiotis *et al.*, 2013; Goodyer *et al.*, 2017).

Of 106 trials that used multivariable methods, 98 trials clearly reported the covariates adjusted in the primary analysis. Among them, 85 trials adjusted for baseline score of the PRO, and three trials modelled the change of PRO from baseline in the primary analysis (Goodacre *et al.*, 2014; Williams *et al.*, 2015; Clarke *et al.*, 2016). The use of random effects for the primary analysis of PROs were clearly specified in 47 trials: 44 used linear mixed models, and the remaining three used repeated measures ANOVA (Peveler *et al.*, 2005), binary logistic mixed model (Pickard *et al.*, 2020), and ordinal logistic mixed model (Cooper *et al.*, 2019). The most common random factors applied in the multivariable methods were therapists, centres (i.e. hospital sites), and individual patients.

Table 3.6 Statistical methods for the primary analysis of PROs that are used as primary outcomes

Statistical methods	**N**	**%**	**Total**
Univariable methods			27
t-test	11	40.7	
unadjusted regression methods	7	25.9	
Wilcoxon rank-sum test (Mann-Whitney U test)	4	14.8	
Chi-squared test	3	11.1	
Kruskal-Wallis test	1	3.7	
log-rank test	1	3.7	
Multivariable methods^			106
linear mixed model	45	42.5	
linear regression	28	27.4	
ANCOVA	13	12.3	
linear regression with GEE	2	1.9	
binary logistic regression	8	7.5	
binary logistic mixed model	1	0.9	
ordinal logistic regression	4	3.8	
ordinal logistic mixed model	1	0.9	
repeated measures ANOVA	6	5.7	
survival analysis	1	0.9	
Repeated measures analysis			39
linear mixed model	23	59.0	
response feature analysis	7	17.9	
repeated measures ANOVA	6	15.4	
GLM with GEE	3	7.7	

^106 trials used multivariable methods for the analysis of PROs, including four trials that used two different methods for the primary analysis of PROs. ANCOVA, analysis of covariance; ANOVA, analysis of variance; GLM, generalized linear model; GEE, generalized estimating equation. Note that the three categories (i.e. univariable methods, multivariable methods, and repeated measures analysis) are not mutually exclusive, and more than one method can be used to analyse the primary endpoint.

3.3.4 Trend of using statistical methods for the primary analysis of PROs

The change of statistical multivariable methods applied for the primary analysis of PROs that are used as primary clinical effectiveness outcomes from 1999 to 2020 is shown in Figure 3.3. In general, there is an increasing trend for using complex/advanced models in most recent years for the analysis of PROs. The linear regression, ANCOVA and repeated measures ANOVA were the most popular regression methods used from 1999 to 2010, used in around 63% of trials averagely, but this popularity dropped to 28.6% during 2011-2015 and 30.0% during 2016-2020. In contrast, the proportion of trials using the linear mixed model and linear regression with GEE for correlated outcomes as primary analysis methods generally increased across the observation period, from 11.1% to 54.0%, which witnessed the linear mixed model becoming the most popular regression methods in recent years. Meanwhile, the use of logistic regression across years remained comparatively stable from 1999 to 2020. The trend of using binary logistic regression for primary PROs slightly decreased from 11.1% to 4.0% over time, and the proportion of trials using ordinal logistic regression for the primary analysis remained small.

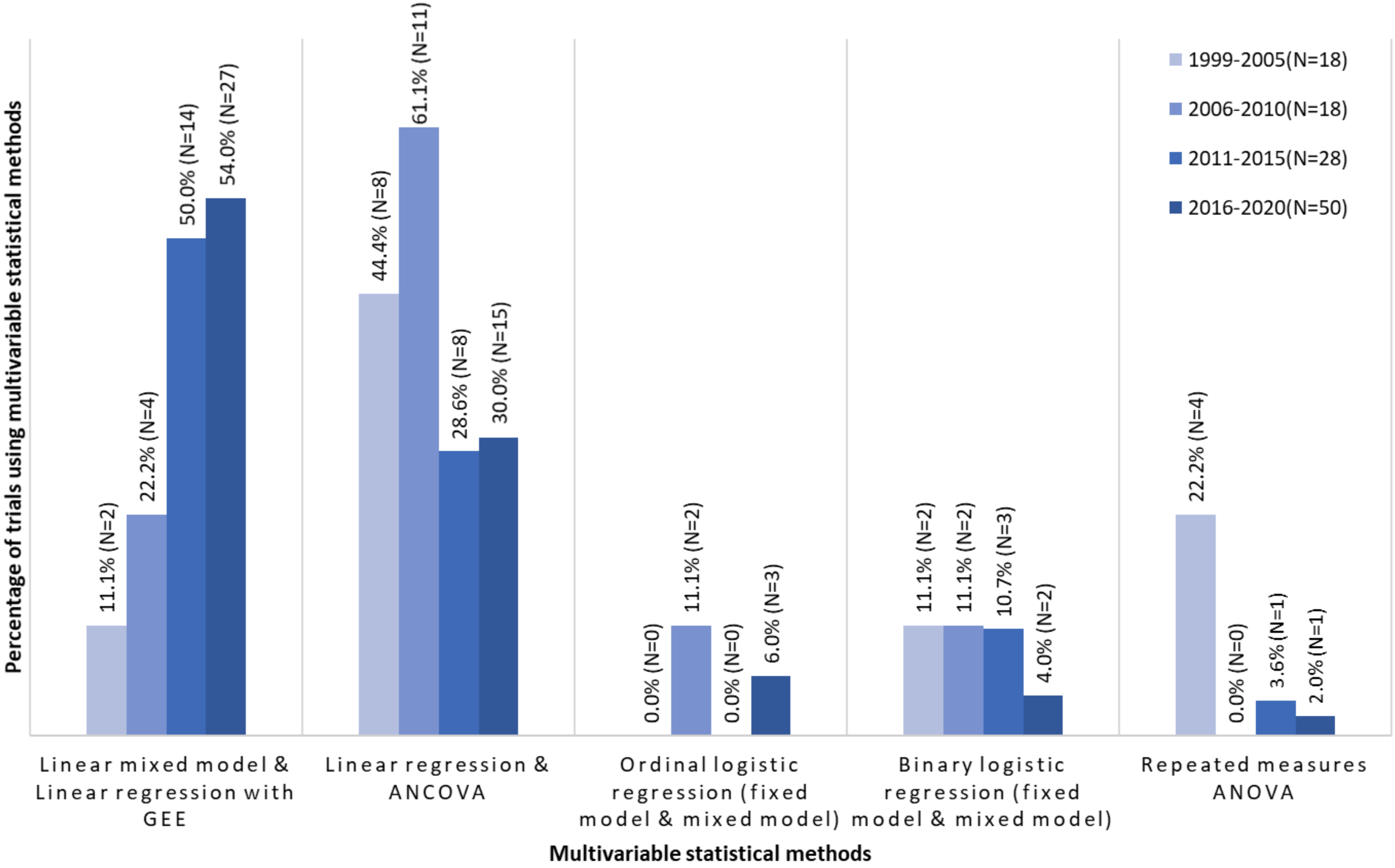


Figure 3.3 Percentage of trials using multivariable methods for the primary analysis of PROs from 1999 to 2020

(N=xx) denotes the number of trials published in the specified period. As the survival analysis was used only in one trial, it is not shown in this graph. As this graph only summarised multivariable methods and one trial could use two or more multivariable methods for the primary analysis, the number of trials summarised in this graph may not equal the total number of included trials. ANCOVA, analysis of covariance; ANOVA, analysis of variance; GEE, generalized estimating equation.

3.4 Discussion

This chapter has systematically conducted a comprehensive review that summarised how frequently PROs have been used and what statistical methods have been applied for the primary analysis of PROs in RCTs published by the UK NIHR HTA Journal between 1997 and 2020.

This review found that 82.8% (251/303) of the included published trials used PROs as clinical outcomes, and 37.6% (114/303) of the trials used PROs as primary outcomes. Multivariable methods that adjusted for additional covariates besides the randomised group were conducted in 106 (93%) of the 114 trials. The linear mixed model was the most used regression method for the primary analysis, and *t*-test or Wilcoxon rank-sum test were the most popular univariable methods for the primary analysis of PROs. This result is consistent with the review conducted by Pe *et al.* (2018) which summarised the statistical methods applied for PROs in oncology. A decrease in the use of binary logistic regression was found in this review, possibly because the dichotomised outcome retains less information from the PROs compared to other score types (Shields *et al.*, 2015).

Ordinal regression, binomial regression and beta regression that were identified from the identified reviews in Chapter 2 were rarely seen applied for the analysis of PROs in this HTA review. Arostegui, Núñez-Antón and Quintana (2007, 2012) recommended using ordinal logistic regression with random effects model, beta-binomial regression or binomial-logit-Normal regression for continuous or ordinal PRO data after testing distributional assumptions. However, only one of the five included trials that used ordinal logistic regression for the primary analysis considered random effects (Cooper *et al.*, 2019). Neither beta-binomial regression nor binomial-logit-Normal regression were used by the 114 trials.

Compared to this HTA review that summarised statistical methods for the analysis of PROs in different disease areas, the identified reviews in Chapter 2 reported a high proportion of cancer studies using time-to-event data. A large proportion of HTA trials used linear mixed models compared to the cancer studies, and the statistical methods for ordered data were barely reported in use in cancer studies. None of the two identified cancer trials from this HTA review conducted ordered logistic regression. One trial used primary survival analysis for quality-adjusted life years estimated by EQ-5D (Russell *et al.*, 2013), and the other used a linear mixed model for the analysis of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30) (Prescott *et al.*, 2007).

The multivariable methods for the analysis of PROs that were used as primary outcomes changed over time. The general linear model (including *t*-test, ANOVA, ANCOVA and MLR) and linear mixed model were widely used over the past two decades, which could possibly result from the frequent use of continuous data type of response variables. This is consistent with the findings from the identified reviews, which summarised statistical methods for analysing PROs in clinical trials in Chapter 2. Repeated measures analysis was popular for trials with more than one follow-up timepoint. The trend of using repeated measures ANOVA in early years was replaced by using linear mixed models in recent years, which can be explained by the increasing complexity of trial designs and from recommendations on using linear mixed models over repeated measures ANOVA for the longitudinal analysis of PROs (Walters, 2009; Arostegui, Núñez-Antón and Quintana, 2012).

Various statistical methods could be applied for the analysis of a PRO. In the HTA review, 4/106 studies conducted more than one multivariable analysis for the primary analysis of PROs. Similarly, the review of cancer trials by Pe *et al.* (2018) found that more than one statistical method were used for the primary analysis of PROs in cancer trials. In addition, there is an obvious growing trend of using complex models such as linear mixed models (with both fixed and random effects) for the primary analysis of PROs over the observed period, while conventional methods (such as the *t*-test, MLR and ANCOVA) have been widely used. However, there is a delay in applying newly developed methods. An example is the linear mixed model that has been applied for the analysis of PROs since 2000 (Pinheiro and Bates, 2000), but becomes a popular method for statistical analysis among the identified trials from around 2014.

Even if methods that can be used to deal with the bounded, skewed, ordinal and multidimensional features of PRO data have been developed, these methods have not been widely applied to PRO data, especially in clinical trials settings where the commonly used methods are still conventional methods such as ANOVA and Chi-squared test. This might be due to the following reasons. When assumptions hold in conventional methods, it is unnecessary to apply complex statistical methods over simple ones if similar and reliable estimates can be produced by both simple and advanced methods. Although the violation exists, such as the violation of Normality assumptions of residuals caused by PRO data, it is less likely to challenge the robustness of general linear models (Walters and Campbell, 2004). This fact might explain why the obvious violation of Normality assumptions did not stop researchers from using linear regression or *t*-test. In addition, some statistical methods are difficult to apply in statistical software and the estimations of some complex methods are difficult to interpret, researchers are therefore reluctant in using newly developed methods when conducting statistical analysis.

PROs tend to generate data with discrete, skewed, and bounded distributions that are not Normally distributed, and the assumptions for statistical methods such as the *t*-test, linear regression and ANCOVA may not be valid. However, Heeren and D’Agostino (1987) have demonstrated the robustness of the two independent samples *t*-test when applied to three-, four-, and five-point ordinal scaled data using assigned scores, in sample sizes as small as twenty subjects per group. Sullivan and D’Agostino (2003) have expanded this work to account for a covariate when the outcome is ordinal in nature. They again assign numeric scores to the distinct response categories and compare means between treatment groups adjusting for a covariate reflecting a baseline assessment measured on the same scale. Their simulation study shows that in the presence of three-, four- and five-point ordinal data and small sample sizes (as low as twenty per group) that both ANCOVA and the two independent sample *t*-test on difference scores are robust and produce actual significance levels close to the nominal significance levels. Furthermore, statistical theory says that if the distribution of an outcome variable is Normal, so will be the distribution of the sample mean for that outcome variable. Much more importantly, even if the distribution of the outcome is not Normal, that of the sample mean will become closer to the Normal distribution as the sample size gets larger. This is a consequence of the Central Limit Theorem (CLT). The Normal distribution is strictly only the limiting form of the sampling distribution as the sample size increases to infinity, but it provides a remarkably good approximation to the sampling distribution even when the sample size is small, and the distribution of the outcome variable is far from Normal (Armitage, Berry and Matthews, 2002). Thus, conventional statistical methods such as the *t*-test, MLR and ANCOVA for analysing PROs are robust to the violation of assumptions for moderate to large sample sizes (Walters, 2009).

A clear classification of the terminology of the statistical methods is desired. It is a historical problem that the names of statistical methods are confusing (e.g. general linear model vs. generalized linear model), and multiple terms can be used to describe the same method, for example, the proportional-odds model, ordered logit model and ordinal logistic regression refer to the same regression technique. In addition, there are various ways to group statistical methods, depending on the study aim. For instance, the linear mixed model is categorised as a method for between-group difference considering the classification of within or between group difference, and it can also be categorised as a model for repeated measures analysis when accounting for time. Thus, researchers should be clear and cautious when describing the exact statistical method for the analysis.

To the best of our knowledge, this study is by far the largest review of trials (with 114 studies) published by the HTA Journal which analysed the frequency of using PROs and the statistical methods for the analysis of PROs. The reviews by Pe *et al.* (2018) (breast cancer); Hamel *et al.* (2017) (lung cancer), and Fiteni *et al.* (2016) (lung cancer) had sample sizes of 66, 33, and 27 articles respectively. Compared to reviews that only concentrate on oncology (Fiteni *et al.*, 2016, 2019; Turner-Bowker *et al.*, 2016; J.-F. Hamel *et al.*, 2017; Pe *et al.*, 2018; Nielsen *et al.*, 2019), this review summarised details in the use and analysis of PROs in RCTs with a range of clinical areas.

It is noteworthy that the proportions of trials using PROs reported in this review represents the average rate of HTA trials focusing on different clinical areas, and when considering specific disease(s) or selecting different database(s), the proportions may vary. For example, Pe *et al.* (2018) identified 3/66 (5%) and 46/66 (70%) randomised controlled trials of locally advanced and metastatic breast cancer using PROs as primary and secondary endpoints respectively. Marandino *et al.* (2018) reviewed 446 cancer trials published in major journals between 2012 and 2016, and found that PRO or QoL was a primary end point in five trials (1.1%), a secondary end point in 195 trials (43.7%), an exploratory endpoint in 36 trials (8.1%), while in the remaining 210 (47.1%) QoL was not listed at all among study end points. Our review found that three of 18 cancer trials (17%) used PROs as primary outcomes, and 13/18 (72%) used PROs as secondary outcomes. Our results showed that PROs were more frequently used for health problems such as mental and musculoskeletal disease.

This study has the following limitations.

First, this review only looked at UK trials funded by the NIHR HTA Programme, which may represent a limitation in terms of the generalisability of the findings. It is possible that statistical methods are used differently in industrial funded trials or in trials in other countries. However, as the NIHR HTA Journal intends to publish all NIHR funded projects, it has less publication bias compared to journals that only publish positive outcomes, and the information related to PROs in other journals are not reported in as much detail. The extracted statistical methods for the analysis of PROs from this review are consistent with those included from other similar reviews (Fiteni *et al.*, 2016, 2019; Turner-Bowker *et al.*, 2016; J.-F. Hamel *et al.*, 2017; Pe *et al.*, 2018; Nielsen *et al.*, 2019).

Second, there might be other appropriate methods for the analysis of PROs that were not included in this review. This review mainly analysed RCTs with PROs that were used as primary outcomes because the primary outcomes and the corresponding statistical methods were more explicitly reported. Methods such as probit model that were identified from the secondary analyses or sensitivity analyses in the 114 trials are not presented. Other potentially appropriate methods that might be available for the analysis of PROs cannot be identified from this literature review.

Third, trials with PROs only used as cost-effectiveness outcomes were excluded. This is because the statistical strategies for clinical effectiveness and cost-effectiveness outcomes may vary, and cost-effectiveness analysis (CEA) produces both cost and clinical effectiveness outcomes. If PROs for CEA were included in this review, the proportion of included trials would increase, as there were some studies using EQ-5D for the primary CEA. It could be argued that the analysis of effectiveness estimated by PROs in CEA also requires appropriate statistical methods, but estimands for a health economic analysis could be different from those for a clinical analysis as they hold different purposes to conduct these analyses. Therefore, we believe it is justified to make this exclusion.

Fourth, we used a broad definition for a PRO and a small number of trials (seven) used PROs that were specifically developed for the trial and were not validated in another external study. The inclusion of such non-validated instruments as primary outcomes should be discouraged, and may have affected the results, although the characteristics of these PROs (Likert or VAS) are similar to those of the PROs that have been formally validated. We believe that it is not unreasonable to assume that the statistical analysis of such outcomes would be similar to the analysis of validated PROs. Another potential limitation is the large time window, 1997-2020, chosen for the review. This may introduce some variability and potential heterogeneity in the trials included in the review, but on the positive side, it allows to test time trends in the type of statistical methods used in the trials.

Last, the information related to PROs and statistical methods were not clearly reported in some trials. Although assumptions have been made based on the context where some required information was not explicitly stated, it is possible that the data extracted was inconsistent with researchers' intention. However, as the data have been extracted for all reports by one reviewer there is consistency in the interpretation and assumptions made.

Insufficient reporting exists in some parameters for the statistical analysis of PROs, such as the pre-specified PRO definition, specific hypothesis and the statistical methods. To produce explicit reports, it is recommended that researchers follow specific guidelines that can instruct the reporting of using PROs in RCT papers and protocols such as the CONSORT-PRO (Calvert *et al.*, 2013), SPIRIT-PRO (Calvert *et al.*, 2018), and the standards for the analysis of PROs in cancer RCTs (Coens *et al.*, 2020). In addition to these guidelines, the following information is recommended to consider for explicit reporting of the PROs and the statistical methods for analysing PROs.

For specifying primary outcomes: What PRO is used as the primary outcome? Who is the informant of the PRO? What outcome is derived from the PRO as the primary outcome, including the score directly generated from the PRO and other outcomes generated from the PRO? What is the specific timepoint for the primary endpoint?

For specifying statistical methods: Are a specific timepoint or time series used for analysis? What are the statistical methods used for the analysis? Are there any extensions such as random effects applied for the analysis? If so, what is it? What covariates are adjusted? Are there any add-on techniques such as bootstrapped CIs or robust SEs applied for the analysis? What is the strategy to deal with missing data? Is there any assumption made for missing data?

In conclusion, the majority of trials funded by the NIHR HTA Programme used PROs as clinical outcomes. Although there is an increasing trend of using complex models (e.g. mixed effects), conventional methods such as linear regression remain widely used for the analysis of PROs, despite the potential violation of their assumptions. Statistical methods developed to address these violations when analysing PROs, such as beta-binomial regression, are not routinely used in practice. Various methods for the analysis of PROs have been identified from this review, but it is still unknown which methods are the most appropriate for the analysis of PRO data in RCT settings.

The next chapter will establish specific research aims and objectives given the evidence extracted from the two literature reviews in Chapter 2 and Chapter 3. As the recommendations by different groups may vary, it is necessary to establish criteria for the evaluation of statistical methods in order to validate the appropriateness of recommended methods. Summary tables of available statistical methods and statistical properties to evaluate statistical methods for the analysis of PROs in RCTs will also be presented in the next chapter.

Chapter 4 Research aims and objectives

4.1 Summary of previous work

The previous chapters introduced the PROs as a widely used measurement to evaluate the clinical effectiveness from the patient perspective; and explained the data features of PROs that are likely to be bounded, discrete, skewed, and multidimensional, which increases the complexity to analyse PRO data in RCT settings. Various statistical methods for the analysis of PROs that have been proposed in theory and have been applied in publicly funded RCTs in the UK as well as potential statistical properties that can be considered for the evaluation of statistical methods are identified and summarised.

The method review in Chapter 2 provides an insight into what statistical methods have been proposed, developed, and adopted in recent years, and what statistical properties can be considered to compare and contrast models. The identified reviews reported the widely use of general linear models (e.g. *t*-test, ANOVA, ANCOVA and linear regression), and their non-parametric counterparts (e.g. Mann-Whitney U-test or Wilcoxon signed rank test), whereas the identified method studies proposed more complex and advanced statistical methods for the analysis of PROs, such as beta-binomial regression.

The HTA review in Chapter 3 summarises statistical methods used in RCTs that are published in the NIHR HTA Journal. It shows that PROs are widely used to measure the clinical effectiveness in publicly funded trials in the UK. Over 80% of RCTs used PROs as clinical outcomes, and around half of them used PROs as primary outcomes. The majority of included trials applied general linear models (e.g. *t*-test, ANOVA, ANCOVA and linear regression) and repeated measures analysis (e.g. repeated measures ANOVA, linear mixed model and GLM with coefficients estimated by GEE) for the primary analysis of PROs. An increasing trend of using linear mixed model over repeated measures ANOVA is seen to analyse repeated measurements (i.e. longitudinal analysis).

An overview of available statistical methods for the analysis of PROs and statistical criteria from different perspectives for the evaluation of statistical methods are extracted from the two reviews and summarised in Table 4.1 and Table 4.2.

Both reviews show that the general linear models, their extension for longitudinal analysis, and their non-parametric counterparts are popularly applied for the primary analysis of PROs in published RCTs. This indicates that researchers tend to analyse PROs as continuous outcome regardless of the likely discrete and skewed nature of PRO data. Although various statistical methods that have been developed to deal with the bounded, skewed, or ordinal features of PRO data are identified in the method review, these methods are not seen applied for the analysis of PROs in the published RCTs.

4.2 Research gap

In the light of the two comprehensive reviews, a research gap between the statistical methods that have been developed by methodologists and statistical methods that have been used for the analysis of PROs in RCTs is identified. Statistical methods such as beta-binomial regression, beta regression, and CLAD regression that are proposed or recommended for the analysis of PROs, are found to be rarely used in practice. Therefore, there is a need to evaluate whether these recommended methods can fit RCT datasets better than classical methods such as MLR.

A set of criteria is to be established and clearly stated before the evaluation of statistical methods, since it is found that studies comparing different statistical methods do not share the same set of criteria for the evaluation of statistical methods. These studies vary from the study design, the study population, the set of statistical methods to compare, and especially the criteria to evaluate these statistical methods. It is unlikely to have different studies using different sets of criteria to reach the same conclusion.

Furthermore, most studies developed and compared statistical methods with evidence from a single dataset, which potentially eliminates the robustness of the conclusion when being extrapolated to other datasets with different types of PROs. Such that multiple RCT datasets that focus on a range of disease areas shall be used to apply different statistical methods for the analysis of PROs.

This research proposes to focus on the SF-36 PRO, i.e. apply different statistical methods to RCTs with SF-36 in the empirical analysis and use the typical distribution of SF-36 dimension scores to instruct simulation analysis, since the SF-36 or SF-6D was found the most used PRO in use with the evidence from the published 114 RCTs in the HTA review.

4.3 Research objectives

Regarding the identified research gap, the following specific aims and objectives are established:

First, in Chapter 5, a set of desired criteria for the evaluation of statistical methods presented in Table 4.2 will be established to filter the statistical methods for the analysis of PROs in RCTs that are identified from previous chapters as shown in Table 4.1. The statistical methods that passed the filtration will be carried forward to the empirical analysis.

Second, the technical details of the filtered statistical methods will be described in Chapter 6, together with the commands to run in the computational software and the possible interpretation of the estimates from each statistical method. An example RCT dataset with SF-36 as its primary outcome will be used to explain the process of fitting different statistical methods to PRO data. The filtered statistical methods will be applied to multiple RCT datasets with SF-36 as clinical outcomes in Chapter 7. The list of statistical methods will then be narrowed down to carry forward to the simulation analysis.

Third, the simulation protocol to compare model performance of the narrowed list of statistical methods that are carried forward from the empirical analysis will be proposed in Chapter 8. The simulation analysis will be conducted using Monte Carlo methods in Chapter 9, to compare the performance measures of these statistical methods in terms of estimating the predefined treatment effect of PROs under a range of scenarios in RCT settings.

Finally, recommendations on what statistical methods are the most appropriate for the analysis of PROs in RCT settings will be made and discussed in Chapter 10, according to the technical details and model theories of included statistical methods, and their model performances in the empirical analysis and the simulation analysis.

Table 4.1 A list of potential statistical methods for the analysis of PROs

Statistical methods classification			**Reference**
Multivariable methods (that allow or adjust for covariates besides randomised group)			
Methods for correlated responses			
		Survival analysis	(Coens *et al.*, 2020)
		Generalized linear mixed model	(Qian *et al.*, 2000; Lee and Daniels, 2008; Zou, Carlsson and Quinn, 2010; Najera-Zuloaga, Lee and Arostegui, 2019)
		GLM with parameters estimated by GEE	
		Repeated measures ANOVA	(Qian *et al.*, 2000; Walters and Campbell, 2004; Parsons, 2013; Zheng, Qin and Tu, 2017)
Methods for uncorrelated responses			
Continuous			
	Normal	MLR (ANOVA / ANCOVA)	(Walters, Campbell and Lall, 2001; Walters and Campbell, 2004, 2005; Arostegui, Núñez-Antón and Quintana, 2007, 2012)
	Normal but censored	Tobit regression	(Austin, Escobar and Kopec, 2000; Pullenayegum *et al.*, 2010)
		CLAD regression	(Austin, 2002; Pullenayegum *et al.*, 2010)
	Skewed	Quantile regression	(Leng and Zhang, 2014)
	Bounded	Beta regression	(Zou, Carlsson and Quinn, 2010; Hunger, Baumert and Holle, 2011)
		Fractional logistic regression	(Meaney and Moineddin, 2014)
Categorical			
	Binary	Binary logistic regression	(J.-F. Hamel *et al.*, 2017)
		Binary probit regression	
	Ordered	Ordered logit	(Qian *et al.*, 2000; Lall *et al.*, 2002; Lee and Daniels, 2008; Manuguerra and Heller, 2010; Arostegui, Núñez-Antón and Quintana, 2012; Parsons, 2013)
		Ordered probit	
		Beta-binomial regression	(Arostegui, Núñez-Antón and Quintana, 2007, 2012; Najera-Zuloaga, Lee and Arostegui, 2018, 2019)
		Binomial-logit Normal regression	
	Unordered	Multinomial logit	
		Multinomial probit	
Count			
		Negative binomial regression	
		Poisson regression	

Continued on the next page

Statistical methods classification		**Reference**
Univariable methods (that do not adjust for covariates)		
	Log-rank test (Kaplan Meier)	(Coens *et al.*, 2020)
	t-test	
	Sign test	(Coens *et al.*, 2020)
	Wilcoxon signed rank test	(Pe *et al.*, 2018)
	Mann-Whitney U test	(Qian *et al.*, 2000)
	Chi-squared test for independence	(Fiero *et al.*, 2019)
	Fisher's exact test	(Fiero *et al.*, 2019)
	Mantel-Haenszel test	(Coens *et al.*, 2020)
	McNemar's test	(Coens *et al.*, 2020)
Add-ons		
	Robust standard errors	(Pullenayegum *et al.*, 2010)
Resampling methods	Bootstrapping	(Walters, Campbell and Lall, 2001; Walters and Campbell, 2004; Moerkerke *et al.*, 2005;
	Permutation	Arostegui, Núñez-Antón and Quintana, 2012; Wang and Tu, 2020)

ANOVA, analysis of variance; ANCOVA, analysis of covariance; CLAD, censored least absolute deviations; GEE, generalized estimating equation; MLR, multiple linear regression.

Table 4.2 A summary of potential characteristics to be considered for the statistical analysis and reporting of PROs

Characteristics domains	Further details	Reference
PRO information		
Specific the PRO concepts		(Pe *et al.*, 2018)
Primary or secondary endpoint?	Strategy for sample size calculation if primary outcome	(Calvert *et al.*, 2013, 2018)
	Specific timepoint(s) of the primary endpoint	(Qian *et al.*, 2021)
	Co-primary PROs?	(Qian *et al.*, 2021)
Describe the PRO		(Calvert *et al.*, 2018)
Specific dimensions of interest		(Calvert *et al.*, 2013, 2018; Fiteni *et al.*, 2016)
Proxy-reported PRO?	If yes, who is the proxy?	(Qian *et al.*, 2021)
Specific research question and hypothesis on PROs		(Calvert *et al.*, 2013; Fiteni *et al.*, 2016; Pe *et al.*, 2018; Coens *et al.*, 2020)
Data structure		
PRO at baseline		(Calvert *et al.*, 2013; Pe *et al.*, 2018; Fiteni *et al.*, 2019)
PRO at follow-up timepoints (for multidimensional PRO results from each dimension and each timepoint)		(Calvert *et al.*, 2013; Fiteni *et al.*, 2016; J. F. Hamel *et al.*, 2017; Pe *et al.*, 2018)
Multiple dimensions		(J. F. Hamel *et al.*, 2017; Pe *et al.*, 2018)
PRO data type (summary score, subscale, single item)		(Qian *et al.*, 2021)
Statistical analysis		
Comparison	Compare between treatment arms	(Coens *et al.*, 2020)
	Within group difference or between group difference	(Nielsen *et al.*, 2019)
Assumptions	Data distribution assumption	(Austin, 2002; Arostegui, Núñez-Antón and Quintana, 2007; Hutton and Stanghellini, 2011)
	Missing data assumption	(Fairclough, 2004)
Ability to adjust …	Confounding factors	(Walters, Campbell and Lall, 2001; Coens *et al.*, 2020)
	Baseline score	(Coens *et al.*, 2020)
	Random effect	(Qian *et al.*, 2021)
Interpretation	Statement of MCID	(Fiteni *et al.*, 2016; J.-F. Hamel *et al.*, 2017)
	Clinically relevant statistical estimates	(Saver, 2011; Arostegui, Núñez-Antón and Quintana, 2012; Calvert *et al.*, 2013; Fiteni *et al.*, 2016; Pe *et al.*, 2018; Coens *et al.*, 2020)

Characteristics domains	**Further details**	**Reference**
Statistical analysis		
Efficiency	Ability to detect group differences when truly exist	(Hutton and Stanghellini, 2011; Saver, 2011)
Dealing with different data characteristics	Bounded data	(Hutton and Stanghellini, 2011)
	Ordinal data	(Lall *et al.*, 2002; Arostegui, Núñez-Antón and Quintana, 2007, 2012; Najera-Zuloaga, Lee and Arostegui, 2018, 2019)
	Clustered data	(Coens *et al.*, 2020)
Model fit	Goodness-of-fit	(Arostegui, Núñez-Antón and Quintana, 2007)
	Account for overdispersion	(Arostegui, Núñez-Antón and Quintana, 2012)
	Procedure to control the Type I error	(Fairclough, 2004; Calvert *et al.*, 2018; Fiteni *et al.*, 2019; Nielsen *et al.*, 2019)
Making prediction		(Austin, 2002)
Multivariate analysis		(Fiteni *et al.*, 2016)
Longitudinal analysis		(Qian *et al.*, 2000)
Robustness / sensitivity analysis	Heteroscedasticity, non-normality of errors, missingness	(Austin, 2002; Shields *et al.*, 2015)
Loss of information		(Shields *et al.*, 2015)
Statistical power		(Shields *et al.*, 2015)
Statistical significance		(Shields *et al.*, 2015)
Missing data management		
Profile of missing data at baseline		(Fiteni *et al.*, 2016; Calvert *et al.*, 2018)
Strategy to handle missing data		(Calvert *et al.*, 2013; Fiteni *et al.*, 2016; Pe *et al.*, 2018; Coens *et al.*, 2020)
Compliance rates		(J. F. Hamel *et al.*, 2017; Pe *et al.*, 2018)
Study population		(J. F. Hamel *et al.*, 2017; Pe *et al.*, 2018)

Chapter 5 Filtration of the identified statistical methods with justifications

5.1 Introduction

A list of statistical methods has been identified from two literature reviews in Chapter 2 and Chapter 3. These two reviews identified 29 statistical methods that can potentially be used for the analysis of PROs in RCT settings. Pragmatically, this list of 29 statistical methods is long with too many methods to compare and evaluate. Therefore, a process of screening and filtering the identified statistical methods is required to reduce the number of methods to a manageable size.

In practice, there are disagreements on what statistical methods should be applied for analysing PROs. For example, Austin (2002) recommended CLAD regression for analysing health utility because of its low prediction error and its robustness to heteroscedasticity and non-Normality of error; whereas Pullenayegum *et al.* (2010) recommended linear regression with robust SEs or nonparametric bootstrap as a simple and valid approach for analysing health utility. As each method has its own strengths and weaknesses in analysing PROs, it may not be straightforward to decide which method is more appropriate than others without clarifying the criteria for assessment, and what is meant by ‘an appropriate statistical method’.

This chapter establishes statistical properties that are appropriate or desired in terms of the statistical methods for the analysis of PROs, and reduces the number of statistical methods to carry forward with a series of justifications.

5.2 Statistical properties that are desired for PRO analysis

The identified statistical properties for PRO analysis in Table 5.1, incorporating the assessment criteria for the evaluation of PRO analysis that were listed in the identified review papers, methodological articles and guidelines, are summarised in Table 4.2. The ability to estimate a treatment effect with its associated CIs is regarded as a key feature that a statistical method should have for analysing PROs in clinical trials (Moher et al., 2010). This is also associated with the ability to adjust for other confounding variables such as baseline scores.

Different statistical methods assume different data distribution assumptions for the outcome, and correspondingly have the ability to deal with different data features such as skewed, ordered, and bounded/censored data. Skewness implies that the PRO data is asymmetrically distributed. The ordered distribution means that the PRO scores only take specific values with natural ranking. Censoring or boundedness from above takes place when subjects with a value at or above some threshold, all take on the value of that threshold, so that the true value might be equal to the threshold, but it might also be higher. In the case of censoring from below, values that fall at or below some threshold can only take on the value of that threshold, so that the true value might be equal to the threshold or below. For example a PRO with scores bounded at 0 and 100, then subjects may have true underlying or latent PRO scores of below 0 or above 100, but the PRO is unable to measure these true values, so true underlying PRO scores of below 0 or above 100 are censored at 0 and 100 respectively.

Some methods account for other levels of complexity such as clustering of outcomes and time effects. Some statistical methods can still be robust regarding the violation of model assumptions. For example, MLR theoretically should be applied for Normally distributed outcome data, but it may also be used when applying to skewed, ordinal or bounded data after relaxing the model assumptions due to the CLT, and it can still produce valid estimates of the population mean (Lumley et al., 2002; Walters and Campbell, 2004, 2005).

Postestimation techniques can be applied after fitting statistical methods to a dataset for making predictions, testing model assumptions, and comparing model fit. For example, residual plots can be produced to test whether the model assumptions are violated. However in some occasions, assumptions of missing data and data distributions may be untestable (Arostegui, Núñez-Antón and Quintana, 2012; Smuk, Carpenter and Morris, 2017). Simulation analysis can be conducted to test the efficiency of the method i.e. whether the estimators (i.e. the statistical methods) can estimate the predefined ‘truth’ (i.e. the accuracy of the estimations) and whether a method is robust in various scenarios (Morris, White and Crowther, 2019; Boulesteix et al., 2020).

Other properties of statistical methods that may be adapted include whether the statistical method can result in a loss of information, whether the statistical method can handle missing data, what missing data management is conducted by the method, and whether the method is able to handle unbalanced design and maintain the intention-to-treat (ITT) population (Altman and Royston, 2006; Fiteni et al., 2016; J. F. Hamel et al., 2017; Pe et al., 2018; Coens et al., 2020).

With the pool of criteria that we identified from the method review (Table 5.1), we selected statistical properties that are desired in terms of an appropriate statistical methods for PRO analysis in RCTs, and can be evaluated before carrying out empirical analysis and simulation analysis. The criteria includes:

1. whether the method can compare two or more treatment arms;
2. whether the method can adjust for confounding factors, including baseline PRO score;
3. whether the estimated treatment effect from the method is of clinical relevance (i.e. the method can produce an estimate of treatment effect and associated CIs);
4. whether a method can handle a bounded/censored scale; and
5. whether recoding the PRO is required to use the statistical method.

Table 5.1 Identified statistical properties for the evaluation of statistical methods

Identified statistical properties	Further explanation
Comparison	Compare between two or more treatment arms Within group difference or between group difference
Assumptions	Data distribution assumption Missing data assumption
Ability to adjust …	Confounding factors Baseline score Random effect
Interpretation	Statement of minimum clinically important difference Clinically relevant statistical estimates
Efficiency	Ability to detect group differences when truly exist
Ability to handle …	Bounded data Censored data Ordinal data Clustered data Missing data Repeated measurements Multiplicity
Model fit	Goodness-of-fit Account for overdispersion Procedure to control the Type I error
Making prediction	Ability to make prediction Precision of the prediction
Robustness / sensitivity analysis	Heteroscedasticity, non-Normality of errors, missingness
Loss of information	
Uncertainty	Statistical significance Statistical power Confidence intervals
Handle unbalanced designs	
Calculate sample size	
Allow for time-varying covariates	
Ability to maintain intention-to-treat population	

Multiplicity is the potential inflation of the Type I error rate due to multiple testing, and Type I error is the probability that one falsely rejects the true null hypothesis. Loss of information occurs when discretising or dichotomising a variable into ordinal or binary data.

5.3 An MCDA framework for the filtration of statistical methods

A multi-criteria decision analysis (MCDA) framework is experimented for the filtration of these statistical methods. MCDA is a popular technique for decision making when it involves the comparison of various options with a range of criteria from different stakeholders with multiple judgements (Velasquez and Hester, 2013). The MCDA allows trade-offs among different options regarding various criteria which can facilitate the decision making process with transparency and comprehensiveness (Dodgson *et al.*, 2009; Diaby and Goeree, 2014). It can help clearly explain rationales for the decision on which statistical method to apply in different scenarios, and further bring consistency and transparency for decision-making.

A quantitative MCDA of statistical methods for PRO analysis is presented in the Appendix B using scoring and weighting systems elicited from the SISAQOL Consortium (Coens *et al.*, 2020). However, as a standard process of establishing expert panel, eliciting scores and weights, and deliberating on the ordering rank was not carried out, this quantitative MCDA is merely presented as a supplement to support the filtration process in this chapter.

5.4 Filtration of statistical methods

The filtration starts from the summary of available statistical methods for the analysis of PROs as shown in Table 4.1. The statistical methods for analysing PROs are classified into univariable methods and multivariable methods according to whether the method is able to adjust for both treatment group and other covariates such as baseline scores. Under the multivariable methods, the GLM which assumes a linear relationship between the dependent variable and explanatory variable through a link function is further categorised into methods for correlated responses, e.g. health utilities at 3, 6 and 12 months post-randomisation, and uncorrelated responses. Techniques for uncorrelated responses are further classified according to the data type i.e. time-to-event, continuous, categorical/ordinal, censored, and bounded that the statistical methods are initially designed to analyse is shown alongside the methods.

The rest of this section justifies the exclusion of univariable methods, statistical methods for correlated responses, multivariable methods that account for count data, unordered data, and binary data. The filtration process is presented by a flow diagram in Figure 5.1, where the methods highlighted in light green are included carried forward for the empirical analysis.

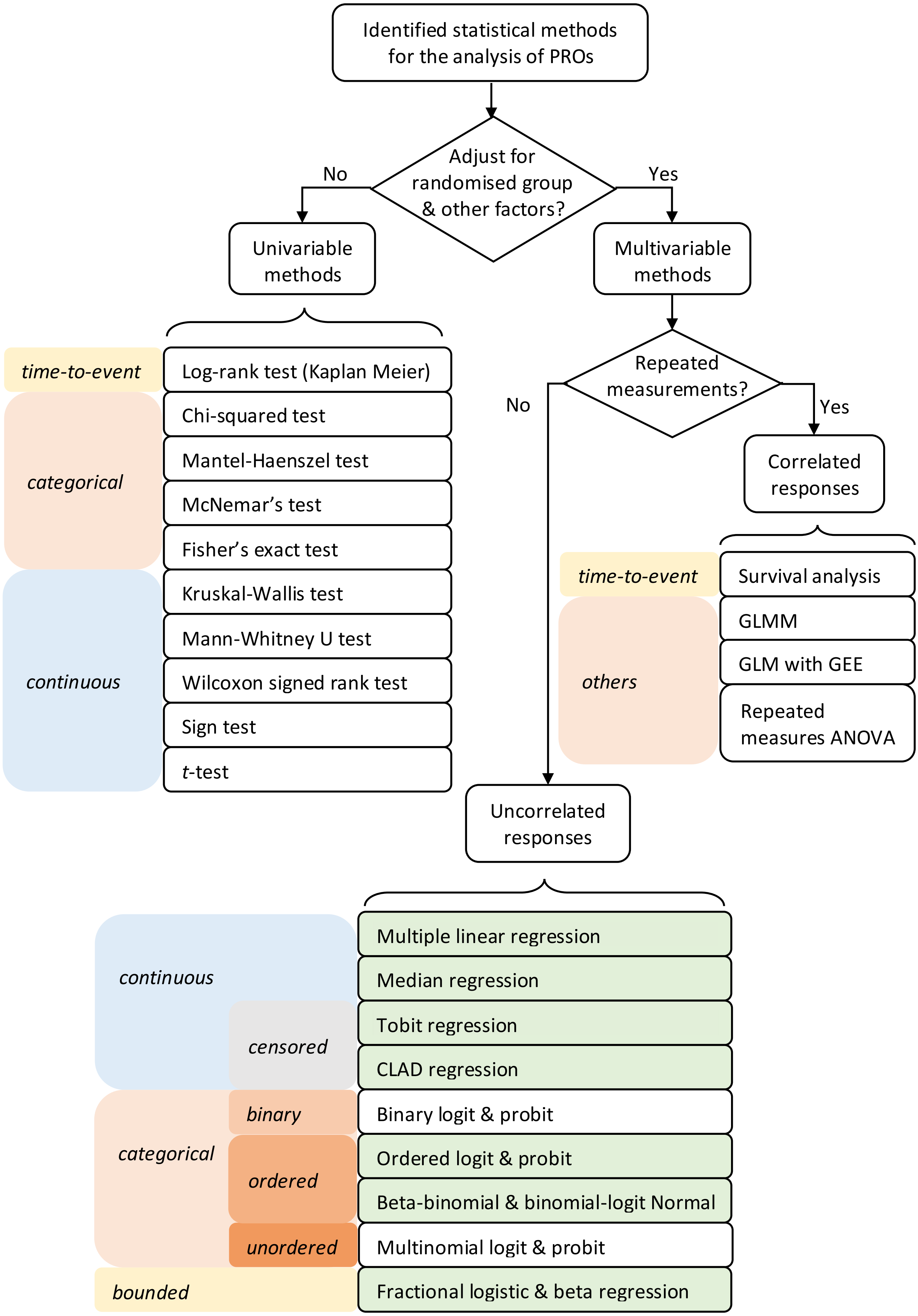


Figure 5.1 Flow diagram for the selection of statistical methods for the analysis of PRO data

5.4.1 Justification for omitting univariable statistical methods

The clinical relevance of the statistical method, i.e. the ability of the method to estimate a treatment effect with its associated CIs, is regarded as a key feature that a statistical method should have for analysing PROs in RCTs. The CONSORT guidelines (Item number 17a) for reporting RCTs says *'For each primary and secondary outcomes, results for each group, and the estimated effect and its precision (as a 95% CIs) should be reported'* (Moher *et al.*, 2010). Therefore, we excluded univariable methods such as Kruskal-Wallis test, Mann-Whitney U test, Wilcoxon signed rank test, and sign test that can only provide p-values but not an effect size.

We excluded the two independent sample *t*-test that produces the effect size and associated CIs because it is analogous to a simple linear regression model with a binary predictor variable (i.e. randomised group). Also, the *t*-test can only compare two treatment groups simultaneously and does not have the ability to adjust for potential confounding factors such as baseline PRO score which the MLR does.

5.4.2 Justification for omitting some multivariable statistical methods

ANOVA and ANCOVA are also analogous to a multiple linear regression model with a binary predictor/explanatory variable (i.e. randomised group). ANOVA and ANCOVA are not included, because though they are similar to linear models, in their purest form, they just provide ratios of sums of squares and F-statistics and p-values, and cannot provide the value of between-group difference (treatment effect) which is considered as one of the key results in data analysis of RCTs.

Statistical methods for count data and unordered categorical data are excluded because they do not reflect the nature of PRO data and they are rarely used in publicly funded RCTs. Therefore, multinomial logit model and multinomial probit model for unordered categorical data, and Poisson regression and negative binomial regression for count data are not included.

Statistical methods for binary data (i.e. proportion of responders) including binary logistic regression and binary probit regression are excluded because it requires a cut-point or threshold for the PRO score to categorise participants as responders or not. In reality, this threshold or cut-point in score is not readily available for most PROs and may vary from trial to trial even with the same PRO, and furthermore the dichotomised outcome gathers less information from the PROs compared to other score types (Altman and Royston, 2006; Shields *et al.*, 2015).

5.4.3 Justification for omitting statistical models for correlated responses

This research proposes to concentrate on the simpler situation where there is a single baseline and a single post-randomisation assessment of outcome, and compare the statistical methods that are suitable for such an analysis.

All RCTs are longitudinal and involve collection of outcome data post randomisation. Participants are randomised to different interventions and then their outcomes are assessed post-randomisation. The simplest RCT design has outcomes and data collected at baseline and one post-randomisation assessment or timepoint, e.g. 6-month post-randomisation, sometimes called a pre-test post-test design. However, in many randomised controlled trials the primary and secondary outcomes are often measured at multiple timepoints, for example, 3, 6, 9 and 12 months post randomisation. Our review of 114 RCTs published in the HTA Journal with a PRO as the primary outcome found that only 4.3% (5/114) of the RCTs had a single post-baseline assessment of the outcome with the majority having two or more assessments and 86.3% (101/114) had a baseline assessment of the PRO (Qian et al 2021). These repeated outcome measurements, on the same individual subject, are likely to be related or correlated. This means that the usual statistical methods for analysing such data that assume independent outcomes may not be appropriate.

There are a number of ways to analyses such repeated measures data. Three broad approaches are (Walters, 2009):

1. Time by time analysis;
2. Response feature analysis, with the use of summary measures (Matthews *et al.*, 1990);
3. Modelling of longitudinal data GLMMs or GLMs with parameters estimated by GEEs (Liang and Zeger, 1986).

In the literature, GLMMs or GLM with parameters estimated by GEEs are sometimes referred to as mixed/random-effects or marginal models respectively.

The analysis of correlated outcomes/responses with a longitudinal model raises a number of issues.

1. How should we treat time (of assessment) in the longitudinal model? Should we treat the time the PRO assessment was completed post-randomisation as a continuous predictor/explanatory variable in the model i.e. time from randomisation (in days or weeks), or a discrete variable (i.e., protocol stipulated follow-up timepoint) e.g. 6, 12, 26, 52 weeks, or as a factor e.g. 1, 2, 3, 4 corresponding to 1st post-randomisation follow-up, 2nd, 3rd, 4th etc.

2. What sort of trend in PRO scores over time should we assume? Should this be monotonic and linear or non-monotonic and non-linear?

3. What sort of correlation between the repeated outcomes/responses should we assume? Should this be a relatively simple correlation structure e.g. random intercept, or a more complex correlation structure e.g. random intercept and slope? If the latter model, should we assume the two random effects are uncorrelated or correlated? And if they are assumed to be correlated, then what covariance structure should we assume for the two random effects (unstructured, identity, or exchangeable)?

The CONSORT statement for RCTs also recommends specifying a primary timepoint for the analysis (Boutron *et al.*, 2008; Moher *et al.*, 2010). *'When outcomes are assessed at several timepoints after randomisation, authors should also indicate the pre-specified timepoint of primary interest'*. Again this potentially suggests a simpler analysis, that does not involve another level of complexity and multiple assumptions regarding the time effects, the pattern of PRO scores over multiple time assessments, the correlation structures between the repeated assessments. i.e. use a specific timepoint without accounting for the correlated responses. Our HTA review found that of the RCTs, only 28.1% (32/114) used the statistical modelling for correlated responses in primary analysis despite 95.6% (109/114) collected two or more post-randomisation assessments of the outcome. In addition, although linear mixed model is reported for the primary analysis in 39.4% (45/114) trials, only 23 trials conducted a repeated measures analysis, and the rest of trials only analysed a baseline and follow-up outcome assessment for primary analysis and applied mixed effects to adjust for other random factors such as clustering by site or centre.

We believe that the analysis of repeated and correlated PROs is very interesting and raises a number of issues that are briefly outlined above but it is beyond the scope of this PhD. As this project only considers the analysis of PROs under the simple scenario i.e. baseline scores with a post-randomisation timepoint, statistical methods that account for correlated responses (i.e. survival analysis, GLM with parameters estimated by GEE, GLMM, and repeated measures ANOVA) are excluded. For similar reasons, log-rank test (Kaplan Meier), the univariable method for analysing time-to-event data, is excluded.

5.4.4 List of statistical methods to carry forward

Statistical methods to carry forward that are highlighted in light green in Figure 5.1 are listed below.

1. Multiple linear regression (MLR)
2. Median regression (Median)
3. Tobit regression (Tobit))
4. Censored absolute least square deviation regression (CLAD)
5. Ordered logit model (OL)
6. Ordered probit model (OP)
7. Beta-binomial regression (BB)
8. Binomial-logit-Normal regression (BLN)
9. Fractional logistic regression (Frac)
10. Beta regression (BR)

MLR is the classical regression method to use. It is widely applied for the analysis of PROs in trials despite the potential violation of model assumptions such as the Normality of residuals and homoscedasticity. Tobit, a type of censored regression model, is designed to estimate linear relationships between the outcome variable and predictor variables when there is censoring in the outcome variable. In comparison, CLAD which holds similar assumptions on latent variable and estimates coefficients by minimising the sum of absolute value of deviations from the regression line is reported to be more robust to the violation of model assumptions than Tobit regression (Austin, 2002). As a derivation of Median, CLAD models the median but not the mean.

Ordinal regression assumes that a continuous latent variable of the ordinal PRO has a linear relationship with independent variables. The classic link functions for ordinal regression are the logit and probit links, corresponding to OL (i.e. proportional-odds model) and OP. The key assumption of ordinal regression is the proportional-odds assumption which assumes each covariate share a constant OR across the cut-points i.e. the influence of each covariate on the response variable is independent of the cut-points (Arostegui, Núñez-Antón and Quintana, 2012).

Binomial regression (including BB and BLN) assumes the discrete PRO is an aggregation of multiple Bernoulli processes on a number of dichotomous items (i.e. possible score values of a PRO) (Liang *et al.*, 2014). The probability of success for each PRO score value is assumed to have a linear relationship with dependent variables through a logit link, and to follow a beta distribution in BB and a logit-Nomal distribution in BLN (Arostegui, Núñez-Antón and Quintana, 2007, 2012; Najera-Zuloaga, Lee and Arostegui, 2018, 2019).

Fractional regression (including BR and Frac) can also account for boundedness by fitting responses scattering between 0 and 1. BR is proposed because of its flexibility to skewed and bounded data, in which case the PROs is assumed to have a beta distribution with predefined boundaries and the logit link function is often used (Ferrari and Cribari-Neto, 2004). Frac shares similar traits as BR, but it does not require data distribution of the responses (Papke, 1996) except for scattering between 0 and 1.

The evaluation of the statistical methods for PRO analysis to carry forward is presented in Table 5.3, using the established set of desired statistical properties.

MLR and Tobit are similar in most domains except that Tobit is able to account for censored data while MLR cannot. This is also the case for Median and CLAD, where both methods produce estimates in medians but CLAD can account for the censored data while Median cannot.

Two ordinal regressions, i.e. OL and OP, have identical performances in set of criteria domains, as they are very similar except that they have different link functions, i.e. the estimates from the OP cannot be explained as ORs as the OL does. They are regarded as partially clinically relevant, because these models can generate estimated effect sizes and associated CIs, but they are not in the original scale (i.e. in mean or median scores of the treatment effect of PROs) as the MLR or Tobit does. Similarly, fractional regression (i.e. Frac and BR) and binomial regression (BB and BLN) all adapted a logit link function, which means their estimates can be explained in ORs but not in means or medians of the treatment difference between treatment arms, and therefore they are assessed as partially clinically relevant.

The outcome responses (i.e. the PRO scores) are required to be ordinal for the application of ordinal regression (OL and OP) and binomial regression (BB and BLN), so the recoding of the PRO scores to a finite sequence of integers is needed to be performed. Similarly fractional regression (Frac and BR) requires the recoding of PRO scores to an interval between 0 and 1, but BR needs the transformation to an open interval i.e. the PRO scores are not allowed to scatter at 0 and 1, whereas Frac allows closed interval where the values of 0 and 1 are permitted. Details of the recoding techniques for these methods will be explained in Chapter 6.

Similar patterns can be seen for two models of binomial regression, BB and BLN, the difference between which lies in the assumptions of probability of success. Both ordinal regression and binomial regression are considered to have the ability to adjust for bounded or censored scale, as these models are developed for ordinal data with finite number of values, the minimum and maximum of which can be regarded as boundaries. Adapting the logistic regression model, ordinal regression and binomial regression necessarily requires recoding the observed response to a binomial form, from 0 to $k-1$, where k represents the number of the ordinal categorical values (Arostegui, Núñez-Antón and Quintana, 2013).

Table 5.2 Evaluation of filtered statistical methods for PRO analysis using the established set of desired statistical properties

Statistical methods	Compare ≥2 treatment arms	Adjust for confounding factors	Be clinically relevant	Handle a bounded/censored scale	Require recoding of PRO
MLR	Y	Y	Y	N	N
Tobit	Y	Y	Y	Y, assume the observed outcome variable is censored	N
CLAD	Y	Y	Y	Y, assume the observed outcome variable is censored	N
Median	Y	Y	Y	N	N
OL	Y	Y	Partially	Y	Y (to $[0, k\text{-}1] \subseteq \mathbb{N}$)
OP	Y	Y	Partially	Y	Y (to $[0, k\text{-}1] \subseteq \mathbb{N}$)
BB	Y	Y	Partially	Y	Y (to $[0, k\text{-}1] \subseteq \mathbb{N}$)
BLN	Y	Y	Partially	Y	Y (to $[0, k\text{-}1] \subseteq \mathbb{N}$)
Frac	Y	Y	Partially	Y, assume the observed outcome variable is bounded in a closed interval	Y (to [0, 1] scale)
BR	Y	Y	Partially	Y, assume the observed outcome variable is bounded in an open interval	Y (to (0, 1) scale)

Partially, provide effect size and CI, but not in original scale. BB, Beta-binomial regression; BLN, binomial-logit-Normal regression; CI, confidence interval; CLAD, censored least absolute deviations; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; N, no; Tobit, Tobit regression; OL, ordered logit model; OP, ordered probit model; Y, yes. k, represents the number of possible categorical values in a domain; $\mathbb{N}$, denotes the non-zero positive natural numbers i.e. 1,2,3… k-1.

5.5 Discussion

In this chapter, the number of statistical methods is reduced to a more manageable number for evaluation using explicit justification and the process is presented in a flow diagram.

An MCDA framework was adapted for the filtration process. Though MCDA has been applied in healthcare settings to support decision making in trading-off various conflicting criteria for the evaluation of health technologies (Marsh *et al.*, 2016; Thokala *et al.*, 2016), this is the first MCDA applied to filter and select statistical methods considering various statistical properties to the best of our knowledge. We need to be aware that outcomes of the filtration are subjective to the evaluation conducted by stakeholders, and different sets of criteria, scoring and weighting systems may generate different scores and ranks of various statistical methods. Another point worth noting is that a method with the ability to do more does not necessarily mean the method is the best. The specific scenarios need to be considered (Thokala *et al.*, 2016).

The decision process of selecting a statistical method to analyse any type of data is based on the objective of the analysis, the nature of the data, the proposal of the analysis and the adherence of the data to the method's assumptions. The filtration in this chapter is purely based on the statistical theory of the identified statistical methods regarding our selected set of statistical properties for evaluation. It can provide an overview of which method may perform better and cut down the number of methods for empirical analysis and simulation, but it cannot guarantee the final choice. For example, we did not select the robustness of violation of model assumptions as one criterion because it cannot be tested before the empirical analysis, however, if we took it into account, univariable methods that do not rely on data assumptions could have better performances.

To make the final decision, other criteria such as the consistency of estimates among different statistical methods and the accuracy and robustness of these estimates also need to be considered to provide a thorough evaluation. In addition, computational difficulties, which may exist in performing complex and novel statistical methods in data analysing software, and interpretability of the estimates, which is important for understanding or explanation to non-statisticians, can also be included.

The next step is to explicitly describe the statistical methods filtered from this chapter, and to conduct empirical analysis and simulation analysis to evaluate how similar estimates from different statistical methods are, how accurate their estimates are from the predefined 'truth', and how robust these estimators are under different scenarios.

Chapter 6 Description of the filtered statistical methods with an example of application to PROs in RCTs

6.1 Introduction

Our previous work in Chapter 5 has filtered the potential statistical methods for the analysis of PROs in RCT settings and decided to carry forward the following list of 10 statistical methods: MLR, Median, Tobit, CLAD, OL, OP, BB, BLN, Frac, and BR.

In this chapter, technical details of the 10 filtered statistical methods are provided to explicitly explain their model functions, estimation methods, model assumptions, and interpretations of their estimated treatment coefficients. Two estimand frameworks are introduced to allow the comparison of estimates on their original scales and on a unified standardised scale. The 10 statistical methods are applied to an RCT dataset with a PRO as clinical outcome measurement to demonstrate how these methods can be fitted, and how their outputs can be interpreted in the computational software Stata/MP 17.0.

Table 6.1 SF-36v2 dimension scores at baseline and 6-month post-randomisation in the LM trial

SF-36 dimension	Control			Intervention			Total		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
At baseline									
PF	143	71.66	26.37	145	67.52	25.28	288	69.57	25.86
RP	143	76.79	25.51	145	72.36	27.62	288	74.56	26.64
BP	143	64.67	26.47	145	61.21	25.56	288	62.93	26.03
GH	143	68.76	20.38	145	63.65	20.36	288	66.18	20.49
VT	143	60.31	20.85	145	58.45	21.42	288	59.38	21.12
SF	142	81.95	26.40	144	82.90	21.96	286	82.43	24.23
RE	143	84.50	21.52	145	82.70	23.43	288	83.59	22.48
MH	143	77.00	18.24	145	75.47	18.34	288	76.23	18.28
At 6-month post-randomisation									
PF	126	70.71	27.28	136	66.03	28.39	262	68.28	27.91
RP	126	73.86	26.39	136	69.85	29.93	262	71.78	28.30
BP	126	61.60	27.44	136	60.46	27.95	262	61.01	27.66
GH	126	64.81	21.06	136	61.89	22.67	262	63.30	21.92
VT	126	58.02	21.74	136	56.42	22.17	262	57.19	21.94
SF	126	81.35	26.02	136	77.85	28.18	262	79.53	27.16
RE	125	86.67	19.37	136	82.72	23.21	261	84.61	21.51
MH	126	75.93	18.74	136	77.31	18.22	262	76.65	18.45

BP, bodily pain; GH, general health; MH, mental health; PF, physical functioning; RE, role limitation - emotional; RP, role limitation - physical; SD, standard deviation; SF, social functioning; VT, vitality.

6.2 Data source

We used secondary data from an RCT named Lifestyle Matters (Mountain *et al.*, 2017), referred as LM in the rest of this thesis, as the example dataset. This trial compared an occupation-based lifestyle intervention to usual care for sustaining and improving the mental well-being of adults aged 65 years or over, using the SF-36 version 2 (SF-36v2) mental health (MH) score at 6-month follow-up as the primary outcome. The randomisation ratio of the LM study was set at 1:1. Baseline measures of this study were collected before randomisation, and SF-36v2 were sent out and collected at 6-month and 24-month post-randomisation.

The primary outcome of the LM trial, the SF-36v2 MH score at 6-month follow-up is chosen as the response variable, and the SF-36v2 MH score at baseline and treatment group is selected as confounding factors. Table 6.1 summarises the SF-36v2 dimension scores at baseline and at 6-month post-randomisation in the LM trial. The distribution of original SF-36v2 MH scores in each treatment arm at baseline and at 6-month follow-up are shown in Figure 6.1.

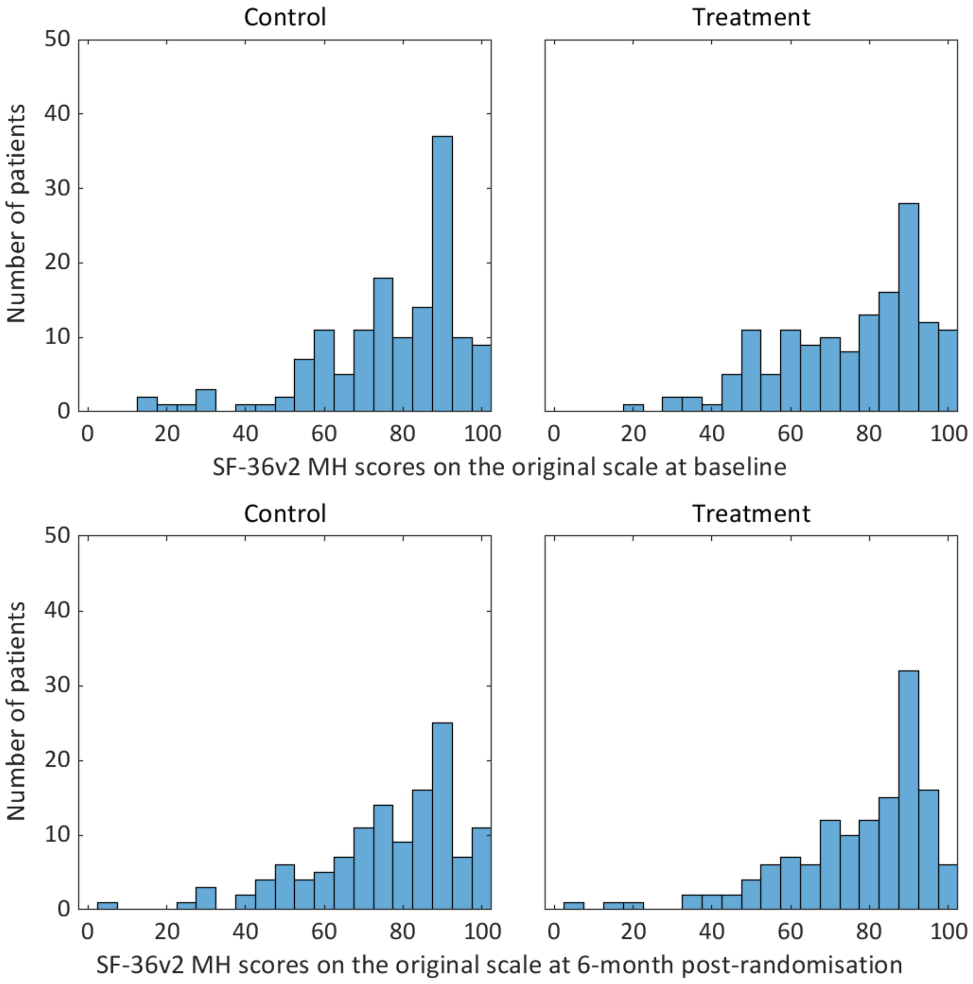


Figure 6.1 Distribution of SF-36v2 MH scores, at baseline and 6-month post-randomisation on the original 0 to 100 scale

6.3 Estimand framework

An estimand is a well-defined and explicit description of precisely what treatment effect is to be estimated in an RCT (Little and Lewis, 2021). An estimand is the target of the estimation for a particular trial objective (ICH, 2021). The estimand framework typically consists of five attributes: the treatment condition of interest; the population of patients targeted by the clinical question; the variable or endpoint to be obtained for each patient; intercurrent event handling; and a population-level summary measure of how outcomes between the different randomised groups will be compared (Lawrance *et al.,* 2020).

For the empirical analysis of the treatment difference in a PRO score at a specific post-randomisation between randomised groups, four attributes of the estimand framework, i.e. population, treatments, outcomes, and intercurrent event handling, are unchanged, but the fifth attribute, the population summary measure, may not be consistent among the different statistical methods. Common population summary measures include the difference in means, risk ratios and ORs.

If the treatment coefficients from the 10 statistical methods are chosen as the population summary measure to compare outcomes between the different groups, then it may not make sense to compare the 10 statistical methods (i.e. estimators) and their associated estimates because their estimated treatment coefficients stand for different things. For instance, MLR produces estimates of treatment effect as a difference in location or in group means while OL produces estimates of the treatment effect as a logOR. However, some of the methods, such as MLR, Tobit, CLAD, and Median, produce estimates that have similar population summary measures that look at differences in location or central tendency e.g. differences in means or medians. Therefore, it may be sensible to compare the treatment coefficient estimates of difference in means or medians between two treatment groups produced by these statistical methods. Again, some of the methods, such as OL, BB, BLN, Frac, and BR, also have similar population summary measures that estimate the logORs for the treatment difference. Therefore, it may be sensible to compare the estimates of log ORs between two treatment groups produced by these statistical methods.

If we need the comparison of estimates from these methods that are based on different scales, a universal population summary measure, which can be compared across these 10 statistical methods, is required. The standardised effect size (SES), which is calculated through a standardisation procedure by dividing the group difference by the pooled standard deviation (SD), can produce estimates with no units of measurement and therefore handle the issue of comparing estimates that are based on different scales (Cohen, 2013; Cook *et al.*, 2014; Rothwell, Julious and Cooper, 2018). The SES is believed as a suitable population summary measure to compare outcomes between the different treatment groups. The other four attributes for the estimand are the same despite whichever of the 10 statistical methods are used, such that the estimand, the SES, is the same, but the estimators (i.e. the 10 statistical methods) will be different and may produce different estimates that will be compared and presented.

Therefore, two estimand frameworks are adopted to compare estimates from different statistical methods that produce treatment estimates on different scales. The first framework is called the scale-based estimand framework that categorises statistical methods according to the scales of their original population summary measures (i.e. means/medians and logORs). The second framework is called the SES estimand framework, which adapts the SES as the population summary measure and allows the comparison of the SES estimates from different statistical methods.

6.3.1 Scale-based estimand framework

In the scale-based estimand framework, the population summary measure is the mean or median in the treatment difference between two treatment arms for statistical methods with estimates on the untransformed scale (MLR, Tobit, CLAD, and Median), and the population summary measure is the logOR of the treatment difference between two treatment arms for statistical methods with estimates are based on the transformed scale (OL, BB. BLN, BR, and Frac). For example, the scale-based estimand framework of the LM trial can be interpreted as:

For statistical methods with the estimates on the untransformed (ordinal measurement) scale:

> *In independently living adults aged 65 years or over with normal cognition, **what is the difference in mean/median score of their mental wellbeing** (as measured with the SF-36v2 mental health dimension scores) between an occupation-based lifestyle intervention in addition to usual care followed by any subsequent therapy/treatment (as needed) compared with usual care treatment group only followed by any subsequent therapy/treatment (as needed) after 6-month post-randomisation or death (whichever occurs first), regardless of study treatment discontinuation?*

For statistical methods with the estimates on the transformed scale:

> *In independently living adults aged 65 years or over with normal cognition, **what is odds ratio for the odds of having of higher/better mental wellbeing scores** (as measured with the SF-36v2 mental health dimension scores) between an occupation-based lifestyle intervention in addition to usual care followed by any subsequent therapy/treatment (as needed) compared with usual care treatment group only followed by any subsequent therapy/treatment (as needed), after 6-month from randomisation or death (whichever occurs first), regardless of study treatment discontinuation?*

As the logit link is used for OL, BB, BLN, BR, and Frac, their estimated treatment coefficient, denoted by $coef(TE)$, is the logOR which can be interpreted as OR through the exponential transformation.

$$OR_{TE} = \exp\,(coef(TE)) \tag{6.1}$$

A special case of these methods is the OP which uses a probit link, denoted by Φ^{-1}. The estimates of OP cannot be transformed to the original scale or to an OR as other methods do, but it can be interpreted as the probability or the effect size in the response.

6.3.2 SES estimand framework

Since the population summary measures of different statistical methods are not comparable, to allow the comparison of estimates from different statistical methods, these statistical methods should estimate the same estimand (i.e. target of estimation). The SES is such an estimand that to unify and to allow the comparison of estimates from different statistical methods. For example, the SES estimand framework of the LM trial for 10 statistical methods can be interpreted as:

> *In independently living adults aged 65 years or over with normal cognition,* ***what is the difference in standardised effect size of their mental wellbeing*** *(as measured with the SF-36v2 mental health dimension scores) between an occupation-based lifestyle intervention in addition to usual care followed by any subsequent therapy/treatment (as needed) compared with usual care treatment group only followed by any subsequent therapy/treatment (as needed) after 6-month post-randomisation or death (whichever occurs first), regardless of study treatment discontinuation?*

The fundamental formula to calculate SES is the estimated values divided by the common within-population SD (Cohen, 2013). In a two-arm RCT, the scale-invariant SES was calculated using the estimated treatment coefficient, denoted by $coef(TE)$, divided by its associated standard error, denoted by $SE(TE)$, after adjusting for the sample size, which is equivalent to the Z-statistics with the adjustment of the sample size in each treatment arm (Rothwell, Julious and Cooper, 2018).

$$SES = \frac{coef(TE)}{SD(TE)} = \frac{coef(TE)}{SE(TE)} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = Z \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{6.2}$$

where Z stands for the Z-statistics; *TE* stands for the treatment effect; and n_1 and n_2 represents the sample size in each treatment arm respectively.

The standard error of the SES, denoted by $SE(SES)$, is based on a Normal approximation of non-central *t*-distribution (Hedges, 1981):

$$SE(SES) = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{SES^2}{2(n_1+n_2)}} \tag{6.3}$$

Therefore, the associated 95% CIs of SES are given using the following formula:

$$SES \pm 1.96\, SE(SES) \tag{6.4}$$

6.4 Description of the filtered statistical methods

This section gives detailed descriptions and explanations of the 10 filtered statistical methods. These methods are described under the GLM framework. Key components of a GLM are (Nelder and Wedderburn, 1972):

1. An outcome or dependent variable, denoted by Y, whose distribution with parameter, denoted by π, is assumed to follow a particular distribution from the exponential family,
2. A set of independent variables, denoted by $\boldsymbol{X}$, that provide a linear predictor, denoted by $\boldsymbol{X\beta}$, for Y, and
3. A link function $g(\cdot)$ that connects the parameter (π) and the linear predictor $(\boldsymbol{X\beta})$, i.e. $g(\pi) = \boldsymbol{X\beta}$.

In the case of analysing SF-36 scores, each dimension at a specific post-randomisation timepoint (Y) is analysed separately, with the independent variables $(\boldsymbol{X})$ being the treatment group and the corresponding baseline score to detect the treatment effect on SF-36 dimension scores. This is because the treatment effect is regarded as the main outcome of clinical trials, and a PRO score is likely to correlate with its baseline score (Vickers and Altman, 2001; Coens *et al.*, 2020). The parameter (π) is defined as the mean value of PRO score (μ) for methods with the identity link for MLR and Tobit , as the median of PRO score, denoted by $Q_{Y|X}(median)$, for Median and CLAD, and as the mean value of the probability of success, denoted by θ for methods with logit or probit link. The probability of success represents the probability that the PRO score, Y, is less than or equal to particular score or category l, i.e. $P(Y \leq l)$ for the ordinal regression methods, or represents the probability of the PRO score being the discrete value or category l, i.e. $P(Y = l)$ for the binomial and fractional regression methods.

The characteristics of these statistical methods are summarised in Table 6.2, including their model assumptions, link function, recoding requirement, and estimation methods. A detailed description of the 10 statistical methods, together with their application to an example RCT dataset using SF-36v2 MH score at a single post-randomisation follow-up, is summarised and presented in the rest of this section. Appropriate recoding techniques were applied for eight dimension scores to accommodate the application of various statistical methods. The Stata codes for the recoding and regression analysis of the SF-36v2 MH scores from the LM trial are presented alongside the explanation of each statistical method in this chapter.

Table 6.2 Summary of 10 statistical methods for the analysis of SF-36 dimension scores under the GLM framework

Statistical methods	**Distribution of the outcome/dependent variable (Y)**	**Link function $g(\cdot)$**	**Model assumption**	**Recoding of PRO needed***	**Interpretation**	**Estimation method**	**Stata command**
Classical model							
MLR	Continuous	Identity $g(\mu) = \mu = \boldsymbol{X\beta}$	Normality (of residuals); homoscedasticity; linearity; independence of outcomes.	No	Mean	OLS or MLE	regress
Median	Continuous	$g\left(Q_{Y\|X}(median)\right) = Q_{Y\|X}(median) = \boldsymbol{X\beta}_{median}$	Linearity; independence of outcomes.	No	Median	LAD	qreg
Censored regression							
Tobit	Observed Y: Continuous and censored; latent Y^*: Continuous	Identity $g(\mu^*) = \mu^* = \boldsymbol{X\beta}$	Normality (of residuals); homoscedasticity; linearity; independence of outcomes.	No	Latent Mean	MLE	tobit
CLAD	Observed Y: Continuous and censored; latent Y^*: Continuous	$g\left(Q_{Y^*\|X}(median)\right) = Q_{Y^*\|X}(median) = \boldsymbol{X\beta}_{median}$	Linearity; independence of outcomes.	No	Latent Median	CLAD	clad
Ordinal regression							
OL	Ordinal	Logit $g(\theta_{il}) = \ln(\frac{\theta_{il}}{1-\theta_{il}})$	Proportional-odds; linearity; independence of outcomes.	Yes (to [0, k-1] ⊆ ℕ)	Odds ratio	MLE	ologit
OP	Ordinal	Probit $g(\theta_{il}) = \Phi^{-1}(\theta_{il})$	Proportional odds; linearity; independence of outcomes.	Yes (to [0, k-1] ⊆ ℕ)	Probability	MLE	oprobit

Continued on the next page

Statistical methods	Distribution of the dependent variable (Y)	Link function $g(\cdot)$	Model assumption	Recoding of PRO needed*	Interpretation	Estimation method	Stata code
Binomial regression							
BB	Beta-binomial i.e. $Y_i \sim Bin(k, \theta_i)$ $\theta_i \sim Beta(\alpha, \gamma)$	Logit $g(\theta_i) = \ln\left(\frac{\theta_i}{1-\theta_i}\right)$	Linearity; independence of outcomes; beta distribution of probability of success	Yes (to [0, k-1] ⊆ ℕ)	Odds ratio	MLE	betabin
BLN	Binomial-logit-Normal i.e. $Y_i \sim Bin(k, \theta_i)$ $\theta_i \sim LN(0,1)$	Logit $g(\theta_i) = \ln\left(\frac{\theta_i}{1-\theta_i}\right)$	Linearity; independence of outcomes; logit-Normal distribution of probability of success	Yes (to [0, k-1] ⊆ ℕ)	Odds ratio	MLE	glm...link(logit) family(binomial N)
Fractional regression							
Frac (logit link)	Recoded Y': continuous and bounded in [0, 1]	Logit $g(\mu_{Y'}) = \ln\left(\frac{\mu_{Y'}}{1-\mu_{Y'}}\right)$	Linearity; independence of outcomes	Yes (to [0, 1] scale)	Odds ratio	Quasi-likelihood estimation	fracreg logit
BR (logit link)	Recoded Y'': continuous and bounded in (0, 1) $Y'' \sim Beta(\mu\varphi, (1-\mu)\varphi)$	Logit $g(\mu_{Y''}) = \ln\left(\frac{\mu_{Y''}}{1-\mu_{Y''}}\right)$	Linearity; independence of outcomes	Yes (to (0, 1) scale)	Odds ratio	MLE	betareg

BB, Beta-binomial regression; BLN, binomial-logit-Normal regression; CI, confidence interval; CLAD, censored least absolute deviations; Frac, fractional logistic regression; GLM, generalized linear model; LAD, least absolute deviations; Median, median regression; MCAR, missing completely at random; MCDA, multi-criteria decision analysis; MLE, maximum likelihood estimation; MLR, multiple linear regression; Tobit, Tobit regression; OL, ordered logit model; OLS, ordinary least squared; OP, ordered probit model; PRO, patient-reported outcomes; k, represents the number of possible categorical values in a domain; μ denotes the mean; μ^* denotes the latent mean; ℕ, denotes the non-zero positive natural numbers i.e. 1,2,3… k-1; φ is the precision parameter for beta distribution; Φ stands for the standard Normal cumulative distribution function. θ, denotes the probability of success or the cumulative response probabilities i.e. $\theta_{il} = P(Y \leq l)$ the probability of a response in category l or below for OL, and $\theta_i = P(Y = l)$ the probability of a response in category l for BB, BLN, and BR. Note that `clad` and `betabin` are user-developed packages in Stata, and therefore installation of the corresponding package is required to run these two commands.

We first start with classical models including MLR, as it is one of the most popular methods that have been applied for PRO analysis, and Median, which does not require Normally distributed residuals. Then, we consider censored regression methods that deal with the boundedness or censoring, including Tobit and CLAD which have similar assumptions to MLR and Median respectively but additionally accounts for the censored feature of PRO scores. Later, we look at methods that assume PRO scores as ordinal outcomes and require recoding techniques to apply, including ordinal regression (OL and OP) and binomial regression (BB and BLN). Finally, fractional regression (Frac and BR) for ratio response outcomes between 0 and 1 are applied after recoding SF-36 dimension scores from their original 0 to 100 interval to a 0 to 1 interval.

6.4.1 Classical model

6.4.1.1 Multiple linear regression (MLR)

In MLR, the relationship between the mean of each SF-36 dimension score and the linear predictor is described using the following equation:

Model

$$Y = \boldsymbol{X\beta} + \varepsilon \tag{6.5}$$

$$\varepsilon \sim N(0, \sigma^2) \tag{6.6}$$

where the Y is a vector of observed PRO score (`mh6`), $\boldsymbol{X}$ is the design matrix that denotes multiple row vectors of independent variables such as baseline score (`mh0`) and treatment group (`group`), and ε is the error term which captures the difference between the linear predictor and the independent variable and is assumed to follow a Normal distribution with the mean of 0, and standard error of σ.

Model assumptions

1. Linearity: There must be a linear relationship between the outcome variable Y and the independent variables $\boldsymbol{X}$.
2. Multivariate Normality: The residuals ε are Normally distributed.
3. No multicollinearity: The independent variables $\boldsymbol{X}$ are not highly correlated with each other.
4. Homoscedasticity: The residuals ε have the same variability or constant variance for all the fitted values of Y.
5. Independence: The observations in the sample are independent.

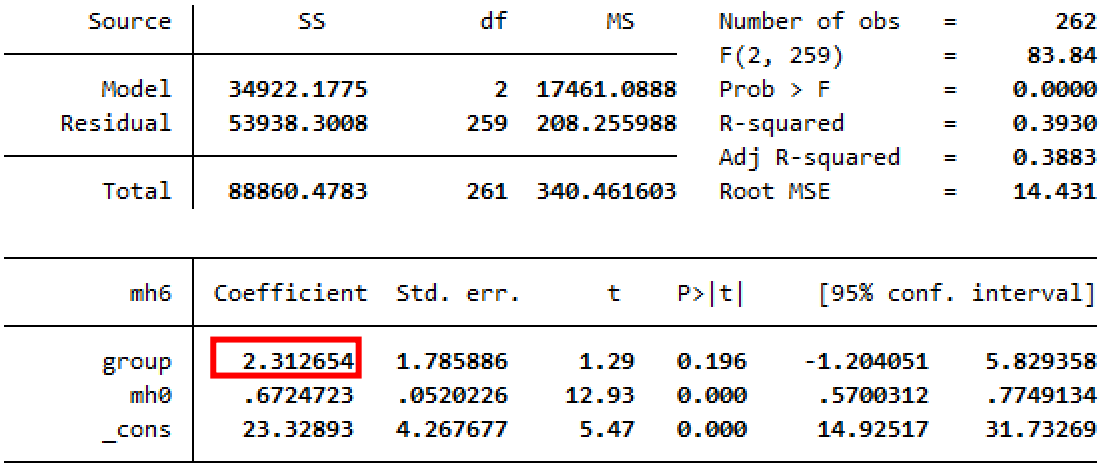
Estimation method: OLS or MLE

OLS is an estimation method of unknown parameters in a linear model by minimising the sum of squares of the differences between the observed dependent variable and the predicted values by the linear model. MLE is an estimation method of unknown parameters of assumed probability distributions in the assumed statistical model by maximising the likelihood function. OLS is identical to the MLE when the Normality assumption of residuals is met.

Stata code

```
regress mh6 group mh0
```

Stata output



```
      Source |       SS           df       MS      Number of obs   =       262
-------------+----------------------------------   F(2, 259)       =     83.84
       Model |  34922.1775         2  17461.0888   Prob > F        =    0.0000
    Residual |  53938.3008       259  208.255988   R-squared       =    0.3930
-------------+----------------------------------   Adj R-squared   =    0.3883
       Total |  88860.4783       261  340.461603   Root MSE        =    14.431

------------------------------------------------------------------------------
         mh6 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |   2.312654   1.785886     1.29   0.196    -1.204051    5.829358
         mh0 |   .6724723   .0520226    12.93   0.000     .5700312    .7749134
       _cons |   23.32893   4.267677     5.47   0.000     14.92517    31.73269
------------------------------------------------------------------------------
```

Interpretation

The mean of SF-36v2 MH score at 6 months for the treatment group is 2.31 points higher than the mean for the control group after allowing or adjusting for the baseline MH score. Alternatively, given two individuals with the same baseline MH score, on average the individual subject in the treatment group will have a 2.31 point higher MH score at 6 months than a subject in the control group.

6.4.1.2 Median regression (Median)

Median is a special case of quantile regression when the quantile level is set at 50th. Quantile regression estimates the conditional median of the response variable and does not assume a particular parametric distribution for the response. It has a similar equation as MLR, but depicts the relationship between the median of dimension scores, denoted by $Q_{Y|X}(median)$ and the linear predictor. It can be used as an alternative to MLR when the conditions of linear regression are not met. Quantile regression is robust to response outliers, i.e. very small or large PRO scores, and it needs sufficient data to run (Rodriguez, Yao and Inc, 2017).

Model

$$Q_{Y|X}(median) = \boldsymbol{X\beta}_{median} \tag{6.7}$$

where 50th conditional quantile of Y is given as a linear function given $\boldsymbol{X}$.

Model assumptions

1. Linearity: There must be a linear relationship between the outcome variable Y and the independent variables $\boldsymbol{X}$.
2. Independence: The observations in the sample are independent.

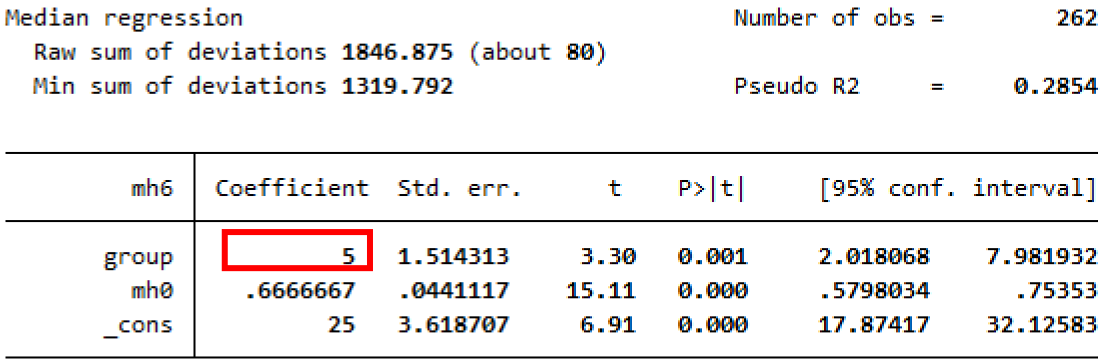
Estimation method: LAD

Median model finds a line through the data that minimises the sum of the absolute residuals, which are the deviations of the data points from the line.

Stata code

```
qreg mh6 group mh0
```

Stata output



```
Median regression                                   Number of obs =        262
  Raw sum of deviations 1846.875 (about 80)
  Min sum of deviations 1319.792                    Pseudo R2     =     0.2854

------------------------------------------------------------------------------
         mh6 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |          5   1.514313     3.30   0.001     2.018068    7.981932
         mh0 |   .6666667   .0441117    15.11   0.000     .5798034      .75353
       _cons |         25   3.618707     6.91   0.000     17.87417    32.12583
------------------------------------------------------------------------------
```

Interpretation

The median of SF-36v2 MH score at 6 months for the treatment group is 5.00 points higher than the median of the control group after allowing or adjusting for the baseline MH score.

6.4.2 Censored regression

Censored regression models assume the boundaries of the dependent variable are due to censoring, i.e. the mean of latent PRO scores, denoted by Y^* for Tobit, or the median of latent PRO scores, denoted by $Q_{Y^*|X}(median)$ for CLAD regression, can exceed the upper and lower boundaries, but they are not observable. The CLAD model is a subset of Median that estimates the median value of the parameters, whereas Tobit regression is an extension of MLR that estimates the mean. Both Tobit and CLAD regression describe the relationship between the latent variable and the linear predictor. The observed dependent variable of the censored regression is defined using the following equations:

$$Y = \begin{cases} a, \ Y^* \leq a \\ Y^*, \ a < Y^* < b \\ b, \ Y^* \geq b \end{cases} \tag{6.8}$$

where a and b denotes the lower and upper bounds of the PRO score respectively.

6.4.2.1 Tobit regression (Tobit)

Type I Tobit model is used for most PRO scenarios where both sides of the PRO scale are bounded or censored to certain scores e.g. SF-36v2 MH score is bounded between 0 and 100.

Model

$$Y^* = \boldsymbol{X\beta} + \varepsilon \tag{6.9}$$

$$\varepsilon \sim N(0, \sigma^2) \tag{6.10}$$

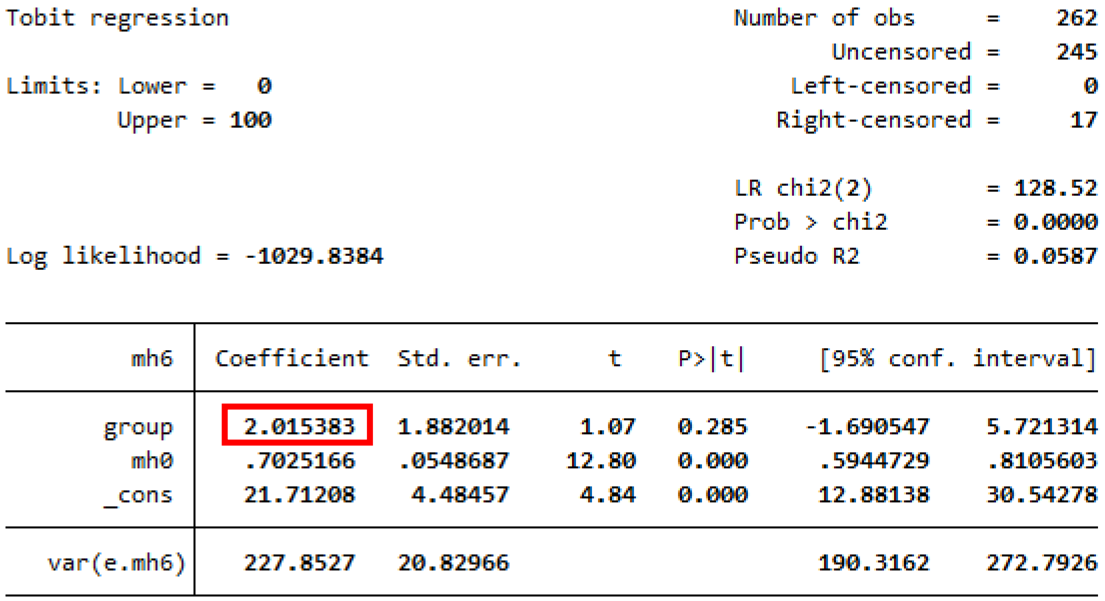
where Y^* denotes the latent variable of the PRO score, which satisfies the classical linear model assumptions; whereas y, the censored outcome of the latent variable Y^* does not have a linear relationship with $\boldsymbol{X}$ and ε.

Model assumptions and estimation methods of Tobit are identical to MLR, with one difference that Y denotes the latent variable but not the observed variable.

Stata code

```
tobit mh6 group mh0, ll(0) ul(100)
```

Stata output



```
Tobit regression                                Number of obs     =        262
                                                   Uncensored     =        245
Limits: Lower =    0                               Left-censored  =          0
        Upper =  100                              Right-censored  =         17

                                                LR chi2(2)        =     128.52
                                                Prob > chi2       =     0.0000
Log likelihood = -1029.8384                     Pseudo R2         =     0.0587

------------------------------------------------------------------------------
         mh6 | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |   2.015383   1.882014     1.07   0.285    -1.690547    5.721314
         mh0 |   .7025166   .0548687    12.80   0.000     .5944729    .8105603
       _cons |   21.71208    4.48457     4.84   0.000     12.88138    30.54278
-------------+----------------------------------------------------------------
  var(e.mh6) |   227.8527   20.82966                      190.3162    272.7926
------------------------------------------------------------------------------
```

Interpretation

The mean of ***uncensored*** *SF-36v2 MH score at 6 months for the treatment group is 2.02 points higher than the mean of the control group after allowing or adjusting for the baseline MH score.*

Tobit regression coefficients are interpreted in the similar manner to MLR coefficients; however, the linear effect is on the uncensored latent variable Y^*, but not on the observed variable y, i.e. $E(Y^*|X)$ is linear to $\boldsymbol{X}$, but $E(Y|X)$ is nonlinear to $\boldsymbol{X}$.

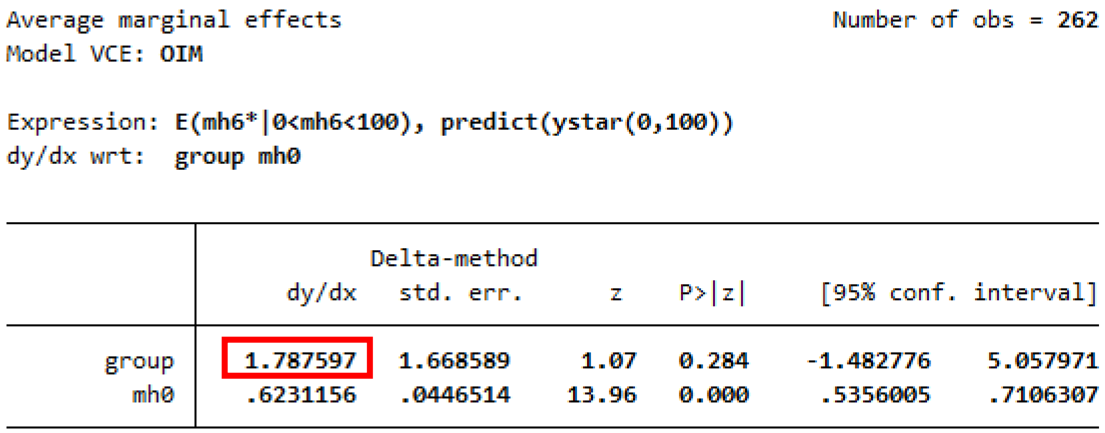
Marginal effect

The estimated treatment effect for the latent variable Y^* that is assumed not to have boundaries. The command `margins` is used to estimate the means of the marginal effects on the expected value of the censored outcome Y. The following code is used to estimate the changes in the conditional expected value of the dependent variable i.e. the change in the observed PRO score that is bounded between 0 and 100.

Stata code

```
margins, dydx(*) predict(ystar(0,100))
```

Stata output



```
Average marginal effects                          Number of obs = 262
Model VCE: OIM

Expression: E(mh6*|0<mh6<100), predict(ystar(0,100))
dy/dx wrt:  group mh0

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |   1.787597   1.668589     1.07   0.284    -1.482776    5.057971
         mh0 |   .6231156   .0446514    13.96   0.000     .5356005    .7106307
------------------------------------------------------------------------------
```

Interpretation

The average marginal effect of SF-36v2 MH score at 6 months for the treatment group is 1.79 points higher than the marginal effect of the control group after allowing or adjusting for the baseline MH score.

6.4.2.2 Censored least absolute deviations regression (CLAD)

CLAD holds similar assumptions on the latent variable as Tobit. i.e. the latent variable Y^* is assumed to have a linear relationship with $\boldsymbol{X}$, but the observed variable Y is non-linear to $\boldsymbol{X}$. As a derivation of Median, CLAD models the median but not the mean (Pullenayegum *et al.*, 2010).

Model

$$Q_{Y^*|X}(median) = \boldsymbol{X\beta}_{median} \tag{6.6}$$

where $Q_{Y^*|X}(median)$ denotes the latent variable of PRO score Y^*, which is assumed to be linear to $\boldsymbol{X}$.

Model assumption

1. Linearity: There must be a linear relationship between the latent outcome variable Y^* and the independent variables $\boldsymbol{X}$.
2. Independence: The observations in the sample are independent.

Estimation method: CLAD

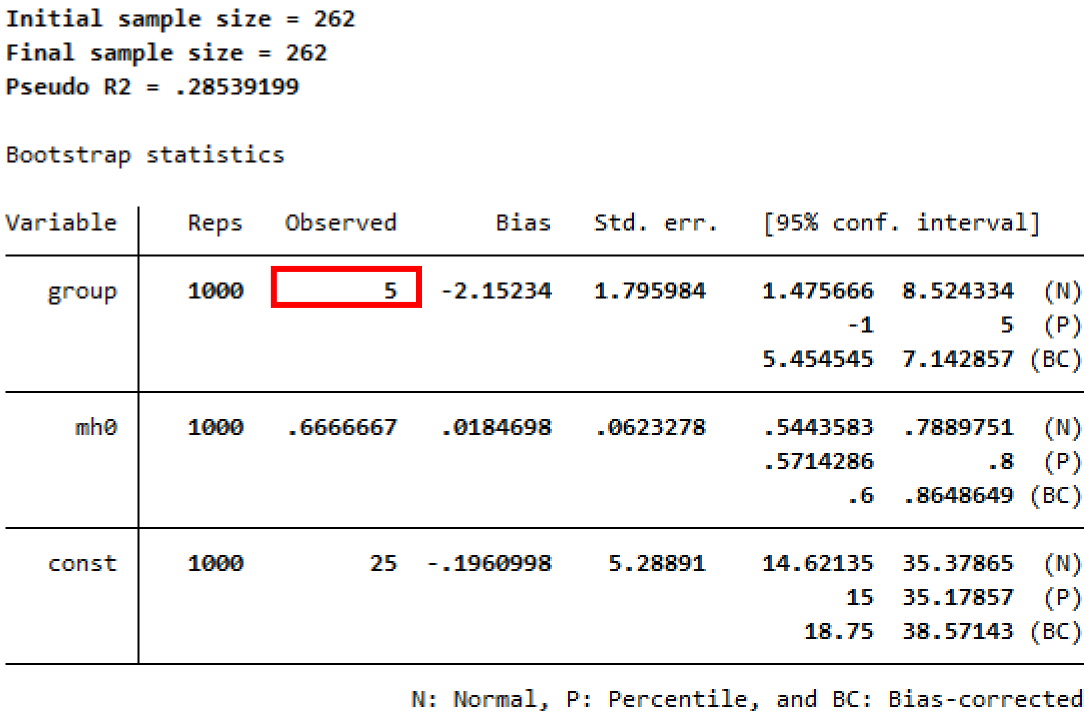
Stata code

```
clad mh6 group mh0, rep(1000) ul(100)
```

Note:

1. A package for CLAD (package `sg153.pkg`) needs to be installed before running the code.
2. The number of iterations are required to be pre-specified to run CLAD in Stata, and we select 1,000 as the repetition number for this analysis.
3. Command clad only functions with lower or upper censoring; you cannot specify censoring at both the lower and upper bound. If nothing is specified for a lower or upper bound, `clad` assumes that the lower limit is zero.

Stata output



```
Initial sample size = 262
Final sample size = 262
Pseudo R2 = .28539199

Bootstrap statistics

    Variable |   Reps   Observed       Bias   Std. err.   [95% conf. interval]
-------------+------------------------------------------------------------------
       group |   1000          5   -2.15234   1.795984    1.475666   8.524334   (N)
             |                                                  -1          5   (P)
             |                                            5.454545   7.142857  (BC)
-------------+------------------------------------------------------------------
         mh0 |   1000   .6666667   .0184698   .0623278    .5443583   .7889751   (N)
             |                                            .5714286         .8   (P)
             |                                                  .6   .8648649  (BC)
-------------+------------------------------------------------------------------
       const |   1000         25  -.1960998    5.28891    14.62135   35.37865   (N)
             |                                                  15   35.17857   (P)
             |                                               18.75   38.57143  (BC)
--------------------------------------------------------------------------------
                                   N: Normal, P: Percentile, and BC: Bias-corrected
```

Interpretation

The median of ***uncensored*** *SF-36v2 MH score at 6 months for the treatment group is 5.00 points higher than the median of the control group after allowing or adjusting for the baseline MH score.*

CLAD, similar to Tobit, depicts the relationship between the latent variable and covariates, and in order to obtain coefficients for the observed PRO score i.e. SF-36v2 MH score on a 0 to 100 scale in our case, marginal effects can be used (Clarke, Gray and Holman, 2002; Pullenayegum *et al.*, 2011). However, the marginal estimations for the CLAD estimator are not available in Stata/MP 17.0.

6.4.3 Ordinal regression

Ordinal regression assumes the observed dependent variable (Y) is ordinal with k possible ordered categories or levels, and the latent dependent variable (Y^*) is a continuous variable with a linear function of a series of independent variables ($\boldsymbol{X}$). The inverse standard Normal cumulative distribution function, denoted by Φ^{-1}, is used for ordered probit model, and the logit link is used for ordered logit model. The cumulative probability if individual i having Y equal or less than the value of a level l is

$$\theta_{il} = P(Y_i \leq l); l = 0,1,\dots,k-1; i = 1,\dots,N \tag{6.12}$$

$$Y^* = \boldsymbol{X\beta} + \varepsilon \tag{6.13}$$

For the OL:

$$g(\theta_{il}) = logit(\theta_{il}) = \ln\left(\frac{\theta_{il}}{1-\theta_{il}}\right) = \boldsymbol{X\beta} \tag{6.14}$$

$$\varepsilon \sim Logistic(0,1) \tag{6.15}$$

For the OP:

$$g(\theta_{il}) = probit(\theta_{il}) = \Phi^{-1}(\theta_{il}) = \boldsymbol{X\beta} \tag{6.16}$$

$$\varepsilon \sim N(0,1) \tag{6.17}$$

The key assumption of both models is the proportional odds assumption, which assumes that each covariate share a constant OR across the cut-points i.e. the influence of each covariate on the response variable is independent of the cut-points (Arostegui, Núñez-Antón and Quintana, 2012).

6.4.3.1 Recoding technique for ordinal and binomial regression

Recoding of PRO scores to ordinal scale is required to run ordinal regression (OL and OP) (Arostegui, Núñez-Antón and Quintana, 2013). In our example, SF-36v2 MH scores at baseline and at 6-month post-randomisation follow-up are recoded into a new ordinal scale with 21 possible ordered categorical values, ranging from 0 to 20.

The observed dependent variable (Y) of the ordinal regression and binomial regression is defined using the following equations:

$$Y = \begin{cases} 0,\ 0 < Y^* \leq j_1 \\ 1,\ j_1 < Y^* \leq j_2 \\ \dots, \\ l_{k-2},\ j_{k-3} < Y^* \leq j_{k-2} \\ l_{k-1},\ j_{k-2} < Y^* \leq 100 \end{cases} \tag{6.18}$$

where j is a set of cut-points, k is the number of possible categorical values, and Y^* denotes the continuous latent dependent variable. In the case of the SF-36v2 MH score,

$$k = 21$$

$$l \in \{0,1,\dots,20\}$$

$$j \in \{0, 2.5, 7.5, 12.5, \dots 92.5, 97.5, 100\}$$

Variable specification (using `mh6` as an example)

- `mh6` is the SF-36v2 MH scores at 6-month post-randomisation in the LM trial;
- `omh6` is the recoded ordinal score of `mh6` to fit ordinal and binomial regression.

Stata code for recoding to ordinal scale

```
gen omh6 = recode(mh0, 2.5, 7.5, 12.5, 17.5, 22.5, 27.5, 32.5, 37.5,
42.5, 47.5, 52.5, 57.5, 62.5, 67.5, 72.5, 77.5, 82.5, 87.5, 92.5,
97.5)
```

The distribution of the recoded SF-36v2 MH scores at baseline and 6-month post-randomisation on a 21-point ordinal scale is shown in Figure 6.2.

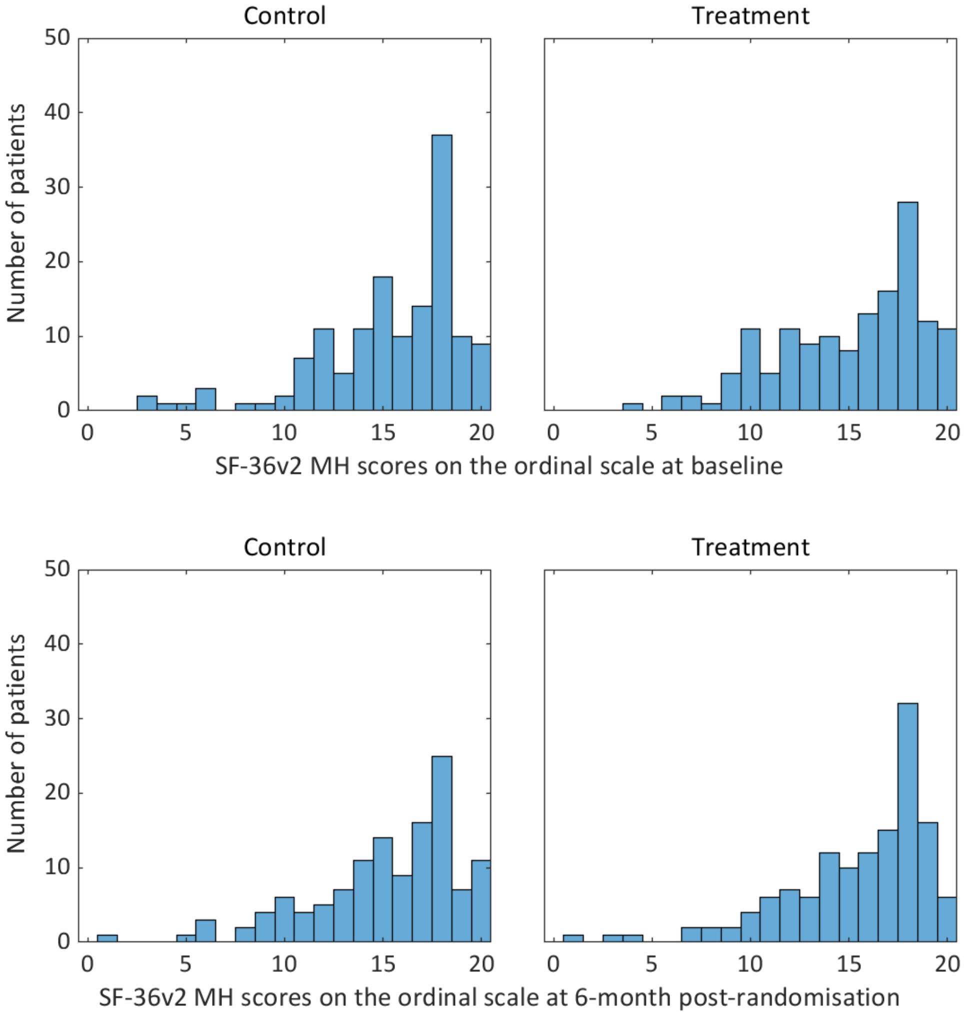


Figure 6.2 Distribution of SF-36v2 MH score on a 21-point ordinal scale from 0 to 20

6.4.3.2 Ordered logit model (OL)

Model

$$logit(\theta_{il}) = \beta_{0l} + \boldsymbol{X_i\beta} \quad l = 0, \dots, k-1; i = 1, \dots, N \tag{6.19}$$

where β_{0l} stands for different intercepts for L categories, which is assumed to be specific for each one of the k equations in the model. β stands for the effect the covariates have on the cumulative probability, which is assumed to be equal for the k equations in the model (proportional odds assumption).

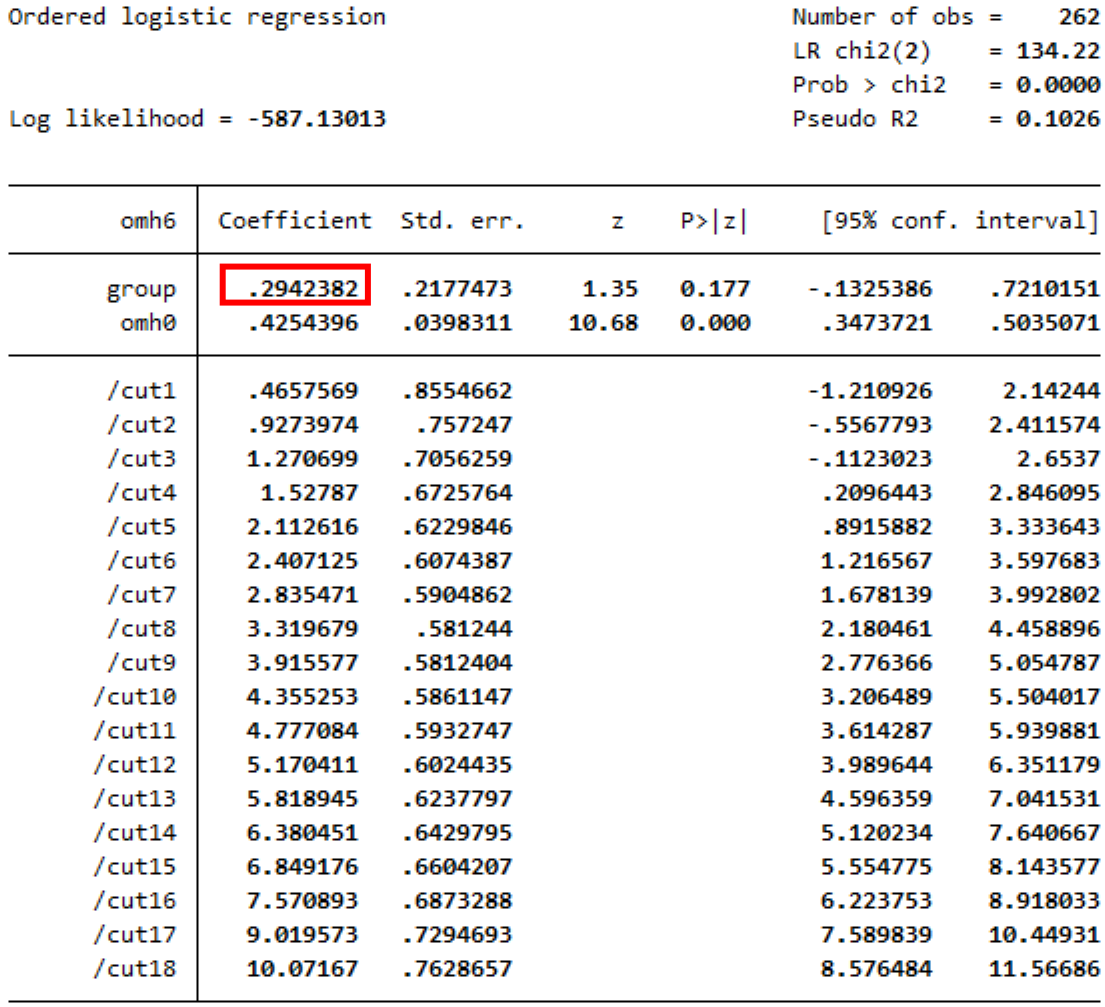
Model assumption: proportional odds assumption.

Estimation method: MLE

Stata code

```
ologit omh6 group omh0
```

Stata output



```
Ordered logistic regression                         Number of obs =    262
                                                    LR chi2(2)    = 134.22
                                                    Prob > chi2   = 0.0000
Log likelihood = -587.13013                         Pseudo R2     = 0.1026

------------------------------------------------------------------------------
        omh6 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |   .2942382   .2177473     1.35   0.177    -.1325386    .7210151
        omh0 |   .4254396   .0398311    10.68   0.000     .3473721    .5035071
-------------+----------------------------------------------------------------
       /cut1 |   .4657569   .8554662                     -1.210926     2.14244
       /cut2 |   .9273974    .757247                     -.5567793    2.411574
       /cut3 |   1.270699   .7056259                     -.1123023      2.6537
       /cut4 |    1.52787   .6725764                      .2096443    2.846095
       /cut5 |   2.112616   .6229846                      .8915882    3.333643
       /cut6 |   2.407125   .6074387                      1.216567    3.597683
       /cut7 |   2.835471   .5904862                      1.678139    3.992802
       /cut8 |   3.319679    .581244                      2.180461    4.458896
       /cut9 |   3.915577   .5812404                      2.776366    5.054787
      /cut10 |   4.355253   .5861147                      3.206489    5.504017
      /cut11 |   4.777084   .5932747                      3.614287    5.939881
      /cut12 |   5.170411   .6024435                      3.989644    6.351179
      /cut13 |   5.818945   .6237797                      4.596359    7.041531
      /cut14 |   6.380451   .6429795                      5.120234    7.640667
      /cut15 |   6.849176   .6604207                      5.554775    8.143577
      /cut16 |   7.570893   .6873288                      6.223753    8.918033
      /cut17 |   9.019573   .7294693                      7.589839    10.44931
      /cut18 |   10.07167   .7628657                      8.576484    11.56686
------------------------------------------------------------------------------
```

Interpretation

Treatment group tends to have higher SF-36v2 MH scores at 6-month follow-up. There are three different population summary measures that are commonly generated from OL: OR, effect size (Chinn, 2000), and probability.

OR: *A person in the treatment group has an increase of 1.34 in the odds of having a SF-36 mental health score one level or higher at 6 months than the odds in the control group after allowing or adjusting for the baseline MH score.* $OR_{TE} = \exp\big(coef(TE)\big) = \exp(0.294) = 1.34$

Effect size (ES): $ES = \frac{coef(TE)}{\sigma} = \frac{0.294}{\pi/\sqrt{3}} = 0.163$

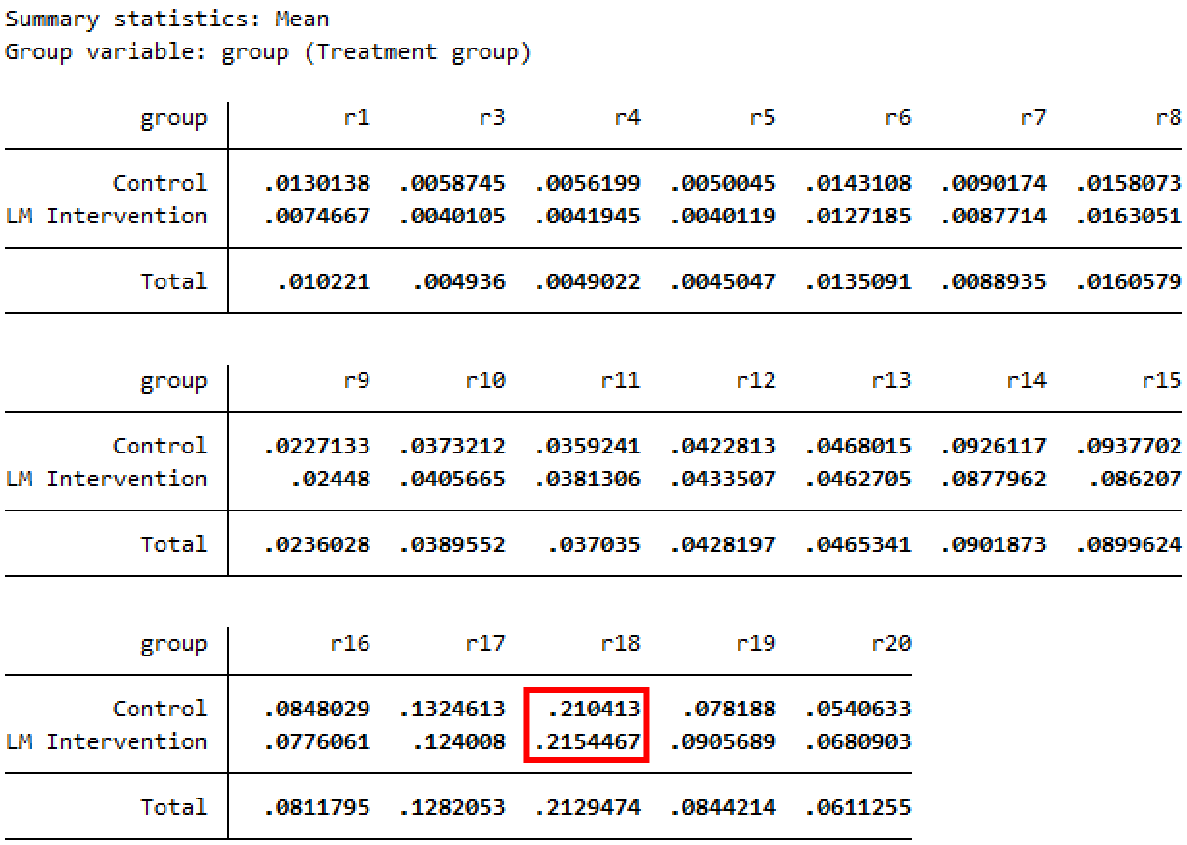
Probability: *The probability of scoring 90.0 at 6 months is 21.5% for the treatment group and 21.0% for the usual care group…*

Stata codes

```
predict r1 r3 r4 r5 r6 r7 r8 r9 r10 r11 r12 r13 r14 r15 r16 r17 r18
r19 r20
```

```
tabstat r1 r3 r4 r5 r6 r7 r8 r9 r10 r11 r12 r13 r14 r15 r16 r17 r18
r19 r20, by(group) stat(mean)
```

Stata output



```
Summary statistics: Mean
Group variable: group (Treatment group)

          group |        r1        r3        r4        r5        r6        r7        r8
----------------+----------------------------------------------------------------------
        Control |  .0130138  .0058745  .0056199  .0050045  .0143108  .0090174  .0158073
LM Intervention |  .0074667  .0040105  .0041945  .0040119  .0127185  .0087714  .0163051
----------------+----------------------------------------------------------------------
          Total |   .010221   .004936  .0049022  .0045047  .0135091  .0088935  .0160579
---------------------------------------------------------------------------------------

          group |        r9       r10       r11       r12       r13       r14       r15
----------------+----------------------------------------------------------------------
        Control |  .0227133  .0373212  .0359241  .0422813  .0468015  .0926117  .0937702
LM Intervention |    .02448  .0405665  .0381306  .0433507  .0462705  .0877962   .086207
----------------+----------------------------------------------------------------------
          Total |  .0236028  .0389552   .037035  .0428197  .0465341  .0901873  .0899624
---------------------------------------------------------------------------------------

          group |       r16       r17       r18       r19       r20
----------------+--------------------------------------------------
        Control |  .0848029  .1324613   .210413   .078188  .0540633
LM Intervention |  .0776061   .124008  .2154467  .0905689  .0680903
----------------+--------------------------------------------------
          Total |  .0811795  .1282053  .2129474  .0844214  .0611255
-------------------------------------------------------------------
```

6.4.3.3 Ordered probit model (OP)

Model

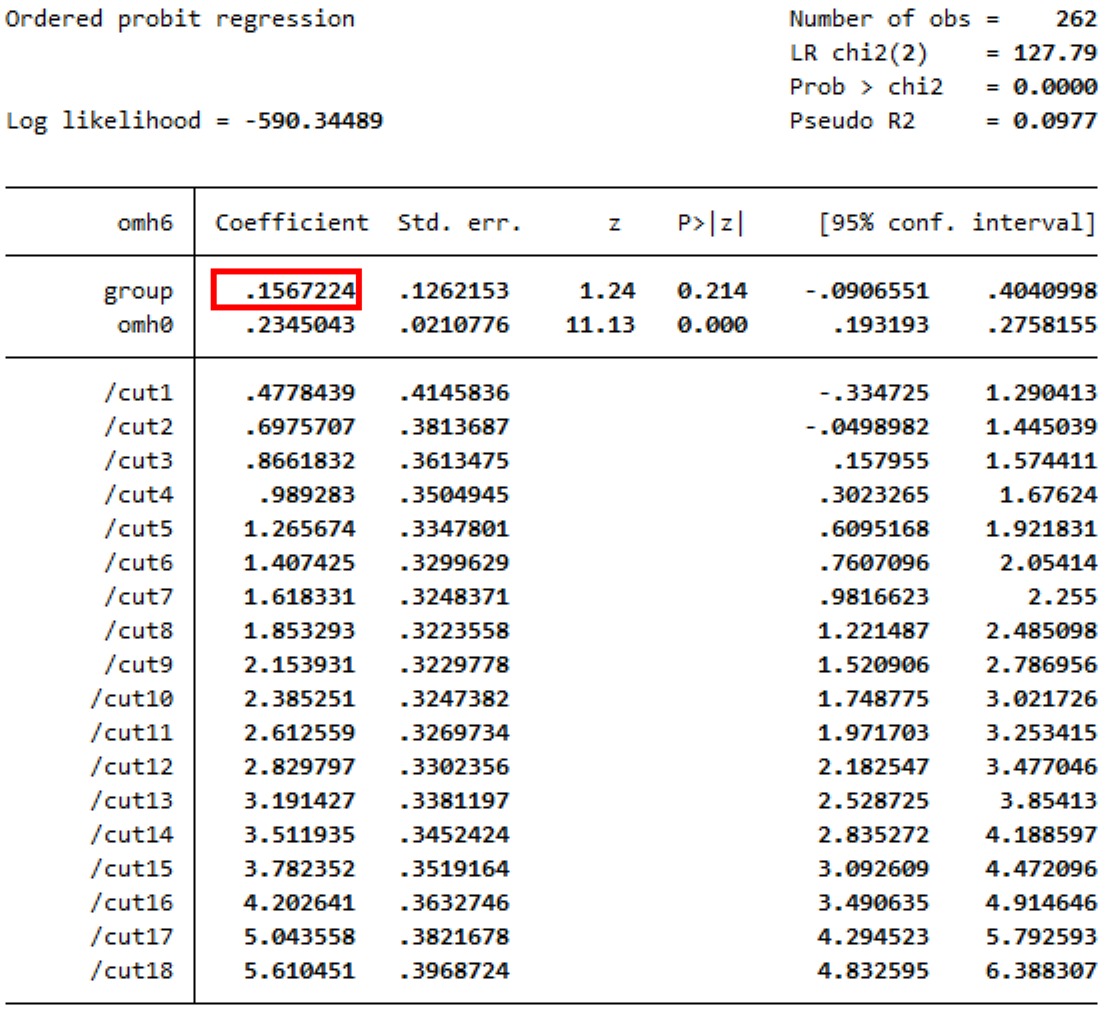
$$\Phi^{-1}(\theta_{il}) = \beta_{0l} + \boldsymbol{X_i\beta} \tag{6.20}$$

The model assumption of OP s identical to OL except that OP uses a probit link function.

Stata code

```
oprobit omh6 group omh0
```

Stata output



```
Ordered probit regression                           Number of obs =     262
                                                    LR chi2(2)    =  127.79
                                                    Prob > chi2   =  0.0000
Log likelihood = -590.34489                         Pseudo R2     =  0.0977

------------------------------------------------------------------------------
        omh6 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |   .1567224   .1262153     1.24   0.214    -.0906551    .4040998
        omh0 |   .2345043   .0210776    11.13   0.000      .193193    .2758155
-------------+----------------------------------------------------------------
       /cut1 |   .4778439   .4145836                      -.334725    1.290413
       /cut2 |   .6975707   .3813687                     -.0498982    1.445039
       /cut3 |   .8661832   .3613475                       .157955    1.574411
       /cut4 |    .989283   .3504945                      .3023265     1.67624
       /cut5 |   1.265674   .3347801                      .6095168    1.921831
       /cut6 |   1.407425   .3299629                      .7607096     2.05414
       /cut7 |   1.618331   .3248371                      .9816623       2.255
       /cut8 |   1.853293   .3223558                      1.221487    2.485098
       /cut9 |   2.153931   .3229778                      1.520906    2.786956
      /cut10 |   2.385251   .3247382                      1.748775    3.021726
      /cut11 |   2.612559   .3269734                      1.971703    3.253415
      /cut12 |   2.829797   .3302356                      2.182547    3.477046
      /cut13 |   3.191427   .3381197                      2.528725     3.85413
      /cut14 |   3.511935   .3452424                      2.835272    4.188597
      /cut15 |   3.782352   .3519164                      3.092609    4.472096
      /cut16 |   4.202641   .3632746                      3.490635    4.914646
      /cut17 |   5.043558   .3821678                      4.294523    5.792593
      /cut18 |   5.610451   .3968724                      4.832595    6.388307
------------------------------------------------------------------------------
```

Interpretation

The population summary measure of OL can be ORs whereas there is no such explanation for OP. There are two different population summary measures that are commonly generated from OL: effect size (Chinn, 2000), and probability. The most common way to interpret ordered probit is the predicted probabilities based on estimates.

Effect size (ES): *ES estimated by the OP (0.156) is similar to the effect size estimated by the OL (0.163).*

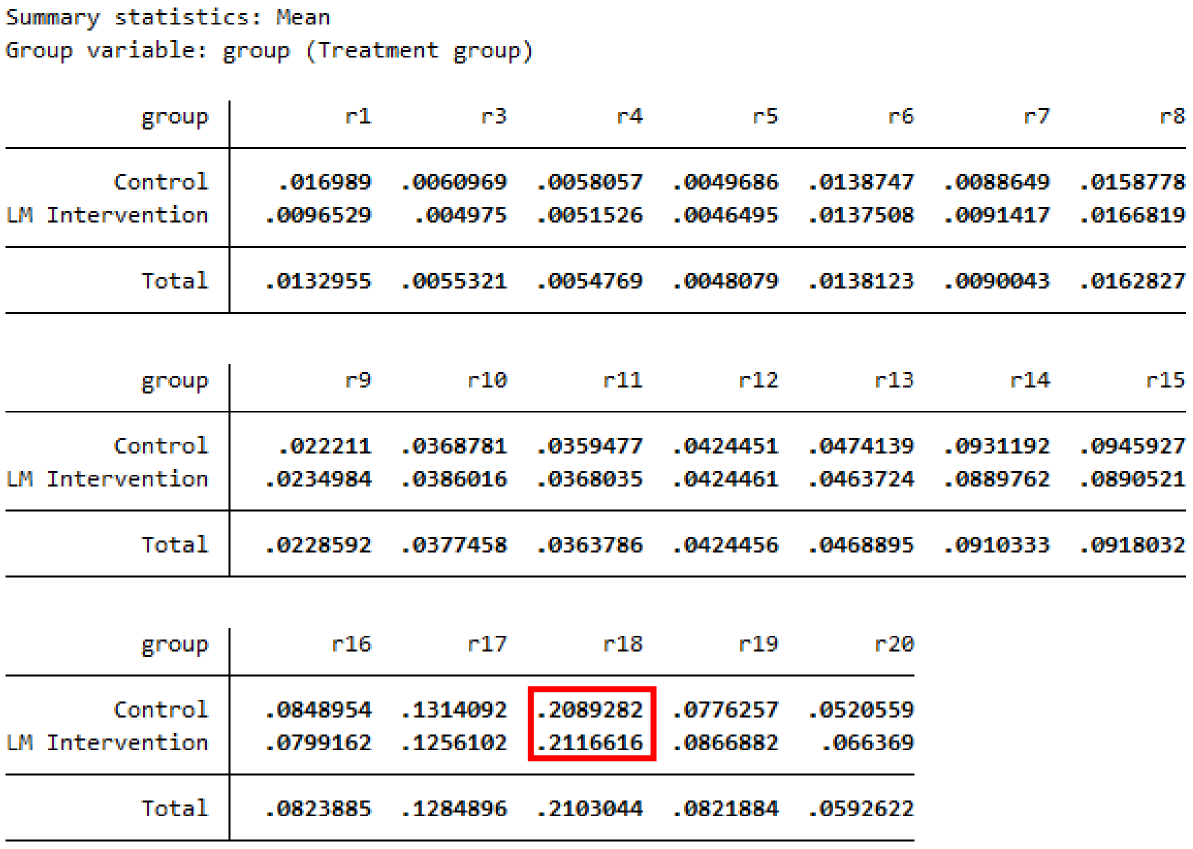
Probability: *The probability of scoring 90.0 at 6 months is 21.2% for the treatment group and 20.9% for the usual care group after allowing or adjusting for the baseline MH score…*

Stata codes

```
predict r1 r3 r4 r5 r6 r7 r8 r9 r10 r11 r12 r13 r14 r15 r16 r17 r18
r19 r20

tabstat r1 r3 r4 r5 r6 r7 r8 r9 r10 r11 r12 r13 r14 r15 r16 r17 r18
r19 r20, by(group) stat(mean)
```

Stata output



```
Summary statistics: Mean
Group variable: group (Treatment group)

           group |        r1        r3        r4        r5        r6        r7        r8
-----------------+----------------------------------------------------------------------
         Control |   .016989  .0060969  .0058057  .0049686  .0138747  .0088649  .0158778
 LM Intervention |  .0096529   .004975  .0051526  .0046495  .0137508  .0091417  .0166819
-----------------+----------------------------------------------------------------------
           Total |  .0132955  .0055321  .0054769  .0048079  .0138123  .0090043  .0162827
----------------------------------------------------------------------------------------

           group |        r9       r10       r11       r12       r13       r14       r15
-----------------+----------------------------------------------------------------------
         Control |   .022211  .0368781  .0359477  .0424451  .0474139  .0931192  .0945927
 LM Intervention |  .0234984  .0386016  .0368035  .0424461  .0463724  .0889762  .0890521
-----------------+----------------------------------------------------------------------
           Total |  .0228592  .0377458  .0363786  .0424456  .0468895  .0910333  .0918032
----------------------------------------------------------------------------------------

           group |       r16       r17       r18       r19       r20
-----------------+--------------------------------------------------
         Control |  .0848954  .1314092  .2089282  .0776257  .0520559
 LM Intervention |  .0799162  .1256102  .2116616  .0866882   .066369
-----------------+--------------------------------------------------
           Total |  .0823885  .1284896  .2103044  .0821884  .0592622
--------------------------------------------------------------------
```

Highlight

Although the coefficients estimated by `ologit` and `oprobit` are different, the predicted probabilities by both methods are similar. For both OL and OP, marginal effect can be produced to show the change in probability when the independent variable increases by one unit, using the `margins` command following the regression command.

`ologit omh6 group omh0` or `oprobit omh6 group omh0`

```
margins, dydx(*)
```

6.4.4 Binomial regression

Binomial regression is a generalisation from logistic regression, which assumes that the conditional distribution of Y_i of the ith subject given the number of success (θ_i) obtained in k binomial trials follows a binomial distribution, i.e. $Y_i \sim Bin(k, \theta_i)$. Unlike ordinal regression, θ_i is thought to differ between individuals i.e. there are individual random effects. In terms of the PROs, the original score Y, is recoded into k discrete values or categories, and the recoded outcome Y^* is on a 0 to $k-1$ interval, θ_i represents the probability of obtaining one point or value, l, on the recoded PRO dimension, Y^*, for a particular subject.

The baseline probability of success (θ_{0i}) with no covariates is a random variable. θ_{0i} is assumed to follow a beta distribution for beta-binomial regression, and a logit-Normal distribution for binomial-logit-Normal regression (Arostegui, Núñez-Antón and Quintana, 2007, 2012; Liang *et al.*, 2014).

$$Y_i \sim Bin(k, \theta_i) \tag{6.21}$$

$$E(Y_i) = k\theta_i \tag{6.22}$$

The estimated regression coefficients from both methods can be interpreted as (log) ORs. Same as the ordinal regression, recoding of the observed SF-36 dimension scores to an ordinal form is required to run binomial regression.

6.4.4.1 Beta-binomial regression (BB)

Model

$$logit(\theta_i) = logit(\theta_{0i}) + \boldsymbol{X_i \beta} \tag{6.23}$$

$$\theta_{0i} \sim Beta(\alpha, \gamma) \tag{6.24}$$

where θ_{0i} represents the baseline probabilities of success with no covariates in the model, and is independent beta random variables.

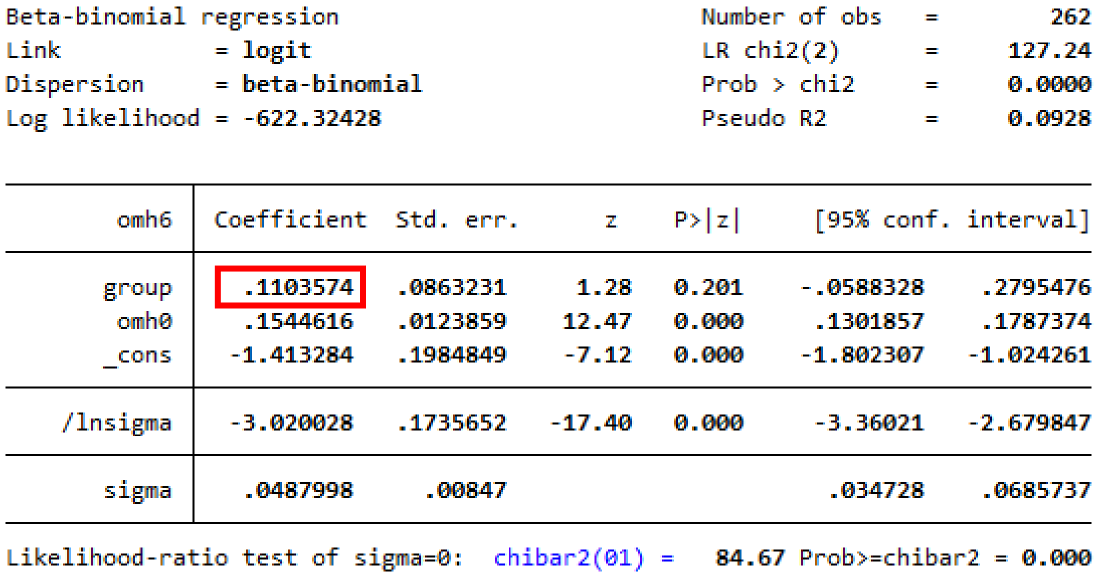
Estimation method: MLE

Stata code

```
gen N = 21 //N is the maximum number of score
replace N=. if omh6 ==.
betabin omh6 group omh0, n(N) link(logit) nolog
```

Note that a package for BB (st0337_1.pkg) needs to be installed before running the codes.

Stata output



```
Beta-binomial regression                        Number of obs     =        262
Link           = logit                          LR chi2(2)        =     127.24
Dispersion     = beta-binomial                  Prob > chi2       =     0.0000
Log likelihood = -622.32428                     Pseudo R2         =     0.0928

------------------------------------------------------------------------------
        omh6 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |   .1103574   .0863231     1.28   0.201    -.0588328    .2795476
        omh0 |   .1544616   .0123859    12.47   0.000     .1301857    .1787374
       _cons |  -1.413284   .1984849    -7.12   0.000    -1.802307   -1.024261
-------------+----------------------------------------------------------------
    /lnsigma |  -3.020028   .1735652   -17.40   0.000     -3.36021   -2.679847
-------------+----------------------------------------------------------------
       sigma |   .0487998     .00847                       .034728    .0685737
------------------------------------------------------------------------------
Likelihood-ratio test of sigma=0:  chibar2(01) =   84.67 Prob>=chibar2 = 0.000
```

Interpretation

Odds ratio: *A person in the treatment group has an increase of 1.12 in the odds of having a one level better (out of 21) or 5.00 points higher SF-36 MH score at 6-month post-randomisation follow-up than the odds in the control group after allowing or adjusting for the baseline MH score.* $OR_{TE} = \exp(coef(TE)) = \exp(0.110) = 1.12$.

6.4.4.2 Binomial-logit-Normal regression (BLN)

Model

$$logit(\theta_i) = \sigma z_i + \boldsymbol{X_i \beta} \tag{6.25}$$

$$z_i \sim N(0,1) \quad \text{i.e. } \theta_{0i} \sim LN(0,1) \tag{6.26}$$

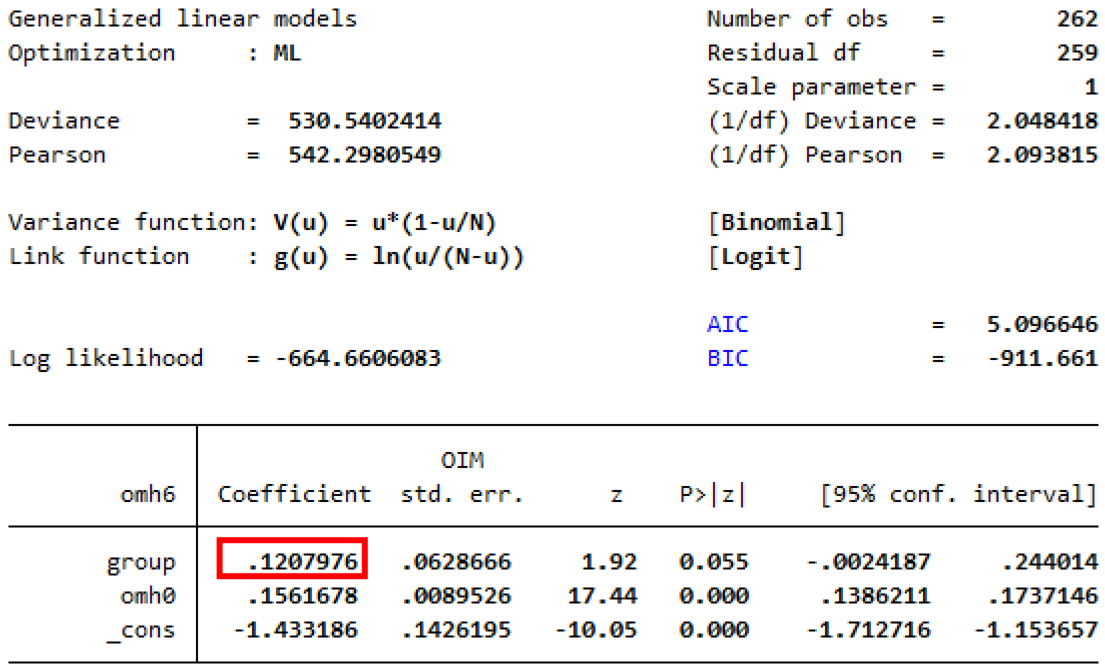
where z_i is assumed to be independent standard Normal random variables. θ_{0i} has a standard logit-Normal distribution (notation LN).

Estimation method: MLE cannot be numerically evaluated, Gauss-Hermite quadrature can be used, leading to maximise marginal likelihood approximation (Arostegui, Núñez-Antón and Quintana, 2012).

Stata code

```
glm omh6 group omh0, link(logit) family(binomial N) nolog
```

Stata output



```
Generalized linear models                         Number of obs     =        262
Optimization     : ML                             Residual df       =        259
                                                  Scale parameter   =          1
Deviance         =  530.5402414                   (1/df) Deviance   =   2.048418
Pearson          =  542.2980549                   (1/df) Pearson    =   2.093815

Variance function: V(u) = u*(1-u/N)               [Binomial]
Link function    : g(u) = ln(u/(N-u))             [Logit]

                                                  AIC               =   5.096646
Log likelihood   = -664.6606083                   BIC               =   -911.661

------------------------------------------------------------------------------
             |                 OIM
        omh6 | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |   .1207976   .0628666     1.92   0.055    -.0024187     .244014
        omh0 |   .1561678   .0089526    17.44   0.000     .1386211    .1737146
       _cons |  -1.433186   .1426195   -10.05   0.000    -1.712716   -1.153657
------------------------------------------------------------------------------
```

Interpretation

A person in the treatment group has an increase of 1.13 in the odds of having a one level better (out of 21) or 5.00 points higher SF-36 MH score at 6-month post-randomisation follow-up than the odds in the control group after allowing or adjusting for the baseline MH score. $OR_{TE} = \exp(coef(TE)) = \exp(0.121) = 1.13$

6.4.5 Fractional regression

Fractional regression can be used to analyse bounded data on a continuous 0 to 1 scale (Papke, 1996). Under this category, we introduce BR and Frac, both of which assume a continuous ratio scattering between 0 and 1. Note that a Frac fits a dependent variable that is greater than or equal to 0 and less than or equal to 1, whereas BR cannot deal with scores at 0 or 1.

$$\text{logit}(\mu_i/(1-\mu_i)) = \boldsymbol{X\beta} \tag{6.27}$$

$$E(y') = \mu \tag{6.28}$$

where y' is the recoded form of y that distributes between 0 and 1.

6.4.5.1 Recoding technique

Recoding of the observed dimension scores to a 0 to 1 scale is necessary to run fractional regression. For Frac, the SF-36 scores on a $[a, b]$ scale need to be transformed to a closed interval [0, 1]. The equation to calculate the transformed score, denoted by Y', is shown below:

$$Y' = (Y - a)/(b - a) \tag{6.29}$$

As BR cannot account for scores at boundaries (i.e. a or b or 0 or 1), the SF-36 scores need to be 'squeezed' to an open interval (0, 1) using sample size, denoted by N, for adjustment. The transformed score, denoted by Y'', is calculated using the following equation (Ferrari and Cribari-Neto, 2004; Smithson and Verkuilen, 2006; Hunger, Baumert and Holle, 2011).

$$Y'' = [Y'(N - 1) + 0.5]/N \tag{6.30}$$

The distributions of the recoded SF-36v2 mental health on percentile scales to run fractional logistic regression and beta regression are shown in Figure 6.3 and Figure 6.4 respectively.

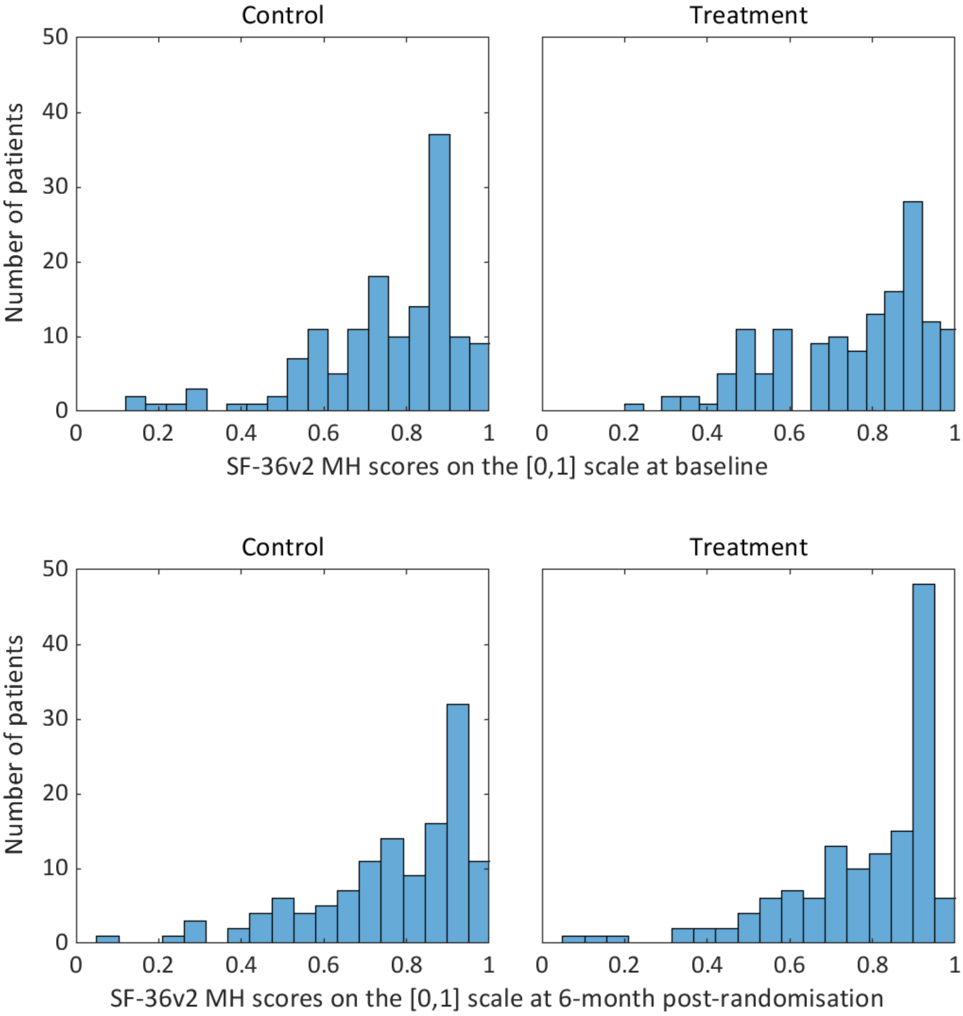


Figure 6.3 Distribution of SF-36v2 MH score on percentile scale [0-1]

Variable specification (using `mh6` as an example)

- `mh6` is the SF-36v2 MH scores at 6-month post-randomisation in the LM trial;
- `fmh6` is the recoded score of `mh6` on a [0,1] scale to fit fractional logistic regression;
- `bmh6` is the recoded score of `mh6` on an (0,1) scale to fit beta regression;
- `SZ` is the sample size of patients;

Stata code for recoding to percentile scale

** To [0,1] scale

```
gen fmh6 = mh6/100
```

** To (0,1) scale

```
egen SZ6 = count(mh6)
gen bmh6 = (fmh6*(SZ6-1)+0.5)/SZ6
```

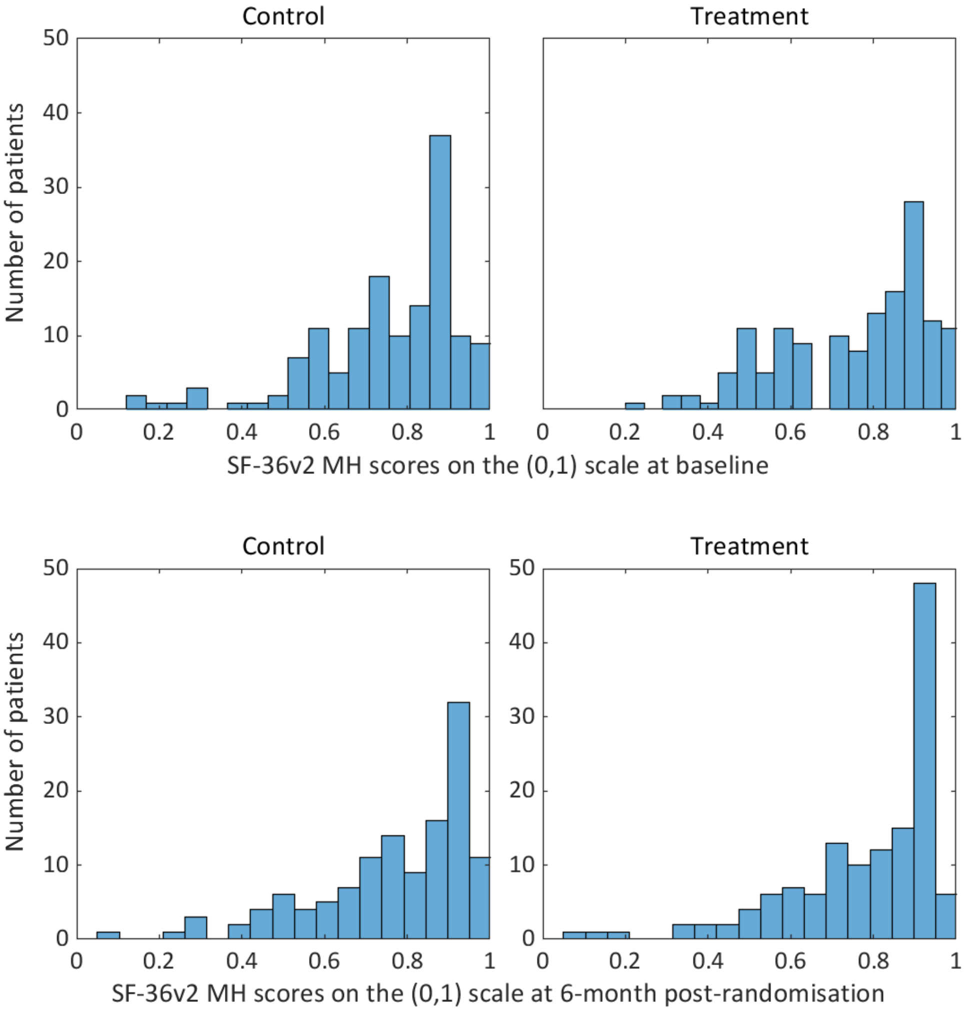


Figure 6.4 Distribution of SF-36v2 mental health score on percentile scale (0-1), squeezed

6.4.5.2 Fractional logistic regression (Frac)

The logit link is adapted for fractional logistic regression and therefore the regression coefficients can be interpreted as logORs.

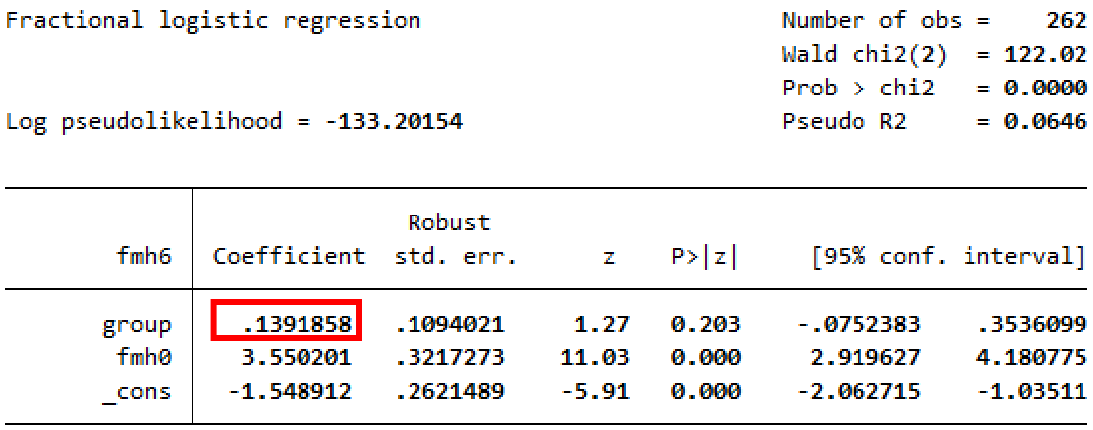
Estimation method: Quasi-likelihood estimation, which does not require the knowledge of the true distribution of the entire model to obtain consistent parameter estimates.

Stata code

```
fracreg logit pkpain12 group kpain0
```

To fit a fractional logistic regression, the response variable is required to be adjusted to a closed interval between 0 and 1. Note that the robust SEs are calculated by default in Stata/MP 17.0.

Stata output



```
Fractional logistic regression                     Number of obs =     262
                                                   Wald chi2(2)  =  122.02
                                                   Prob > chi2   =  0.0000
Log pseudolikelihood = -133.20154                  Pseudo R2     =  0.0646

------------------------------------------------------------------------------
             |               Robust
        fmh6 | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       group |   .1391858   .1094021     1.27   0.203    -.0752383    .3536099
        fmh0 |   3.550201   .3217273    11.03   0.000     2.919627    4.180775
       _cons |  -1.548912   .2621489    -5.91   0.000    -2.062715    -1.03511
------------------------------------------------------------------------------
```

Interpretation

A person in the treatment group has an increase of 1.15 in the odds of having one level better (out of 21) or 5.00 points higher SF-36 MH score at 6-month post-randomisation follow-up than the odds in the control group after allowing or adjusting for the baseline MH score. $OR_{TE} = \exp\big(coef(TE)\big) = \exp(0.139) = 1.15$

6.4.5.3 Beta regression (BR)

BR is proposed because of its flexibility to skewed and bounded data, in which case the PROs is assumed to have a beta distribution with predefined boundaries and the logit link function is often used (Ferrari and Cribari-Neto, 2004). The logit link is adapted because the resulting regression coefficients can be interpreted as logORs.

Model assumption

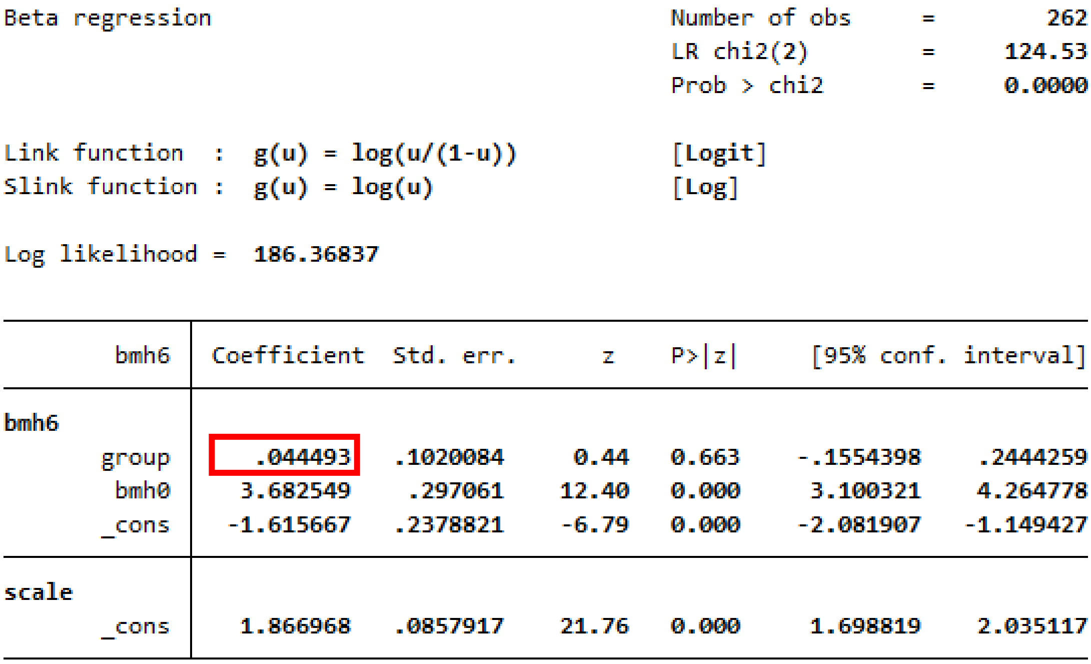
$$y' \sim beta(\mu\varphi, (1-\mu)\varphi) \tag{6.36}$$

Estimation method: MLE

Stata code and output

```
betareg kpain12beta2 group kpain0
```

Stata output



```
Beta regression                                 Number of obs     =        262
                                                LR chi2(2)        =     124.53
                                                Prob > chi2       =     0.0000

Link function   :  g(u) = log(u/(1-u))          [Logit]
Slink function  :  g(u) = log(u)                [Log]

Log likelihood =  186.36837

------------------------------------------------------------------------------
        bmh6 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
bmh6         |
       group |    .044493   .1020084     0.44   0.663    -.1554398    .2444259
        bmh0 |   3.682549    .297061    12.40   0.000     3.100321    4.264778
       _cons |  -1.615667   .2378821    -6.79   0.000    -2.081907   -1.149427
-------------+----------------------------------------------------------------
scale        |
       _cons |   1.866968   .0857917    21.76   0.000     1.698819    2.035117
------------------------------------------------------------------------------
```

Interpretation

A person in the treatment group has an increase of 1.05 in the odds of having one level better (out of 21) or 5.00 points higher SF-36 MH score at 6-month post-randomisation follow-up than the odds in the control group after allowing or adjusting for the baseline MH score. $Odds\ Ratio_{TE} = exp\big(coef(TE)\big) = exp(0.044) = 1.05$

A summary of how to interpret the estimated treatment coefficients from the included 10 statistical methods is presented in Table 6.3. The estimate produced by Tobit is slightly smaller than MLR, while the estimates from both Median and CLAD are much larger than the estimates by MLR and Tobit. OL produces larger estimated ORs than other transformed scale-based methods, and the rest of these methods tend to produce similar estimates except that the estimates from BR is relatively small.

Table 6.3 Summary of how to interpret the estimated treatment coefficient from the statistical methods evaluated based on the SF-36 MH score at 6-month follow-up in the LM trial

(a) Statistical methods that used the untransformed scales of measurement

Statistical methods	Coef	Interpretation
MLR	2.31	The mean of the MH score at 6 months for the treatment group is 2.31 points higher than the mean for the control group, after adjusting for baseline MH score.
Tobit	2.02	The mean of the *uncensored* MH score at 6 months for the treatment group is 2.02 points higher than the mean for the control group, after adjusting for baseline MH score.
Median	5.00	The median of the MH score at 6 months for the treatment group is 5.00 points higher than the median for the control group, after adjusting for baseline MH score.
CLAD	5.00	The median of the *uncensored* MH score at 6 months for the treatment group is 5.00 points higher than the median for the control group, after adjusting for baseline MH score.

(b) Statistical methods that used transformed scales of measurement

Statistical methods	Coef	Odds ratios	Interpretation
OL	0.29	1.34	The treatment group has an increase of 1.34 in the odds of having one level or 5.00 points higher MH score at 6 months than the odds in the control group, after adjusting for baseline MH score.
OP	0.16	NA	(Marginal effects need to be calculated to generate the probability) The probability of scoring 90.0 at 6 months is 21.2% for the treatment group and 20.9% for the usual care group, after adjusting for baseline MH score…
BB	0.11	1.12	The treatment group has an increase of 1.12 in the odds of having one level or 5.00 points higher MH score at 6 months than the odds in the control group, after adjusting for baseline MH score.
BLN	0.12	1.13	The treatment group has an increase of 1.13 in the odds of having one level or 5.00 points higher MH score at 6 months than the odds in the control group, after adjusting for baseline MH score.
Frac	0.14	1.15	The treatment group has an increase of 1.15 in the odds of having one level or 5.00 points higher MH score at 6 months than the odds in the control group, after adjusting for baseline MH score.
BR	0.04	1.05	The treatment group has an increase of 1.05 in the odds of having one level or 5.00 points higher SF-36 MH score at 6 months than the odds in the control group, after adjusting for baseline MH score.

BB, beta-binomial regression; BLN, binomial-logit-Normal regression; BR, beta regression; CLAD, censored absolute least deviations regression; Coef, treatment coefficient; Frac, fractional logistic regression; Median, median regression; MH, SF-36 mental health dimension; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

6.5 Discussion

In this chapter, we explicitly described 10 statistical methods that we have identified and filtered from previous chapters. We started with the explanation of the theory under the GLM framework of these statistical methods including model formula, model assumptions, and estimation methods. We provided an example of the empirical analysis by using SF-36v2 MH sores at 6-month post-randomisation in the LM trial to explain how to fit these methods in computer statistical software (Stata/MP 17.0) and how to interpret the estimates generated by different statistical methods (i.e. estimators).

In addition, we introduced the concept of the estimand framework and explained the described the estimand framework for the example dataset. We proposed the use of two different estimand frameworks, i.e. the scale-based estimand framework and the SES estimand framework. The scale-based estimand framework present the estimates from different statistical methods on their original scales, but some estimates cannot be compared directly as they are not on the same scale. Therefore, we adapted the SES as the population summary measure of the treatment effect in the SES estimand framework, which is not as frequently seen as other population summary measures of the treatment effect such as means, risks, or ORs (Walters, Campbell and Lall, 2001).

In practice, the concept of SES has been applied in various scenarios in trials using PROs and their related studies. This includes summary studies such as meta-analysis in literature reviews that compare PROs on different scales, sample size calculation in trial designs that use PRO as primary outcomes, and trials with PROs that used the effect size as the measurement of treatment effectiveness (Parsons *et al.*, 2014; Bell *et al.*, 2017; Clare *et al.*, 2019; Brealey *et al.*, 2020; Vanderhout *et al.*, 2022). For linear models, the effect size is a ratio of estimated coefficient over SD of the estimate. The standardisation procedure is completed by the Z statistics, adjusting for sample size. Therefore, when estimating the same treatment effect using different methods, the SES assesses the statistical power of these methods for a given sample size. For instance, if the data is ordinal and the model assumption of OL is satisfied, the OL is likely to have higher power than other statistical methods, and thus be the most appropriate method for analysing the data. In theory, the most appropriate method is more likely to capture the 'truth' than other methods. However, as the dimension scores of SF-36 have different categories and distribution patterns, it is therefore difficult to assign them to a certain type of distribution and to decide what statistical methods to use for analysis.

In the next chapter, the empirical analysis will be conducted by applying these 10 statistical methods to eight dimensions in both versions of SF-36 using RCT datasets.

Chapter 7 Results of empirical analysis of the filtered statistical methods for the analysis of SF-36 in RCTs

7.1 Introduction

In Chapter 6, the technical details and practical applications of the 10 filtered statistical methods have been described and explained using an example dataset. This chapter aims to conduct empirical analysis to test the ability of these methods to run in different real world data scenarios, to compare the consistency in their estimates, and to evaluate their model fit and post-estimation statistics.

In the rest of this chapter, we fit the 10 filtered statistical methods to PROs in multiple RCTs, interpret their estimated treatment coefficients and SESs, and present the change of model fit statistics with an increase in the number of possible categorical values.

7.2 Methods

A series of secondary analyses of PROs at multiple post-randomisation time-points were conducted by applying the 10 statistical methods to RCT datasets that used SF-36 as clinical outcomes.

7.2.1 Description of the SF-36 PRO

The SF-36 is a widely used PRO to measure QoL from patients' perspectives in clinical trials. The SF-36 consists of eight health dimension scores and one health transition item using 36 items on different ordinal categorical scales. The eight dimension scores include physical functioning (PF), role limitation - physical (RP), bodily pain (BP), general health (GH), vitality (VT), social functioning (SF), role limitation - emotional (RE), and mental health (MH) (Ware, Kosinski and Gandek, 1993).

The original version of SF-36 was initially released in 1992 (Ware and Sherbourne, 1992), with its validity and reliability tested in the sequent two years (McHorney, Ware and Raczek, 1993; McHorney *et al.*, 1994). Modifications of the original SF-36 (SF-36v1) include the RAND 36-item (Hays, Sherbourne and Mazel, 1993), a publicly available version with slightly different scoring methods to the original version (Ware, Kosinski and Gandek, 1993); the SF-36v2 (Ware, 2000), an upgraded version with improvements in wording and in different ordinal categorical scales of some items to improve the internal reliability consistency and reduce ceiling and floor effects (Jenkinson *et al.*, 1999); the SF-6D, a popular preference-based measure that produces utility scores for use in economic evaluation (Brazier, Roberts and Deverill, 2002); and other shorter versions that use only eight or 12 items instead of all 36 items (Laucis, Hays and Bhattacharyya, 2015).

The main difference between the SF-36v1 and SF-36v2 is the change in wording and in ordinal categorical scales of some items, which accordingly changes the number of possible values in some dimensions. For example, the RE dimension in SF-36v1 is composed of three binary items, the crude score of which ranges from three to six with four possible values; after scattering the four values equally in a 0 to 100 scale, the possible values for RE is 0, 33.3, 66.7, and 100. Whereas the RE dimension in SF-36v2 is composed of three items on five-point Likert-scale, the crude score of which ranges from three to 15 with 13 possible values. The scoring strategies for the eight dimension scores in both versions of SF-36 are shown in Table 7.1.

Two types of scoring mechanisms are commonly seen to produce SF-36 dimension scores: the original scoring and the norm-based scoring. The original scoring anchors each scale from 0, representing the worst score on all items to 100, representing the best score on all items. Norm-based scoring linearly rescales the eight dimension scores to achieve a mean score of 50 and a SD of 10 in the reference population (i.e. the US general population) (Ware, 2000; Maruish, 2011). While the described methods apply to both scoring approaches, we use the original, 0 to 100, scoring here for simplicity.

Table 7.1 Different scoring strategies for eight dimensions in SF-36v1 and SF-36v2

SF-36 Dimension	No. items	SF-36v1			SF-36v2		
		No. levels	Crude scores	No. values after recoding	No. levels	Crude scores	No. values after recoding
PF	10	21	4 to 24	21	21	4 to 24	21
RP	4	5	4 to 8	5	17	4 to 20	17
BP	2	10	2 to 11	10	10	2 to 11	10
GH	5	21	4 to 24	21	21	4 to 24	21
VT	4	21	4 to 24	21	17	4 to 20	17
SF	2	10	2 to 11	9	9	2 to 10	9
RE	3	4	3 to 6	4	13	3 to 15	13
MH	5	26	5 to 30	26	21	4 to 24	21

BP, bodily pain; GH, general health; MH, mental health; PF, physical functioning; RE, role limitation - emotional; RP, role limitation - physical; SF, social functioning; VT, vitality.

7.2.2 Description of datasets

The nine RCTs analysed in this chapter are Acupuncture (Thomas *et al.*, 2006), FED (Gariballa *et al.*, 2006), IPSU (Jha *et al.*, 2018), Knee Replacement (Mitchell *et al.*, 2005), Leg Ulcer (Morrell *et al.*, 1998), NAMEIT (Jack, Prestele and Bakshi, 2000), COPD (Waterhouse *et al.*, 2010), LM (Mountain *et al.*, 2017), and PLINY (Mountain *et al.*, 2014). Their trial characteristics are presented in Table 7.2. A total number of 2,045 patients were randomised and 1,569 of them were analysed. Of the nine trials, three used SF-36 dimension scores as primary outcomes: the Acupuncture trial used SF-36v1 BP score at 12-month post-randomisation; LM and PLINY used SF-36v2 MH score at 6-month post-randomisation.

7.2.3 Statistical methods for empirical analysis

The 10 statistical methods were applied for the empirical analysis of the treatment difference of SF-36 dimension scores using trial data: MLR, Median, Tobit, CLAD, OL, OP, BB, BLN, Frac, and BR.

The estimated treatment coefficients by different statistical methods may not be comparable as they are theoretically based on different scales. Therefore, the two estimand frameworks that were described in Chapter 6 Section 6.3 were adapted to generate the scale-based population summary measures (i.e. means/medians and logORs) and the SES population summary measure.

Model assumptions were tested in trials using SF-36 dimension scores as primary outcomes (i.e. Acupunture, LM, and PLINY), including the Normality of residuals and the homoscedasticity assumptions for MLR and Tobit regression, and proportional odds assumptions for ordinal regression methods.

Different recoding techniques are required to run ordinal regression, binomial regression, and fractional regression. We used the recoding techniques for SF-36v1 dimension scores proposed by Arostegui, Núñez-Antón and Quintana (2013) for this empirical analysis, but we slightly tweaked this recoding techniques to fit our data. This is because our dataset also contains SF-36v2, which requires different recoding techniques in some dimensions that have been improved with a different number of possible values. In addition, the scorings for the BP and SF in SF-36v1 in our RCT datasets did not follow the standard scoring strategy (Ware, Kosinski and Gandek, 1993) and the raw scores are not retrievable for recalculation of the standardised dimension scores. The detailed recoding techniques for SF-36 dimension scores are explained in Appendix C.1.

7.2.4 Model fit estimated by Akaike information criterion

The Akaike information criterion (AIC) (Akaike, 1974) values were produced when fitting different statistical methods to compare the model fit, using the following equation:

$$AIC = 2s - 2\ln(\hat{L}) \tag{7.5}$$

where $\hat{L}$ is the maximum likelihood for the model, and s is the number of estimated parameters.

A lower AIC value represents a better model fit. The value s for different statistical methods with the same set of independent variables can be different, and the AIC values cannot be calculated for CLAD and Median since they both are quantile regression methods which use LAD or CLAD for estimation and not maximum likelihood. As the comparison of AICs require the methods to model the same response variable (Akaike, 1974), these methods are categorised into different groups according to their distributional assumptions and recoding techniques on SF-36 dimension scores in the scatterplots.

Table 7.2 Characteristics of the nine RCT datasets with SF-36 for empirical analysis

Trial name	Trial population	SF-36 version	Follow-up timepoints (months)	Sample size at baseline			Max N (PRO) ^			Reference
				Total	Control	Treat ment	Total	Control	Treat ment	
Acupuncture*	Adults with non-specific low back pain	v1	3, 12, 24	239	80	159	217	71	146	(Thomas *et al.*, 2006)
FED	Older hospitalised patients (aged>=65) with acute illness	v1	1.5, 6	445	222	223	225	119	106	(Gariballa *et al.*, 2006)
IPSU	Women with urinary incontinence and sexual dysfunction	v1	6	107	53	56	66	35	31	(Jha *et al.*, 2018)
Knee Replacement	Osteoarthritis patients undergoing total knee replacement	v1	3	115	58	57	114	57	57	(Mitchell *et al.*, 2005)
Leg Ulcer	Patients with venous leg ulcers	v1	3, 12	233	113	120	233	113	120	(Morrell *et al.*, 1998)
NAMEIT	Patients with early severe rheumatoid arthritis	v1	2, 4, 6, 8, 10, 12	222	110	112	222	110	112	(Jack, Prestele and Bakshi, 2000)
COPD	Patients with chronic obstructive pulmonary disease	v2	2, 6, 12, 18	239	129	110	174	93	81	(Waterhouse *et al.*, 2010)
Lifestyle Matters*	Independently living older people (aged 65 or more)	v2	6, 24	288	143	145	262	126	136	(Mountain *et al.*, 2017)
PLINY*	Independently living older people (aged 75 or more)	v2	6	157	79	78	56	30	26	(Mountain *et al.*, 2014)
Total				2045	987	1060	1569	754	815	

* Trials using SF-36 dimensions as primary outcomes.

^ Max N is the maximum sample size for the baseline and post-randomisation follow-up correlations.

PRO, patient-reported outcomes.

Scatterplots were generated to compare the estimated treatment coefficients from different statistical methods, using consistent markers for each method and consistent colours for each trial. Estimated treatment coefficients from MLR and BB were used as the reference benchmark for statistical methods that produce estimates on the untransformed and transformed scales respectively, since MLR is the most commonly used methods for analysing PROs (Qian *et al.*, 2021), and BB was reported to render satisfactory results in various situations for PRO analysis (Arostegui, Núñez-Antón and Quintana, 2012). The SESs from different methods were displayed in scatterplots, using MLR as the reference benchmark for all included methods regardless of their scales, as the SES is believed comparable among statistical methods on different scales under the estimand framework. Effect size plots were graphed for SESs with its associated CIs from 10 statistical methods, together with two horizontal lines representing the clinical and statistical significance. We also graphed the change of AICs against a different number of possible categorical values in SF-36, using a series of scatterplots with fitted lines. The statistical package Stata/MP 17.0 is used for statistical analysis and MATLAB R2023a is used for data visualisation. The Stata codes to apply recoding techniques and regression analysis are summarised in Appendix C.3.

7.2.5 Ethics approval

This empirical analysis conducts secondary analyses of previously collected trial data. The data have been previously obtained from RCTs run within the School of Health and Related Research (ScHARR), University of Sheffield, where the ethics approvals were received from the appropriate NHS Research Ethics Committee and individual informed consent was obtained to take part in the original trial. This project uses non-personal robustly anonymised, existing data from which the original participants cannot be identified, and this project will not involve recruiting new participants. The ethics application has been approved by the ScHARR Research Ethics Committee (Reference Number 036168). The approval letter is available in Appendix C.2.

7.3 Results

A total of 1,760 estimates of treatment coefficients were calculated across nine trials from 10 statistical methods. Each method produced 176 estimates of treatment effect for eight SF-36 dimensions across the nine RCTs with 22 post-randomisation timepoints in total. In general, in the three trials using SF-36 dimension scores as primary outcomes (i.e. Acupuncture, LM and PLINY), skewed distributions of residuals and heteroscedasticity were seen using SF-36 dimension scores post estimation of MLR and Tobit. For estimating treatment coefficients of the group difference in the SF-36 eight dimensions in these three trials, 6/24 and 5/24 of the estimation violated the proportional odds assumption when applying OL and the OP respectively. A summary of post-estimation plots for MLR and Tobit are shown in the Appendix C.3.

7.3.1 Estimated treatment coefficients under the scale-based estimand framework

Figure 7.1 shows the scatterplots of estimated treatment coefficients from the untransformed scale-based methods (i.e. Tobit, Median, and CLAD) against MLR and the transformed scale-based methods (i.e. OL, BLN, Frac, and BR) against BB. Estimates from statistical methods on the untransformed scale deviated from MLR. When the magnitude of the estimated treatment effects were large, they tended to produce higher estimates than MLR. Tobit especially produced much larger magnitude of estimates than MLR for LegUlcer and NAMEIT. CLAD and Median shared a similar pattern when the estimates were small, i.e. they tended to produce estimates scattering at zero and to take values that are multiplier of the difference between two neighbouring categorical values. Estimates from methods on the transformed scales are presented using logORs, except for OP. The logORs estimated from BLN and Frac were shown similar to BB. The OL produced higher absolute estimates than other methods, which was obvious in PLINY that presented averagely higher treatment estimates than other trials.

7.3.2 Estimated SESs under the SES estimand framework

Overall, the estimated SESs in our datasets were small (i.e. absolute value less than 0.2), except for the SESs of some dimensions being large or very large in PLINY (i.e. absolute value between 0.5 and 1.4). Median and CLAD failed to converge on one and 10 occasions respectively, especially under the scenarios for analysing SF-36 dimensions with less than 10 possible values.

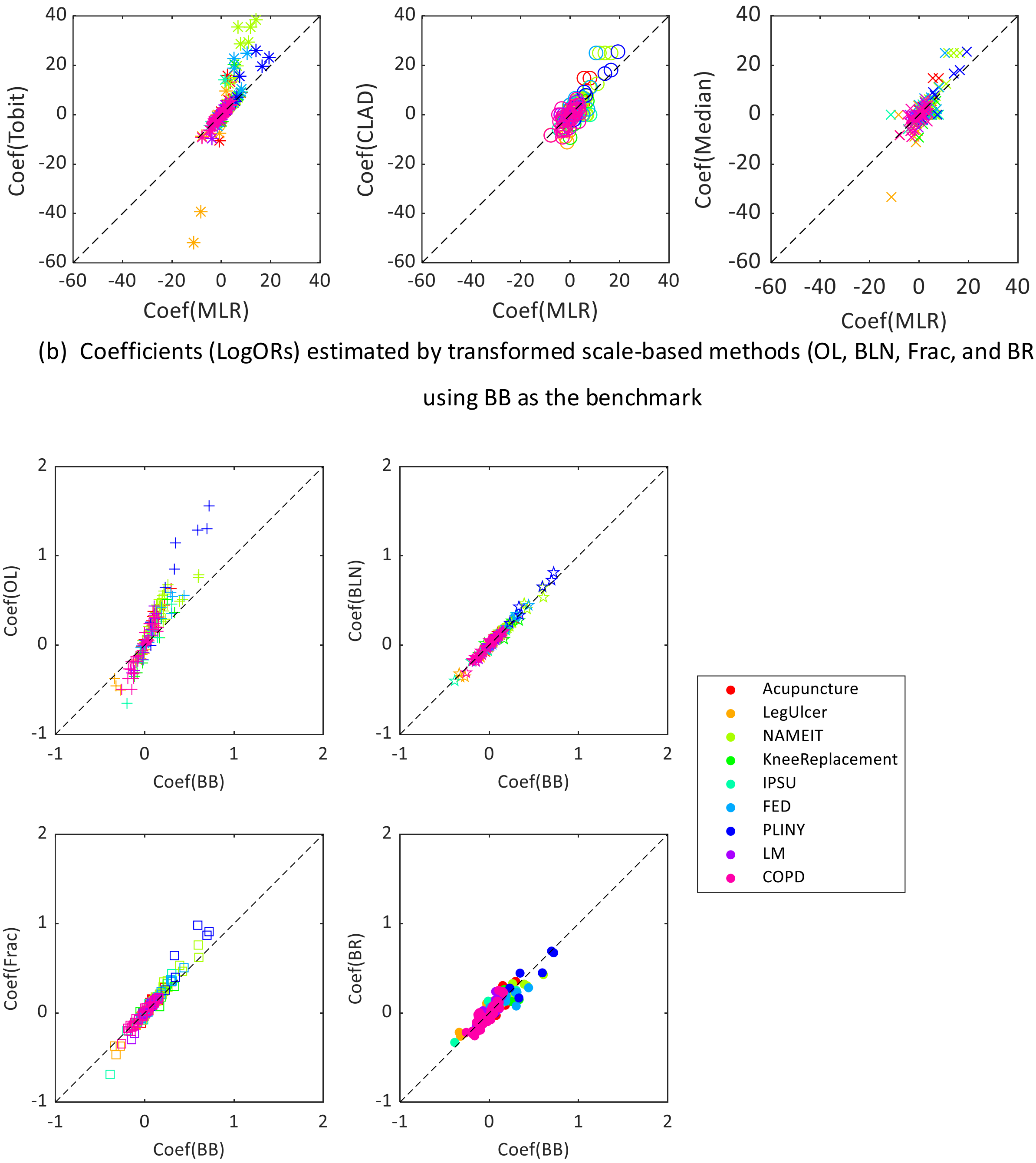
When estimating the treatment coefficient of the same response variable using different methods, SESs with different directions were produced from the 10 statistical methods, but there was no case where these methods produced statistically significant estimates with different directions. Moreover, SESs with the same direction may have different magnitude of effect size. We found that 16/176 analyses of the same response did not have similar magnitude of the estimated standardised effect from all 10 methods, i.e. 16 analyses of the same response estimated from these methods had statistically significant SESs in different ranges of effect size. This mainly appeared in the SESs estimated from BLN that tended to produce large and significant SES compared to other methods.

Figure 7.2 presents the SESs estimated from the 10 statistical methods against MLR across nine trials. For statistical methods that used the untransformed scale of measurement, Tobit has almost identical pattern against MLR, whereas both Median and CLAD tended to produce estimates close to zero and showed less consistency to MLR than Tobit does. OL that produced higher estimated treatment coefficients (i.e. logORs) showed similar SESs as other methods after standardisation. Conversely, although the BLN produced similar estimated coefficients (i.e. logORs) as BB, the SESs from BLN are larger than other methods after standardisation. The figure also shows that the Tobit, BB, OP, OL, and Frac have stronger agreement with MLR, i.e. the difference between each of these four methods against MLR is associated with less bias and narrower 95% CIs than rest of the methods.

Figure 7.3 shows the SESs with associated 95% CIs estimated from 10 statistical methods in three trials (Acupuncture, PLINY and LM) that used SF-36 dimension scores as primary outcomes. Two horizontal lines are drawn in the plot, representing the SES having no effect or no difference between two treatment groups (i.e. y = 0), and clinical significance (i.e. y = MCID/SD). When analysing the treatment effect of the same response, the use of different statistical methods may draw different results in terms of statistical significance and/or clinical significance. For example, SESs estimated from CLAD and Median are shown statistically significant for BP scores at 12-month post-randomisation in Acupuncture and for MH scores at 6-month post-randomisation in LM, whereas this is not the case for MLR and Tobit. In PLINY, most methods produced statistically significant estimates except for MLR, CLAD and Median. Additional effect size plots for the rest of the SF-36 dimension scores in three trials (i.e. COPD, LM, and PLINY) are generated and presented in Appendix C.3.

7.3.3 Change of model fit with the number of possible categorical values in SF-36 dimension scores

Figure 7.4 is a series of scatterplots on how the model fit, measured by the AIC statistic, changed when applying the 10 statistical methods to analyse different number of possible categorical values (levels) of dimension scores in SF-36, using data from the nine RCTs. When fitting SF-36 dimension score with higher levels, the AIC statistics of Tobit, ordinal, and binomial regression became larger, representing a poorer fit, whereas the AICs for the MLR became smaller, representing a better fit. The scatterplot shows that AIC values for Frac were less sensitive to the change in dimension levels.


(a) Coefficients (means or medians) estimated by untransformed scale-based methods (Tobit, CLAD, and Median) using MLR as the benchmark

(b) Coefficients (LogORs) estimated by transformed scale-based methods (OL, BLN, Frac, and BR) using BB as the benchmark

Figure 7.1 Estimated coefficients under the scale-based estimand framework using nine RCT datasets with SF-36 as clinical outcomes

BB, beta-binomial regression; BLN, binomial-logit-Normal regression; BR, beta regression; CLAD, censored absolute least deviations regression; Coef, treatment coefficient; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

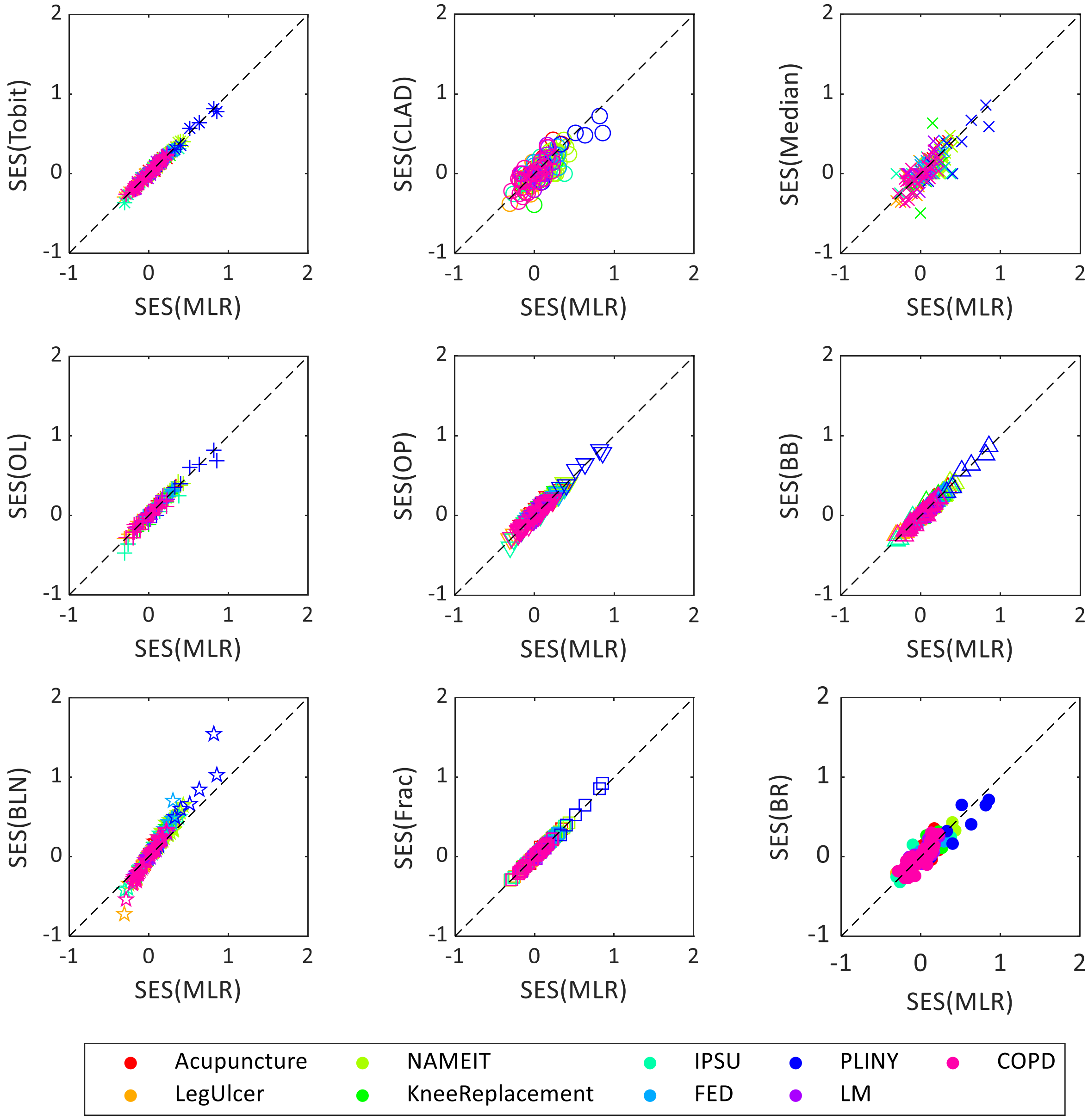
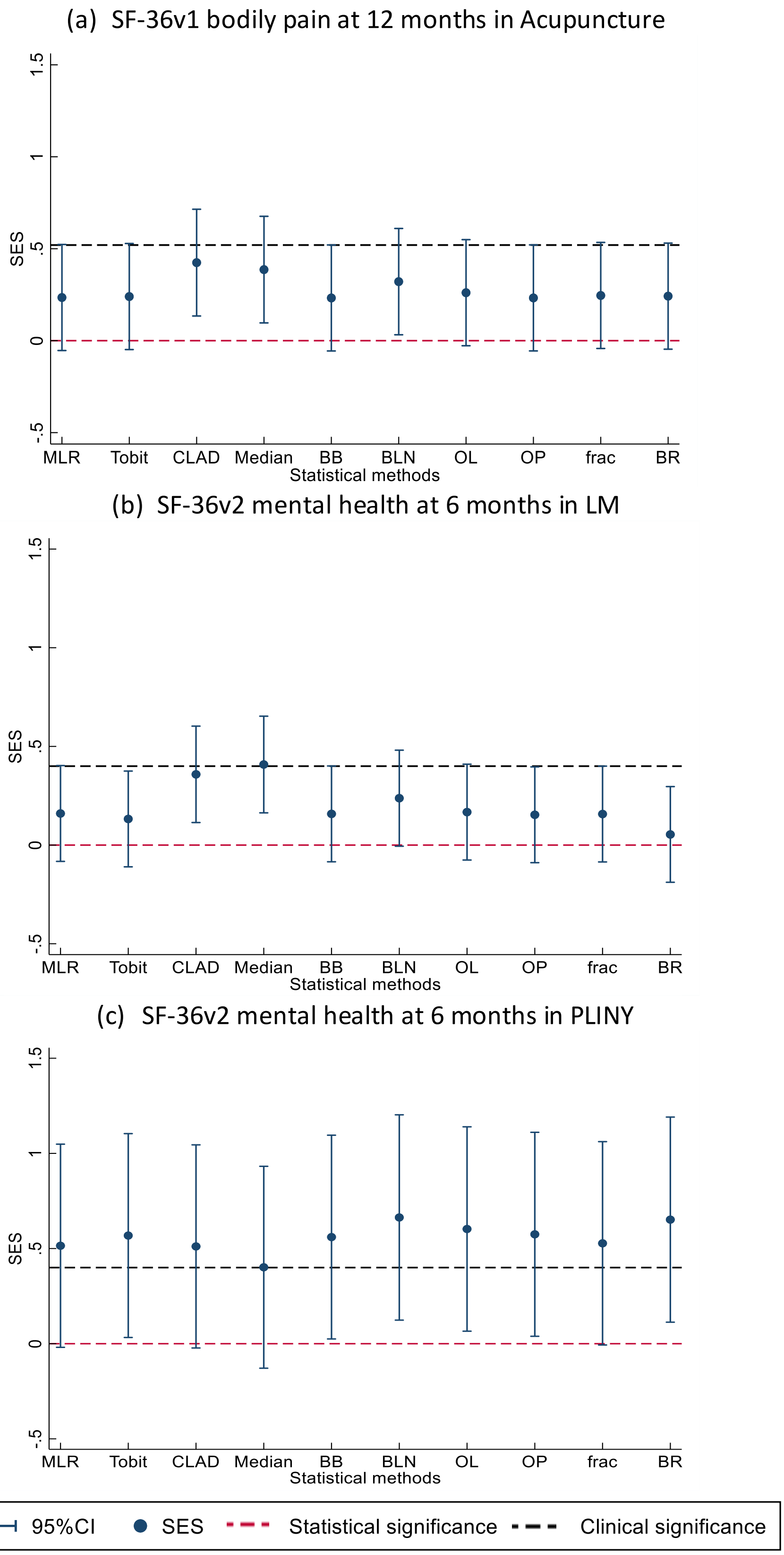


Figure 7.2 SES estimated from ten different statistical methods against MLR using nine RCT datasets with SF-36v2 as clinical outcomes

The x-axis of the scatterplots is the SES estimated by MLR, and the y-axis is the SES estimated by other statistical methods. The black dash line represents the method that produces the same standardised effect size as MLR.

BB, beta-binomial regression; BLN, binomial-logit-Normal regression; BR, beta regression; CLAD, censored least absolute deviations regression; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; OP, ordered probit model; Tobit, Tobit regression.


(a) SF-36v1 bodily pain at 12 months in Acupuncture

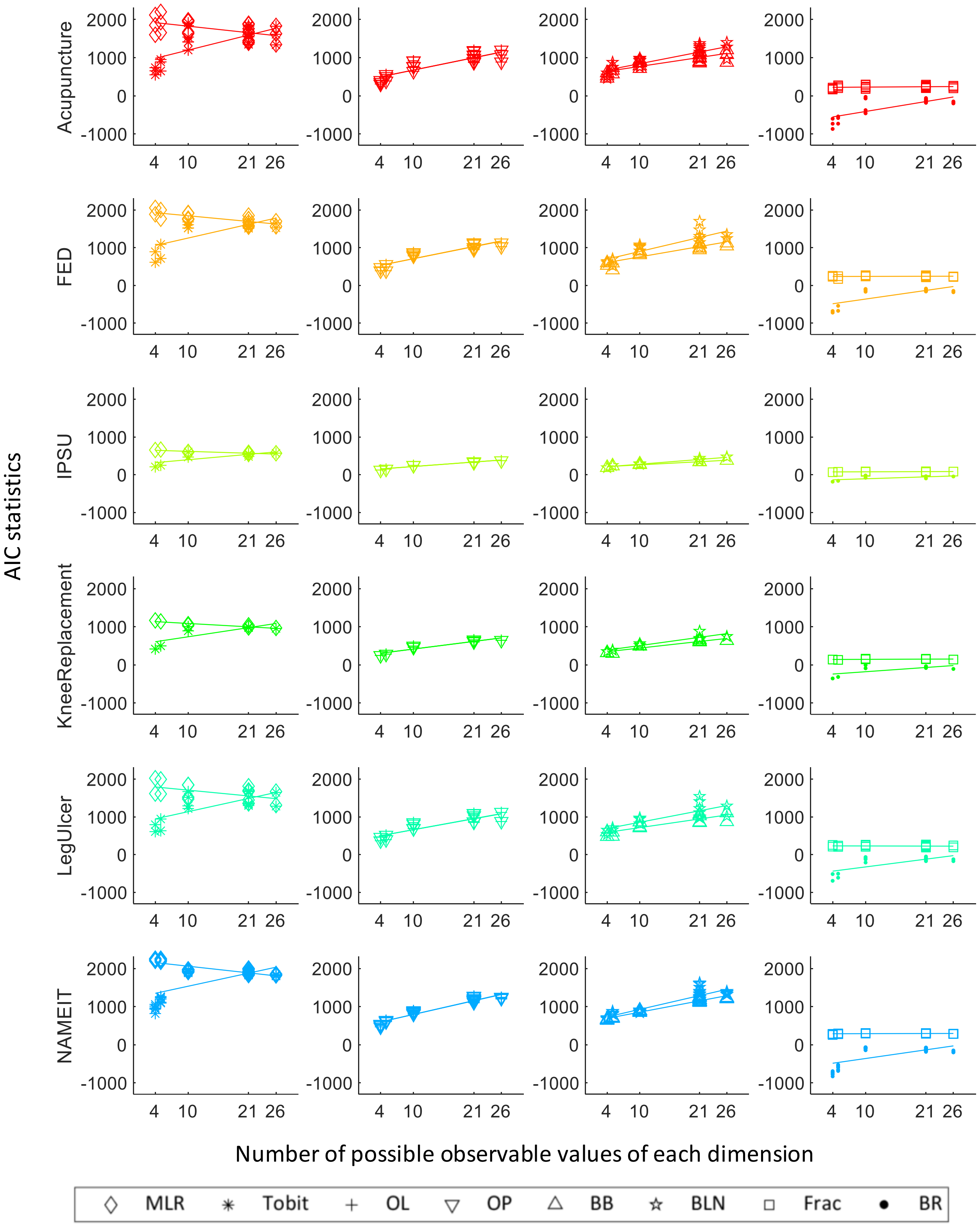
(b) SF-36v2 mental health at 6 months in LM

(c) SF-36v2 mental health at 6 months in PLINY

Figure 7.3 SES with 95% CIs from ten different statistical methods for Acupuncture, LM and PLINY that used SF-36 dimension scores as primary outcomes

BB, beta-binomial regression; BLN, binomial-logit-Normal regression; BR, beta regression; CLAD, censored least absolute deviations regression; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; OP, ordered probit model; Tobit, Tobit regression.

(a) SF-36 version 1



(b) SF-36 version 2

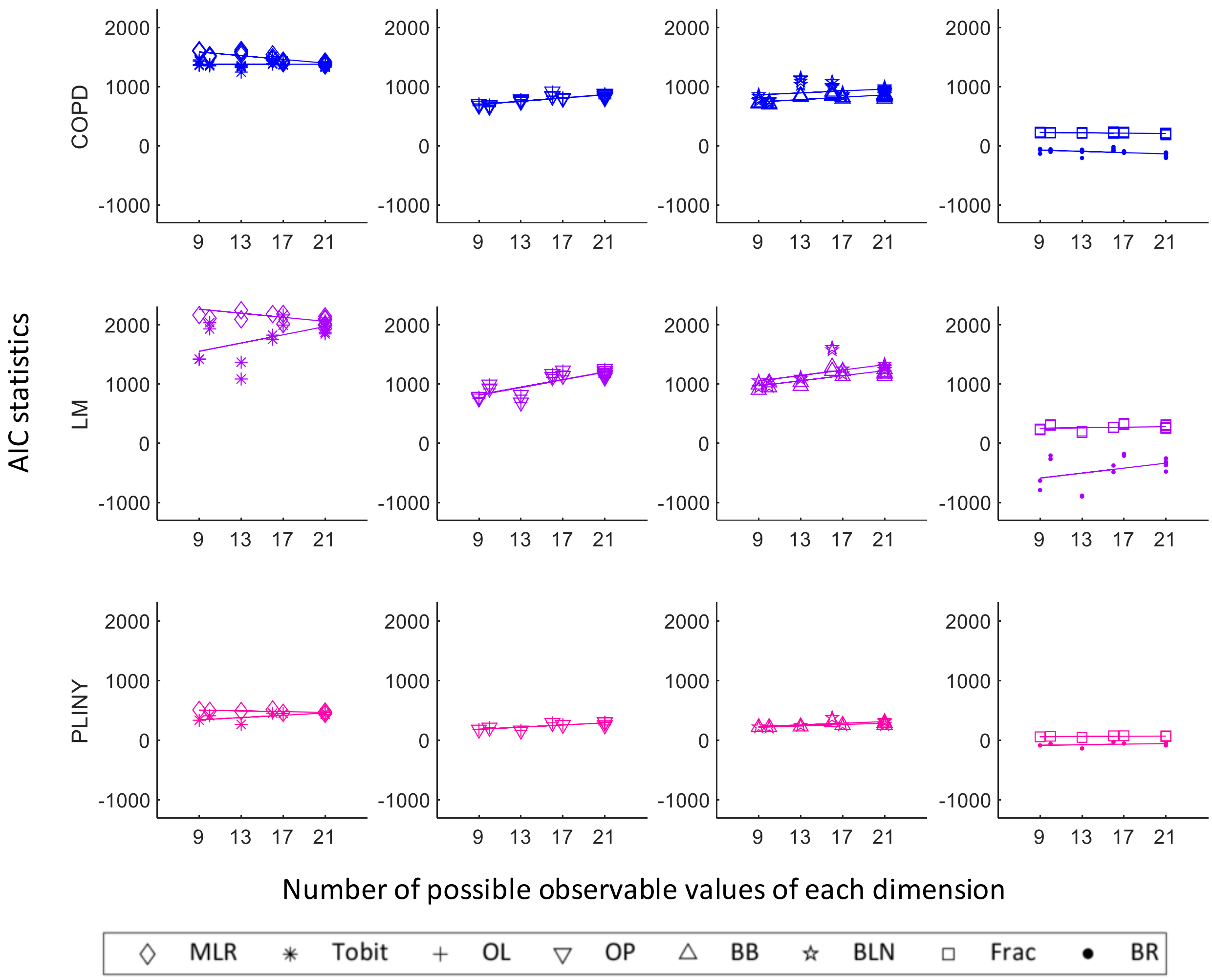


Figure 7.4 Scatterplot of AIC statistics for different statistical methods against the number of possible observable values of SF-36 dimensions in nine RCT datasets

AIC, Akaike information criterion; BB, beta-binomial regression; BLN, binomial-logit-Normal regression; BR, beta regression; CLAD, censored least absolute deviations regression; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; OP, ordered probit model; Tobit, Tobit regression. Note that Median and CLAD do not have AIC scores, and thus are not compared in this figure.

7.4 Discussion

This chapter applied 10 statistical methods for the analysis of PROs to nine RCT datasets in various clinical areas using both versions of the SF-36. A total of 1,760 estimates of treatment coefficients for SF-36 dimension scores were calculated across nine trials with 22 post-randomisation time-points using 10 methods: MLR, Median, Tobit, CLAD, BB, BLN, OL, OP, Frac, and BR.

Our empirical analysis shows that SESs estimated from different methods are generally consistent, using MLR as the reference benchmark, although the estimated treatment coefficients by different methods vary. For example, the magnitude of estimated treatment coefficients from Tobit is larger than MLR in some trials, but the estimated SESs from Tobit and MLR are almost identical. This may result from the fact that the Tobit accounts for the censoring or boundedness of the response variable and assumes the latent response variable having a wider scale than the observed response variable that is used for estimation by MLR. However, adjusting the SD of treatment estimates offsets the large magnitude of the estimated coefficients, and thus results in the agreement in SES between Tobit and MLR.

It is, therefore, possible for different combinations of values to produce the same effect size on the standardised scale (Cook *et al.*, 2014). An example for the transformed scale-based methods is OL, which generally produced higher estimates than other methods, resulting in higher estimated OR. However, the SEs estimated from OL were also higher than other methods, offsetting the high values of estimates when calculating the SESs. Conversely, BLN produced slightly higher SES estimates, whereas the treatment coefficients from it were similar to other methods. A possible explanation for large estimates from OL is that the estimated coefficients from OL stands for the probability that the PRO score, Y, is less than or equal to particular score or category l, i.e. $P(Y \leq l)$, while the estimated coefficients from other methods that produced logORs such as BB and Frac stands for the probability of the PRO score being the discrete value or category l, i.e. $P(Y = l)$ for the binomial and fractional regression methods. Naturally, the magnitude of the previous estimand is supposed to be higher than the latter. However, after the standardisation procedure the scale and meaning of the original estimand in each method does not matter, and the estimated SESs from these methods with different meanings in their original scale are shown similar.

CLAD, another method that can account for boundedness of the outcome, was found to be inefficient compared to other methods, as it took longer to run in Stata/MP 17.0 and failed to converge on some occasions. For statistical methods that produced estimates on the untransformed scale, quantile regressions (Median and CLAD) showed more variation. This may be because they adopt a different estimation method (i.e. LAD or CLAD) in comparison to those that use MLE. As was found in another study comparing Tobit, Median, and CLAD using the HUI (Austin, 2002), Median and CLAD tend to produce estimates with similar patterns and their estimates tend to be shrunk to zero compared to MLR and Tobit.

SESs from BR scattered more than Frac using MLR as the reference benchmark, MLR. This may be caused by the required ‘squeezing’ procedure in BR, which can reduce the estimation precision (Hunger, Baumert and Holle, 2011). The requirement on data distributed between 0 and 1 by fractional regression methods makes them more seemingly suitable for the analysis of health utility scores than other included statistical methods. However, it is worth noting that, in scenarios where health utilities index scatter on slightly different scales, e.g. SF-6D scattering between 0.291 and 1 for the UK value set (Brazier, Roberts and Deverill, 2002), the two fractional regression methods may not be straightforward (Hunger, Baumert and Holle, 2011; Kharroubi, 2020).

Generally, when increasing the number of possible observable values, the AIC for statistical methods with logit or probit link increased; it decreased, however, for MLR. Interestingly, Tobit, an extension of MLR designed to adapt for censored outcomes, generated lower AIC values (i.e. better model fit) when analysing outcomes with a small number of possible observable values. This shows an adverse trend compared to MLR and requires further investigation.

Regarding the results of this empirical analysis, the following six statistical methods are carried forward for the simulation analysis.

1. Multiple linear regression (MLR)
2. Median regression (Median)
3. Tobit regression (Tobit)
4. Ordered logit model (OL)
5. Beta-binomial regression (OP)
6. Fractional logistic regression (Frac)

The following summary of rationales for why other statistical methods are excluded are covered in previous paragraphs in this discussion and presented below:

- CLAD regression (CLAD): CLAD is not shown efficient in this empirical analysis, given it takes comparatively long time to run, it requires bootstrapping to generate CIs, and it does not provide p-values directly. Moreover, it is a user-built code in both Stata and R, and it is found not to converge on some occasions even when increasing the number of iterations.
- Ordered probit model (OP): OP produces almost identical SES estimates and model fit statistics to ordered logit model, but its estimated treatment effect cannot be explained in (log)ORs as OL does which decreases the interpretability (or the clinical relevance) of OP.
- Binomial-logit-Normal regression (BLN): BLN produces slightly higher SES estimates than other methods, but its estimated treatment coefficients are similar to other methods. It has

similar model assumptions except for the distribution of the random variable, but in comparison to BB, the command for running BLN in the computational software is constructed using the general code under the GLM framework since there is no established and validated commands available for running BLN.

- Beta regression (BR): Unlike Frac, BR cannot account for scores at boundaries, i.e. values at 0 or 1, and therefore requires the 'squeezing' process to recode dimension scores, the compressing process of which is likely to bias the estimations and reduce the precision.

This empirical analysis has the following limitations:

First, we included nine trials that focused on different disease areas and populations, which can be seen as a source of heterogeneity. However, this study does not intend to compare the size of treatment estimates across different trials but to compare whether different statistical methods can produce similar estimates under two estimand frameworks. Therefore, the results of this chapter should not be influenced by the magnitude of effect sizes and heterogeneity in trials.

Second, this chapter focused on dimension scores in SF-36v1 and SF-36v2, and extrapolation to other versions of SF-36 and other types of PROs may require further validation. However, the SF-36 is a widely used generic PRO which shares similar data features (i.e. discrete, bounded, and skewed) with other PROs, and it may be more prone to ceiling effects or less responsive to subtle changes in some dimensions that are not targeted than a disease-specific measure.

Third, the regression models were kept in simple and similar forms, i.e. only the treatment group and the baseline score of the corresponding dimension were included as independent variables, since the aim of this chapter is to compare different statistical methods but not to identify or determine the best model. Other potential effects such as time and clustering, i.e. hospital sites or centres, were not considered. As the majority of the statistical methods included are under the GLM framework, they can be extended for longitudinal analysis by using GLMM with coefficients estimated by MLE or GLM with coefficient estimated by GEE (Walters, 2009).

Fourth, our empirical analysis was based on the real case data such that the 'truth' of the treatment effect is unknown (Morris, White and Crowther, 2019; Boulesteix *et al.*, 2020). Therefore, we were not able to evaluate which statistical methods have less bias than other methods using results from this empirical analysis. Using real data to compare statistical methods in this chapter can show how robust the methods could be when applied to real case data. However, it still needs further investigation on how close the estimates produced by these methods are to the predefined 'truth', and which method remains robust when analysing different dimension scores of the SF-36 and when model assumptions are violated.

The next chapter will develop a protocol for simulation analysis to evaluate the narrowed list of six statistical methods in terms of their estimation accuracy and model robustness in different scenarios.

Chapter 8 Protocol of simulation analysis of statistical methods for the analysis of PROs in RCT settings

8.1 Introduction

In the previous chapter (Chapter 7), a series of empirical analyses were conducted by applying identified statistical methods to RCT datasets that used PROs as clinical outcomes. However, as the 'truth' is unknown for real world datasets (Boulesteix *et al.*, 2020), it still needs further investigation on how close the estimates produced by these methods are to the predefined 'truth', and whether the performance of these methods remain robust when analysing different dimension scores of PROs and when model assumptions are violated. Therefore, simulation analysis is needed to evaluate the statistical methods in terms of their accuracy and robustness using a number of performance measures such as bias, coverage of the 95% CIs, Type I error rate and power. This chapter establishes a simulation protocol to guide the simulation analysis in the next chapter. The five key steps proposed by Morris et al (Morris, White and Crowther, 2019), including aims, data-generating mechanism (DGM), estimands, methods, and performance measure (ADEMP), are adapted to develop this protocol.

8.2 Aims

This simulation study aims to evaluate whether the estimators (i.e. statistical methods) can estimate the predefined 'truth', i.e. compare the performance of the statistical methods in estimating the treatment effect under a range of scenarios including no treatment effect and different values of true treatment effects, using a number of performance measures. The statistical methods that were narrowed down from the previous chapters include MLR, Tobit, Median, Frac, OL, and BB.

8.3 Data-generating mechanism

Multiple DGMs are proposed to ensure the coverage of different scenarios, by varying the number of observations and predefined treatment difference, such that each DGM provides us with empirical results for a specific scenario (Morris, White and Crowther, 2019). Monte Carlo methods will be applied to generate data from pre-specified distributions.

The SF-36 is used as the representative of PROs as it is found as the most used PROs in publicly funded RCTs in the UK (Qian *et al.*, 2021). Because the layout and scoring of SF-36v1 and SF-36v2 partly overlapped with each other and SF-36v1 contains more variability in the number of possible values that a dimension score can have, the distribution of the SF-36v1 dimension scores that were collected in our RCTs (as described in Chapter 7) are used as the real-world dataset to guide the construction of DGMs.

The eight dimension scores in SF-36v1 have a different number of possible categorical values after discretisation. We aim to apply a similar DGM to construct simulated datasets for three different dimensions with a range of ordinal categorical values i.e. role limitation - emotional (RE) ($n = 4$), bodily pain (BP) ($n = 10$), and mental health (MH) ($n = 26$), where n is the number of possible ordinal categorical values. BP is selected because it was used as the primary outcome in the Acupuncture trial (Thomas *et al.*, 2006), and RE and MH are selected because they are the dimensions with the lowest and highest number of possible ordinal categorical values in SF-36v1. The possible ordinal categorical values of these three dimensions are presented in Table 8.1.

Table 8.1 Possible ordinal categorical values of the three dimensions in SF-36v1 (RE, BP, MH)

Dimensions	Possible ordinal categorical values
RE ($n = 4$)	0, 33.3, 66.6, 100
BP ($n = 10$)	0, 11.1, 22.2, 33.3, 44.4, 55.6, 66.7, 77.8, 88.9, 100
MH ($n = 26$)	0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, 60, 64, 68, 72, 76, 80, 84, 88, 92, 96, 100

n represents the number of possible ordinal categorical values of a dimension score. BP, bodily pain; MH, mental health; RE, role limitation - emotional.

We also plan to adapt a variation of predefined treatment difference to see whether these methods may fail when having different location shifts. This is to be conducted by adding different magnitudes of predefined values to the observations in the treatment group, and then conduct the analysis using these methods to measure whether they are able to detect the predefined 'truth'. As these statistical methods are developed to solve different problems, results on a better performance model may vary according to different scenarios.

Since one DGM may favour certain methods over others, we consider different combination of parameter values to observe if a method can cover a wide spectrum of potentially plausible situations. The dilemma of this is finding the appropriate distribution with predefined parameters that can depict the distribution of SF-36 dimension scores, which is typically bounded, skewed, and ordinal.

The Normal distribution, denoted by $Normal(\mu, \sigma^2)$, is proposed to generate the SF-36 dimension scores. Of the three parameters, treatment difference (θ) is varied to test the performance of six included statistical methods under multiple scenarios, while mean (μ) and SD (σ) are kept the same. We use random sample generated under $Normal(50, 22^2)$ for the base-case DGM, using evidence from our Acupuncture trial (Thomas *et al.*, 2006) where the average mean score of the SF-36v1 BP score for the control group at the primary endpoint (i.e. 12-month post-randomisation) is 58, with a SD of 22. The distribution of the SF-36v1 BP score at 12-month post-randomisation is shown in Figure 8.1.

As the statistical methods to be evaluated in this simulation produce estimates on different scales, the SES is used to compare the group difference produced by the different statistical methods. To simulate different magnitudes of the effect size, we derived the value of 0, 0.2, 0.5, 0.8, and 1.0 for the SES from different degrees of the Cohen's classified effect size (Sawilowsky, 2009; Cohen, 2013), representing no effect, small effect, median effect, large effect, and very large effect. With the value of SD fixed at 22, the treatment difference (d) is set at 0, 4.4, 11, 17.6, and 22 respectively. Table 8.2 presents the parameter specification using Normal distribution. The same set of parameters are used for all the three levels in SF-36v1 (i.e. RE, BP, and MH). The seeds and streams for the random-number generator are set the same to ensure that the exact same set of Normal distributions are used to produce scores for these three dimensions.

An issue of generating data from the Normal distribution is that the simulated dimension scores may go beyond the boundaries of 0 and 100 for the SF-36 BP score. The following strategy is used to deal with this issue. Firstly, simulated scores exceeding the lower and upper bounds will be rounded to the values at boundaries, i.e. 0 or 100, and secondly the values on the continuous scale are discretised onto a categorical scale between 0 and 100. The discretisation techniques are determined by the number of possible ordinal categorical values of the simulated dimension score. For example, RE scores with four possible values will be discretised into 0, 33.3, 66.6, or 100. The discretisation techniques for these dimension scores in this simulation analysis are shown in Table 8.3.

When adding a positive value to the treatment group, the discretisation techniques will underestimate the treatment difference if the sample is right bounded (i.e. censored to the upper bound at 100), and overestimate the difference if the sample is left bounded (i.e. censored to the lower bound at 0). However, we can never know the exact 'truth' after discretising the sample, except when the pre-specified group difference is set at zero, i.e. simulation under the null hypothesis. Thus, the pre-specified treatment difference to generate the sample before discretisation, and the observed treatment difference for the generated sample after discretisation will be recorded and presented.

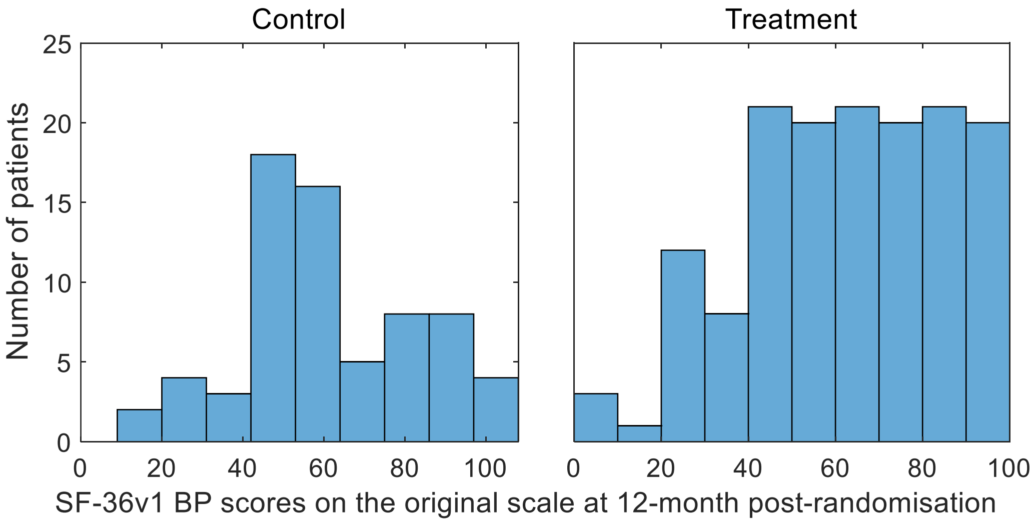


Figure 8.1 Distribution of the SF-36v1 bodily pain scores at 12-month post-randomisation from the Acupuncture trial

Table 8.2 Parameter specification for five DGMs using Normal distribution

DGM Number	Normal distribution generator (control group)		Additional value add to treatment group	
	Mean (μ)	SD (σ)	Group difference (d)	SES
1	50	22	0.0	0.0
2			4.4	0.2
3			11.0	0.5
4			17.6	0.8
5			22.0	1.0

DGM, data-generating mechanism; SD, standard deviation; SES, standardised effect size.

Table 8.3 Discretisation techniques for the three dimensions in SF-36v1 (RE, BP, MH)

RE ($n = 4$)			BP ($n = 10$)			MH ($n = 26$)		
Possible scores	Discretisation techniques		Possible scores	Discretisation techniques		Possible scores	Discretisation techniques	
0	$(-\infty,$	16.65]	0	$(-\infty,$	5.55]	0	$(-\infty,$	2]
33.3	(16.65,	49.95]	11.1	(5.55,	16.65]	4	(2,	6]
66.6	(49.95,	83.25]	22.2	(16.65,	27.75]	8	(6,	10]
100	(83.25,	$+\infty)$	33.3	(27.75,	38.85]	12	(10,	14]
			44.4	(38.85,	49.95]	16	(14,	18]
			55.6	(49.95,	61.05]	20	(18,	22]
			66.7	(61.05,	72.15]	24	(22,	26]
			77.8	(72.15,	83.25]	28	(26,	30]
			88.9	(83.25,	94.35]	32	(30,	34]
			100	(94.35,	$+\infty)$	36	(34,	38]
						40	(38,	42]
						44	(42,	46]
						48	(46,	50]
						52	(50,	54]
						56	(54,	58]
						60	(58,	62]
						64	(62,	66]
						68	(66,	70]
						72	(70,	74]
						76	(74,	78]
						80	(78,	82]
						84	(82,	86]
						88	(86,	90]
						92	(90,	94]
						96	(94,	98]
						100	(98,	$+\infty)$

n represents the number of possible ordinal categorical values of a dimension score. BP, bodily pain; MH, mental health; RE, role limitation - emotional.

8.3.1 Number of observations of a simulated dataset

The number of observations (n_{obs}) (i.e. sample size) of each simulated dataset is determined using evidence from the 114 identified HTA trials that used PROs as primary clinical outcomes (Qian *et al.*, 2021). The overall sample size of these HTA trials ranges from 65 up to 7677, with the 5th, 50th, and 95th percentile being 102, 387, and 1,084 respectively. Given the right skewed distribution of their sample size, we decide to use 1600 as the maximum number of observations, and 100 as the minimum number of observations in each simulated dataset. The number of observations of the simulated dataset is set at 100, 200, 400, 800, 1,200, and 1,600 under the null and alternative hypothesis. For each simulated dataset, half of the sample is assigned to the treatment group and half to the control group.

8.3.2 Number of repetitions

The number of repetitions (n_{sim}) is set at 5,000 for each scenario under both the null hypothesis and alternative hypothesis. The justification for the number of repetitions is described in Section 8.6.2.

8.4 Estimands

An estimand is a clear and explicit description of precisely what treatment effect is to be estimated in an RCT. It is made up of five connected attributes: the population, the treatments (you want to compare), the outcome or endpoint, how to account for intercurrent events and a population-level summary measures of how the outcome between the different treatment conditions will be compared.

For the simulations four of the elements (population, treatments, outcomes, how to account for intercurrent events were unchanged, but the fifth the population-level summary measure of the how the outcome between the different treatment conditions will be compared was changed for some of the simulations. We adapt two estimand frameworks in this simulation study to compare estimates from the statistical methods that produce different types of outcomes, e.g. the difference in means from the MLR and OR from the ordered logit model.

The first framework is called the scale-based estimand framework. The population-level summary measure of how the outcome between the different treatment groups used the mean or median in the treatment difference (θ) for methods that produce estimates on the untransformed scale, and used the logOR of the treatment difference for the population summary measure for methods that produce estimates on a transformed scale.

The second framework is called the SES estimand framework, which unifies the estimated treatment difference from different methods using a standardisation procedure. The SES and its associated standard error , denoted by $SE(SES)$, based on the Normal approximation of non-central *t*-distribution (Hedges, 1981) has been defined in Chapter 6 using the following formula.

$$SES = Z \times \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \frac{coef(TE)}{SE(TE)} \times \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{8.1}$$

$$SE(SES) = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{SES^2}{2(n_1+n_2)}} \tag{8.2}$$

where Z stands for the Z-statistics; TE stands for the treatment effect; $coef(TE)$ stands for the estimated values for the treatment effect parameter; $SE(TE)$ stands for the SE of the treatment effect estimates; $SE(SES)$ represents the SE of the SES; and n_1 and n_2 represents the number of observations (i.e. sample size) in each treatment group respectively.

8.5 Methods of analysis

Given the results from our previous empirical analyses in Chapter 7, six statistical methods are taken forward for this simulation analysis. These methods are categorised according to the type of estimates that they produce:

- Untransformed scale-based methods:
  - Multiple linear regression (MLR)
  - Tobit regression (Tobit)
  - Median regression (Median)
- Transformed scale-based methods:
  - Fractional logistic regression (Frac)
  - Ordered logit model (OL)
  - Beta-binomial regression (BB)

Since the DGMs are simulated under the RCT settings, no other factors are introduced to influence the treatment difference. Therefore, these methods are applied to the simulated datasets by estimating the treatment effect without adjusting for covariates. The statistical package Stata/MP 17.0 is used for simulation analysis and MATLAB R2023a is used for data visualisation, and the Stata codes for conducting the simulation analysis are available in Appendix D.2.

8.6 Performance measures

The performance measures of interest are listed below. The estimated treatment coefficient from each statistical method is denoted using $\hat{\theta}_{i,}$ and the pre-specified group difference is denoted using θ. For untransformed scale-based methods (i.e. MLR, Tobit, and Median), θ is the predefined treatment difference in mean or median. For transformed scale-based methods (i.e. Frac, OL, and BB), θ is supposed to be the logOR equivalence of the effect size, which is not available in this study. Under the null hypothesis, theta equals zero for all included statistical methods.

8.6.1 Estimation of performance

The main performance measures are bias, coverage of 95% CI, power and Type I error. Other performance measures listed below are considered and compared jointly in the simulation analysis. Table 8.4 describes the notation that is kept consistent in the following chapters.

Table 8.4 Description of notation for the simulation analysis

Notation	Description
x	An estimand, and also the true value of the estimand, which is denoted as θ for MLR, Tobit, and Median, and as logOR for Frac, OL, and BB under the scale-based estimand framework; Under the SES estimand framework, it is denoted as SES for all methods.
n_{obs}	Sample size of a simulated dataset.
n_{sim}	Number of repetition used; the simulated sample size.
$i = 1, \ldots, n_{sim}$	Indexes the repetitions of the simulation.
$\hat{x}$	The estimation of x.
$\hat{x}_i$	The estimate of x from the ith repetition.
$\bar{x}$	The mean of $\hat{x}_i$ across repetition.
$Var(\hat{x})$	The true variance of $\hat{x}$, which can be estimated with large n_{sim}.
$Var(\hat{x}_i)$	An estimated of $Var(\hat{x})$ from the ith repetition.
α	The nominal significance level.
p_i	The p-value returned by the ith repetition.

BB, beta-binomial regression; Frac, fractional logistic regression; logOR, log odds ratio; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

Bias ($\boldsymbol{\delta}$) is defined as the average difference between the estimated values and the predefined 'truth'. Due to the discretisation techniques attached to the DGMs, we can never know the exact 'truth', except when the pre-specified group difference is zero. Different biases are generated for the two estimand frameworks: the first is the bias in means or medians and logORs under the scale-based estimand framework, denoted as δ_θ for MLR, Tobit, and Median and as δ_{OR} for Frac, OL, and BB; and the second is the bias in SES under the SES estimand framework, denoted as δ_{SES} for all methods.

$$\delta_x = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{x}_i - x \tag{8.3}$$

where x represents the predefined 'truth' of the target estimand. It is defined as θ for MLR, Tobit, and Median, and defined as logOR for Frac, OL, and BB under the scale-based estimand framework; and defined as SES under the SES estimand framework. $\hat{x}_i$ denotes the estimated values from each method; and n_{sim} is the number of repetitions of the simulation study.

Empirical standard error ($\boldsymbol{EmpSE_x}$) is the precision measure of the estimated values to the average ($\bar{x}$) for each method. The predefined 'truth' (x) is not required to generate this measure. **Relative % increase in precision (B vs A)** compares the precision between different methods using EmpSE.

$$EmpSE_x = \sqrt{\frac{1}{n_{sim}-1}\sum_{i=1}^{n_{sim}}(\hat{x}_i - \bar{x})^2} \tag{8.4}$$

$$Precision(B\ vs\ A) = 100\left(\left(\frac{EmpSE_x(A)}{EmpSE_x(B)}\right)^2 - 1\right) \tag{8.5}$$

Mean squared error ($\boldsymbol{MSE}$) is calculated as the sum of the squared bias and variance of $\hat{x}_i$, and it is reported to be more sensitive to the number of observations than bias or EmpSE (Morris, White and Crowther, 2019).

$$MSE_x = \frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}(\hat{x}_i - x)^2 \tag{8.6}$$

Average Model SE ($\boldsymbol{ModSE}$) is the root of the average squared model SEs. **The relative % error in ModSE** measures whether the ModSE is overestimated or underestimated in comparison to EmpSE.

$$ModSE_x = \sqrt{\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\widehat{Var}\,(\hat{x}_i)} \tag{8.7}$$

$$Relative\ \%error\ in\ ModSE = 100\left(\frac{ModSE}{EmpSE} - 1\right) \tag{8.8}$$

Coverage of CI is defined as the probability that a CI contains the predefined 'truth' (x). For a $(1-\alpha)$ CI, the coverage is expected to be exactly $(1-\alpha)$ of the intervals containing x, else it is regarded as under- or over-coverage. Coverage for each DGM and statistical methods will be estimated as:

$$Coverage_x = \frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}1\left(\hat{x}_{low,i} \le x \le \hat{x}_{upp,i}\right) \tag{8.9}$$

Type I error is defined as the 'false positive rate', i.e. the probability to falsely reject the true null hypothesis that $x = 0$, using an α significance level of 0.05 or 5%. **Power**, denoted as $Power_x$, is defined as the 'true positive rate', i.e. the probability to correctly reject the null hypothesis, when the alternative hypothesis is true.

$$Power_x = \frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}1(p_i \le \alpha) \tag{8.10}$$

8.6.2 Justification for the number of repetitions

Three key performance measures for the simulation study are the accuracy measured by bias, precision measured by coverage of 95% CI, and the Type I error (under the null hypothesis scenarios) and power (under the alternative hypothesis scenarios). Monte Carlo standard error (SE) which quantifies simulation uncertainty and estimates the SE of a certain performance measure is used to justify the number of repetitions (Morris, White and Crowther, 2019).

Assuming we are going to calculate a 95% CI for each of our estimates $\hat{x}_i$, then the coverage of the 95% CIs for these estimates to be close to the nominal 95% or 0.95 level, i.e. for 95 out of the 100 simulated datasets and statistical methods, the CI should include the true value of the treatment effect parameter x. Therefore, assuming the coverage is close to 95% or 0.95, with 5,000 simulations, the Monte Carlo SE of the estimate would be around 0.0031 leading to approximate 95% CI for the coverage of 0.944 to 0.956, which we believe is a sufficient level of precision for the coverage performance measure (Table 8.5). Similarly, with 5,000 simulations, the Monte Carlo SE of the estimated Type I error (assuming the Type I error is close to 0.05) will also be around 0.0031, leading to approximate 95% CI for the Type I error of 0.044 to 0.056, which we believe is a sufficient level of precision for the Type I error performance measure (Table 8.5).

8.6.3 Graphs to present for visualisation

Exploratory analysis will be carried out, mainly by graphs for each DGM, estimand, number of iterations, and method (Morris, White and Crowther, 2019). The following graphs and tables will be generated for each level under five DGMs to visualise the results:

1. A summary table of the predefined parameter values and the observed parameter values.
2. An example of distributions under the five DGMs, with predefined and observed treatment differences marked.
3. A summary table of the estimated treatment difference by six statistical methods.
4. A summary table of the number and percentage of missing values of estimated treatment coefficient from each statistical method.
5. Distributions of estimated treatment coefficients and SESs.
6. Scatterplots of the estimated treatment coefficients and associated SEs for each method.
7. Scatterplots of the estimated treatment coefficients and SESs for one method vs another.
8. Line plots of bias, MSE, coverage, Type I error, and power under the two estimand frameworks, and additional line plots of EmpSE, ModSE, and relative % error in ModSE under the scale-based estimand framework, against the change of sample sizes for different levels, with predefined ‘truth’ marked.

8.7 Summary

This chapter provides a practical plan for conducting simulation analysis to compare the accuracy and robustness of six statistical methods i.e. MLR, Tobit, Median, Frac, OL, and BB that were narrowed down from Chapter 7. Multiple scenarios for different types of PRO data i.e. with 4, 10, and 26 possible ordinal categorical values will be simulated using random-data generator from Normal distribution, and rescored by discretisation techniques. Two estimand frameworks are proposed to establish the target estimands in this simulation analysis. Various performance measures will be considered jointly to compare and contrast these statistical methods. Following this protocol, the next chapter will present the results of this proposed simulation analysis.

Table 8.5 Monte Carlo SE of coverage and power with different number of simulations

(a) Coverage

n_{sim}	Coverage				
	Coverage	1-coverage	Monte Carlo SE (Coverage)	Approx 95%CI	
				-2SE	+2SE
1,000	0.95	0.05	0.0069	0.936	0.964
2,000	0.95	0.05	0.0049	0.940	0.960
3,000	0.95	0.05	0.0040	0.942	0.958
4,000	0.95	0.05	0.0034	0.943	0.957
5,000	0.95	0.05	0.0031	0.944	0.956
6,000	0.95	0.05	0.0028	0.944	0.956
7,000	0.95	0.05	0.0026	0.945	0.955
8,000	0.95	0.05	0.0024	0.945	0.955
9,000	0.95	0.05	0.0023	0.945	0.955
10,000	0.95	0.05	0.0022	0.946	0.954

(b) Power

n_{sim}	Power				
	Power	1-power	Monte Carlo SE (Power)	Approx 95%CI	
				-2SE	+2SE
1,000	0.05	0.95	0.0069	0.036	0.064
2,000	0.05	0.95	0.0049	0.040	0.060
3,000	0.05	0.95	0.0040	0.042	0.058
4,000	0.05	0.95	0.0034	0.043	0.057
5,000	0.05	0.95	0.0031	0.044	0.056
6,000	0.05	0.95	0.0028	0.044	0.056
7,000	0.05	0.95	0.0026	0.045	0.055
8,000	0.05	0.95	0.0024	0.045	0.055
9,000	0.05	0.95	0.0023	0.045	0.055
10,000	0.05	0.95	0.0022	0.046	0.054

Chapter 9 Results of simulation analysis of statistical methods for the analysis of PROs in RCT settings

9.1 Introduction

Following the simulation protocol in Chapter 8, the simulation analysis is conducted to evaluate the performance of the six statistical methods (i.e. MLR, Tobit, Median, Frac, OL, and BB) for the analysis of PROs in RCT settings. This chapter reports the characteristics of the simulated datasets under five proposed DGMs, and evaluates the performance measures of the six statistical methods for the analysis of PROs with different ordinal categorical values (levels) under each DGM.

9.2 Characteristics of the simulated datasets

Following the five proposed DGMs, each simulation generated one simulated dataset assuming an underlying latent Normal distribution to randomly generate the outcome. The base case (DGM 1) has a mean of 50 and SD of 22. The simulated latent Normally distributed PRO scores were then 'discretised' into an outcome with a discrete number of levels or scores (e.g. 4, 10 and 26 levels). Under each level, the mean PRO score in the control group was the same for all DGMs, but the mean PRO score in the treatment group varied by the five predefined treatment differences (i.e. SES of 0, 0.2, 0.5, 0.8, and 1.0). With 5,000 simulations and six sets of sample sizes (n_{obs}) per simulation (i.e. 100, 200, 400, 800, 1,200, and 1,600), a total number of (5,000 simulations × 6 n_{obs}) = 30,000 simulated datasets were produced under each level, resulting in (30,000 simulated datasets × 6 methods × 5 DGMs × 3 levels) = 2,700,000 estimates in total (n_{total}).

Table 9.1 shows the parameter specification under the Normal distribution, the observed means in the control and treatment groups, and the observed treatment difference after applying the discretisation techniques, taking the average of the six n_{obs}. A more detailed table on the parameters for each n_{obs} is available in the Appendix D (Table D.1). Figure 9.1 presents example distributions of five DGMs for different levels using scores generated from the same latent Normal distribution, with the observed average treatment difference and SDs marked. The example dataset used 400 observations since it is the closest to the median sample size of the 114 trials in the review on statistical methods that were used for analysing PROs published in the HTA Journal in Chapter 3 (Qian *et al.*, 2021).

The definition of the estimand framework is outlined in Chapter 6 and Chapter 8. Two estimand frameworks, for the population-level summary measure of treatment effect, are proposed for this simulation. The first is called the scale-based estimand framework that has two categories according to whether the treatment effect estimates from these six methods are in the form of differences in group means or medians (MLR, Tobit, and Median) or logORs (Frac, OL, and BB). The second is called the SES estimand framework that uses the SES to unify the estimates of these six methods by dividing the estimate of the treatment effect by their associated SDs. This allows the comparison across these six methods. Table 9.2 provides the mean estimates from six different statistical methods under the five DGMs for each level using all observed estimates ($n_{total} = 2{,}700{,}000$), taking the average of the six n_{obs}. A more detailed table on the parameters for each n_{obs} are available in the Appendix D (Table D.2). Estimates for these methods under each DGM tend to decrease with an increase in the number of possible ordinal categorical scores, except for Median and BB. For example, under DGM 5 where the treatment difference is predefined as 22-point on the original latent Normally distributed PRO scale, the average treatment difference in MLR is 21.3 for level 4, 21.0 for level 10, and 21.0 for level 26. The average logOR estimates from OL are approximately two times higher than the logOR estimates from other methods.

Table 9.3 summarises the missing values due to non-convergence for the six statistical methods under the five DGMs, taking the average of the six n_{obs}. A more detailed table on the parameters for each n_{obs} are available in the Appendix D (Table D.3). Except for Median and BB, non-convergence was not seen for the other included methods. BB showed a large number of missing estimates when analysing simulated PRO data with four possible ordinal categorical values (level 4), taking up to around 74% of the total estimates by BB for level 4 (111,002/150,000), and with a higher predefined treatment difference in level 4, BB tended to have more missing estimates (Table D.3). Median regression also produced around 0.1% missing values when analysing level 4. The degree of missingness for Median and BB decreased when analysing PRO scores with larger number of ordinal categorical values (e.g. level 10 or 26).

Table 9.1 Comparison of predefined parameter values and observed parameter values

DGM	Predefined means			Observed means								
				Level 4			Level 10			Level 26		
	Control	Treatment	Group Difference	Control	Treatment	Group Difference	Control	Treatment	Group Difference	Control	Treatment	Group Difference
1	50	50	0.00	49.99	50.01	0.02	50.03	50.05	0.02	49.98	50.00	0.02
2		54.4	4.40		54.36	4.37		54.36	4.33		54.29	4.31
3		61	11.00		60.83	10.84		60.77	10.74		60.68	10.70
4		67.8	17.80		67.18	17.20		67.02	16.99		66.92	16.94
5		72	22.00		71.31	21.33		71.06	21.03		70.95	20.97

Each cell contains up to a maximum of 30,000 estimates (= 5,000 simulations × 6 n_{obs}). DGM, data-generating mechanism.

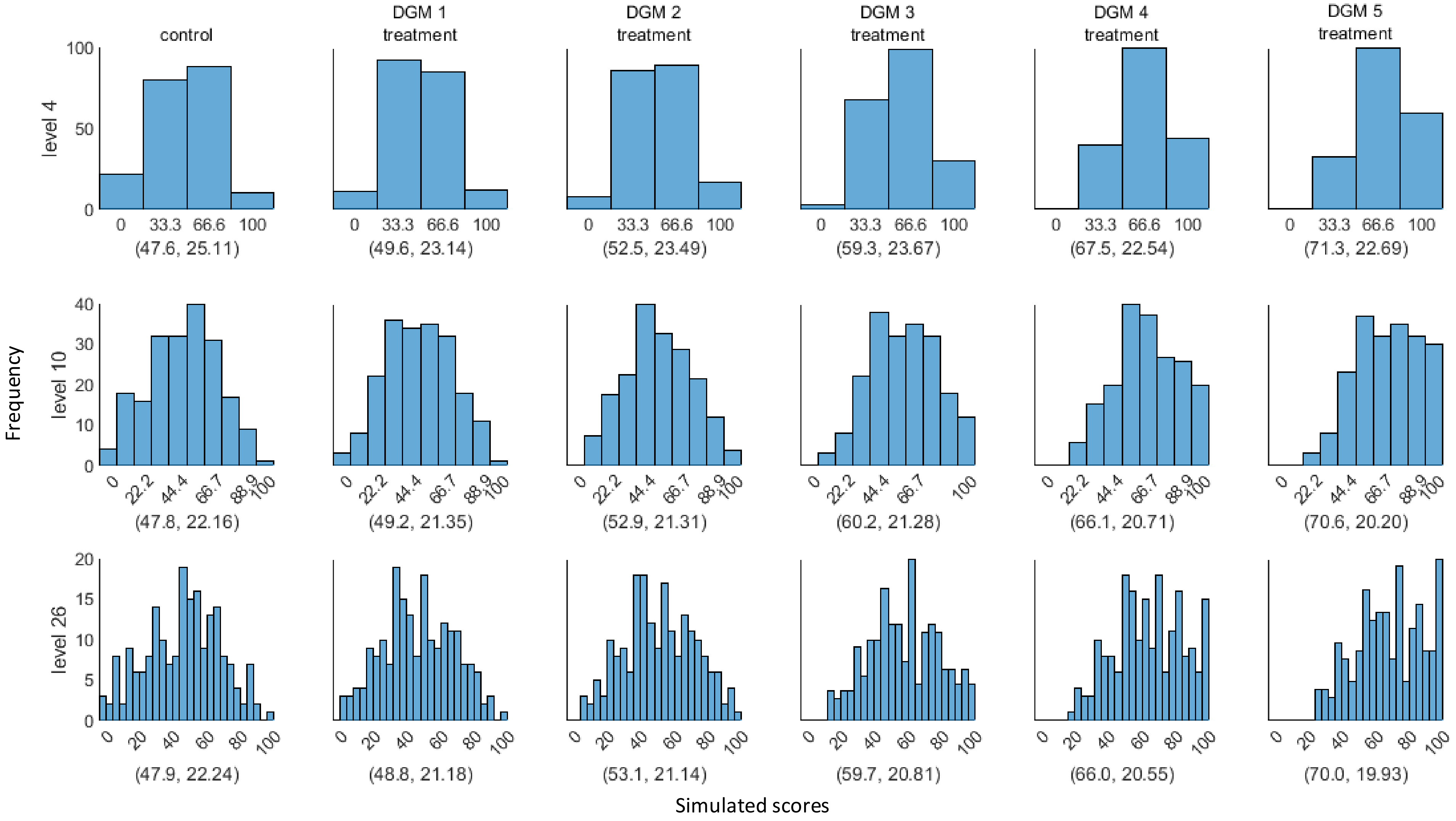


Figure 9.1 Example distributions of the simulated dataset using the five DGMs under three levels (sample size = 400)

The values in the bracket represent the mean and standard deviation of the displayed distribution. The first column represents the score distribution of the control group. DGM 1-5 represents the score distributions of the treatment group using the predefined SESs of 0, 0.2, 0.5, 0.8, and 1.0. Three rows represent three levels (i.e. 4, 10, and 26).

Table 9.2 Estimated treatment difference by six statistical methods for each level and each DGM (Ntotal = 2,700,000)

		Scale-based estimand framework						SES estimand framework					
Level	DGM	Mean or median			logORs			SES					
		MLR	Tobit	Median	Frac	OL	BB	MLR	Tobit	Median	Frac	OL	BB
4	1	0.019	0.020	0.235	0.001	0.002	0.001	0.001	0.001	0.004	0.001	0.001	0.001
	2	4.373	5.011	15.368	0.176	0.346	0.140	0.184	0.184	0.271	0.184	0.180	0.137
	3	10.845	12.626	16.247	0.441	0.864	0.344	0.459	0.455	0.454	0.455	0.440	0.337
	4	17.198	20.612	16.252	0.718	1.393	0.533	0.735	0.718	0.494	0.718	0.676	0.523
	5	21.327	26.250	16.254	0.913	1.754	0.658	0.920	0.887	0.460	0.884	0.814	0.645
10	1	0.019	0.017	0.096	0.001	0.002	0.001	0.001	0.001	0.003	0.001	0.001	0.001
	2	4.325	4.519	5.281	0.174	0.343	0.157	0.199	0.199	0.203	0.199	0.194	0.199
	3	10.736	11.305	10.862	0.437	0.856	0.388	0.496	0.496	0.403	0.493	0.475	0.495
	4	16.994	18.159	16.788	0.709	1.370	0.618	0.796	0.789	0.648	0.779	0.733	0.786
	5	21.030	22.791	21.526	0.898	1.714	0.772	0.998	0.982	0.799	0.963	0.888	0.977
26	1	0.017	0.016	0.033	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	2	4.313	4.444	4.494	0.173	0.343	0.169	0.201	0.201	0.160	0.200	0.196	0.202
	3	10.704	11.096	10.852	0.435	0.854	0.421	0.501	0.502	0.389	0.497	0.479	0.500
	4	16.944	17.770	17.644	0.707	1.367	0.676	0.804	0.800	0.639	0.787	0.740	0.793
	5	20.968	22.238	21.943	0.895	1.711	0.849	1.008	0.997	0.796	0.972	0.897	0.984

Each cell contains up to a maximum of 30,000 estimates (= 5,000 simulations × 6 n_{obs}). A total of 2,700,000 estimates are produced from the six statistical methods under five DGMs for three levels (= 30,000 estimates × 5 DGMs × 6 methods × 3 levels).

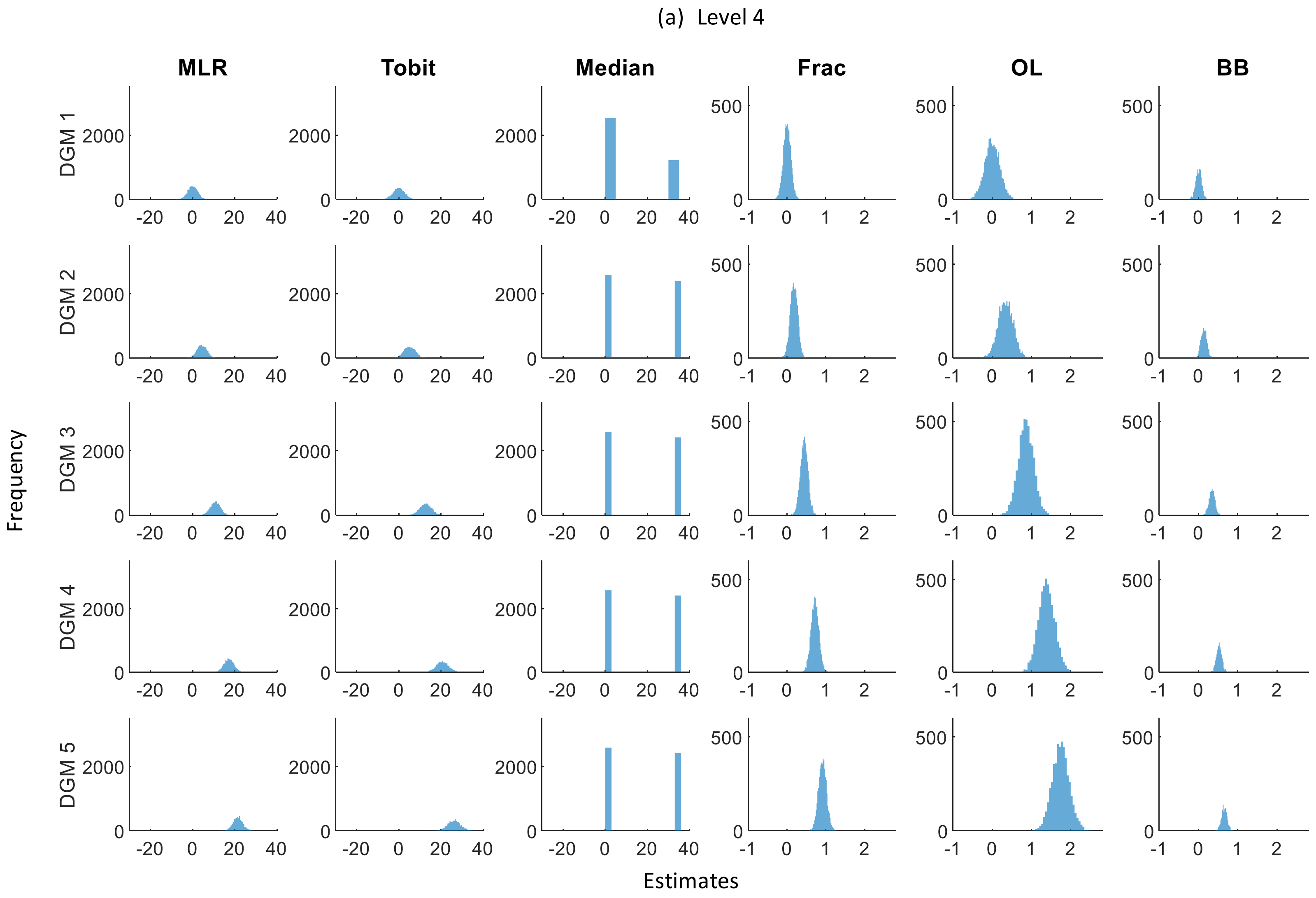
BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; SES, standardised effect size; Tobit, Tobit regression.

Table 9.3 Number and percentage of missing values due to non-convergence for each level under five DGMs (Ntotal = 2,700,000)

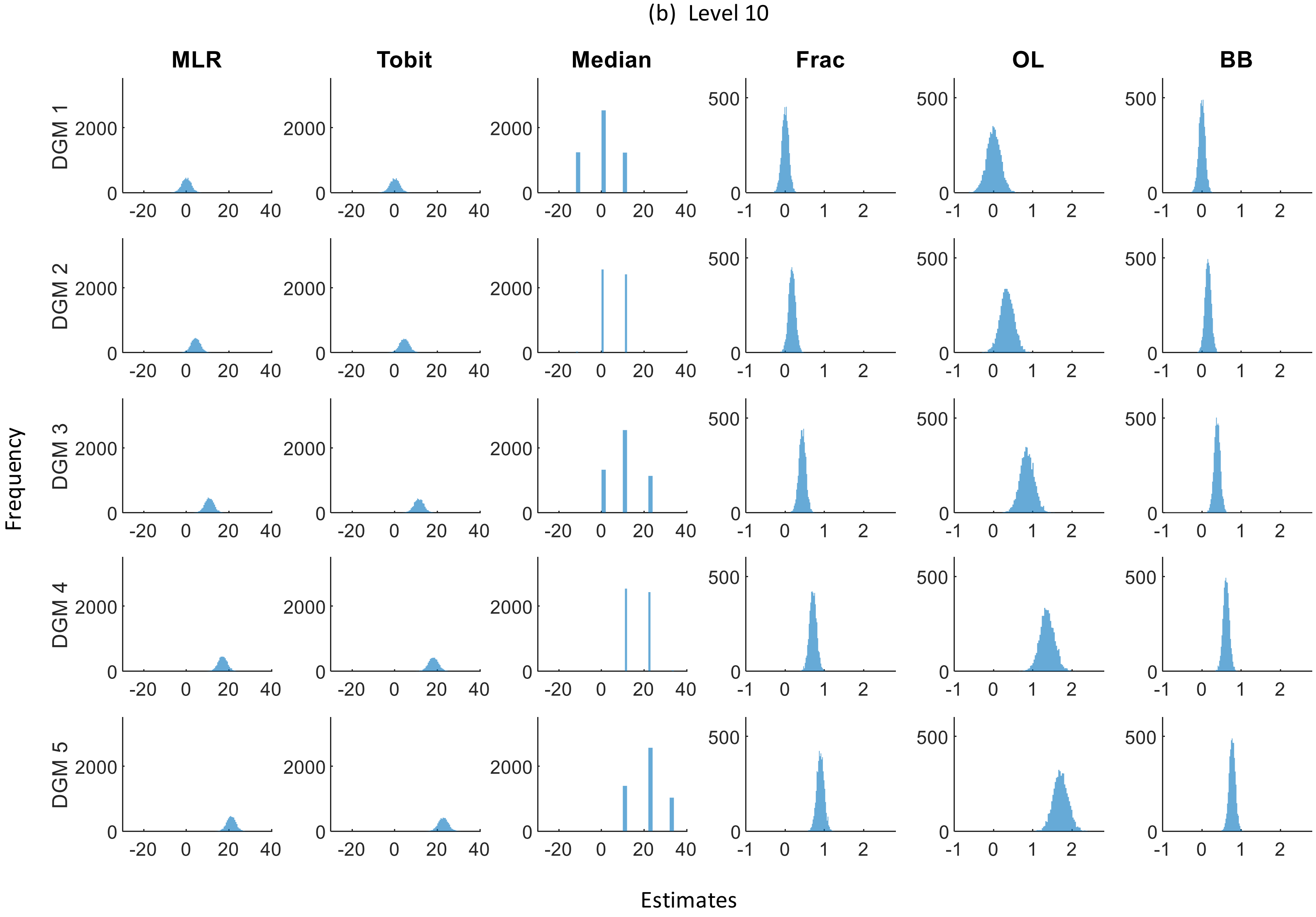
Level	DGM	MLR		Tobit		Median		Frac		OL		BB		Total	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%
4	1	0	0.0	0	0.0	66	0.2	0	0.0	0	0.0	21,396	71.3	21,462	11.9
	2	0	0.0	0	0.0	38	0.1	0	0.0	0	0.0	21,781	72.6	21,819	12.1
	3	0	0.0	0	0.0	38	0.1	0	0.0	0	0.0	22,275	74.3	22,313	12.4
	4	0	0.0	0	0.0	40	0.1	0	0.0	0	0.0	22,408	74.7	22,448	12.5
	5	0	0.0	0	0.0	31	0.1	0	0.0	0	0.0	23,142	77.1	23,173	12.9
	Total	0	0.0	0	0.0	213	0.1	0	0.0	0	0.0	111,002	74.0	111,215	12.4
10	1	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	2	0.0	2	0.0
	2	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	0.0	1	0.0
	3	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4	0.0	4	0.0
	4	0	0.0	0	0.0	1	0.0	0	0.0	0	0.0	1	0.0	2	0.0
	5	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	0.0	1	0.0
	Total	0	0.0	0	0.0	1	0.0	0	0.0	0	0.0	9	0.0	10	0.0
26	1	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	2	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	3	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	4	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	5	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0

Each cell contains up to a maximum of 30,000 estimates (= 5,000 simulations × 6 n_{obs}). A total of 2,700,000 estimates are produced from the six statistical methods under five DGMs for three levels (= 30,000 estimates × 5 DGMs × 6 methods × 3 levels).

BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.



Continued on the next page



Continued on the next page

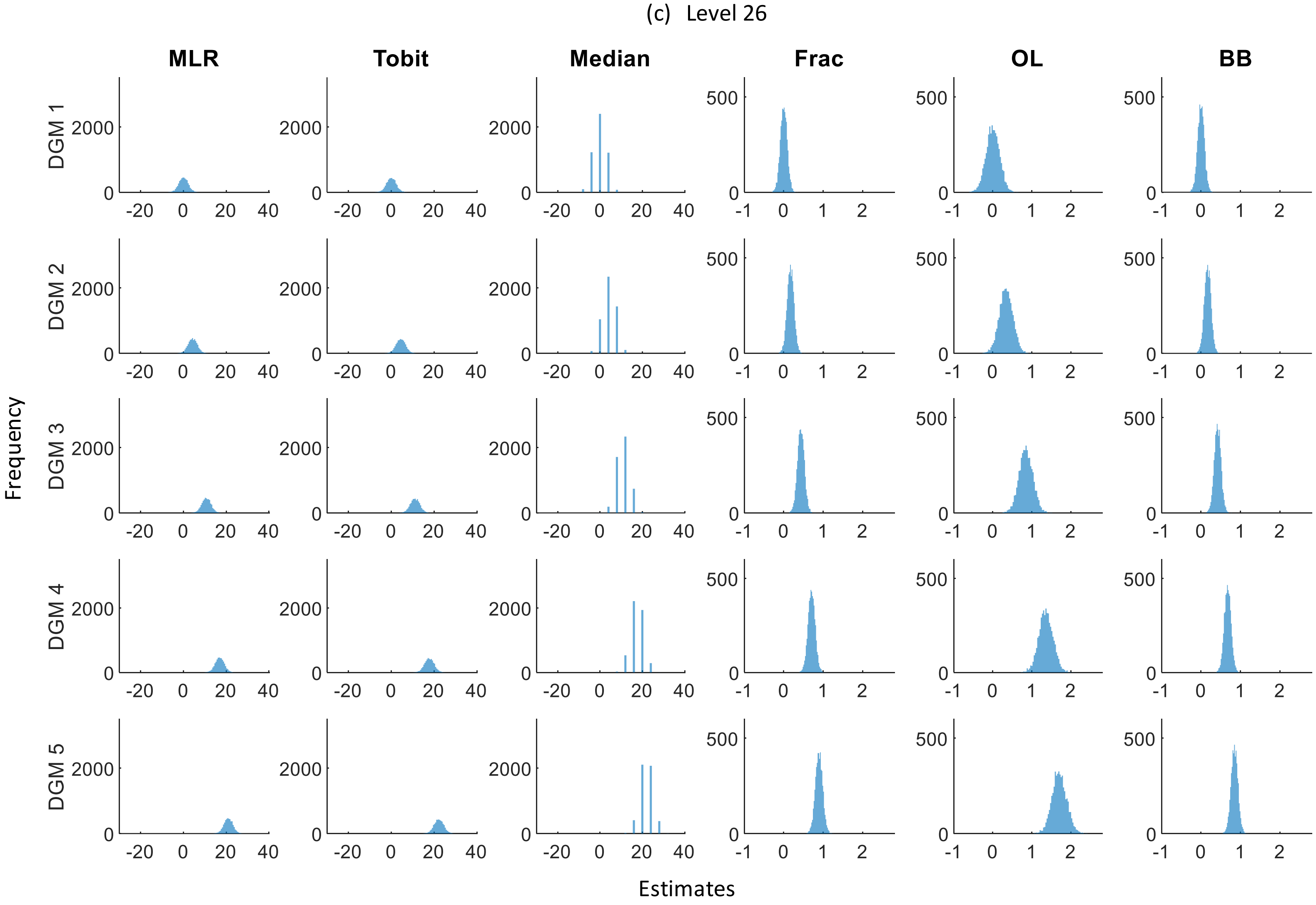
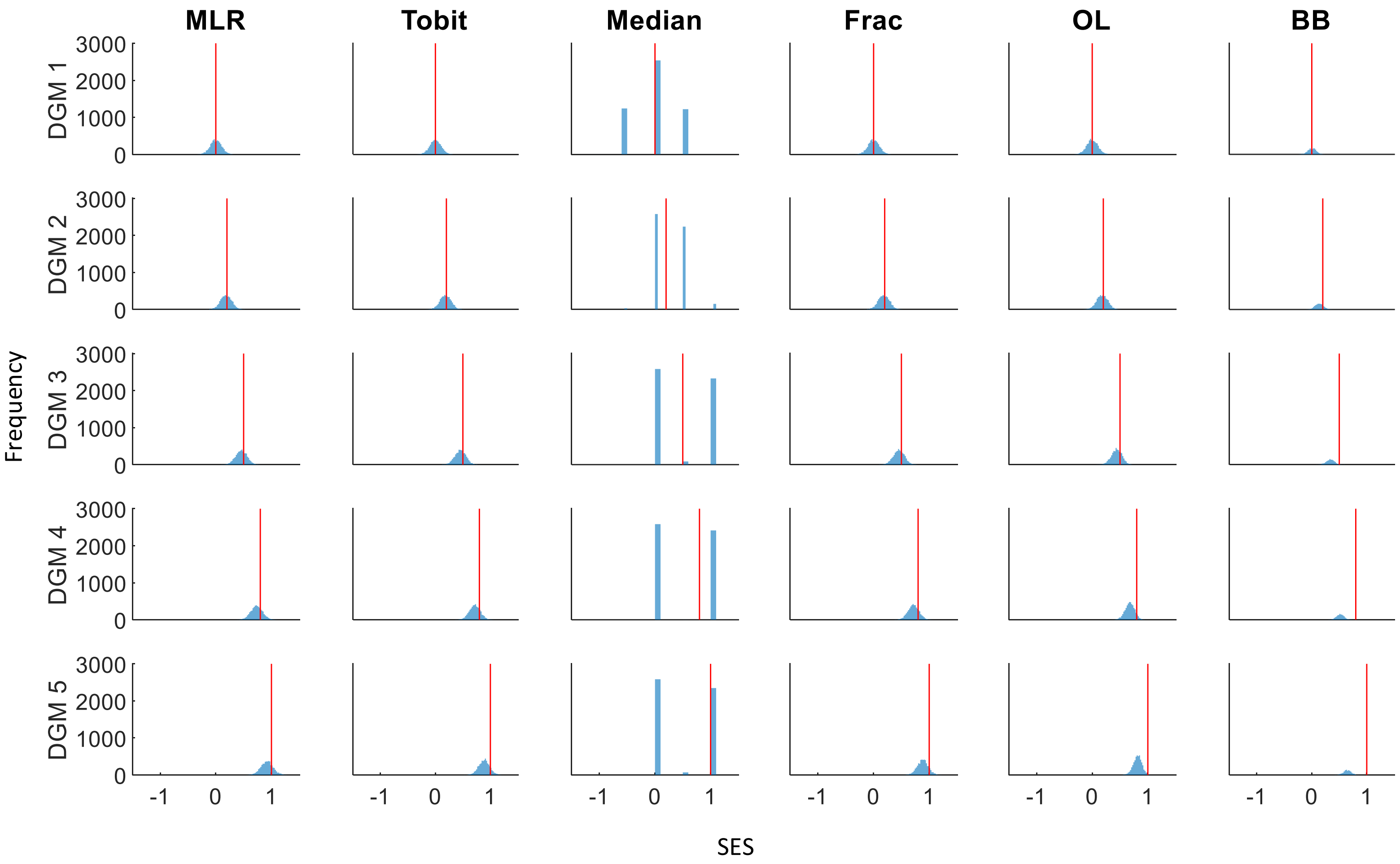


Figure 9.2 Histograms of theta or logOR by six different statistical methods for each level under five DGMs (sample size = 400)

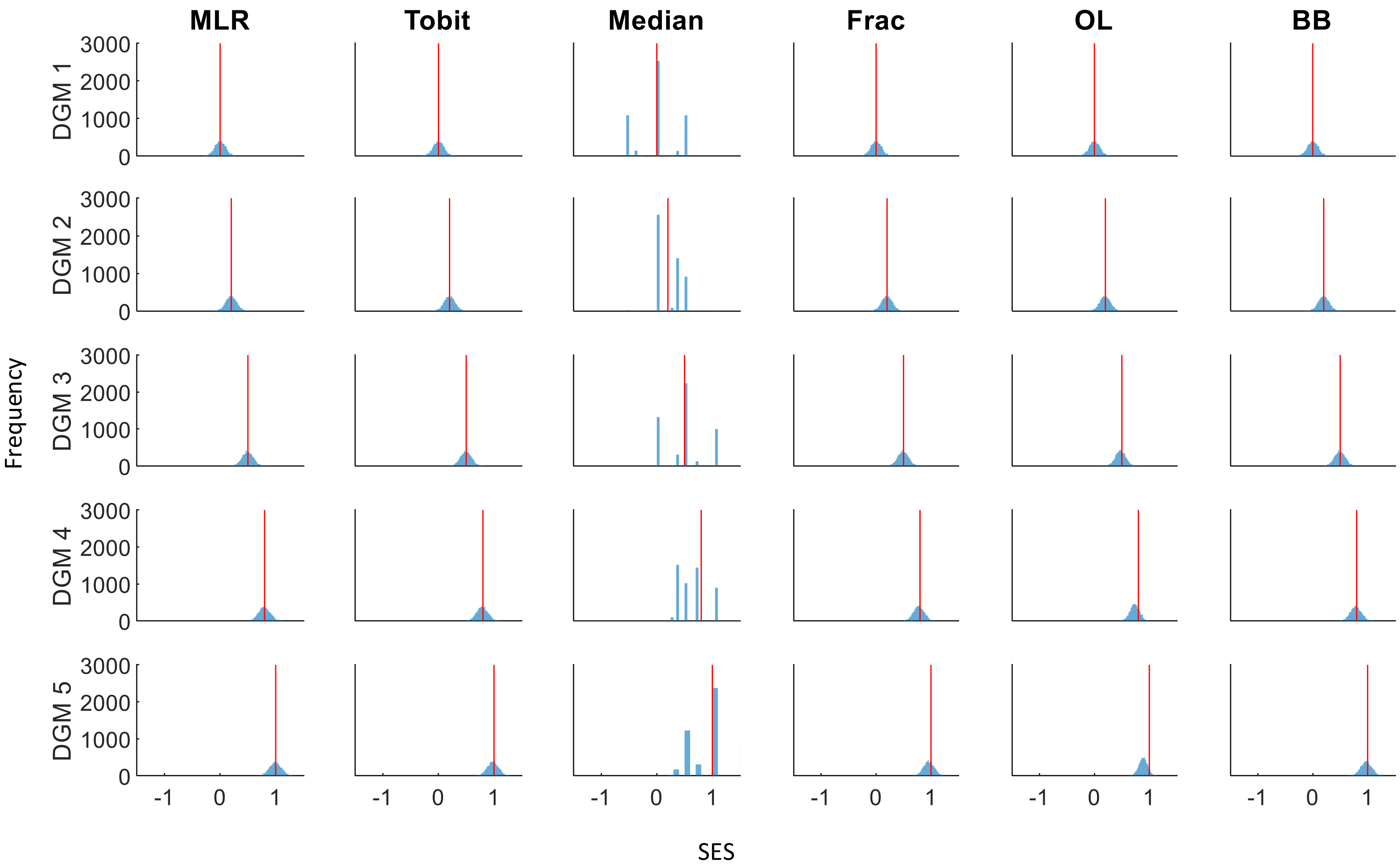
BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; OR, odds ratio; Tobit, Tobit regression.

(a) Level 4



Continued on the next page

(b) Level 10



Continued on the next page

(c) Level 26

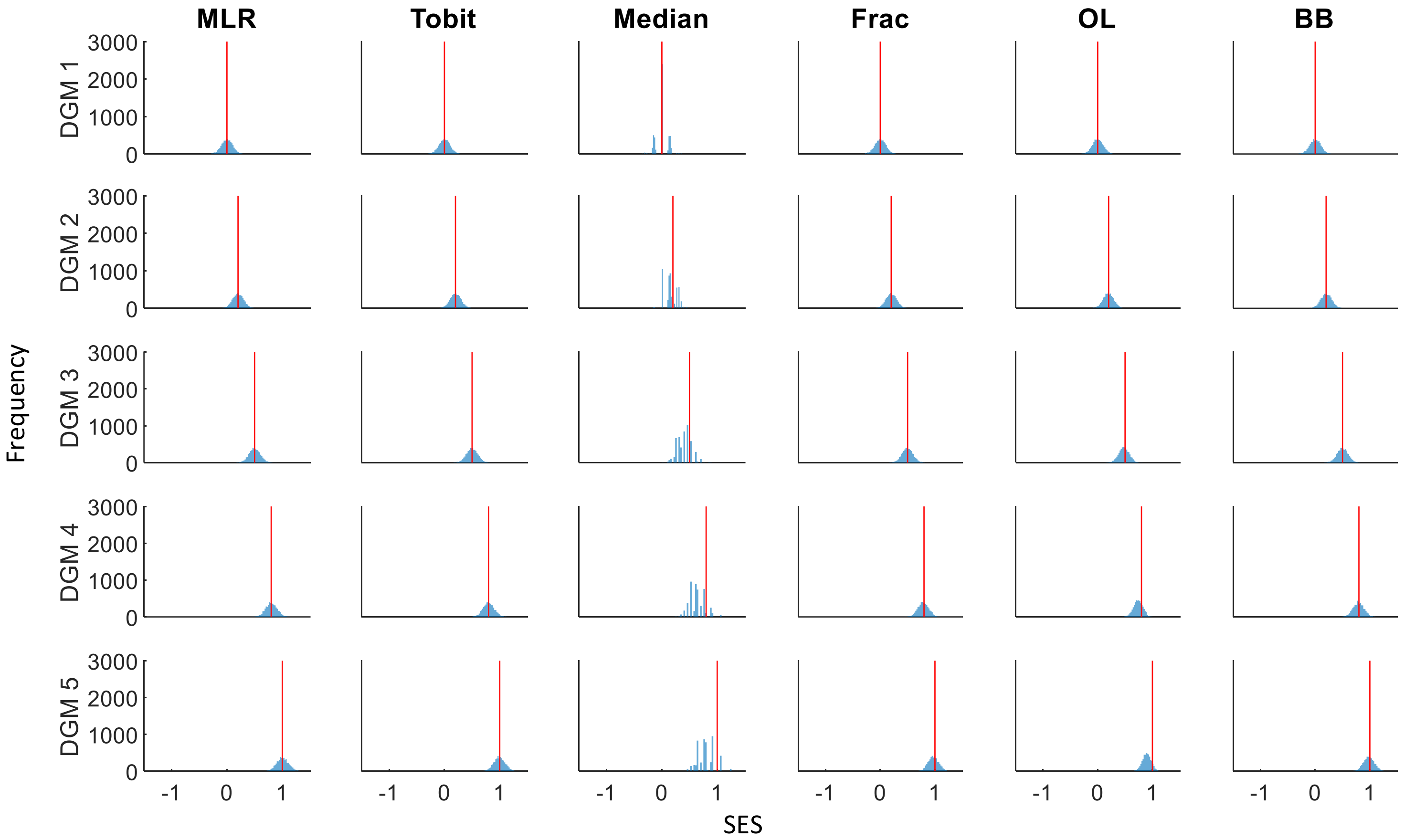
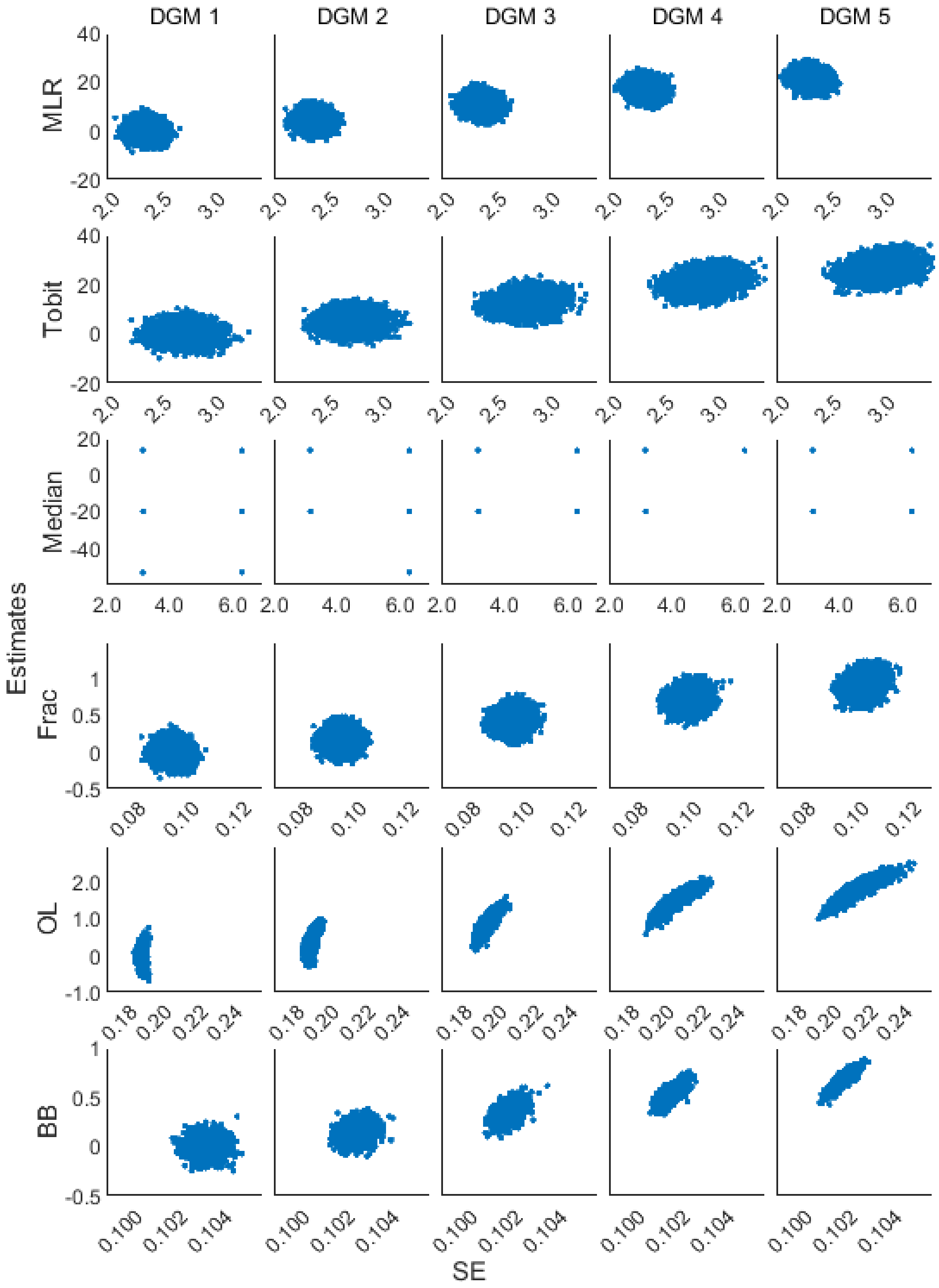


Figure 9.3 Histograms of SES by six different statistical methods under the five DGMs for each level (sample size = 400)

The red reference line represents the predefined treatment effect in SES.

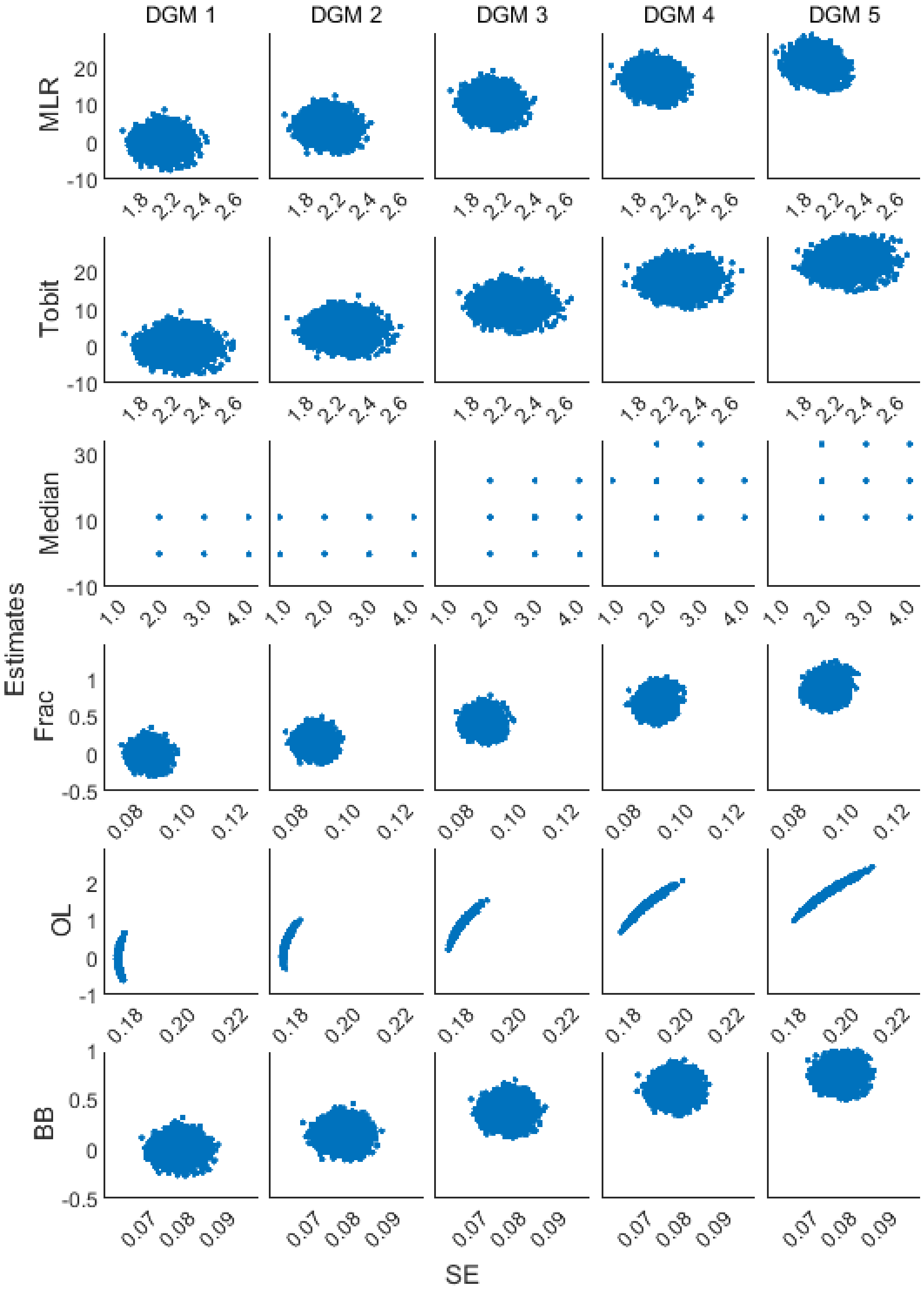
BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; SES, standardised effect size; Tobit, Tobit regression.

(a) Level 4



Continued on the next page

(b) Level 10



Continued on the next page

(c) Level 26

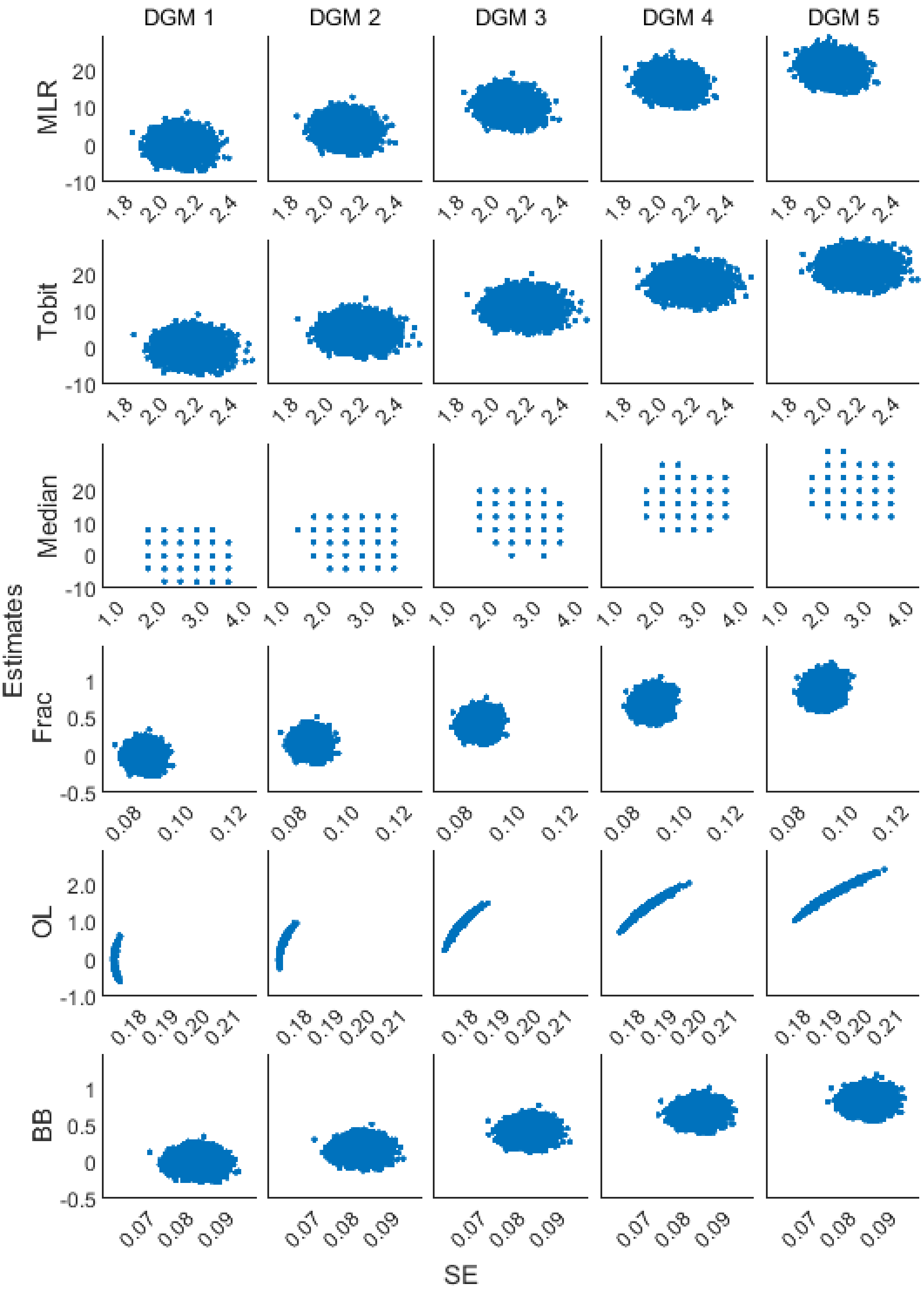
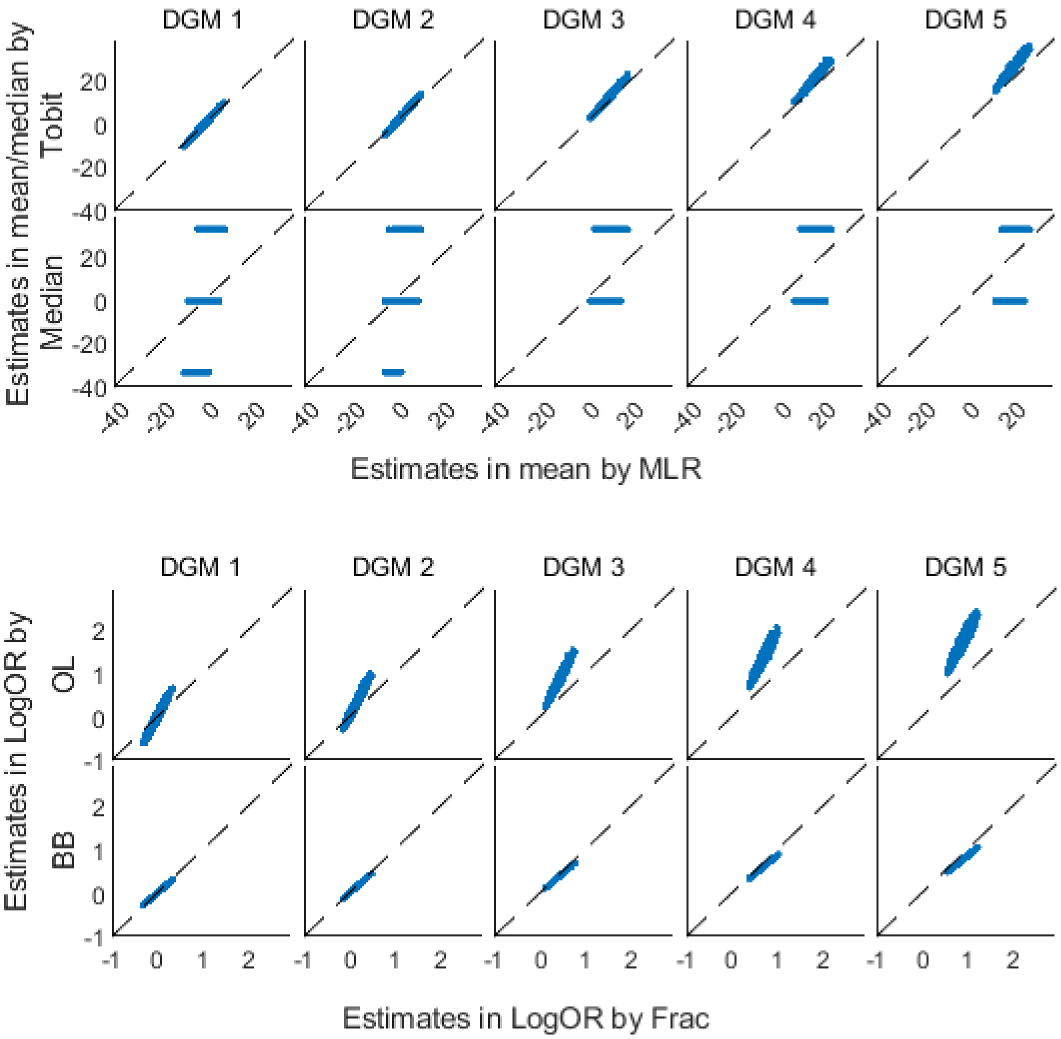


Figure 9.4 Scatterplots of SE vs. theta or logORs for each DGM (sample size = 400)

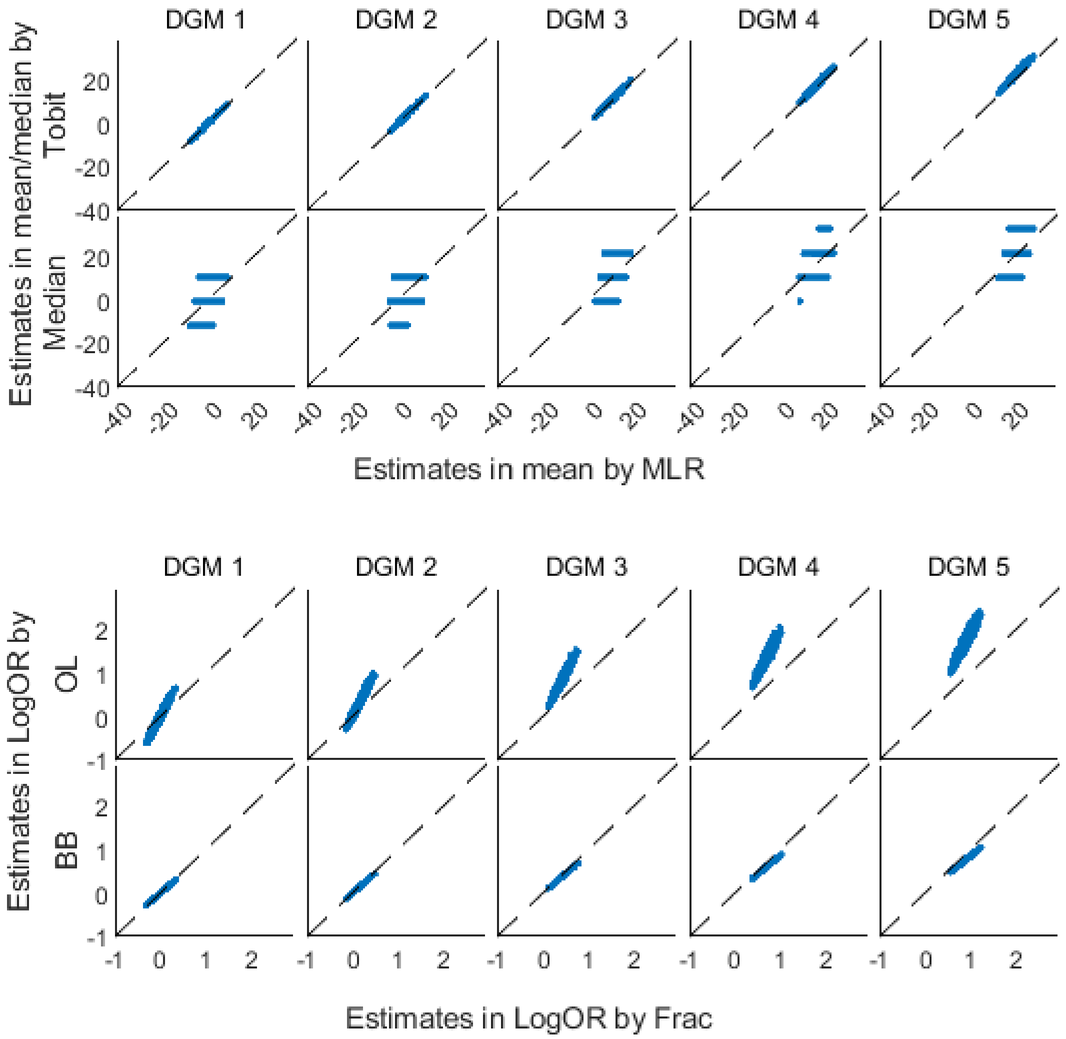
BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; OR, odds ratio; SE, standard error; Tobit, Tobit regression.

(a) Level 4



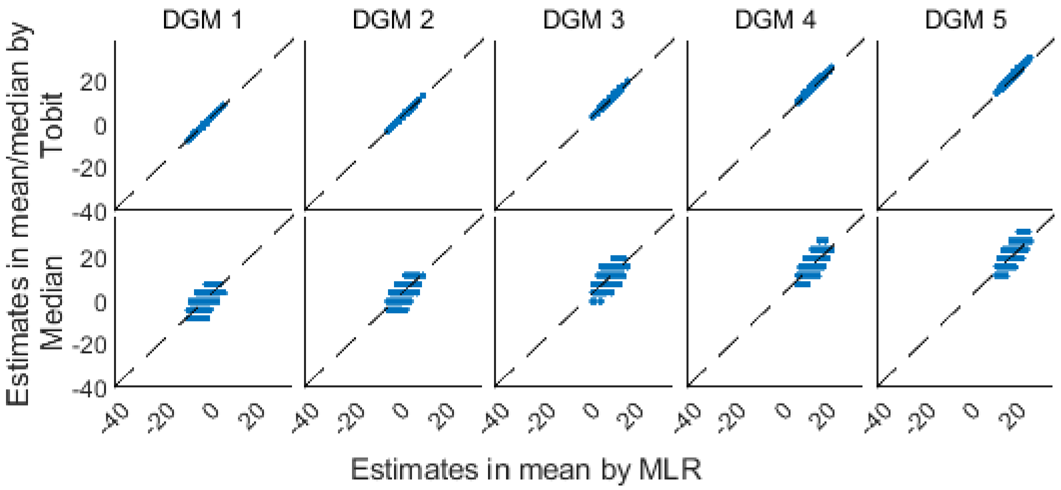
Continued on the next page

(b) Level 10



Continued on the next page

(c) Level 26



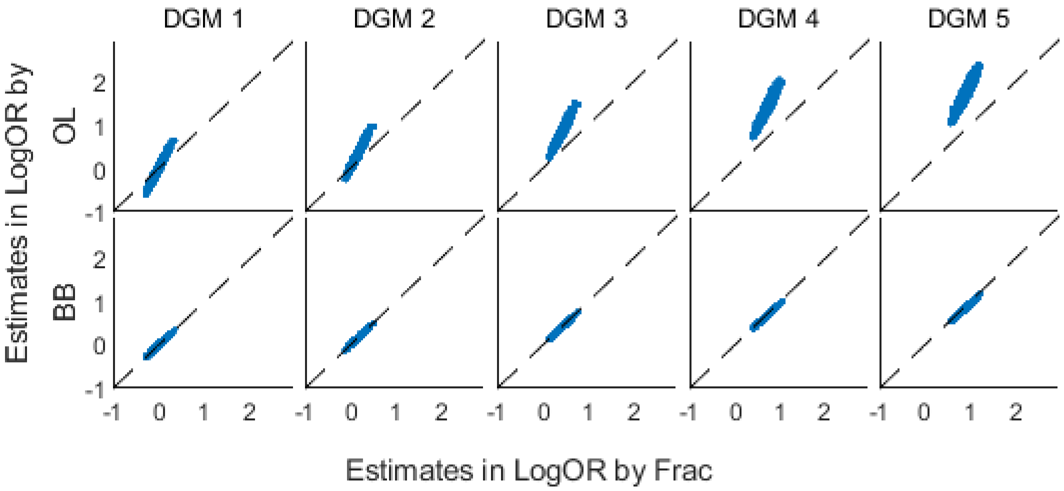
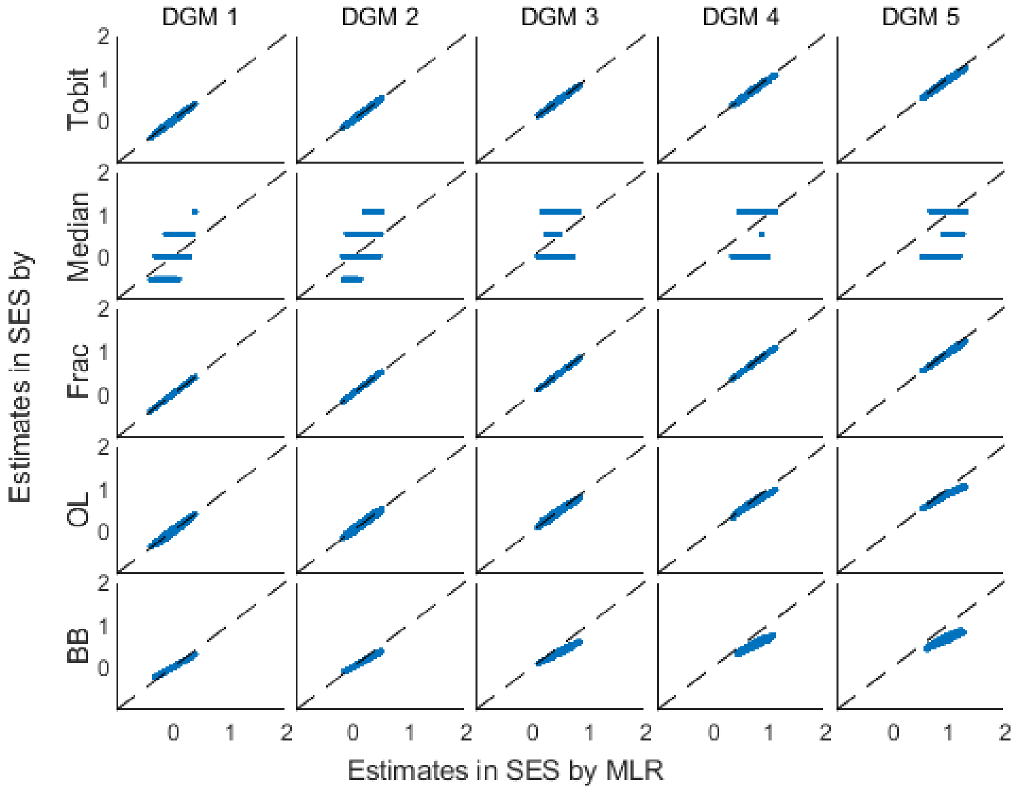


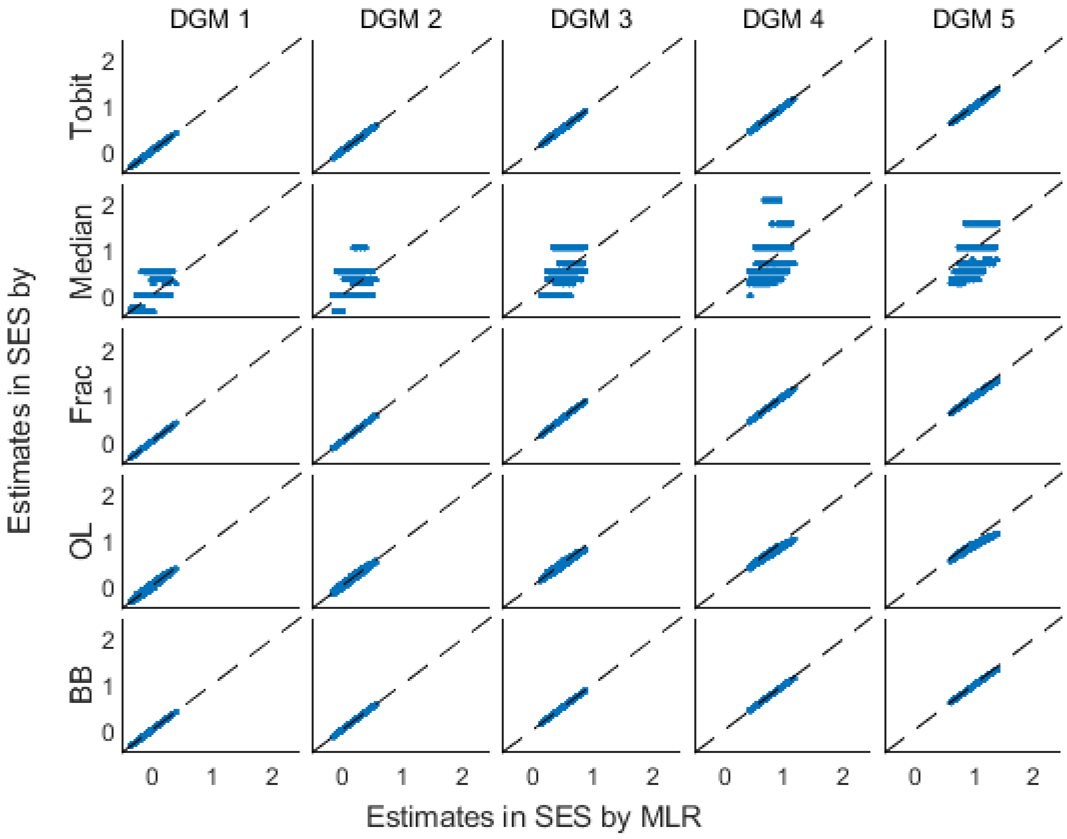
Figure 9.5 Scatterplots of theta using MLR as baseline for Tobit and Median, and of LogOR using Frac as baseline for OL and BB (sample size = 400)

BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; LogOR, log odds ratio; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

(a) Level 4



(b) Level 10



Continued on the next page

(c) Level 26

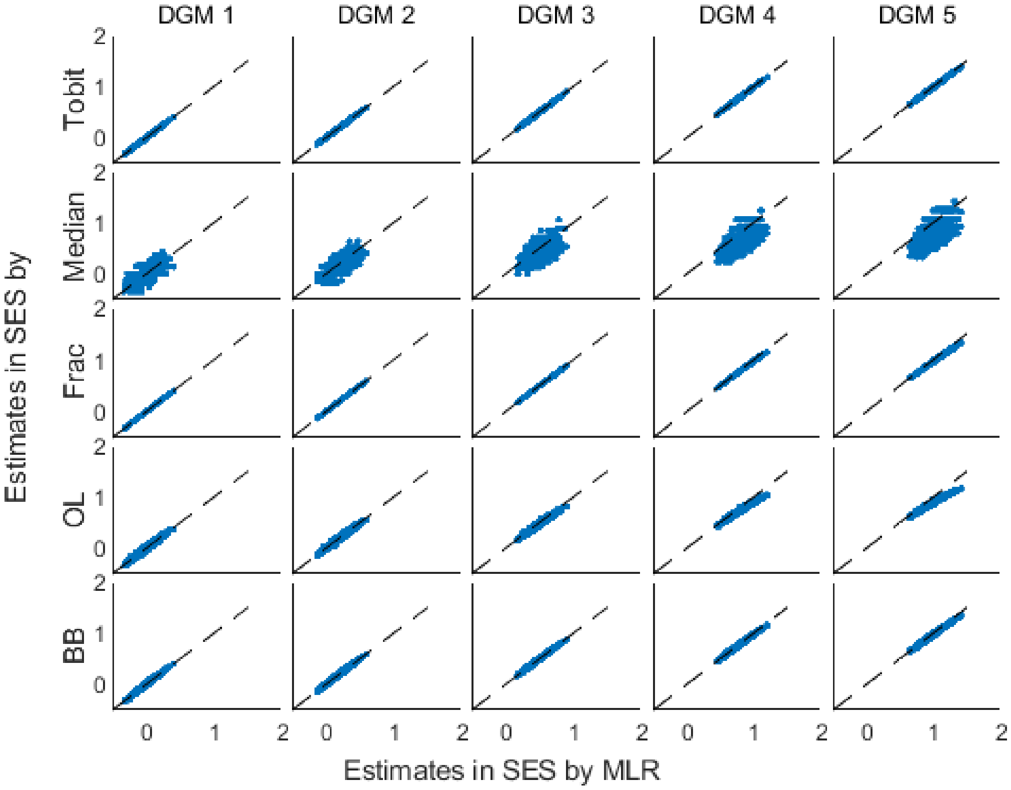


Figure 9.6 Scatterplots of SES using MLR as baseline (sample size = 400)

BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; SES, standardised effect size; Tobit, Tobit regression.

9.3 Evaluation of statistical methods

The six statistical methods are evaluated through exploratory analysis and performance measure analysis to compare their ability to estimate the predefined treatment effect under a range of scenarios that includes no treatment (i.e. treatment difference of zero) and a range of assumed true treatment difference (i.e. small, median, large, and very large values).

9.3.1 Exploratory analysis

A series of figures are graphed for exploratory analysis to identify outliers in this simulation analysis. Figure 9.2 and Figure 9.3 present the distributions of the estimated treatment effect coefficient ($\hat{x}_i$) and the estimated $\widehat{SES}_i$ produced by the six statistical methods under the five DGMs for each level. Figure 9.2 shows the estimates produced by these methods without any transformation under the scale-based estimand framework, so that the estimates from MLR, Tobit, and Median and from Frac, OL, and BB are not comparable. Figure 9.3 allows the comparison of estimates by these methods under the SES estimand framework as these estimates are standardised. Median only produced certain estimates when analysing each level. These estimates are about multipliers of the gap between two nearby categories. For example, the estimates were around multiples of 11.1 for level 10 (Figure 9.2). No bivariate outliers were seen in the scatterplots of estimated treatment coefficients ($\hat{x}_i$) against their associated SEs ($\widehat{SE}(\hat{x}_i)$) by different statistical methods under the five DGMs (Figure 9.4). Most methods tended to present no relationship or a positive relationship between $\hat{x}$ and $\widehat{SE}(\hat{x})$, whereas the Median had a striped pattern.

Figure 9.5 shows the scatterplots of the estimated treatment coefficient by Tobit and Median against MLR, and estimated logORs by OL and BB against Frac under the scale-based estimand framework. Figure 9.6 presents the scatterplots of estimated SESs by different statistical methods against MLR under the SES estimand framework. When the magnitude of the predefined 'truth' increased, Tobit tended to produce numerically larger estimates than MLR when the possible ordinal categorical value was small, but its estimates became similar to MLR after standardisation (Figure 9.6). Similarly, the estimated logORs from OL tended to be numerically larger than Frac and BB, and this trend was more obvious with higher predefined 'truth'. However, after standardisation, the difference of the estimated SES of these methods decreased.

9.3.2 Analysis of performance measures

Various performance measures are compared under the two predefined estimand frameworks. The key performance measures include the bias that measures the accuracy of these statistical methods for estimating the true treatment effect, the coverage of 95% CIs for including the true treatment effect and the power or Type I error that measures the precision or the robustness of these methods. Performance statistics such as mean squared errors, empirical standard errors, and model-based SEs are also presented.

Given the simulated datasets were generated from the Normal distribution where the treatment difference was in means, the predefined ‘truth’ for statistical methods that produced estimates on the logOR scale (i.e. Frac, OL, and BB) is unknown under the scale-based estimand framework, except for under the null hypothesis where the predefined ‘truth’ is zero. Therefore, statistical methods that produced estimates in logORs were only compared under the null hypothesis (i.e. DGM 1) under the scale-based framework. Since the predefined SESs are known, the SES estimates from the six statistical methods were comparable under the SES estimand framework.

9.3.2.1 Under the scale-based estimand framework

Under the scale-based estimand framework, estimates ($\hat{x}_i$) for MLR, Tobit, and Median are means or medians, denoted by $\hat{\theta}_i$, and estimates for Frac, OL, and BB are logORs , denoted by $\widehat{logORs}_i$.

9.3.2.1.1 Bias

Bias measures the difference between the estimates and the predefined ‘truth’. Figure 9.7 presents the change in bias from the original scale-based methods and the transformed scale-based methods separately under the null hypothesis (DGM 1). When the predefined ‘truth’ is zero, both the original scale-based methods and the transformed scale-based methods were able to produce estimates close to the predefined ‘truth’. Their estimates fluctuated and gradually converged to the dash line (bias = 0) with the increase in n_{obs}. Median presented larger bias than other methods especially for small number of levels. Since these methods produced estimates on different scales, their degree of biases were not comparable under the scale-based estimand framework.

Figure 9.8 presents the change in bias estimated from the untransformed (i.e. original) scale-based methods under the alternative hypothesis (DGM 2-5). For MLR, Tobit, and Median, the bias was smaller for larger number of levels and higher n_{obs}. When analysing a small number of levels, especially level 4, Tobit tended to overestimate the treatment difference, and MLR tended to underestimate the treatment difference, but Tobit was more biased than MLR. However, when analysing a higher number of levels, the bias from Tobit became smaller than MLR.

9.3.2.1.2 Mean squared error

Under the null hypothesis, the MSE estimated by the MLR and Tobit in the original-scale based methods, and BB and Frac in the transformed scale-based methods, overlapped with each other, decreasing with the increase in n_{obs}, and coverages to the dash line representing MSE = 0; whereas their counterparts, OL and Median, presented comparatively high MSE respectively (Figure 9.9). Under the alternative hypotheses, MLR and Tobit showed similar trend, with Tobit having slightly larger MSE for level 4 but smaller MSE for level 26 while the predefined 'truth' was large (i.e. under DGM 4 & 5). This indicates that Tobit was less precise than MLR for small number of levels (i.e. level 4), but more precise than MLR for large number of levels (i.e. level 26). Despite the MSE of Median dropped dramatically with the increase in the number of levels, Median had the worst precision in comparison with MLR, and Tobit in all scenarios (Figure 9.10).

9.3.2.1.3 Empirical standard error and average model standard error

EmpSE measures how precise the estimates are to the average estimates of each statistical method (Figure 9.11), and the average ModSE measures the square root of the average squared model SEs (Figure 9.12). Comparing these two performance measures using the relative % error in ModSE can measure to what degree the ModSE is overestimated or underestimated (Figure 9.13).

The EmpSE for MLR and Tobit remained similar under different scenarios, whereas Median was associated with higher EmpSE, i.e. less precision to the average estimates, when the number of levels is small. As shown in Figure 9.11(a) Median tended to have less EmpSE and converged to the trend for MLR and Tobit, when analysing outcome data with more discrete values (i.e. level 10 and level 26). The EmpSEs for the transformed scale-based methods increased for a higher predefined treatment effect, especially for OL as shown in Figure 9.11(b). This means that OL tended to have less precision to its average estimates than Frac and BB, which were associated with EmpSE approaching zero.

The ModSE became small with the increase in n_{obs}. The ModSE for each method remained similar across different DGMs for MLR and Tobit, whereas the Median was associated with higher ModSE under DGM 1 and 2 (Figure 9.12(a)). The ModSE of the transformed scale-based methods also showed a decrease with higher n_{obs}, but their ModSE became higher with the increase in predefined 'truth', especially for OL (Figure 9.12(b)).

For the original scale-based methods (Figure 9.13(a)), the relative % error in ModSE for MLR and Tobit was above the reference line y = 0 when the n_{obs} was set at 200 or 400, indicating that their ModSEs were overestimated by approximately 0.5% - 3%. The relative % error in ModSE for MLR and Tobit fluctuated below the reference line y = 0 when the n_{obs} was higher than 400 for most scenarios, meaning that their ModSEs were underestimated. The relative % error in ModSE for Tobit had an obvious decrease with the increase in predefined treatment difference. The relative % error in ModSE for Frac and OL shared a similar pattern with MLR and Tobit, with an exception that the relative % error in ModSE for OL was overestimated for most scenarios (Figure 9.13(b)). These trends indicate biases in the estimation of ModSEs for these methods, but they were much less than the bias in the estimation of ModSEs for Median in comparison with MLR and Tobit at all three levels, and for BB in comparison with Frac and BB at level 4.

9.3.2.1.4 Coverage of 95% CIs

Figure 9.14 and Figure 9.15 show the coverage of 95% CIs for $\hat{x}_i$ with the change in n_{obs}, where the dashed reference line represents coverage = 0.95. This measures the probability that a CI would include the predefined 'truth'. Under the null hypothesis (DGM 1), most methods were able to produce the estimates of zero, except for Median. Under the alternative hypothesis (DGM 2-5), the coverage of MLR and Tobit deviated from the reference line with the increase in the predefined 'truth' (the assumed true treatment effect) and the increase in n_{obs}, and this phenomenon was more obvious for Tobit at level 4. However, for larger number of levels (i.e. for level 10 and 26), Tobit had less deviations from the reference than MLR did. Median had better performance in coverage when the simulated PRO scores became closer to continuous data (level 26), but its coverage was still not as good as MLR or Tobit under most scenarios.

9.3.2.1.5 Type I error and power

Figure 9.16 shows the Type I error for $\hat{x}_i$ with the change in n_{obs} under the null hypothesis (DGM 1), where the dash line represents the Type I error of 0.05. It measures the probability of incorrectly rejecting the true null hypothesis. The lines of MLR and Tobit, Frac and OL overlapped with each other at y = 0.05 when the significance level was set at $\alpha = 0.05$. The Median had Type I error at around 0.5 under DGM 1, indicating that Median produced far more false positives than MLR and Tobit. At level 4, BB had less Type I error at around 0.01, approximately five times smaller than Frac and OL.

Figure 9.17 shows the power measured by $\hat{\theta}_i$ with the change in n_{obs} under the alternative hypothesis (DGM 2-5). The power measures the probability of rejecting the null hypothesis when it is false. The increasing trend in power of MLR and Tobit overlapped with each other, while the power of Median was lower than MLR and Tobit under all DGMs for level 4 and under DGM 2-5 for level 10 and 26. The power of transformed scale-based methods overlapped in most scenarios, except for DGM 2 under level 4 where the power of BB increased slower with the increase in n_{obs}.

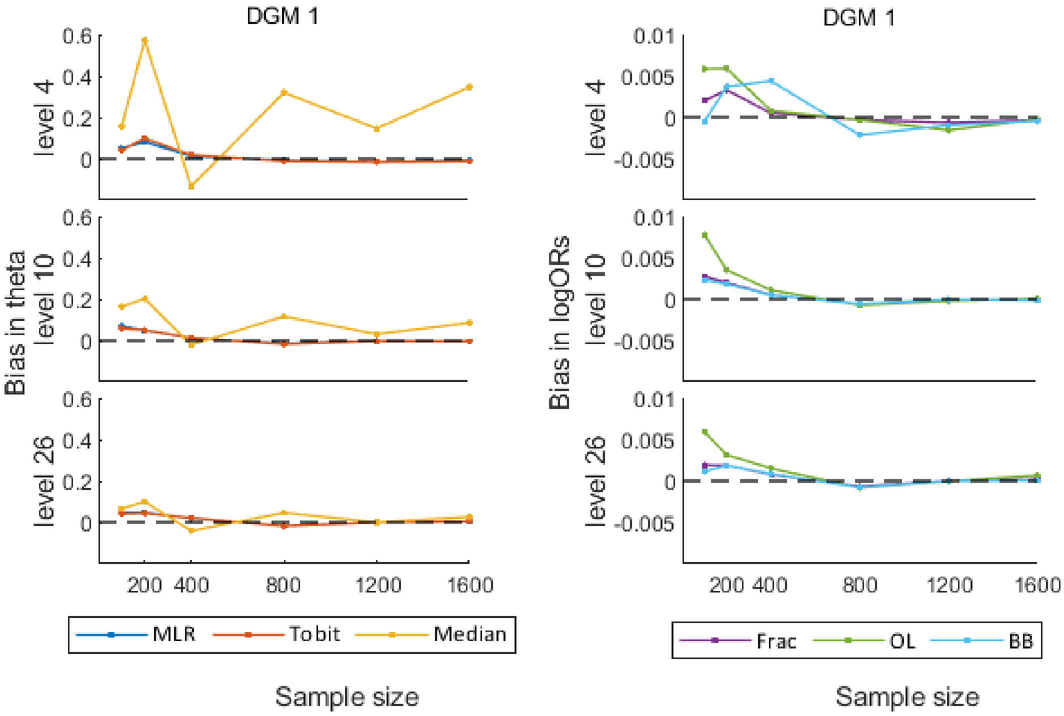


Figure 9.7 Line plots of bias under the scale-based estimand framework with the change of sample size for different levels under the null hypothesis (DGM 1)

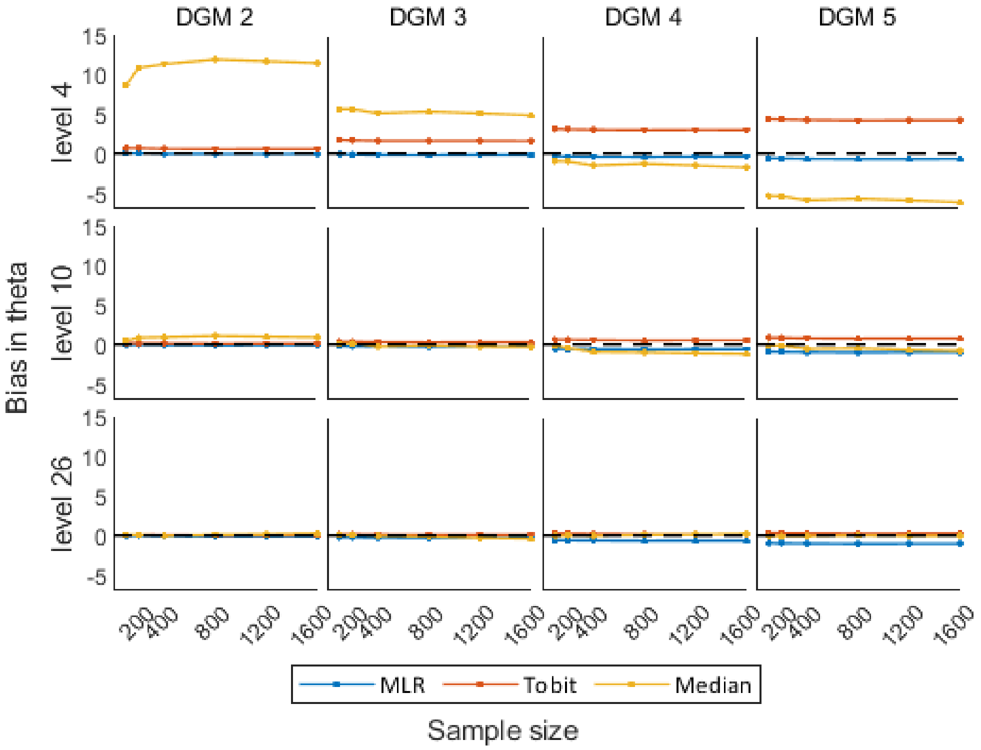


Figure 9.8 Line plots of bias under the scale-based estimand framework with the change of sample size for different levels under the alternative hypotheses (DGM 2-5)

Theta presents the original estimates under the scale-based estimand framework. BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

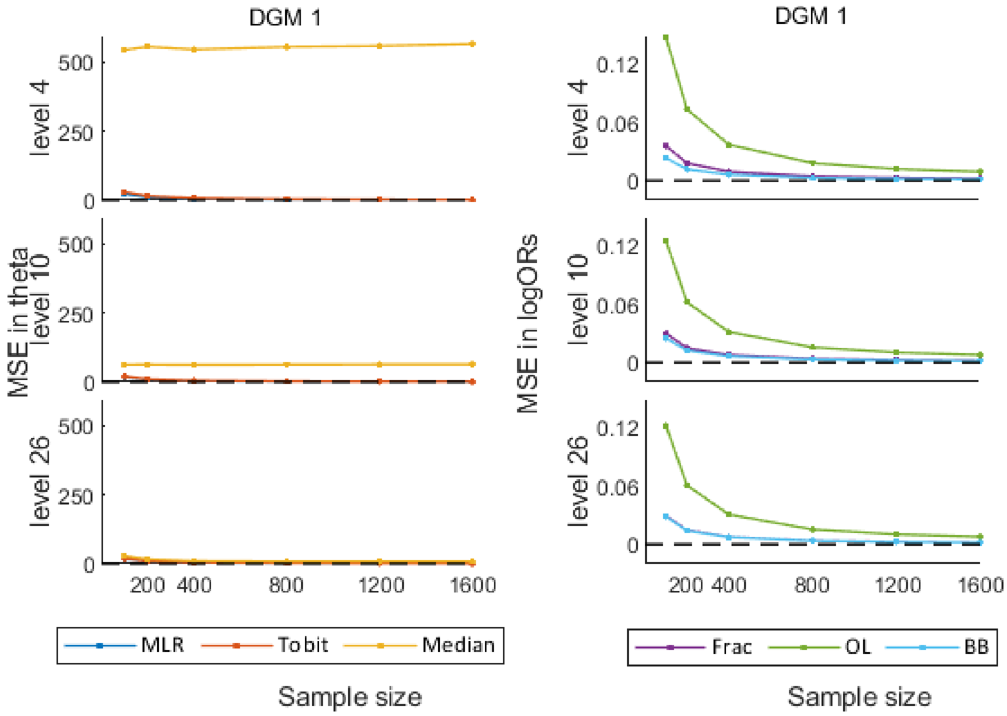


Figure 9.9 Line plots of mean squared error (MSE) under the scale-based estimand framework with the change of sample size for different levels under the null hypothesis (DGM 1)

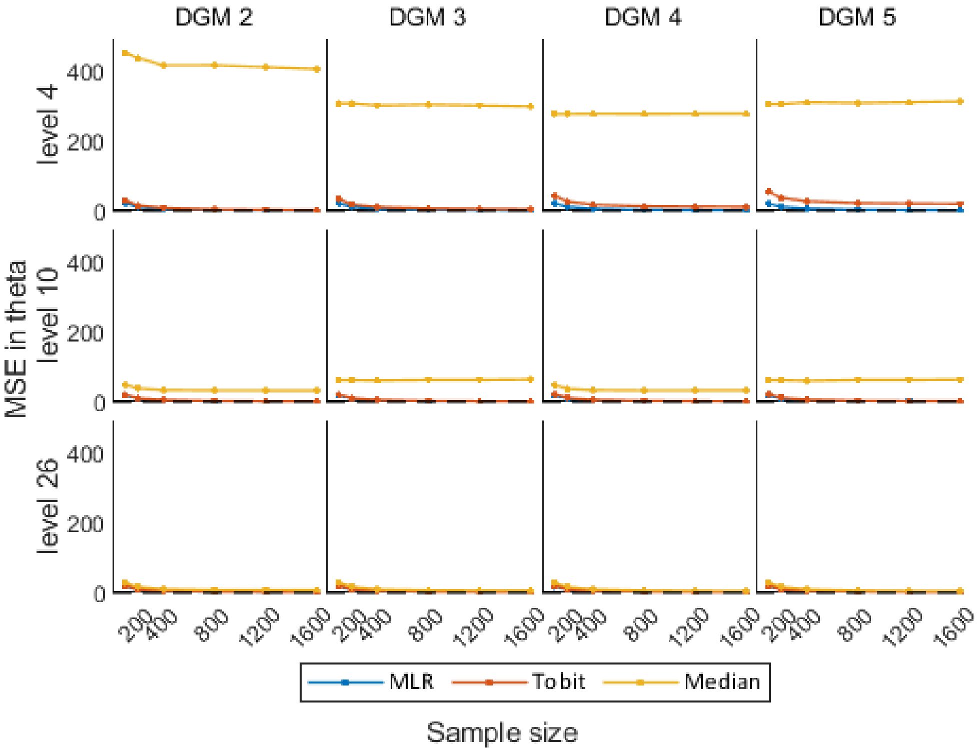
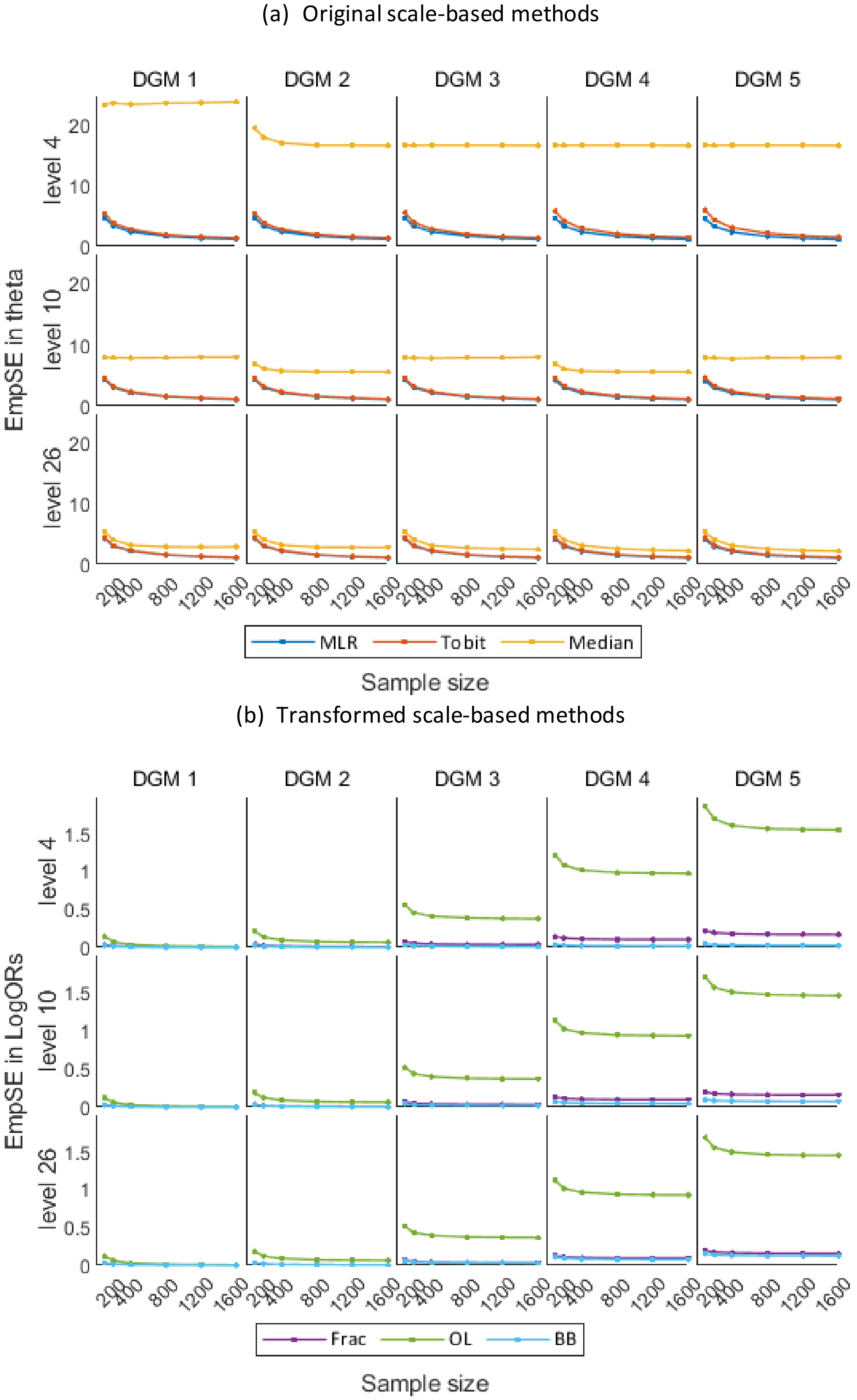


Figure 9.10 Line plots of mean squared error (MSE) under the scale-based estimand framework with the change of sample size for different levels under the alternative hypotheses (DGM 2-5)

Theta presents the original estimates under the scale-based estimand framework. BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

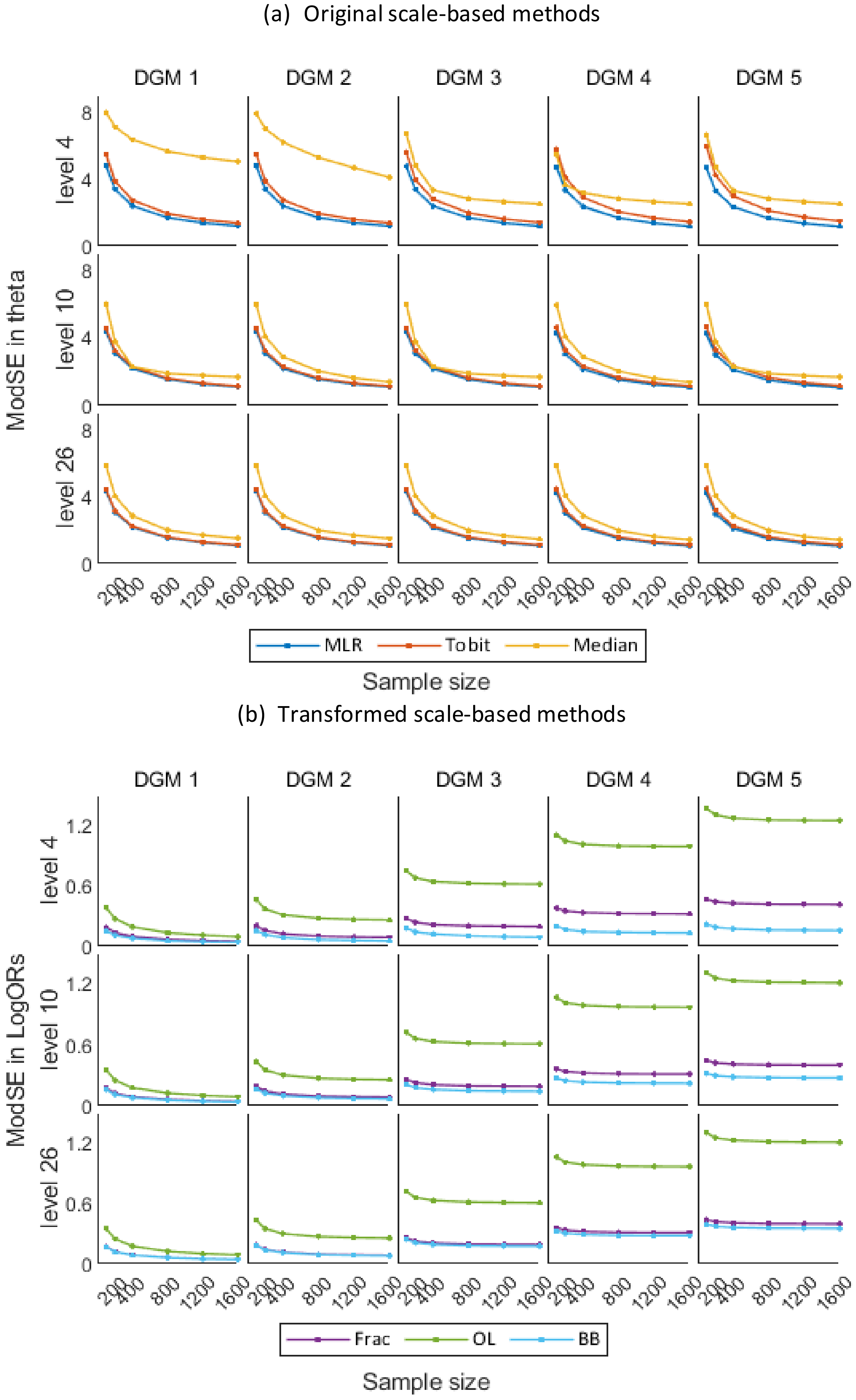


(a) Original scale-based methods

(b) Transformed scale-based methods

Figure 9.11 Line plots of empirical standard error (EmpSE) under the scale-based estimand framework with the change of sample size for different levels

Theta presents the original estimates under the scale-based estimand framework. BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; LogOR, log odds ratio; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

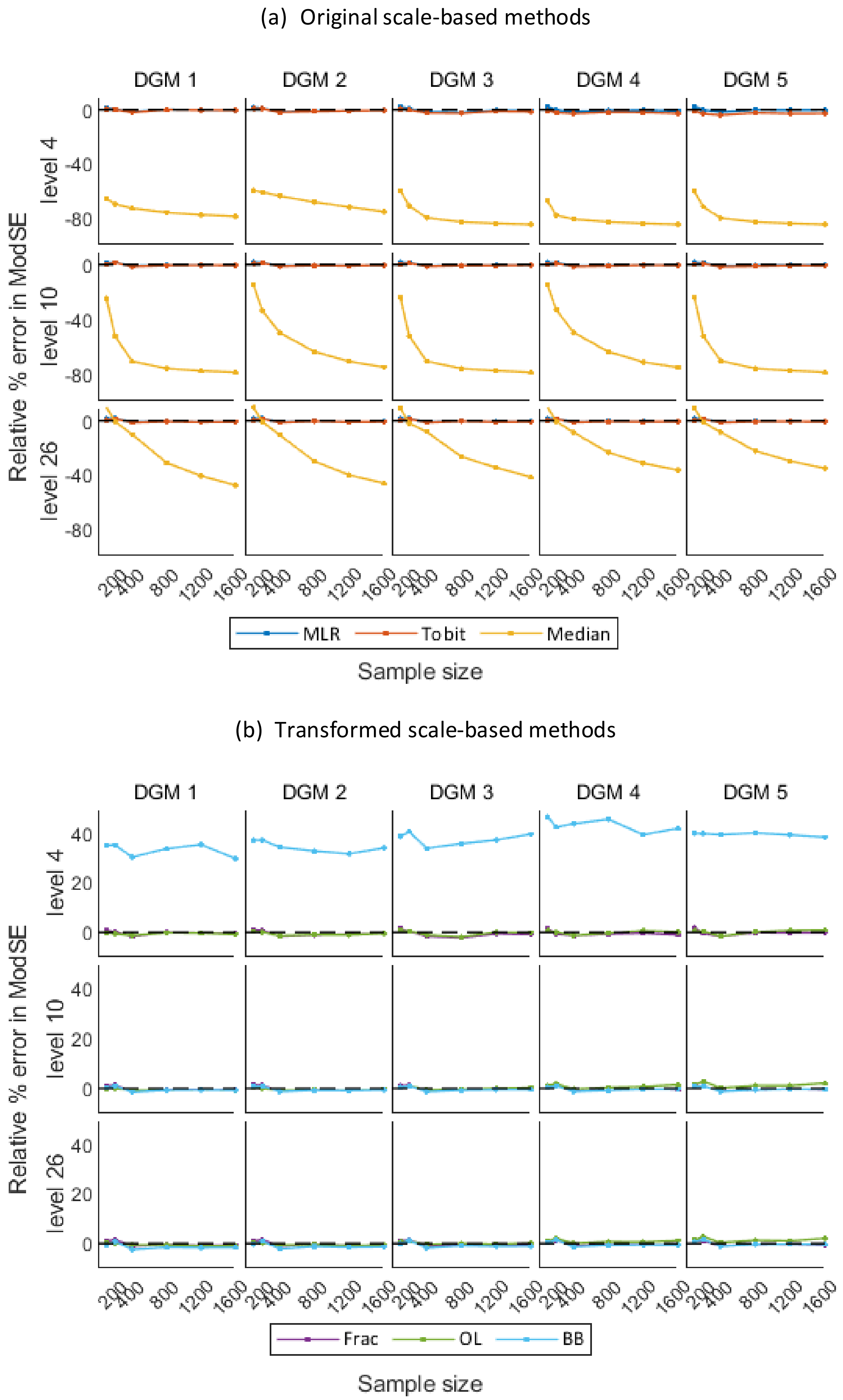


(a) Original scale-based methods

(b) Transformed scale-based methods

Figure 9.12 Line plots of average model standard error (ModSE) under the scale-based estimand framework with the change of sample size for different levels

Theta presents the original estimates under the scale-based estimand framework. BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.



(a) Original scale-based methods

(b) Transformed scale-based methods

Figure 9.13 Line plots of relative % error in ModSE under the scale-based estimand framework with the change of sample size for different levels

BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

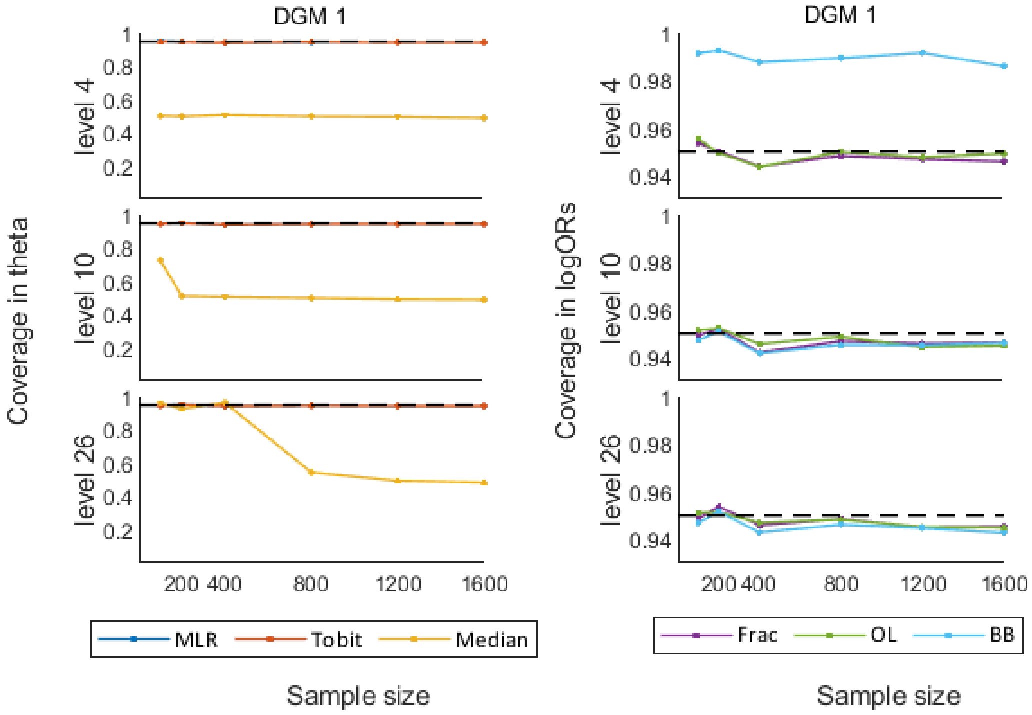


Figure 9.14 Line plots of coverage for theta or LogORs with the change of sample size for different levels under the null hypothesis (DGM 1)

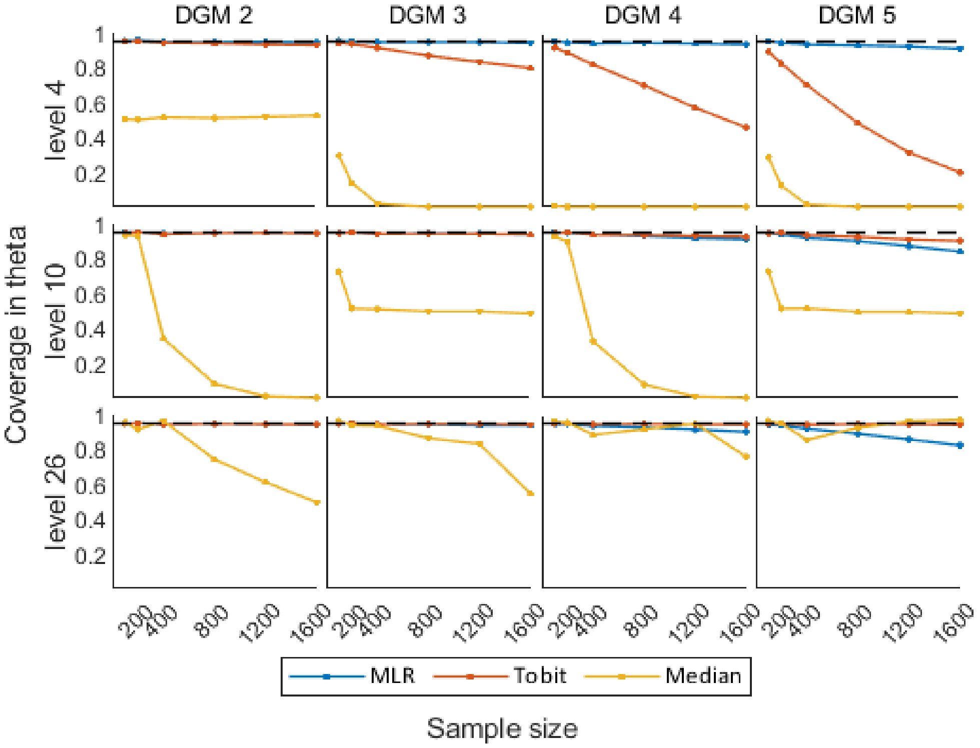


Figure 9.15 Line plots of coverage with the change of sample size for different levels under the alternative hypotheses (DGM 2-5)

Theta presents the original estimates under the scale-based estimand framework. BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

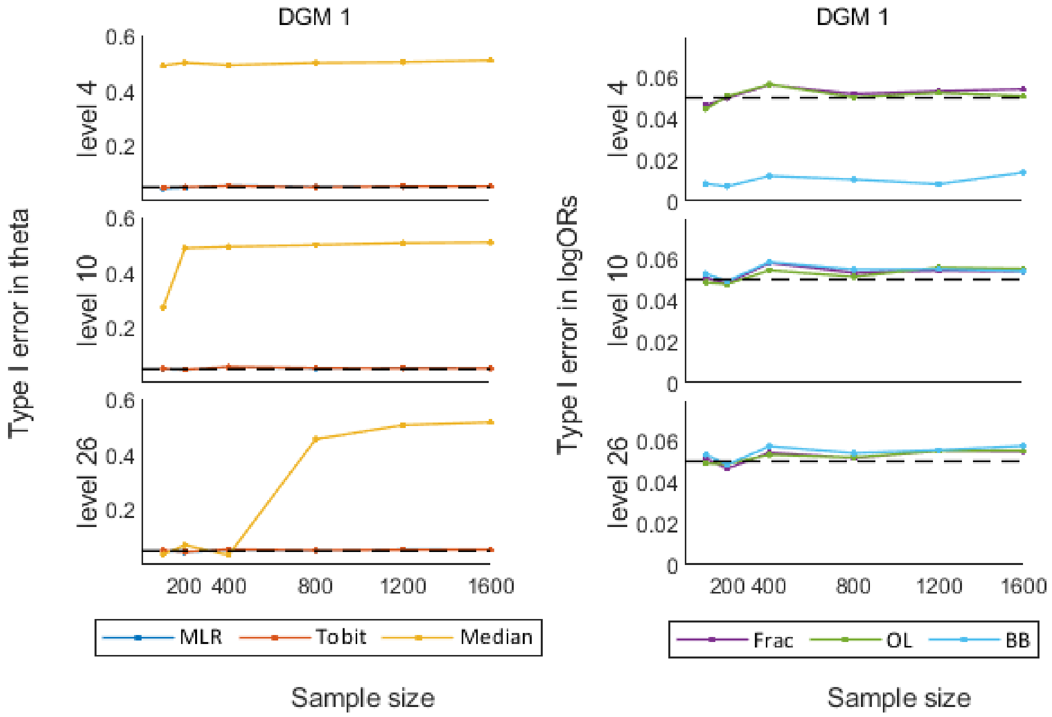
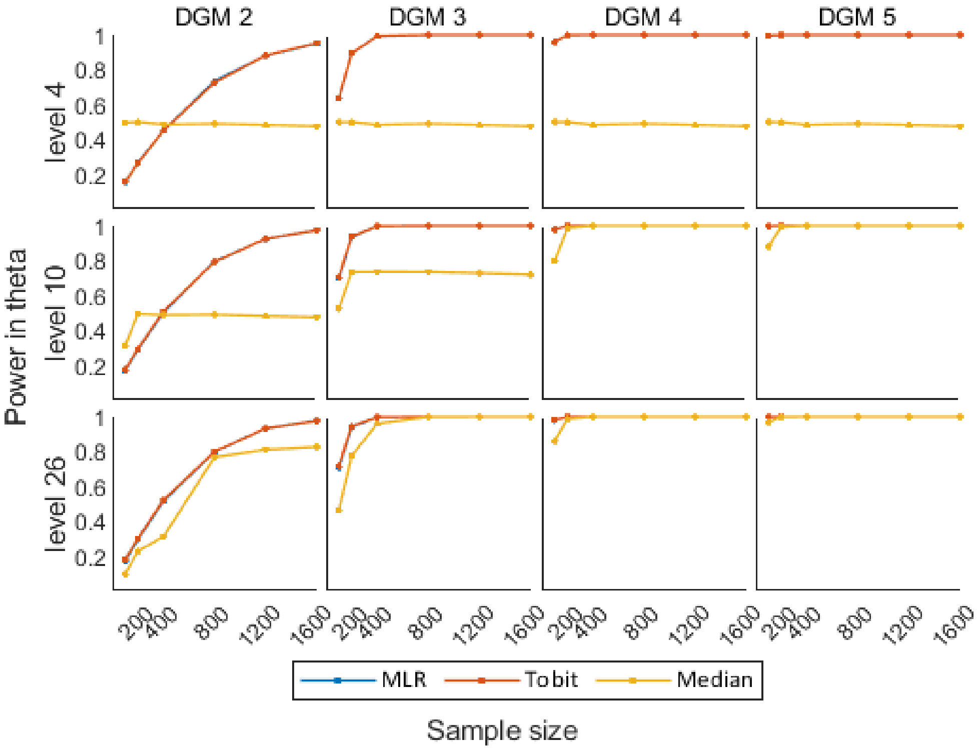


Figure 9.16 Line plots of Type I error for theta or logOR with the change of sample size for different levels under the null hypothesis (DGM 1)

Theta presents the original estimates under the scale-based estimand framework. BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

(a) MLR, Tobit, and Median



(b) Frac, OL, and BB

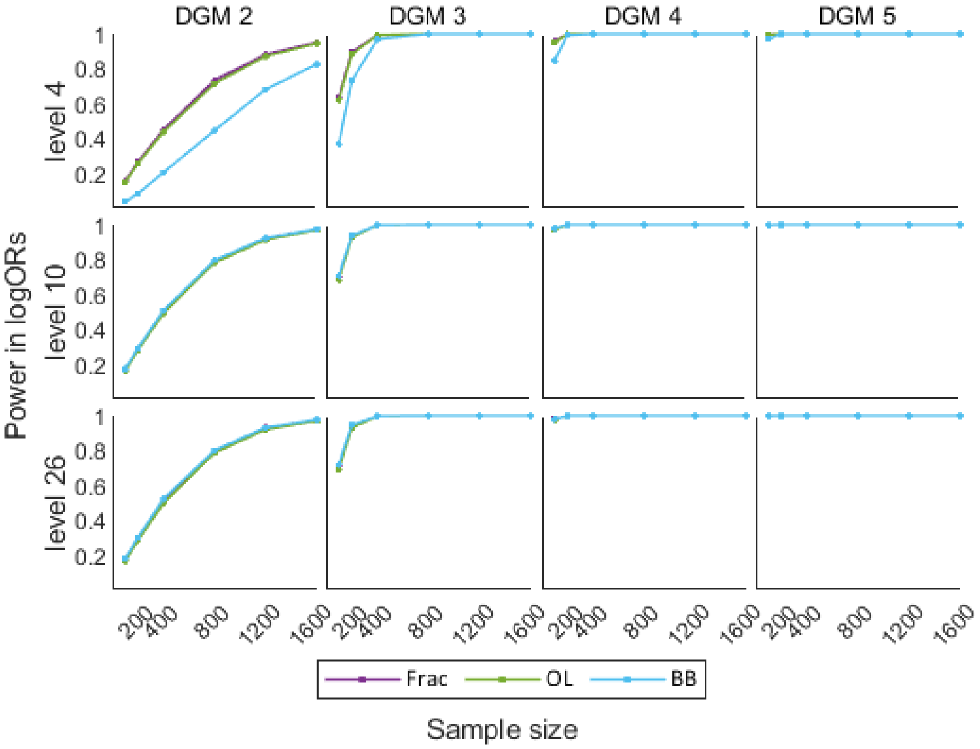


Figure 9.17 Line plots of power for theta or logOR with the change of sample size for different levels under the alternative hypotheses (DGM 2-5)

BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

9.3.2.2 Under the SES estimand framework

9.3.2.2.1 Bias

The estimated SESs in treatment difference for all six methods are compared using the predefined 'truth' in SES, i.e. 0, 0.2, 0.5, 0.8, and 1.0 for five DGMs. Under the null hypothesis, these six methods were able to produce estimated SESs at or around zero. For the majority of the methods, the bias in SESs tended to increase with an increase in the predefined SES, and dimensions with a smaller number of possible values tended to have larger bias. Except for Median, all methods tended to underestimate the predefined SES, and the degree of underestimation was greater in smaller levels. This is shown as the line of each method gradually deviating from the reference line in Figure 9.18. The MLR had the smallest magnitude of bias, followed by Tobit and Frac under level 4 or BB under level 10 and 26. Ignoring Median, BB had the highest bias for level 4, but being replaced by OL for level 10 and 26.

9.3.2.2.2 Mean squared error

Under level 4, Median and BB had larger MSE than other methods, with their trend lines above the majority. When fitting outcome data with more discrete values (i.e. level 10 and 26), the MSEs for Median and BB became smaller and closer to other statistical methods (Figure 9.19). The MSE of MLR, Tobit, Frac, and OL almost overlapped with each other in most scenarios except that the MSE of OL was slightly larger under DGM 4 and 5. Therefore, Median and BB were less precise than other methods when the data was less continuous, and Median and OL were less precise when the predefined treatment effect was large. Median was associated with larger numerically MSEs than other methods.

9.3.2.2.3 Coverage of 95% CIs

Figure 9.20 shows the line plots of coverage measured by SES with the change of n_{obs} for different levels, where the dashed reference line represents coverage = 0.95. Under the null hypothesis, most methods were able to produce the coverage of 0.95, except for Median under all three levels and for BB under level 4. For these methods, the coverage deviated from the reference line with the increase in the predefined 'truth' and the increase in the n_{obs}, and this phenomenon was the most obvious for OL under DGM 4 and 5. With the increase in the number of possible values (i.e. at level 10 and 26), the coverage of OL increased but still remained far from the reference line in DGM 4 and 5.

9.3.2.2.4 Type I error and power

Figure 9.21 presents the Type I error under the null hypothesis (DGM 1) and power under a variety of assumed non-zero true treatment effects measured by SES (DGM 2-5) with the change of n_{obs} for different levels. Under the null hypothesis, most methods can correctly produce the Type I error of 0.05, shown as the lines of these methods overlapped at 0.05, whereas Median had much higher Type I error than other methods. Under the alternative hypotheses, the lines of the majority methods except for Median overlapped and converged to the power of 1 with the increase in n_{obs}, at around 1,600 for DGM 2, 400 for DGM 3, and 200 for DGM 4 and 5. These methods are more likely to have higher power at a given n_{obs} when the predefined treatment effect increased. BB showed similar pattern to other methods under the alternative hypothesis, with an exception that it had less power than the rest of methods at level 4 under DGM 2. The power of Median is less than other methods in most scenarios. Its power remained at 0.5 for all five DGMs at level 4, but it turned to have similar pattern with other methods at level 10 and level 26 when the predefined treatment effect increased (DGM 4 and 5).

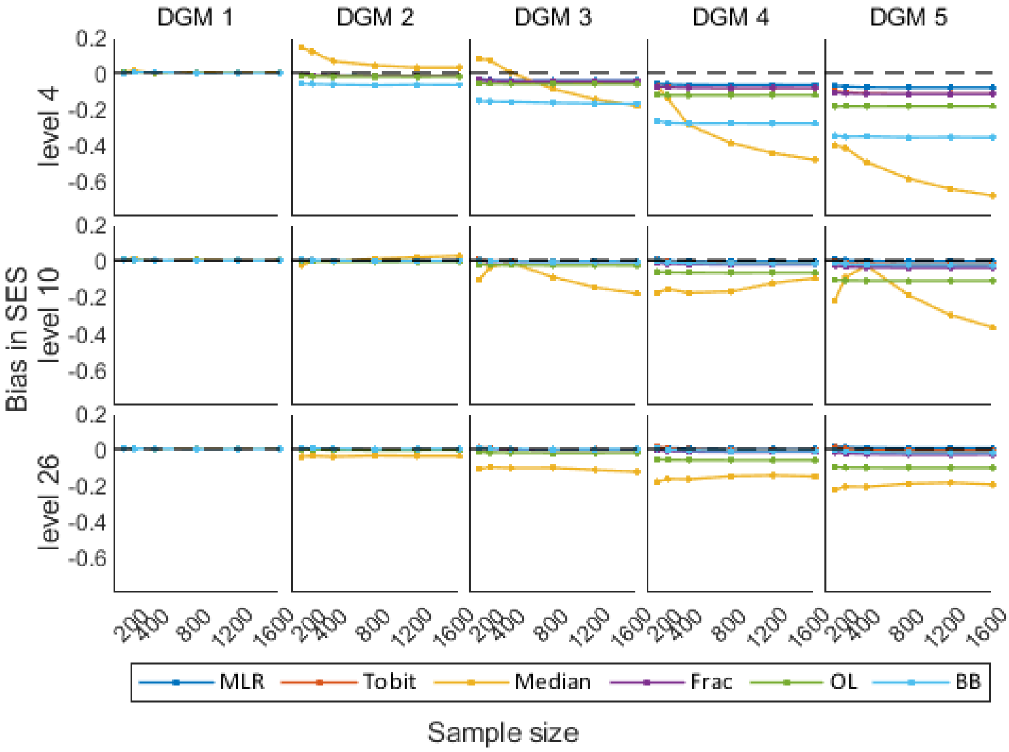


Figure 9.18 Line plots of bias under the SES estimand framework with the change of sample size for different levels

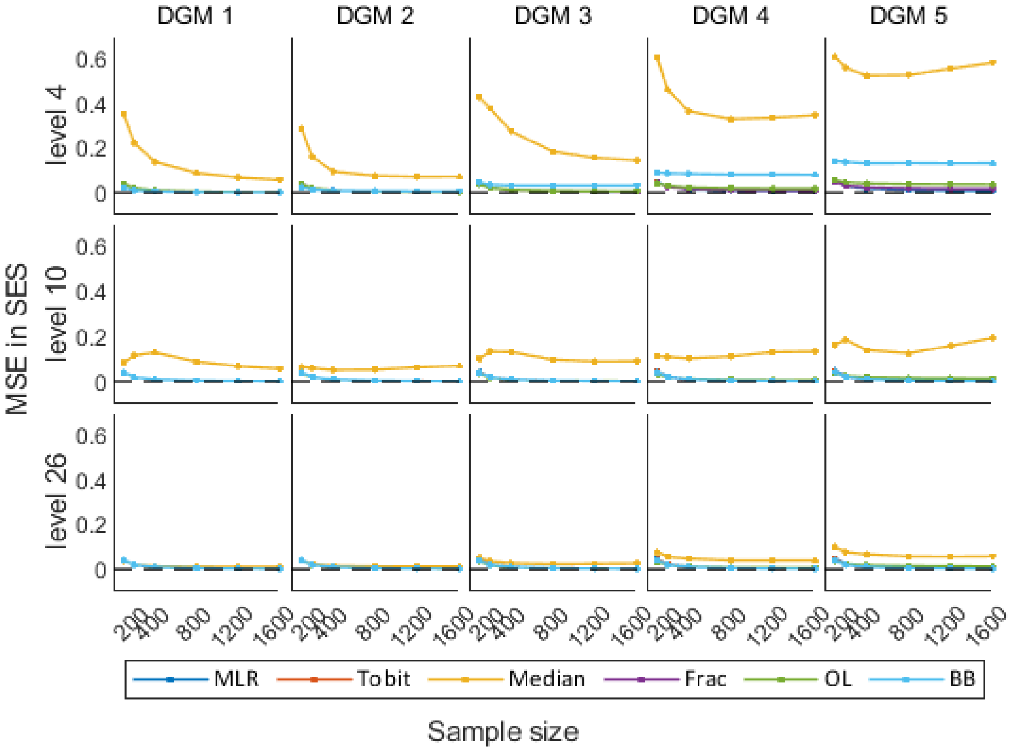


Figure 9.19 Line plots of mean squared error (MSE) under the SES estimand framework with the change of sample size for different levels

BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; SES, standardised effect size; Tobit, Tobit regression.

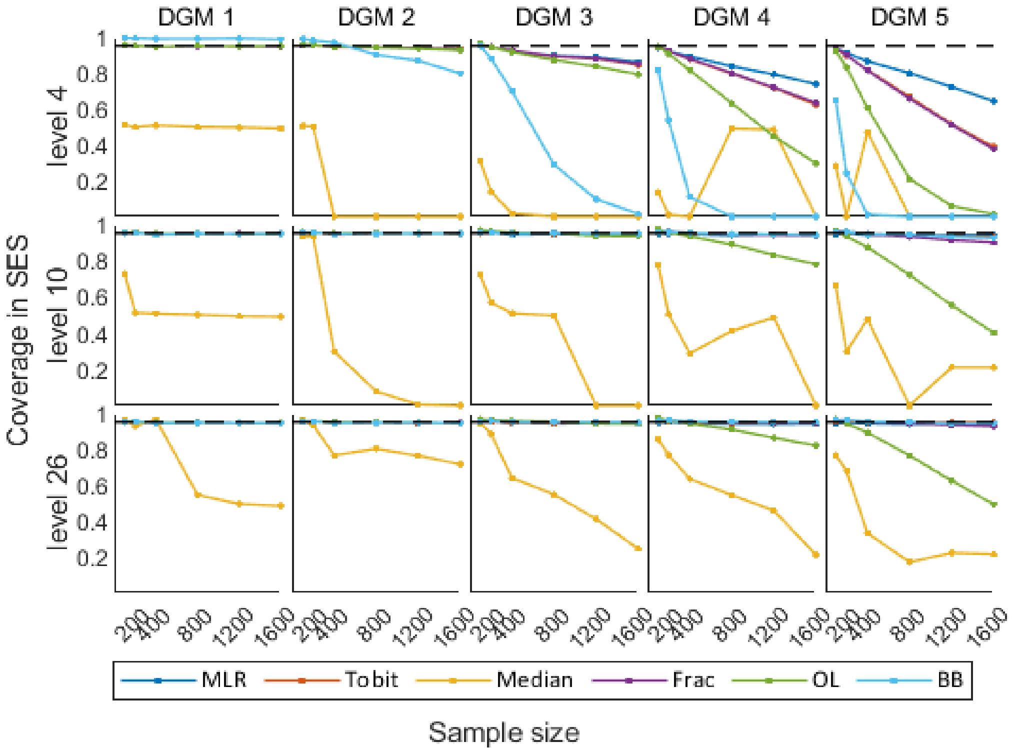


Figure 9.20 Line plots of coverage for SES with the change of sample size for different levels

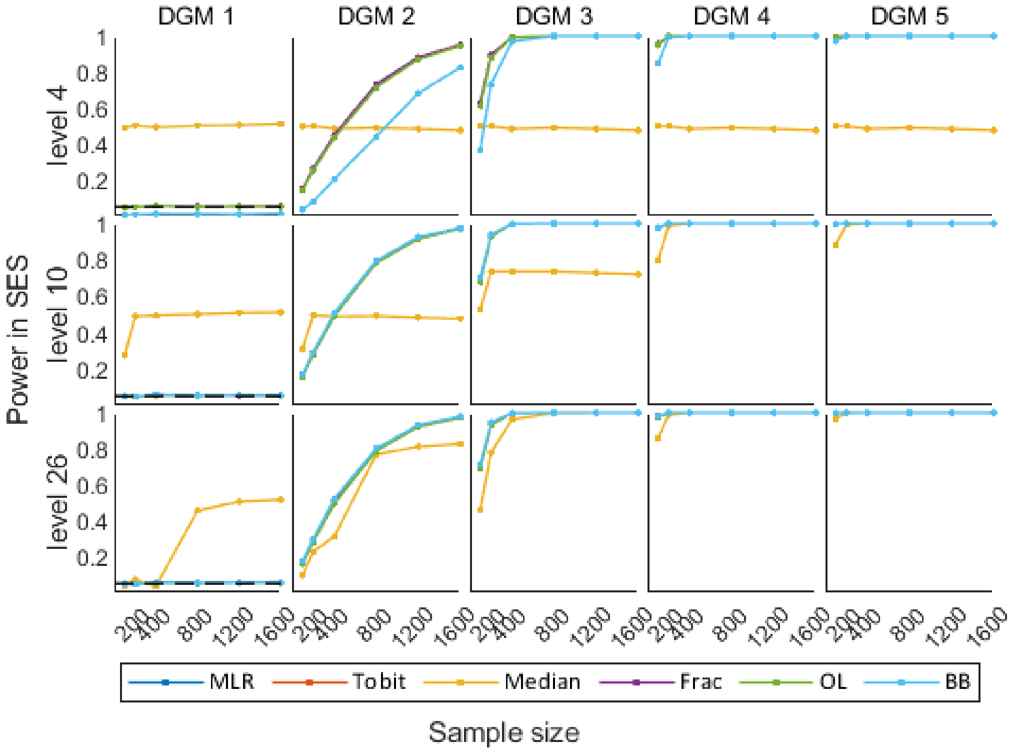


Figure 9.21 Line plots of Type I error and power for SES with the change of sample size for different levels

BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; SES, standardised effect size; Tobit, Tobit regression.

9.4 Discussion

This chapter presents the results of a simulation analysis that was conducted following the simulation protocol established in Chapter 8. It compared the performance of six statistical methods (i.e. MLR, Tobit, Median, Frac, OL, and BB) in estimating the predefined treatment effect of PROs under two estimand frameworks considering a range of scenarios in RCT settings, using the Monte Carlo methods.

The key performance measures of the six statistical methods were compared, i.e. bias, coverage, and Type I error under the null hypothesis of no treatment difference and power under a variety of assumed non-zero true treatment effects, and other performance measures including the MSE, EmpSE, ModSE and relative % error in ModSE were presented.

Among the untransformed scale-based methods (i.e. MLR, Tobit, and Median), MLR performed better than other methods in terms of analysing simulated PRO datasets in RCT settings with a wide range of possible dimension levels. It was associated with little bias in the estimate, small mean squared error, and appropriate coverage of 95% CIs compared to Median and Tobit under most of the simulated scenarios. Tobit had slightly smaller bias with its coverage of 95% CIs closer to the 0.95 reference line than MLR when the predefined treatment difference is large (i.e. DGM 5) for level 10 and 26 under the scale-based estimand framework. However, it had larger bias and worse coverage than MLR when the number of possible categorical values was small (i.e. level 4). Median showed extremely large bias and errors, associated with low power and coverage compared to other statistical methods for most scenarios especially under level 4. The Type I error of Median was found to increase with the sample size under the null hypothesis, which can be explained by the decrease in SE with the increase in sample size, while the biased estimates from Median remain the same due to the discretisation techniques for generating the simulated dataset.

Among the transformed scale-based method (i.e. Frac, BB, and OL) that require the application of discretisation techniques, Frac generally had better performance. It was associated with coverage around two times closer to 0.95 than OL and BB at level 4. Under the SES estimand framework, it had little bias, small errors, and appropriate coverage among the six methods, performing the second best after MLR. OL showed larger magnitude of bias under most of scenarios, and it had much worse coverage than its counterpart with the increase in the sample sizes and predefined treatment difference, which dropped to 0 for level 4, and to around 0.5 for level 10 and 26 under the SES estimand framework. BB produced over 70% missing estimates due to non-convergence when analysing outcomes with a small number of categorical values (i.e. level 4) in this simulation. However, BB had slightly better performance than Frac and OL in other scenarios. The use of BB for PROs with a small number of possible values in other software needs to be further investigated and discussed.

When analysing PRO with four possible ordinal categorical values (i.e. level 4), the missingness caused by the non-convergence of BB was widely seen in Stata. A few options were tried to loosen the criteria for convergence such as decreasing the tolerance level and increasing the number of simulations, or opting for `difficult` option in Stata, but convergence was still not achieved. In addition, due to the non-concaved log likelihood estimation of BB in Stata, it took the maximum number of simulations by default of 300 in Stata, resulting in slow estimation procedure under level 4. Further investigation was made on this issue by retrieving a few non-converged datasets, rerunning them in Stata, and fitting them in an alternative computational software, i.e. R. The results show that Stata still failed to produce converged estimates for BB due to the non-concave log likelihood estimation, but R produced converged estimates for BB using the same datasets. We believe that it is not sensible to merely use R to fit BB and to use Stata to fit other methods, as their estimations may not be consistent (Hodges *et al.*, 2022). After further consideration, we decided not to rerun the entire simulation in R because the command for BB in R is also a user-developed package as the `betabin` command in Stata, such that it would be hard to decide the results of which software package to present if they produced different estimates.

Previous studies (Austin, Escobar and Kopec, 2000; Pullenayegum *et al.*, 2010; Meaney and Moineddin, 2014) have conducted simulation analyses for the comparison of statistical methods for the analysis of PROs, but these studies focused on various groups of methods, proposed different DGMs, and made inconsistent recommendations on what statistical methods are more appropriate to use. This study compared six commonly used or proposed statistical methods for the analysis of PROs under multiple scenarios, with a thorough comparison in the performance measures of these included methods. The outcomes can be extrapolated to other popular PROs similar to SF-36, such as the BDI, HADS, and potentially preference-based PROs such as SF-6D, and EQ-5D.

This simulation study has the following limitations:

First, this simulation study considered 90 scenarios, i.e. five DGMs (five predefined treatment differences) to produce PRO scores under three different number of levels (4, 10, and 26), using six different sample sizes n_{obs}. These scenarios are not able to represent all possible distributions of PRO that would appear in an RCT setting. However, the selection of these parameters was evidence-based, and the same set of parameters were used to compare performance measures across PRO data with different levels. The parameters for the DGMs used the observed mean and SD of the SF-36 pain score that was used as the primary outcome in the Acupuncture trial (Thomas *et al.*, 2006). The categorical values of PROs to simulate (level 4, 10, and 26) were based on the minimum, median, and maximum number of categories of the SF-36v1 dimension scores, and the SF-36 or SF-6D was found to be the most popular PRO in use with the evidence from the published trials in the Health Technology Assessment Journal (Qian *et al.*, 2021). The predefined treatment effect was derived from different degrees of the Cohen’s classified effect size (Sawilowsky, 2009; Cohen, 2013).

Second, the use of Normal distribution assuming an underlying latent variable to generate simulated datasets may favour original scale-based methods such as MLR and Tobit (Morris, White and Crowther, 2019). Alternatively, other distributions could be used to produce the simulated datasets (Meaney and Moineddin, 2014; Najera-Zuloaga, Lee and Arostegui, 2018), for example, the use of beta-binomial distribution to generate PRO data is seen in the simulation analysis comparing two approaches to achieve the BB by Najera-Zuloaga *et al* (Najera-Zuloaga, Lee and Arostegui, 2018), but similarly it would favour the transformed scale-based methods such as BB and OL.

If one believes that the PRO is measuring an underlying latent continuous variable, then use of the continuous Normal distribution for the simulations (albeit followed by a 'discretisation' process to render the outcome more like those observed) may be a reasonable and sensible assumption. However, if one believes the PRO is measuring an underlying latent discrete variable then the beta-binomial distribution may be a more appropriate underlying distribution to base the simulations on. Latent variables are abstract concepts that cannot be directly measured. So unfortunately, we cannot empirically test whether the underlying latent variable for the PRO is continuous or discrete and pragmatically we assumed the former (i.e. it was continuous). The Normal distribution is preferred to the beta-binomial distribution in this simulation analysis not only because of its simplicity and its wide application (Yensy, 2021), but also because we believe that the PRO is by nature continuous.

Third, for the treatment effect we assumed a 'location shift' on the underlying latent continuous Normally distributed PRO, i.e. the underlying distribution for the outcome is 'shifted' to the right so the mean of the new distribution is x-points higher (i.e. 4.4, 11, 17.8 or 22). Again, this assumption may favour statistical methods that assume the outcome is continuous.

This simulation study only looked into the simple 'location shift', i.e. the mean and the SD of the Normal distribution was set at 50 and 22 for all scenarios. As described in Chapter 8, the mean and SDs were set using the evidence from the Acupuncture trial. If the mean was set at a higher value, the location shift would be more influenced by the ceiling effect given the current set of predefined treatment differences, i.e. there would be more values censored at the upper bound, and thus setting the mean at a higher value will make the observed treatment effects farther from the predefined value. If the mean was set at a lower value, there would be more space to move up on the scale, such that the location shift would be less influenced by the ceiling effect given the current set of predefined treatment differences, i.e. the observed treatment effects would be closer to the predefined value. If the baseline PROs were more scattered (i.e. having a comparatively high SD), with the 'location shift' to the right of the scale, the new distribution could be less scattered due to the ceiling effect depending on the degree of 'shift', which would be shown as having a smaller SD than the predefined value. This occurred in the example simulation datasets (Figure 9.1), where the observed SDs in the control and treatment groups were different from each other due to a ceiling effect.

In addition to the ‘location shift’, there could also be a change in the shape of the distribution. For example, when simulating from Normal distributions, the SD could be different in the two groups due to the effect of treatment. However, this simulation study only considered the simple scenario where the SD of two groups were assumed the same. The discretisation procedure is the only factor that may change the shape of the simulated PROs in this simulation besides the ceiling effect.

Fourth, the discretisation techniques were required in this study to simulate the ordinal characteristics of the PRO data. The simulated datasets in this simulation analysis were discretised into equally spaced values, such that whether the conclusion from this simulation is generalizable to non-equally spaced PROs needs to be further investigated. However, there is no standard way to recode the PRO data, particularly for the BB and Frac models. Arostegui *et al.* (2013) have compared three different ways to recode SF-36v1 dimension scores, and proposed an optimal discretisation approach based on the goodness-of-fit. However, their proposed methods are data-dependent and may not be applicable for equally spaced PRO scores, which constrains its generalisability.

Also, the discretisation of the Normally distributed latent PRO into 4, 10, or 26 levels or discrete values may mean that all the statistical methods produce a slightly biased estimate of the true treatment effect (as the observed differences in mean scores from the simulations show). Therefore, this discretisation procedure makes it impossible to observe the exact predefined treatment difference of 4.4, 11, 17.8, 22 points on the underlying Normally distributed scale, except when the predefined treatment difference is set at zero, i.e. under the null hypothesis. Hence, when analysing under the alternative hypothesis, i.e. the predefined treatment difference is not set at zero, the accuracy of these methods, estimated by the bias performance measure, may not be reliable.

Chapter 10 Discussion and conclusion

This thesis, entitled ‘comparison of different statistical methods for the analysis of patient-reported outcomes (PROs) in randomised controlled trials (RCTs)’, aimed to identify, describe, and compare different statistical methods that can be used for the analysis of PROs in RCT settings and make recommendations for the most appropriate statistical methods of analysis.

The identification of available statistical methods for PRO analysis was carried out by conducting two reviews on the statistical methods that have been developed for analysing PROs (Chapter 2), and the statistical methods that have been applied for the analysis of PROs in the UK’s publicly funded RCTs (Chapter 3). The majority of publicly funded RCTs were found to use a PRO as one of their clinical outcomes, and over a third of the RCTs reported using a PRO as the primary outcome. Classical statistical methods such as the *t*-test, multiple linear regression (MLR), and analysis of covariance (ANCOVA) and their extensions to deal with correlated responses, such as mixed models, are widely used for the statistical analysis of PROs. This is despite the fact that complex statistical methods to deal with the bounded, skewed, and ordinal properties of the PROs have been developed and are ready to use.

Chapter 4 specified the research aim and objectives and listed the identified statistical methods that can be considered for the analysis of PROs and the criteria to consider when conducting statistical analysis of PROs in RCTs. Chapter 5 then defined an appropriate statistical method for analysis of PROs in RCTs as one that:

1. Can compare two or more treatment arms;
2. Can adjust for confounding factors, including baseline PRO score;
3. Can produce an estimate of treatment effect and associated confidence intervals (CIs);
4. Can handle a bounded/censored scale;
5. Requires the least amount of recoding to use the statistical method.

The identified list of 29 statistical methods was filtered in Chapter 5 with a series of justifications, and 10 statistical methods remained for further comparison, which includes MLR, Tobit regression (Tobit), Median regression (Median), censored least absolute deviations regression (CLAD), ordered logit model (OL), ordered probit model (OP), beta-binomial regression (BB), binomial-logit-Normal regression (BLN), fractional logistic regression (Frac), and beta regression (BR).

The description of the 10 filtered statistical methods were presented under the generalized linear model (GLM) framework in Chapter 6. An example of an RCT dataset that used Short Form-36 (SF-36) mental health scores as the primary outcome to explain the application of these methods in the computational software Stata and the interpretation of the estimates from each statistical method was also presented. These statistical methods were then applied to nine RCT datasets with SF-36 as clinical outcomes in Chapter 7, and their estimates were compared under two estimand frameworks. Based on their model performances, the 10 statistical methods were narrowed down to six methods, including MLR, Tobit, Median, OL, BB, and Frac.

The comparison of the narrowed list of statistical methods was achieved using Monte Carlo methods to evaluate their accuracy of estimating the predefined treatment effect of PROs under a range of scenarios in RCT settings (Chapter 8 and Chapter 9). The key performance measures i.e. bias, precision, coverage, and power or Type I error of the different methods under multiple scenarios were compared and discussed.

In this chapter, recommendations on the most appropriate statistical methods for the analysis of PROs in RCT settings will be proposed and discussed under different scenarios with the evidence from their technical details (Chapter 6), and their model performances in the empirical analysis (Chapter 7) and the simulation analysis (Chapter 9).

10.1 Recommendations on what statistical methods to use

MLR is recommended as the universal statistical method for the analysis of PROs in RCT settings under the scale-based estimand framework and the standardised effect size (SES) estimand framework. This recommendation is a trade-off on various aspects of the model performance, and is made on the premise that the same statistical method is expected to be used for the analysis of all dimension scores in a multi-dimension PRO such as SF-36. MLR is also recommended by other studies that compared different sets of statistical methods for PRO analysis (Walters and Campbell, 2005; Pullenayegum *et al.*, 2010; Coens *et al.*, 2020). From a medical statistician's point of view, MLR requires no transformation of the response variable, it produces point estimates that are based on the untransformed scale of measurement and are easy to interpret, and is a robust method when faced with the violation of model assumptions (Lumley *et al.*, 2002; Collister *et al.*, 2021), particularly when the population mean and difference in population means between the randomised groups is an appropriate population level summary measure of the treatment effect. From a health economist's point of view, the mean treatment difference in a PRO is commonly used for the calculation of incremental cost-effectiveness ratio (ICER), which represents the additional cost of one unit increase in a PRO to inform the results of a cost-effectiveness analysis, than other estimands such as medians or ORs (Bang and Zhao, 2012).

Other statistical methods can be considered under different scenarios including the types of targeted population summary measures, the assumptions of the PRO distributions, and the number of possible categorical values of a PRO dimension.

Tobit is recommended to analyse PROs with no less than 10 possible categorical values if the average treatment difference is the targeted population summary measure. When analysing a small number of levels, especially level 4, Tobit tended to overestimate the treatment difference, and MLR tended to underestimate the treatment difference, but Tobit was more biased than MLR. The undercoverage of Tobit under level 4 can result from bias, heteroscedasticity, or non-Normality (Pullenayegum *et al.*, 2010). Tobit is shown to have better model performance for the analysis of outcome data with more discrete values (i.e. level 10 and 26), which is evident by previous studies that found Tobit to be consistent and efficient under the Normality assumption of residuals and homoscedasticity (Austin, Escobar and Kopec, 2000; Wilhelm, 2008).

Although Median, a non-parametric statistical method, theoretically makes no assumption about the distributions of the outcome variable, it has been found to fail when the outcome variable is discrete (Padellini and Rue, 2018). Some degree of smoothness should be artificially imposed to apply quantile regression to ordered data, such as adding a uniformly distributed noise to the ordered data (Machado and Santos Silva, 2005). This explains why the simple median regression produced unsatisfactory estimates in striped patterns with poor performance when analysing the small number of possible levels in the empirical analysis and the simulation analysis. CLAD, a censored form of Median, can be used as a substitute to Tobit (Austin, 2002), which is also based on the premise that the latent PRO scores exceeding the low or high boundaries is possible and meaningful (Sullivan, 2011). Similar to Median, it is supposed to have better model performance when facing the non-Normally distributed residuals and homoscedasticity, but it was found to be relatively inefficient in our empirical analysis. This is because CLAD takes extra time to run compared to other methods as it uses bootstrapping to generate CIs, it does not provide p-values directly, and it may not converge on some occasions even when increasing the number of iterations. A study comparing Tobit, Median, and CLAD using the HUI scores found that Median and CLAD tended to produce estimates with similar patterns and their estimates tend to be shrunk to zero compared to MLR and Tobit (Austin, 2002). We therefore do not recommend the use of Median or CLAD for the analysis of PRO data, especially when the number of possible levels is small.

Frac is recommended as the universal statistical method for the analysis of PROs in RCT settings if the (log) odds ratio (ORs) is the preferred population summary measure. In the simulation analysis, the power performance measures of MLR and Frac were found to be similar, which is evident in a simulation study comparing MLR, BR, and Frac in two sample design settings (Meaney and Moineddin, 2014). Among the transformed scale-based methods (i.e. Frac, BB, and OL), it was associated with coverage around two times closer to 0.95 than OL and BB at level 4. Under the SES estimand framework, it had little bias, small errors, and appropriate coverage among the six methods, performing the second best after MLR. However, Frac requires the recoding of PRO scores to apply compared to MLR, which may make it less attractive to use.

BR was not considered for the simulation analysis as, despite it producing similar estimates to Frac, it is not able to account for scores at boundaries and thus requires the ‘squeezing’ of the dimension scores, the compressing process of which is likely to bias the estimations and reduce the precision (Hunger, Baumert and Holle, 2011). This was evident in the empirical analysis where the estimates from BR were more scattered than Frac using MLR as the reference benchmark.

OL and OP, the statistical methods designed for analysing ordered outcomes, were shown to have almost identical SES values and model fit statistics in the empirical analysis. The estimated treatment effect from an OP cannot be explained in (log)ORs, and hence it is less preferable to use. The simulation analysis found that OL tended to generate numerically large bias and small coverage compared to other methods that produce estimates of (log)ORs, especially when the true treatment effect was large (i.e. SES at 0.8 or 1.0). The poor performance and large estimated values of OL may result from the violation of model assumptions in proportional odds, or the different interpretation of the estimated logORs, which needs to be further investigated.

BB is recommended to analyse PROs with no less than 10 possible categorical values if the (log)OR is the preferred population summary measure. BB had slightly better performance than Frac and OL in scenarios where PRO scores have no less than 10 possible categorical values. Similar to Frac, BB requires the recoding or PROs to apply, which may make it less attractive to use. Furthermore, BB produced over 70% missing estimates due to non-convergence when analysing outcomes with a small number of categorical values (i.e. level 4) in the simulation analysis, which is likely to associate with the limitation of the user built code in the computational programming software Stata.

Similarly, BLN was not carried forward to the simulation analysis as it produced similar estimates as BB and there was no command available for running BLN in Stata except for building the code under the general GLM framework using `glm` command. BB and BLN may be preferred to other statistical methods such as Frac or OL as they can account for the ordinal and discrete feature of the PRO scores without the requirement for the distributional or proportional odds assumptions. However, the practical application of BB and BLN requires the availability of commands in computational software.

In summary, considering a simple RCT setting where there is a single baseline and a single post-randomisation assessment of outcome, for the analysis of a multi-dimensional PRO, MLR is recommended as the universal statistical method for the analysis of PROs if the population summary measure of the treatment effect is the difference in group means or the SES. Tobit is recommended as an alternative method to MLR if the number of possible categorical values is 10 or more for a multi-dimensional PRO. Similarly, if the log(OR) is the targeted population summary measure, Frac is recommended as the universal statistical method, and BB is recommended as an alternative method to Frac if the number of possible categorical values is 10 or more in a multi-dimensional PRO.

For the analysis of a unidimensional PRO, if the population summary measure of the treatment effect is the difference in group means or the SES, MLR is recommended as the universal statistical methods for analysis, and Tobit can be used as an alternative for the number of categories in a dimension is 10 or more. If the log(OR) is the targeted population summary measure, Frac is recommended as the universal statistical method, and BB can be used as an alternative when the number of categories in a dimension is 10 or more.

It is worth noting that the application of Tobit regression requires an additional premise that the PRO scores exceeding the lower and upper limits is believed to be meaningful. Both Frac and BB require the application of different recoding techniques to analyse PRO data.

10.2 Main findings and comparison to other work

The strategy for literature search in the method review has been updated using EMBASE, MEDLINE, and EconLit to identify literature that developed, reviewed, and recommended statistical methods for the analysis of PROs between 1 September 2021 and 1 May 2023. No new literature meeting the inclusion criteria, outlined in Chapter 2 was found.

Existing guidance, including the FDA Guideline (FDA, 2009), CONSORT-PRO Extension (Calvert *et al.*, 2013), and SPIRIT-PRO Extension (Calvert *et al.*, 2018), mainly focus on the collection and the reporting of PROs in RCTs. In terms of the statistical analysis of PROs, they provide guidance on what components to report and consider such as the targeted dimensions, the specification of the primary endpoint, and the statistical approaches to deal with the missing data, but they do not provide guidance on what specific statistical methods should be considered to analyse PROs under different scenarios.

The SISAQOL Consortium gathered experts and stakeholders with diverse backgrounds to ratify the statements proposed by four working groups with focuses on research objectives, statistical methods, missing data, and statistical terms (Coens et al., 2020). They recommended the use of Cox proportional-hazards test for evaluating time-to-event data and linear mixed model for evaluating the magnitude of event at a time and a response trajectory over time, and the use of linear regression for evaluating magnitude of event at a specific timepoint. Although their recommendations focused on the analysis of PROs in cancer trials and were purely based on experts' opinions without using evidence from data analysis, their statement is partially in line with our recommendation to use MLR to estimate the treatment effect of PROs for a single baseline and a single post-randomisation assessment of outcome. However, the other statistical methods that are recommended in this thesis such as Tobit, Frac, and BB are neither listed nor assessed by the SISAQOL Consortium.

The existing literature that compared different statistical methods for the analysis of PROs using simulation methods did not reach consensus on the appropriate statistical methods for the analysis of PRO data. For example, Austin (2002) compared CLAD, Tobit, MLR, Median, and other statistical methods for analysing health utility data measured by HUI scores. They found that CLAD and Median produced similar results, and CLAD was recommended because of its low prediction error and its robustness to heteroscedasticity and non-Normality of errors. Pullenayegum et al. (2010) compared MLR, Tobit, and CLAD for the analysis of health utility score measured by EQ-5D in terms of their bias and coverage of CIs, and recommended MLR with robust SEs or the non-parametric bootstrap as a simple and valid approach. Pullenayegum et al. (2011) compared Tobit and CLAD when facing utility decrement, and concluded that these two methods should not be used under this circumstance, and stated that both methods are not appropriate for the analysis of utility decrement. Hunger, Baumert and Holle (2011) compared MLR and BR for analysing SF-6D, and suggested that BR, especially with prevision covariates is a possible supplement to the methods currently used in the analysis of health utility data. Meaney and Moineddin (2014) compared MLR, BR, and Frac to estimate covariate effects on (0,1) response data in terms of bias, variance, Type I error and power, and found these measures were very similar. Kharroubi (2020) compared MLR, BR for analysing SF-6D, and found that BR perform better than MLR in predictive ability. Arostegui, Núñez-Antón and Quintana (2007, 2012) recommended the use of OL with random effects model, BB or BLN for continuous or ordinal PRO data after testing distributional assumptions.

The findings in this thesis partly agreed with previous studies, but different recommendations were drawn because different criteria were considered. For example, we suggest the use of MLR as a universial statistical methods for the analysis of PROs, and Tobit as an altenative to MLR. We do not recommend the use of CLAD and Median due to their poor model performances (e.g. large bias, and undercoverage), which agrees with Pullenayegum *et al.* (2011) but partly disagrees with Austin (2002). Our findings agree with Hunger, Baumert and Holle (2011) that BR produces scattered estimates compared to its counterpart Frac, a possible explanation of which is the ‘squeezing’ process of BR reduced the precision in estimates. This thesis did not propose the use of BLN since the BLN produced higher estimates than other methods, and there is no available code for running BLN in Stata. The BB failed to converge for PRO scores with a small number of possible categorical values, and thus is recommended as an alternative to Frac for scores with over 10 levels if the log (ORs) is the targeted population summary measure.

The recommendation of the most appropriate statistical methods to use in this thesis is based on the results of the technical details of these methods, and their model performance in the empirical analysis and simulation analysis, with the aim of estimating the treatment effect of the latent continuous PROs with equally spaced ordinal categorical scores ranging from 0 to 100 under a two-arm balanced design.

We agree with Bottomley *et al.* (2018) that establishing predefined criteria to assess statistical methods used in PRO analysis is crucial for making scientifically informed choices. The choice of statistical methods for analysing PROs depends on multiple factors, including the nature of the outcome variable, the adherence of the data to the method assumptions, and the criteria set for method evaluation by different stakeholders. The study design and the research question are the fundamental factors and can influence the selection of a statistical method, for example whether the PRO is believed to be measuring an underlying latent variable, and whether the underlying latent variable is believed to be fundamentally continuous or discrete; whether the study design requires the adjustment for clustering, time effects, or unbalanced data; and whether the PRO analysis is to be used for making predictions, estimating a treatment effect, or measuring influencing factors. These factors vary study-by-study and need to be carefully considered when researchers are making the decision on what statistical methods to use.

Finally, we do not recommend the choice of statistical methods simply based on their degree of statistical significance. Our comparison of SES estimates with their associated 95% CIs in the empirical analysis suggested that the choice of statistical methods for data analysis might result in different conclusions drawn from the hypothesis tests in terms of the clinical and statistical significance. It is important to understand different statistical methods before selecting and applying them (Calvert *et al.*, 2018).

10.3 Strengths and limitations

This thesis compared various statistical methods for the analysis of PROs in RCTs using an established set of criteria through their theoretical background, empirical analysis, and simulation analysis. Similar analyses have been carried out by previous studies (Austin, Escobar and Kopec, 2000; Pullenayegum *et al.*, 2010; Meaney and Moineddin, 2014) for the comparison of statistical methods for the analysis of PROs, but these studies focused on different sets of statistical methods, proposed different DGMs, and made inconsistent recommendations on what statistical methods are more appropriate to use.

This thesis conducted a review of publicly funded RCTs published in the HTA Journal, that is, as far as we know, the largest review of trials with 114 studies regarding the statistical methods that have been applied for the primary analysis of PROs in clinical trials. This review analysed the frequency of using PROs and the statistical methods for the analysis of PROs. Together with the method review, it summarised 29 available statistical methods for the analysis of PROs. Two estimand frameworks were proposed in this thesis, i.e. the scale-based estimand framework that allows the presentation of an estimate on its original scale and the SES estimand framework that allows the comparison of estimates from statistical methods on different scales. This thesis compared 10 different statistical methods for the analysis of PROs to nine RCT datasets in various clinical areas using the PRO scores from both versions of the SF-36, and it simulated the distribution of PROs under different scenarios and compared six statistical methods. Recommendations were provided on what statistical methods can be used for the analysis of PROs under different scenarios in RCT settings.

A novel approach to decision making on the statistical methods for PRO analysis using multi-criteria decision analysis (MCDA) was proposed at the filtration stage of the statistical methods. Although it was not fully performed in this thesis in terms of organising expert panel, establishing criteria, eliciting scores and weights, and deliberating on the ordering rank, it provides a possible solution to make the choice of statistical analysis more transparent and consistent.

This thesis has the following limitations:

First, this thesis focused on the dimension scores of SF-36 which were found to be the most used PROs together with SF-6D in UK's publicly funded RCTs (Qian *et al.*, 2021). The results of the simulation analysis were based on the data-generating mechanisms (DGMs) derived from the SF-36 distribution observed in the RCT datasets, with outcomes that have equally spaced scores with the boundaries at 0 and 100, the extrapolation of which to other PROs may require further validation. However, the SF-36 shares similar data features (i.e. discrete, bounded, and skewed) with other PROs, and it may be more prone to ceiling effects or less responsive to subtle changes in some dimensions that are not targeted than a disease-specific measure. Therefore, we believe that the outcomes can be extrapolated to other popular PROs similar to SF-36, such as the Beck Depression Inventory (BDI), Hospital Anxiety and Depression Scale (HADS), European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30), and potentially preference-based PROs such as SF-6D, and EuroQol-5 Dimensions (EQ-5D).

Second, the use of SES as the target population summary measure was proposed in this thesis to unify and compare the estimates from different statistical methods. The application of the SES is based on the premise that the SES derived from the different statistical methods are believed to be comparable which might not be agreed by some researchers. If the SES is not believed as a suitable summary measure to compare outcomes between the different treatment groups, then it may not make sense to compare the estimators on different scales and their associated estimates. Although the SES has been vastly used as the population summary measure in this thesis, it is only seen used in publicly funded clinical trials for sample size calculation, but not for the estimation of treatment effect in PROs (Qian *et al.*, 2021).

Third, we focused on the simple scenarios of PRO analysis where there is a single baseline and a single post-randomisation assessment of outcome in RCT settings. The statistical methods adapted in this thesis were kept in simple and similar forms in RCT settings. In the empirical analysis, we only adjusted for the corresponding baseline score of a PRO dimension to estimate the treatment effect. The baseline score was recorded the same way as the response variable of PRO scores for ordinal and binomial regression into ordinal categorical values, but it was treated as continuous covariate in these regression methods that were included for empirical analysis. In the simulation analysis, the simulated dataset was generated assuming two balanced treatment arms, and the treatment effect was estimated without adjusting for covariates. Although RCTs have charming features that they reduce biases and can inform causality, they cannot bypass the possible systematic error from sources such as invalid measurements e.g. poorly or mistakenly filled PRO forms, ceiling or floor effects of the PRO scales, publication bias or selective reporting of statistical analyses.

Other potential statistical methods that account for time effects, clustering, multivariate analysis, and the effect of missing values, are not included for comparison in this study. These statistical methods were excluded in order to narrow the scope of this thesis to a manageable size due to the workload and timeframe. Accounting for longitudinal data involves another level of complexity and multiple assumptions regarding the time effects, the pattern of PRO scores over multiple time assessments, the correlation structures between the repeated assessments. Trending statistical techniques to deal with time effects or correlated responses include using mixed models or using generalized estimating equation (GEE) for parameter estimation. Other methods such as response feature analysis that uses summary measures and time by time analysis have also been widely used (Walters, 2009). Clustering factor in trials with PROs may include practitioners or hospital sites, the ignorance of which may overestimate the study significance and decrease the study power. Flight *et al.* (2016) compared five approaches for the analysis of individually RCTs using four RCTs that were used in the empirical analysis with clustering in one arm and recommended treating each participant in the unclustered arm as a single cluster. Similarly, the inappropriate handling of missing values in PROs can potentially bias the results of the data analysis. Rombach *et al.* (2018) compared three methods that can be used to deal with missingness in longitudinal PRO data including maximum likelihood, multiple imputation, and inverse probability weighting using simulation analysis, and recommended multiple imputation when additional post randomization information is available.

Additionally, this thesis focuses on the statistical methods under the 'frequentist' theory. Although a few papers on Bayesian statistics (Manuguerra and Heller, 2010; Gheorghe *et al.*, 2017; Lim *et al.*, 2023) and machine learning techniques (Matsangidou *et al.*, 2021; Polce *et al.*, 2021; Katakam *et al.*, 2022; Martin *et al.*, 2022) have been identified in our reviews, these methods were not included for comparison in this thesis.

Furthermore, add-on techniques such as the bootstrapping and the robust SEs were not applied or compared in this thesis. In the HTA review, 6/114 studies applied bootstrapping and 9/114 studies applied robust SEs in the primary analysis. Theoretically, the application of these add-on techniques will not change the point estimate of the treatment effect, but they may influence the p-value, Type I error, and the coverage of 95% CIs. MLR with bootstrapping is recommended to use by Arostegui, Núñez-Antón and Quintana (2012) since bootstrapping is able to detect the statistical significance of an estimation whereas MLR cannot. Similarly, Pullenayegum *et al.* (2011) recommended using MLR with bootstrapping or robust SEs as a simple and valid method for analysing health utility data. In contrast, Walters and Campbell (2005) compared the estimations of SF-36 outcome by bootstrapping and *t*-test, and did not recommend MLR with bootstrapping as the results generated by both methods are similar, indicating that bootstrapping does not add value to the estimation procedure.

Finally, the simulation analysis considered 90 scenarios, and it is not able to represent all possible scenarios of PRO that would appear in an RCT setting. The recommendation to use Tobit or BB for 10 possible categorical values or more is limited to the DGMs that only simulated outcomes with 4, 10, and 26 levels were considered. Therefore, it is possible that Tobit starts performing better than MLR at any other number of categories between 5 and 9. Furthermore, the results of the simulation analysis are based on the potential limited capacity of the user built `betabin` command in Stata. Our investigation on the non-converged estimates by BB under level 4 found that it failed to converge in Stata while converged in R for the same dataset, therefore, the simulation results on BB might change when using a different computational software and the conclusion of this study might be altered. However, even if the estimation procedures of these statistical packages might be different for these statistical methods, they are believed to produce very similar and robust estimates with the commands for other statistical methods that use non user-built command.

10.4 Future research

Possible future research can focus on various topics stemming from this thesis, including the statistical methods for PRO analysis in more complex scenarios, the application of simulation analysis under multiple scenarios in R, and the establishment of a MCDA framework for decisions on statistical methods.

This thesis focused on the simple situation where there is a single baseline and a single post-randomisation assessment of an outcome, and compared the statistical methods that are suitable for such an analysis. There exists an opportunity to explore statistical methods for correlated responses, strategies to deal with missing values, and techniques to estimate clustering effect for the analysis of PROs.

Future research could be extended to replicate the simulation analysis using other computational software, such as R, to see whether the outputs differ from Stata. Our investigation on the non-converged estimates by BB under level 4 found that it failed to converge in Stata while converged in R for the same dataset. Due to the time limit of this thesis, the entire simulation was not rerun in R. With the availability of this information, it would be possible to compare BB to other statistical methods under level 4, and to compare whether the estimates from these statistical methods running in Stata and R are consistent.

In addition, the simulation analysis in this thesis only used the Normal distribution to generate PRO scores and the discretisation techniques were used, resulting in the unknown ‘truth’ under the alternative hypothesis. These choices may result in some of the statistical methods performing better than others. Other distributions such as the beta distribution or beta-binomial distribution can be considered to generate the PRO scores in future research to investigate whether the model performance of each statistical method may change. Furthermore, the location shift parameter i.e. the pre-specified treatment difference was set fixed with the standard deviation (SD) of zero, which may not reflect the reality. It would be interesting to know how these methods would perform with the change in the SD to the pre-specified treatment difference.

Further potential research lies in the establishment of an MCDA framework for the selection of statistical methods for general data analysis. As the MCDA allows trade-offs among different options (i.e. statistical methods) regarding various criteria (i.e. method properties), it will increase the transparency and consistency of the method selection process.

10.5 Conclusions

This thesis identified, described, and compared different statistical methods that can be used for the analysis of PROs in RCT settings. Recommendations have been made for the most appropriate statistical methods of analysis. The recommendations are based on the evidence identified from a series of qualitative and quantitative research including two literature reviews on the statistical methods being developed and used in practice, an empirical analysis and a simulation analysis for the comparison of model performances of different statistical methods. Considering a single baseline and a single post-randomisation assessment of an outcome, MLR is recommended as the universal statistical method for the analysis of PROs in RCT settings if the population summary measure of the treatment effect is the mean or SES. Tobit is recommended as an alternative method for the analysis of PROs with 10 or more possible number of values. Frac is recommended as the universal statistical methods if the log(OR) is the targeted population summary measure, and BB is recommended as an alternative method to Frac if there are 10 or more possible number of values in a PRO dimension. Future research on the comparison and recommendation of statistical methods for PRO analysis should consider in more complex scenarios, the application of the simulation analysis in R, and an additional establishment of a MCDA framework for making decisions on statistical methods.

Bibliography

Ahn, J. and Ahn, H.S. (2020) ‘Bayesian analysis of longitudinal quality of life measures with informative missing data using a selection model’, *Statistical Methods in Medical Research*, 29(5), pp. 1354–1367. Available at: https://doi.org/10.1177/0962280219862001.

Akaike, H. (1974) ‘A New Look at the Statistical Model Identification’, *IEEE Transactions on Automatic Control*, 19(6), pp. 716–723. Available at: https://doi.org/10.1109/TAC.1974.1100705.

Akobeng, A.K. (2005) ‘Understanding randomised controlled trials’, *Archives of Disease in Childhood*. BMJ Publishing Group, pp. 840–844. Available at: https://doi.org/10.1136/adc.2004.058222.

Altman, D.G. (1996) ‘Better reporting of randomised controlled trials: The CONSORT statement’, *British Medical Journal*. BMJ Publishing Group, pp. 570–571. Available at: https://doi.org/10.1136/bmj.313.7057.570.

Altman, D.G. and Royston, P. (2006) ‘The cost of dichotomising continuous variables’, *British Medical Journal*, p. 1080. Available at: https://doi.org/10.1136/bmj.332.7549.1080.

Altun, E. and Turkan, S. (2016) ‘Analysis of Better Life Index of OECD Countries with Multivariate Adaptive Regression Splines Model’, *International Journal of Statistics and Economics*, 17(3), pp. 62–70.

Anota, A. *et al.* (2017) ‘Investigating methodology of patient-reported outcomes data analysis in breast cancer randomized clinical trials.’, *Quality of life research*, 26(1), pp. 117–118. Available at: https://link.springer.com/article/10.1007/s11136-017-1682-x.

Armitage, P., Berry, G. and Matthews, J.N.S. (2002) *Statistical methods in medical research.* 4th editio. Oxford, Blackwell Science.

Arostegui, I., Núñez-Antón, V. and Quintana, J.M. (2007) ‘Analysis of the short form-36 (SF-36): the beta-binomial distribution approach’, *Statistics in Medicine*, 26(6), pp. 1318–1342. Available at: https://doi.org/10.1002/sim.2612.

Arostegui, I., Núñez-Antón, V. and Quintana, J.M. (2012) ‘Statistical approaches to analyse patient-reported outcomes as response variables: An application to health-related quality of life’, *Statistical Methods in Medical Research*, 21(2), pp. 189–214. Available at: https://doi.org/10.1177/0962280210379079.

Arostegui, I., Núñez-Antón, V. and Quintana, J.M. (2013) ‘On the recoding of continuous and bounded indexes to a binomial form: an application to quality-of-life scores’, *Journal of Applied Statistics*, 40(3), pp. 563–582. Available at: https://doi.org/10.1080/02664763.2012.749845.

Austin, P.C. (2002) ‘A comparison of methods for analyzing health-related quality-of-life measures’, *Value in Health*, 5(4), pp. 329–337. Available at: https://doi.org/10.1046/j.1524-4733.2002.54128.x.

Austin, P.C., Escobar, M. and Kopec, J.A. (2000) ‘The use of the Tobit model for analyzing measures of health status’, *Quality of Life Research*, 9(8), pp. 901–910. Available at: https://doi.org/10.1023/A:1008938326604.

Austin, P.C. and Schull, M.J. (2003) ‘Quantile regression: A statistical tool for out-of-hospital research’, *Academic Emergency Medicine*, 10(7), pp. 789–797. Available at: https://doi.org/10.1197/aemj.10.7.789.

Banerjee, S. *et al.* (2013) ‘Study of the use of antidepressants for depression in dementia: The HTA-SADD trial- A multicentre, randomised, double-blind, placebo-controlled trial of the clinical effectiveness and cost-effectiveness of sertraline and mirtazapine’, *Health Technology Assessment*, 17(7), pp. 1–43. Available at: https://doi.org/10.3310/hta17070.

Bang, H. and Zhao, H. (2012) ‘Median-based incremental cost-effectiveness ratio (ICER)’, *Journal of Statistical Theory and Practice*, 6(3), pp. 428–442. Available at: https://doi.org/10.1080/15598608.2012.695571.

Bath, P.M. *et al.* (2018) ‘Triple versus guideline antiplatelet therapy to prevent recurrence after acute ischaemic stroke or transient ischaemic attack: The TARDIS RCT’, *Health Technology Assessment*, 22(48), pp. 1–75. Available at: https://doi.org/10.3310/hta22480.

Beard, D.J. *et al.* (2020) ‘Total versus partial knee replacement in patients with medial compartment knee osteoarthritis: The topkat rct’, *Health Technology Assessment*, 24(20). Available at: https://doi.org/10.3310/hta24200.

Beck, A.T., Steer, R.A. and Carbin, M.G. (1988) ‘Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation’, *Clinical Psychology Review*, 8(1), pp. 77–100. Available at: https://doi.org/10.1016/0272-7358(88)90050-5.

Bedson, E. *et al.* (2014) ‘Folate augmentation of treatment - Evaluation for depression (folated): Randomised trial and economic evaluation’, *Health Technology Assessment*, 18(48), pp. 1–159. Available at: https://doi.org/10.3310/hta18480.

Bell, M.L. *et al.* (2017) ‘Statistical controversies in cancer research: using standardized effect size graphs to enhance interpretability of cancer-related clinical trials with patient-reported outcomes’, *Changes in serum IL-8 levels reflect and predict response to anti-PD-1 treatment in melanoma and NSCLC*, 28(2), pp. 1730–1733. Available at: https://doi.org/10.1093/annonc/mdx064.

Bottomley, A. *et al.* (2018) ‘Moving forward toward standardizing analysis of quality of life data in randomized cancer clinical trials’, *Clinical Trials*, 15(6), pp. 624–630. Available at: https://doi.org/10.1177/1740774518795637.

Bottomley, A., Jones, D. and Claassens, L. (2009) ‘Patient-reported outcomes: Assessment and current perspectives of the guidelines of the Food and Drug Administration and the reflection paper of the European Medicines Agency’, *European Journal of Cancer*, 45(3), pp. 347–353. Available at: https://doi.org/10.1016/j.ejca.2008.09.032.

Boulesteix, A.-L. *et al.* (2020) ‘Introduction to statistical simulations in health research’, *BMJ Open*, 10, p. 39921. Available at: https://doi.org/10.1136/bmjopen-2020-039921.

Boutron, I. *et al.* (2008) ‘Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: Explanation and elaboration’, *Annals of Internal Medicine*. American College of Physicians, pp. 295–309. Available at: https://doi.org/10.7326/0003-4819-148-4-200802190-00008.

Brazier, J. *et al.* (2016) ‘Introduction to the measurement and valuation of health’, in *Measuring and Valuing Health Benefits for Economic Evaluation (2 ed.)*, pp. 7–30. Available at: https://doi.org/10.1093/med/9780198725923.001.0001.

Brazier, J., Roberts, J. and Deverill, M. (2002) ‘The estimation of a preference-based measure of health from the SF-36’, *Journal of Health Economics*, 21(2), pp. 271–292. Available at: https://doi.org/10.1016/S0167-6296(01)00130-8.

Brealey, S. *et al.* (2020) ‘Surgical treatments compared with early structured physiotherapy in secondary care for adults with primary frozen shoulder: The UK frost three-arm RCT’, *Health Technology Assessment*, 24(71), pp. 1–161. Available at: https://doi.org/10.3310/hta24710.

Brittenden, J. *et al.* (2015) ‘Clinical effectiveness and cost-effectiveness of foam sclerotherapy, endovenous laser ablation and surgery for varicose veins: Results from the comparison of LAser, Surgery and foam Sclerotherapy (CLASS) randomised controlled trial’, *Health Technology Assessment*, 19(27), pp. 1–341. Available at: https://doi.org/10.3310/hta19270.

Brombin, C. and Di Serio, C. (2016) 'Evaluating treatment effect within a multivariate stochastic ordering framework: Nonparametric combination methodology applied to a study on multiple sclerosis', *Statistical Methods in Medical Research*, 25(1), pp. 366–384. Available at: https://doi.org/10.1177/0962280212454203.

Calvert, M. *et al.* (2013) 'Reporting of Patient-Reported Outcomes in Randomized Trials', *JAMA*, 309(8), p. 814. Available at: https://doi.org/10.1001/jama.2013.879.

Calvert, M. *et al.* (2018) 'Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols', *JAMA*, 319(5), p. 483. Available at: https://doi.org/10.1001/jama.2017.21903.

Chalder, M. *et al.* (2012) 'A pragmatic randomised controlled trial to evaluate the cost-effectiveness of a physical activity intervention as a treatment for depression: The treating depression with physical activity (TREAD) trial', *Health Technology Assessment*, 16(10). Available at: https://doi.org/10.3310/hta16100.

Chinn, S. (2000) 'A simple method for converting an odds ratio to effect size for use in meta-analysis', *Statistics in Medicine*, 19(22), pp. 3127–3131. Available at: https://doi.org/10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M.

Clare, L. *et al.* (2019) 'Goal-oriented cognitive rehabilitation for early-stage alzheimer's and related dementias: The GREAT RCT', *Health Technology Assessment*, 23(10), pp. 1–244. Available at: https://doi.org/10.3310/hta23100.

Clarke, C.E. *et al.* (2016) 'Clinical effectiveness and cost-effectiveness of physiotherapy and occupational therapy versus no therapy in mild to moderate Parkinson's disease: a large pragmatic randomised controlled trial (PD REHAB)', *Health technology assessment (Winchester, England)*, 20(63), pp. 1–96. Available at: https://doi.org/10.3310/hta20630.

Clarke, P., Gray, A. and Holman, R. (2002) 'Estimating Utility Values for Health States of Type 2 Diabetic Patients Using the EQ-5D (UKPDS 62)', *Medical Decision Making*, 22(4), pp. 340–349. Available at: https://doi.org/10.1177/0272989x0202200412.

Cocks, K., Tharmanathan, P. and Smith, A. (2013) 'Analysis of longitudinal oncology quality of life (QoL) data - are we getting it right?', *Clinical Trials Methodology Conference: Methodology Matters Edinburgh, UK* [Preprint]. Available at: https://doi.org/10.1186/1745-6215-14-S1-P105.

Coens, C. *et al.* (2020) 'International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium', *The Lancet Oncology*, 21(2), pp. e83–e96. Available at: https://doi.org/10.1016/S1470-2045(19)30790-9.

Cohen, J. (2013) *Statistical Power Analysis for the Behavioral Sciences*, *Statistical Power Analysis for the Behavioral Sciences*. Routledge. Available at: https://doi.org/10.4324/9780203771587.

Collister, D. *et al.* (2021) 'Patient reported outcome measures in clinical trials should be initially analyzed as continuous outcomes for statistical significance and responder analyses should be reserved as secondary analyses', *Journal of Clinical Epidemiology*, 134, pp. 95–102. Available at: https://doi.org/10.1016/J.JCLINEPI.2021.01.026.

Conigliani, C., Manca, A. and Tancredi, A. (2015) *Prediction of patient-reported outcome measures via multivariate ordered probit models*, *J. R. Statist. Soc. A*. Available at: www.euroqol.org (Accessed: 12 July 2020).

Cook, J.A. *et al.* (2014) 'Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review', *Health technology assessment (Winchester, England)*, 18(28). Available at: https://doi.org/10.3310/HTA18280.

Coomans, M.B. *et al.* (2020) 'Research objectives, statistical analyses and interpretation of health-related quality of life data in Glioma research: A systematic review', *Cancers*, pp. 1–12. Available at:

https://doi.org/10.3390/cancers12123502.

Cooper, K. *et al.* (2019) 'Laparoscopic supracervical hysterectomy compared with second-generation endometrial ablation for heavy menstrual bleeding: The HEALTH RCT', *Health Technology Assessment*, 23(53). Available at: https://doi.org/10.3310/hta23530.

D'Silva, A. *et al.* (2018) 'Associations of objectively assessed physical activity and sedentary time with health-related quality of life among lung cancer survivors: A quantile regression approach', *Lung Cancer*, 119, pp. 78–84. Available at: https://doi.org/10.1016/j.lungcan.2018.03.010.

Dawson, J., Fitzpatrick, R. and Carr, A. (1996) 'Questionnaire on the perceptions of patients about shoulder surgery', *Journal of Bone and Joint Surgery - Series B*, 78(4), pp. 593–600. Available at: https://doi.org/10.1302/0301-620x.78b4.0780593.

Dennis, M. *et al.* (2006) 'FOOD: A multicentre randomized trial evaluating feeding policies in patients admitted to hospital with a recent stroke', *Health Technology Assessment*, 10(2), pp. 1–91. Available at: https://doi.org/10.3310/hta10020.

Dennis, M. *et al.* (2020) 'Fluoxetine to improve functional outcomes in patients after acute stroke: The focus rct', *Health Technology Assessment*, 24(22), pp. 1–94. Available at: https://doi.org/10.3310/hta24220.

Diaby, V. and Goeree, R. (2014) 'How to use multi-criteria decision analysis methods for reimbursement decision-making in healthcare: A step-by-step guide', *Expert Review of Pharmacoeconomics and Outcomes Research*, 14(1), pp. 81–99. Available at: https://doi.org/10.1586/14737167.2014.859525.

Dodgson, J.S. *et al.* (2009) 'Multi-criteria analysis: a manual.' Available at: https://doi.org/10.1002/mcda.399.

Dolan, P. (1997) 'Modeling Valuations for EuroQol Health States', *Medical Care*, 35(11), pp. 1095–1108. Available at: https://doi.org/10.1097/00005650-199711000-00002.

Fairclough, D.L. (2004) 'Patient reported outcomes as endpoints in medical research', *Statistical Methods in Medical Research*, pp. 115–138. Available at: https://doi.org/10.1191/0962280204sm357ra.

FDA (2009) 'Guidance for Industry Use in Medical Product Development to Support Labeling Claims Guidance for Industry', *Clinical/Medical Federal Register*, (December), pp. 1–39.

Ferrari, S.L.P. and Cribari-Neto, F. (2004) 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics*, 31(7), pp. 799–815. Available at: https://doi.org/10.1080/0266476042000214501.

Fiero, M.H. *et al.* (2019) 'US Food and Drug Administration review of statistical analysis of patient-reported outcomes in lung cancer clinical trials approved between January, 2008, and December, 2017', *The Lancet Oncology*, pp. e582–e589. Available at: https://doi.org/10.1016/S1470-2045(19)30335-3.

Fiteni, F. *et al.* (2016) 'Methodology of health-related quality of life analysis in phase III advanced non-small-cell lung cancer clinical trials: A critical review', *BMC Cancer*, 16(1). Available at: https://doi.org/10.1186/s12885-016-2152-1.

Fiteni, F. *et al.* (2019) 'Health-related quality of life as an endpoint in oncology phase i trials: A systematic review', *BMC Cancer*, 19(1). Available at: https://doi.org/10.1186/s12885-019-5579-3.

Fitzpatrick, R. *et al.* (1998) 'Evaluating patient-based outcome measures for use in clinical trials', *Health Technology Assessment*, 2(14). Available at: https://doi.org/10.3310/hta2140.

Flight, L. *et al.* (2016) 'Recommendations for the analysis of individually randomised controlled trials with clustering in one arm - A case of continuous outcomes', *BMC Medical Research Methodology*, 16(1). Available at: https://doi.org/10.1186/s12874-016-0249-5.

Francis, N.A. *et al.* (2016) 'A randomised placebo-controlled trial of oral and topical antibiotics for children with clinically infected eczema in the community: The ChildRen with eczema, antibiotic management (CREAM) study', *Health Technology Assessment*, 20(19). Available at: https://doi.org/10.3310/hta20190.

Francis, N.A. *et al.* (2020) 'C-reactive protein point-of-care testing for safely reducing antibiotics for acute exacerbations of chronic obstructive pulmonary disease: The PACE RCT', *Health Technology Assessment*, 24(15). Available at: https://doi.org/10.3310/hta24150.

Fu, H. *et al.* (2012) 'A Bayesian approach to the statistical analysis of device preference studies'. Available at: https://doi.org/10.1002/pst.522.

Gariballa, S. *et al.* (2006) 'A Randomized, Double-Blind, Placebo-Controlled Trial of Nutritional Supplementation During Acute Illness', *American Journal of Medicine*, 119(8), pp. 693–699. Available at: https://doi.org/10.1016/J.AMJMED.2005.12.006.

Gazzard, G. *et al.* (2019) 'Selective laser trabeculoplasty versus drops for newly diagnosed ocular hypertension and glaucoma: The LiGHT RCT', *Health Technology Assessment*, 23(31), pp. 1–101. Available at: https://doi.org/10.3310/hta23310.

Ge, X., Peng, Y. and Tu, D. (2020) 'A threshold linear mixed model for identification of treatment-sensitive subsets in a clinical trial based on longitudinal outcomes and a continuous covariate', *Statistical Methods in Medical Research*, 29(10), pp. 2919–2931. Available at: https://doi.org/10.1177/0962280220912772.

Gheorghe, M. *et al.* (2017) 'Health losses at the end of life: a Bayesian mixed beta regression approach', *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 180(3), pp. 723–749. Available at: https://doi.org/10.1111/rssa.12230.

Gilet, H. *et al.* (2011) 'PCN199 An Evaluation of Statistical Methods Used to Analyse Patient-Reported Outcomes (PRO) Data in Published Metastatic Cancer Studies', *Value in Health*, 14(7), p. A471. Available at: https://doi.org/10.1016/j.jval.2011.08.1298.

Glazener, C. *et al.* (2016) 'Clinical effectiveness and cost-effectiveness of surgical options for the management of anterior and/or posterior vaginal wall prolapse: Two randomised controlled trials within a comprehensive cohort study - results from the PROSPECT Study', *Health Technology Assessment*, 20(95). Available at: https://doi.org/10.3310/hta20950.

Goodacre, S. *et al.* (2014) 'The 3Mg trial: a randomised controlled trial of intravenous or nebulised magnesium sulphate versus placebo in adults with acute severe asthma', *Health Technology Assessment*, 18(22). Available at: https://doi.org/10.3310/hta18220.

Goodyer, I.M. *et al.* (2017) 'Cognitive-behavioural therapy and short-term psychoanalytic psychotherapy versus brief psychosocial intervention in adolescents with unipolar major depression (IMPACT): A multicentre, pragmatic, observer-blind, randomised controlled trial', *Health Technology Assessment*, 21(12), pp. 1–93. Available at: https://doi.org/10.3310/hta21120.

Grant, A. *et al.* (2008) 'The effectiveness and cost-effectiveness of minimal access surgery amongst people with gastro-oesophageal reflux disease – a UK collaborative study. The reflux trial', *Health Technology Assessment*, 12(31), pp. 1–218. Available at: https://doi.org/10.1136/bmj.a2664.

Hamel, J.-F. *et al.* (2017) 'A systematic review of the quality of statistical methods employed for analysing quality of life data in cancer randomised controlled trials', *European Journal of Cancer*, 83, pp. 166–176. Available at: https://doi.org/10.1016/j.ejca.2017.06.025.

Hamel, J.F. *et al.* (2017) 'What are the appropriate methods for analyzing patient-reported outcomes in randomized trials when data are missing?', *Statistical Methods in Medical Research*, 26(6), pp. 2897–2908. Available at: https://doi.org/10.1177/0962280215615158.

Hays, R.D., Sherbourne, C.D. and Mazel, R.M. (1993) 'The rand 36-item health survey 1.0', *Health Economics*, 2(3), pp. 217–227. Available at: https://doi.org/10.1002/hec.4730020305.

Hedges, L. V. (1981) ‘Distribution Theory for Glass’s Estimator of Effect Size and Related Estimators’, *Journal of Educational Statistics*, 6(2), p. 107. Available at: https://doi.org/10.2307/1164588.

Heeren, T. and D’Agostino, R. (1987) ‘Robustness of the two independent samples t-test when applied to ordinal scaled data’, *Statistics in Medicine,* 6(1), pp. 79–90. Available at: https://doi.org/10.1002/sim.4780060110.

Herdman, M. *et al.* (2011) ‘Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L)’, *Quality of Life Research,* 20(10), pp. 1727–1736. Available at: https://doi.org/10.1007/s11136-011-9903-x.

Hewison, J. *et al.* (2006) ‘Amniocentesis results: Investigation of anxiety. The ARIA trial’, *Health Technology Assessment,* 10(50). Available at: https://doi.org/10.3310/hta10500.

Hodges, C.B. *et al.* (2022) ‘Researcher degrees of freedom in statistical software contribute to unreliable results: A comparison of nonparametric analyses conducted in SPSS, SAS, Stata, and R’, *Behavior Research Methods*, 1, pp. 1–25. Available at: https://doi.org/10.3758/S13428-022-01932-2/FIGURES/4.

Hong, H. *et al.* (2013) *A Bayesian missing data framework for multiple continuous outcome mixed treatment comparisons., AHRQ Systematic Reviews Original Methods Research Reports*. Agency for Healthcare Research and Quality (US). Available at: https://pubmed.ncbi.nlm.nih.gov/23367529/ (Accessed: 14 July 2020).

Hunger, M., Baumert, J. and Holle, R. (2011) ‘Analysis of SF-6D Index Data: Is Beta Regression Appropriate?’, *Value in Health,* 14(5), pp. 759–767. Available at: https://doi.org/10.1016/J.JVAL.2010.12.009.

Hutton, J.L. and Stanghellini, E. (2011) ‘Modelling bounded health scores with censored skew-normal distributions’, *Statistics in Medicine,* 30(4), pp. 368–376. Available at: https://doi.org/10.1002/sim.4104.

ICH (2021) ‘E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials’, *FDA Guidance Documents*, 9(November), pp. 1–19. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9r1-statistical-principles-clinical-trials-addendum-estimands-and-sensitivity-analysis-clinical (Accessed: 23 April 2023).

Jack, D.S., Prestele, H. and Bakshi, R. (2000) *A double-blind, randomised, controlled study to compare Methotrexate plus Cyclosporine A/Neoral vs. Methotrexate plus Placebo in subjects with early severe Rheumatoid Arthritis*. Basel, Switzerland.

Jenkinson, C. *et al.* (1999) ‘Assessment of the SF-36 version 2 in the United Kingdom’, *Journal of Epidemiology and Community Health,* 53(1), pp. 46–50. Available at: https://doi.org/10.1136/jech.53.1.46.

Jha, S. *et al.* (2018) ‘Impact of pelvic floor muscle training on sexual function of women with urinary incontinence and a comparison of electrical stimulation versus standard treatment (IPSU trial): a randomised controlled trial’, *Physiotherapy,* 104(1), pp. 91–97. Available at: https://doi.org/10.1016/j.physio.2017.06.003.

Kaciroti, N.A. *et al.* (2006) ‘A Bayesian Approach for Clustered Longitudinal Ordinal Outcome With Nonignorable Missing Data’, *Journal of the American Statistical Association*, 101(474), pp. 435–446. Available at: https://doi.org/10.1198/016214505000001221.

Katakam, A. *et al.* (2022) ‘Development of machine learning algorithms to predict achievement of minimal clinically important difference for the KOOS-PS following total knee arthroplasty’, *Journal of Orthopaedic Research,* 40(4), pp. 808–815. Available at: https://doi.org/10.1002/jor.25125.

Kennedy, A.D.M. *et al.* (2003) ‘A multicentre randomised controlled trial assessing the costs and benefits of using structured information and analysis of women’s preferences in the management of menorrhagia’, *Health Technology Assessment,* 7(8). Available at: https://doi.org/10.3310/hta7080.

Kennedy, T.M. *et al.* (2006) 'Cognitive behavioural therapy in addition to antispasmodic therapy for irritable bowel syndrome in primary care: Randomised controlled trial', *Health Technology Assessment*. National Co-ordinating Centre for HTA. Available at: https://doi.org/10.3310/hta10190.

Kerry, S. *et al.* (2000) 'Routine referral for radiography of patients presenting with low back pain: Is patients' outcome influenced by GPs' referral for plain radiography?', *Health Technology Assessment*, 4(20). Available at: https://doi.org/10.3310/hta4200.

Kharroubi, S.A. (2020) 'Analysis of SF-6D Health State Utility Scores: Is Beta Regression Appropriate?', *Healthcare*, 8(4), p. 525. Available at: https://doi.org/10.3390/healthcare8040525.

Kyte, D. *et al.* (2016) 'International Society for Quality of Life Research commentary on the draft European Medicines Agency reflection paper on the use of patient-reported outcome (PRO) measures in oncology studies', *Quality of Life Research*, 25(2), pp. 359–362. Available at: https://doi.org/10.1007/s11136-015-1099-z.

Lall, R. *et al.* (2002) 'A review of ordinal regression models applied on health-related quality of life assessments', *Statistical Methods in Medical Research*, pp. 49–67. Available at: https://doi.org/10.1191/0962280202sm271ra.

Lamb, S.E. *et al.* (2010) 'A multicentred randomised controlled trial of a primary care-based cognitive behavioural programme for low back pain. the back skills training (BeST) trial', *Health Technology Assessment*, 14(41), pp. 1–281. Available at: https://doi.org/10.3310/hta14410.

Langhorne, P. *et al.* (2017) 'A very early rehabilitation trial after stroke (AVERT): a Phase III, multicentre, randomised controlled trial', *Health Technology Assessment*, 21(54), pp. 1–119. Available at: https://doi.org/10.3310/hta21540.

Laucis, N.C., Hays, R.D. and Bhattacharyya, T. (2015) 'Scoring the SF-36 in Orthopaedics: A Brief Guide', *Journal of Bone and Joint Surgery*, 97(19), pp. 1628–1634. Available at: https://doi.org/10.2106/JBJS.O.00030.

Lawrance, R. *et al.* (2020) 'What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials?', *Journal of Patient-Reported Outcomes*, 4(1), pp. 1–8. Available at: https://doi.org/10.1186/S41687-020-00218-5/FIGURES/3.

Lee, K. and Daniels, M.J. (2008) 'Marginalized models for longitudinal ordinal data with application to quality of life studies', *STATISTICS IN MEDICINE Statist. Med*, 27, pp. 4359–4380. Available at: https://doi.org/10.1002/sim.3352.

Lee, K. and Daniels, M.J. (2013) 'Causal inference for bivariate longitudinal quality of life data in presence of death by using global odds ratios', *Statistics in Medicine*, 32(24), pp. 4275–4284. Available at: https://doi.org/10.1002/sim.5857.

Lee, K., Daniels, M.J. and Sargent, D.J. (2010) 'Causal effects of treatments for informative missing data due to progression/death', *Journal of the American Statistical Association*, 105(491), pp. 912–929. Available at: https://doi.org/10.1198/jasa.2010.ap08739.

Leng, C. and Zhang, W. (2014) 'Smoothing combined estimating equations in quantile regression for longitudinal data', *Statistics and Computing*, 24(1), pp. 123–136. Available at: https://doi.org/10.1007/s11222-012-9358-0.

Lewis, S.W. *et al.* (2006) 'Randomised controlled trials of conventional antipsychotic versus new atypical drugs, and new atypical drugs versus clozapine, in people with schizophrenia responding poorly to, or intolerant of, current drug treatment', *Health Technology Assessment*, 10(17), pp. 1–94. Available at: https://doi.org/10.3310/hta10170.

Liang, K.Y. and Zeger, S.L. (1986) 'Longitudinal data analysis using generalized linear models', *Biometrika*, 73(1), pp. 13–22. Available at: https://doi.org/10.1093/biomet/73.1.13.

Liang, Y. *et al.* (2014) 'Modeling Bounded Outcome Scores Using The Binomial-Logit-Normal Distribution', *Chilean Journal of Statistics*, 5(2), pp. 3–14. Available at: http://www.soche.cl/chjs (Accessed: 17 August 2021).

Lim, W. *et al.* (2023) 'Bayesian semiparametric joint modeling of a count outcome and inconveniently timed longitudinal predictors', *Statistical Methods in Medical Research* [Preprint]. Available at: https://doi.org/10.1177/09622802231154325.

Little, P. *et al.* (2009) 'Dipsticks and diagnostic algorithms in urinary tract infection: Development and validation, randomised trial, economic analysis, observational cohort and qualitative study', *Health Technology Assessment*, 13(19). Available at: https://doi.org/10.3310/hta13190.

Little, P. *et al.* (2014) 'PRImary care Streptococcal Management (PRISM) study: In vitro study, diagnostic cohorts and a pragmatic adaptive randomised controlled trial with nested qualitative study and cost-effectiveness study', *Health Technology Assessment*, 18(6), pp. 1–101. Available at: https://doi.org/10.3310/hta18060.

Little, R.J. and Lewis, R.J. (2021) 'Estimands, Estimators, and Estimates', *JAMA - Journal of the American Medical Association*. American Medical Association, pp. 967–968. Available at: https://doi.org/10.1001/jama.2021.2886.

Lumley, T. *et al.* (2002) 'The importance of the normality assumption in large public health data sets', *Annual Review of Public Health*, 23(1), pp. 151–169. Available at: https://doi.org/10.1146/annurev.publhealth.23.100901.140546.

Lv, Y. *et al.* (2019) 'Quantile regression and empirical likelihood for the analysis longitudinal data with monotone missing responses due to dropout, with applications to quality of life measurements from clinical trials', *Statistics in Medicine* [Preprint]. Available at: https://doi.org/10.1002/sim.8152.

Machado, J.A.F. and Santos Silva, J.M.C. (2005) 'Quantiles for counts', *Journal of the American Statistical Association*, 100(472), pp. 1226–1237. Available at: https://doi.org/10.1198/016214505000000330.

Majewska, R. *et al.* (2019) 'Pns322 Review of Methodological Approaches To Analyse Longitudinal Utility Data', *Value in Health*, 22(November), p. S818. Available at: https://doi.org/10.1016/j.jval.2019.09.2222.

Manuguerra, M. and Heller, G.Z. (2010) 'Ordinal Regression Models for Continuous Scales', *The International Journal of Biostatistics*, 6(1). Available at: https://doi.org/10.2202/1557-4679.1230.

Marandino, L. *et al.* (2018) 'Deficiencies in health-related quality-of-life assessment and reporting: a systematic review of oncology randomized phase III trials published between 2012 and 2016'. Available at: https://doi.org/10.1093/annonc/mdy449.

Marinacci, C. *et al.* (2001) 'Application of random effect ordinal regression model for outcome evaluation of two randomized controlled trials ‡', *STATISTICS IN MEDICINE Statist. Med*, 20, pp. 3769–3776. Available at: https://doi.org/10.1002/sim.1170.

Marsh, K. *et al.* (2016) 'Multiple Criteria Decision Analysis for Health Care Decision Making - Emerging Good Practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force', *Value in Health*, 19(2), pp. 125–137. Available at: https://doi.org/10.1016/j.jval.2015.12.016.

Martin, R.K. *et al.* (2022) 'Predicting subjective failure of ACL reconstruction: a machine learning analysis of the Norwegian Knee Ligament Register and patient reported outcomes', *Journal of ISAKOS*, 7(3), pp. 1–9. Available at: https://doi.org/10.1016/j.jisako.2021.12.005.

Maruish, M.E. (2011) *User's manual for the SF-36v2 Health Survey.* 3rd ed. Lincoln: QualityMetric Inc.

Matsangidou, M. *et al.* (2021) 'Machine Learning in Pain Medicine: An Up-To-Date Systematic

Review', *Pain and Therapy*. Pain Ther, pp. 1067–1084. Available at: https://doi.org/10.1007/s40122-021-00324-2.

Matthews, J.N.S. *et al.* (1990) 'Analysis of serial measurements in medical research', *British Medical Journal*, 300(6719), pp. 230–235. Available at: https://doi.org/10.1136/bmj.300.6719.230.

McHorney, C.A. *et al.* (1994) 'The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups', *Medical Care*, 32(1), pp. 40–66. Available at: https://doi.org/10.1097/00005650-199401000-00004.

McHorney, C.A., Ware, J.E. and Raczek, A.E. (1993) 'The MOS 36-item short-form health survey (Sf-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs', *Medical Care*, 31(3), pp. 247–263. Available at: https://doi.org/10.1097/00005650-199303000-00006.

Meaney, C. and Moineddin, R. (2014) 'A Monte Carlo simulation study comparing linear regression, beta regression, variable-dispersion beta regression and fractional logit regression at recovering average difference measures in a two sample design', *BMC Medical Research Methodology*, 14(1), pp. 1–22. Available at: https://doi.org/10.1186/1471-2288-14-14/TABLES/6.

Michaels, J.A. *et al.* (2006) 'Randomised clinical trial, observational study and assessment of cost-effectiveness of the treatment of varicose veins (REACTIV trial)', *Health Technology Assessment*, 10(13), pp. 1–114. Available at: https://doi.org/10.3310/hta10130.

Mihaylov, S. *et al.* (2008) 'Stepped treatment of older adults on laxatives. The STOOL trial', *Health Technology Assessment*, 12(13). Available at: https://doi.org/10.3310/hta12130.

Mishra, K.K. and Ghosh, S.K. (2009) *Bayesian Regression Models for the Quality Adjusted Lifetime Data with Zero Time Duration Health States, Statistical Theory and Practice*.

Mitchell, C. *et al.* (2005) 'Costs and effectiveness of pre- and post-operative home physiotherapy for total knee replacement: Randomized controlled trial', *Journal of Evaluation in Clinical Practice*, 11(3), pp. 283–292. Available at: https://doi.org/10.1111/J.1365-2753.2005.00535.X.

Moerkerke, B. *et al.* (2005) 'Permutation based methods for comparing quality of life between observed treatments', *Statistics in Medicine*, 24(24), pp. 4055–4066. Available at: https://doi.org/10.1002/sim.2395.

Moher, D. *et al.* (2010) 'CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials', *BMJ*, 340(mar23 1), pp. c869–c869. Available at: https://doi.org/10.1136/bmj.c869.

Mokkink, L.B. *et al.* (2010) 'The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study', *Quality of Life Research*, 19(4), pp. 539–549. Available at: https://doi.org/10.1007/s11136-010-9606-8.

Molassiotis, A. *et al.* (2013) 'The effectiveness and cost-effectiveness of acupressure for the control and management of chemotherapy-related acute and delayed nausea: Assessment of Nausea in Chemotherapy Research (ANCHoR), a randomised controlled trial', *Health Technology Assessment*, 17(26), pp. 1–114. Available at: https://doi.org/10.3310/hta17260.

Montedori, A. *et al.* (2011) 'Modified versus standard intention-to-treat reporting: Are there differences in methodological quality, sponsorship, and findings in randomized trials? A cross-sectional study', *Trials*, 12, p. 58. Available at: https://doi.org/10.1186/1745-6215-12-58.

Morrell, C.J. *et al.* (1998) 'Cost effectiveness of community leg ulcer clinics: Randomised controlled trial', *British Medical Journal*, 316(7143), pp. 1487–1491. Available at: https://doi.org/10.1136/bmj.316.7143.1487.

Morrell, C.J. *et al.* (2000) 'Costs and benefits of community postnatal support workers: A randomised controlled trial', *Health Technology Assessment*, 4(6). Available at: https://doi.org/10.3310/hta4060.

Morris, T.P., White, I.R. and Crowther, M.J. (2019) 'Using simulation studies to evaluate statistical methods', *Statistics in Medicine*, 38(11), pp. 2074–2102. Available at: https://doi.org/10.1002/sim.8086.

Mountain, G. *et al.* (2017) 'A preventative lifestyle intervention for older adults (lifestyle matters): a randomised controlled trial', *Age and Ageing*, 46(4), pp. 627–634. Available at: https://doi.org/10.1093/ageing/afx021.

Mountain, G.A. *et al.* (2014) '"Putting Life in Years" (PLINY) telephone friendship groups research study: pilot randomised controlled trial', *Trials*, 15(1), p. 141. Available at: https://doi.org/10.1186/1745-6215-15-141.

Najera-Zuloaga, J., Lee, D.J. and Arostegui, I. (2018) 'Comparison of beta-binomial regression model approaches to analyze health-related quality of life data', *Statistical Methods in Medical Research*, 27(10), pp. 2989–3009. Available at: https://doi.org/10.1177/0962280217690413.

Najera-Zuloaga, J., Lee, D.J. and Arostegui, I. (2019) 'A beta-binomial mixed-effects model approach for analysing longitudinal discrete and bounded outcomes', *Biometrical Journal*, 61(3), pp. 600–615. Available at: https://doi.org/10.1002/bimj.201700251.

Nelder, J.A. and Wedderburn, R.W.M. (1972) 'Generalized Linear Models', *Journal of the Royal Statistical Society. Series A (General)*, 135(3), p. 370. Available at: https://doi.org/10.2307/2344614.

NHS Digital (2022) *Patient Reported Outcome Measures (PROMs) - NHS Digital, Online*. Available at: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/patient-reported-outcome-measures-proms (Accessed: 26 May 2020).

Nielsen, L.K. *et al.* (2019) 'Methodological aspects of health-related quality of life measurement and analysis in patients with multiple myeloma', *British Journal of Haematology*, pp. 11–24. Available at: https://doi.org/10.1111/bjh.15759.

O'Kelly, M. *et al.* (2019) 'PNS294 REVIEW OF ANALYTICAL METHODS FOR ANALYSIS OF PATIENT REPORTED OUTCOME (PRO) DATA IN THE PRESENCE OF CENSORING DUE TO DEATH', *Value in Health*, 22, p. S813. Available at: https://doi.org/10.1016/j.jval.2019.09.2194.

Orgeta, V. *et al.* (2015) 'Individual cognitive stimulation therapy for dementia: A clinical effectiveness and cost-effectiveness pragmatic, multicentre, randomised controlled trial', *Health Technology Assessment*, 19(64), pp. 7–73. Available at: https://doi.org/10.3310/hta19640.

Padellini, T. and Rue, H. (2018) 'Model-aware Quantile Regression for Discrete Data'. Available at: https://doi.org/10.48550/arXiv.1804.03714.

Page, M.J. *et al.* (2021) 'The PRISMA 2020 statement: An updated guideline for reporting systematic reviews', *The BMJ*. Available at: https://doi.org/10.1136/bmj.n71.

Papke, L.E. (1996) 'Econometric methods for fractional response variables with an application to 401 (k) plan participation rates', *Journal of Applied Econometrics*, 11(6), pp. 619–632. Available at: https://doi.org/10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1.

Parsons, N. *et al.* (2014) 'Standardised effect sizes in clinical research', *The Bone & Joint Journal*, 96-B(7), pp. 853–854. Available at: https://doi.org/10.1302/0301-620x.96b7.34109.

Parsons, N.R. (2013) 'Proportional-odds models for repeated composite and long ordinal outcomescales', *Statistics in Medicine*, 32(18), pp. 3181–3191. Available at: https://doi.org/10.1002/sim.5756.

Pe, M. *et al.* (2018) 'Statistical analysis of patient-reported outcome data in randomised controlled trials of locally advanced and metastatic breast cancer: a systematic review', *The Lancet Oncology*, 19(9), pp. e459–e469. Available at: https://doi.org/10.1016/S1470-2045(18)30418-2.

Peveler, R. *et al.* (2005) 'A randomised controlled trial to compare the cost-effectiveness of tricyclic antidepressants, selective serotonin reuptake inhibitors and lofepramine.', *Health Technology*

Assessment, 9(16). Available at: https://doi.org/10.3310/hta9160.

Pickard, R. *et al.* (2020) ‘Open urethroplasty versus endoscopic urethrotomy for recurrent urethral stricture in men: The open rct’, *Health Technology Assessment*, 24(61), pp. 1–110. Available at: https://doi.org/10.3310/hta24610.

Pinheiro, J. and Bates, D. (2000) *Mixed-Effects Models in S and S-PLUS*, *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag. Available at: https://doi.org/10.1007/b98882.

Polce, E.M. *et al.* (2021) ‘Development of supervised machine learning algorithms for prediction of satisfaction at 2 years following total shoulder arthroplasty’, *Journal of Shoulder and Elbow Surgery*, 30(6), pp. e290–e299. Available at: https://doi.org/10.1016/j.jse.2020.09.007.

Prescott, R.J. *et al.* (2007) ‘A randomised controlled trial of postoperative radiotherapy following breast-conserving surgery in a minimum-risk older population. The PRIME trial’, *Health Technology Assessment*, 11(31). Available at: https://doi.org/10.3310/hta11310.

Pullenayegum, E.M. *et al.* (2010) ‘Analysis of health utility data when some subjects attain the upper bound of 1: Are tobit and CLAD models appropriate?’, *Value in Health*, 13(4), pp. 487–494. Available at: https://doi.org/10.1111/j.1524-4733.2010.00695.x.

Pullenayegum, E.M. *et al.* (2011) ‘Calculating Utility Decrements Associated With an Adverse Event: Marginal Tobit and CLAD Coefficients Should Be Used With Caution’, *Medical Decision Making*, 31(6), pp. 790–799. Available at: https://doi.org/10.1177/0272989x11393284.

Qian, W. *et al.* (2000) ‘Analysis of messy longitudinal data from a randomized clinical trial’, *Statistics in Medicine*, 19(19), pp. 2657–2674. Available at: https://doi.org/10.1002/1097-0258(20001015)19:19<2657::AID-SIM557>3.0.CO;2-3.

Qian, Y. *et al.* (2021) ‘Comprehensive review of statistical methods for analysing patient-reported outcomes (PROs) used as primary outcomes in randomised controlled trials (RCTs) published by the UK’s Health Technology Assessment (HTA) journal (1997-2020)’, *BMJ Open*, p. 51673. Available at: https://doi.org/10.1136/bmjopen-2021-051673.

Quintana, M. *et al.* (2019) ‘Bayesian model of disease progression in GNE myopathy’, *Statistics in Medicine*, 38(8), pp. 1459–1474. Available at: https://doi.org/10.1002/sim.8050.

Ribaudo, H.J. and Thompson, S.G. (2002) ‘The analysis of repeated multivariate binary quality of life data: A hierarchical model approach’, *Statistical Methods in Medical Research*, pp. 69–83. Available at: https://doi.org/10.1191/0962280202sm272ra.

Rodriguez, R.N., Yao, Y. and Inc, S.I. (2017) ‘Five Things You Should Know about Quantile Regression’, in *Proceedings of the SAS global forum 2017 conference*, pp. 2–5. Available at: https://support.sas.com/resources/papers/proceedings17/SAS0525-2017.pdf (Accessed: 8 March 2022).

Rombach, I. *et al.* (2018) ‘Comparison of statistical approaches for analyzing incomplete longitudinal patient-reported outcome data in randomized controlled trials’, *Patient Related Outcome Measures*, Volume 9, pp. 197–209. Available at: https://doi.org/10.2147/prom.s147790.

Rothwell, J.C., Julious, S.A. and Cooper, C.L. (2018) ‘A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal’, *Trials*, 19(1). Available at: https://doi.org/10.1186/s13063-018-2886-y.

Russell, I. *et al.* (2013) ‘Cancer of oesophagus or gastricus: New assessment of technology of endosonography (COGNATE): Report of pragmatic randomised trial’, *Health Technology Assessment*, 17(39), pp. 1–13. Available at: https://doi.org/10.3310/hta17390.

Sajobi, T.T. *et al.* (2018) ‘Scoping review of response shift methods: current reporting practices and recommendations’, *Quality of Life Research*, pp. 1133–1146. Available at: https://doi.org/10.1007/s11136-017-1751-x.

Santer, M. *et al.* (2018) ‘Adding emollient bath additives to standard eczema management for children with eczema: The BATHE RCT’, *Health Technology Assessment*, 22(57), pp. 1–116. Available at: https://doi.org/10.3310/hta22570.

Saver, J.L. (2011) ‘Optimal End Points for Acute Stroke Therapy Trials’, *Stroke*, 42(8), pp. 2356–2362. Available at: https://doi.org/10.1161/STROKEAHA.111.619122.

Sawilowsky, S.S. (2009) ‘New Effect Size Rules of Thumb’, *Journal of Modern Applied Statistical Methods*, 8(2), pp. 597–599. Available at: https://doi.org/10.22237/jmasm/1257035100.

Schober, P. and Vetter, T.R. (2018) ‘Repeated measures designs and analysis of longitudinal data: If at first you do not succeed-try, try again’, *Anesthesia and Analgesia*, 127(2), pp. 569–575. Available at: https://doi.org/10.1213/ANE.0000000000003511.

Sharples, L. *et al.* (2018) ‘Amaze: A double-blind, multicentre randomised controlled trial to investigate the clinical effectiveness and cost-effectiveness of adding an ablation device-based maze procedure as an adjunct to routine cardiac surgery for patients with pre-existing atrial fibrillation’, *Health Technology Assessment*, 22(19). Available at: https://doi.org/10.3310/hta22190.

Shawo, L. *et al.* (2020) ‘An extended stroke rehabilitation service for people who have had a stroke: The extras rct’, *Health Technology Assessment*, 24(24), pp. 1–202. Available at: https://doi.org/10.3310/hta24240.

Shields, A. *et al.* (2015) ‘Patient-reported outcomes for US oncology labeling: Review and discussion of score interpretation and analysis methods’, *Expert Review of Pharmacoeconomics and Outcomes Research*, 15(6), pp. 951–959. Available at: https://doi.org/10.1586/14737167.2015.1115348.

Simpson, W.M. *et al.* (1999) ‘A randomised controlled trial of different approaches to universal antenatal HIV testing: Uptake and acceptability’, *Health Technology Assessment*, 3(4).

Smithson, M. and Verkuilen, J. (2006) ‘A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables.’, *Psychological Methods*, 11(1), pp. 54–71. Available at: https://doi.org/10.1037/1082-989X.11.1.54.

Smuk, M., Carpenter, J.R. and Morris, T.P. (2017) ‘What impact do assumptions about missing data have on conclusions? A practical sensitivity analysis for a cancer survival registry’, *BMC Medical Research Methodology*, 17(1), pp. 1–9. Available at: https://doi.org/10.1186/s12874-017-0301-0.

Spitzer, R.L., Kroenke, K. and Williams, J.B.W. (1999) ‘Validation and utility of a self-report version of PRIME-MD: The PHQ Primary Care Study’, *Journal of the American Medical Association*, 282(18), pp. 1737–1744. Available at: https://doi.org/10.1001/jama.282.18.1737.

Sprigg, N. *et al.* (2019) ‘Tranexamic acid to improve functional status in adults with spontaneous intracerebral haemorrhage: The TICH-2 RCT’, *Health Technology Assessment*, 23(35), pp. vii–48. Available at: https://doi.org/10.3310/hta23350.

Steen, K. Van, Curran, D. and Molenberghs, G. (2001) ‘Sensitivity analysis of longitudinal binary quality of life data with drop-out: An example using the EORTC QLQ-C30’, *Statistics in Medicine*, 20(24), pp. 3901–3920. Available at: https://doi.org/10.1002/sim.1081.

Sullivan, L.M. and D’Agostino, R.B. (2003) ‘Robustness and power of analysis of covariance applied to ordinal scaled data as arising in randomized controlled trials’, *Statistics in Medicine*, 22, pp. 1317–1334. Available at: https://doi.org/10.1002/sim.1433.

Sullivan, P.W. (2011) ‘Are Utilities Bounded at 1.0? Implications for Statistical Analysis and Scale Development’, *Medical Decision Making*, 31(6), pp. 787–789. Available at: https://doi.org/10.1177/0272989X11400755.

Symmons, D. *et al.* (2005) ‘Aggressive versus symptomatic therapy in established rheumatoid arthritis’, *Health Technology Assessment*, 9(34).

The EuroQol Group (1990) ‘EuroQol - a new facility for the measurement of health-related quality of life’, *Health policy*, 16(3), pp. 199–208. Available at: https://doi.org/10.1016/0168-8510(90)90421-9.

Thokala, P. *et al.* (2016) ‘Multiple criteria decision analysis for health care decision making - An introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force’, *Value in Health*, 19(1), pp. 1–13. Available at: https://doi.org/10.1016/j.jval.2015.12.003.

Thomas, K.J. *et al.* (2006) ‘Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain’, *British Medical Journal*, 333(7569), pp. 623–626. Available at: https://doi.org/10.1136/bmj.38878.907361.7C.

Townsend, J. *et al.* (2004) ‘Routine examination of the newborn: The EMREN study. Evaluation of an extension of the midwife role including a randomised controlled trial of appropriately trained midwives and paediatric senior house officers’, *Health Technology Assessment*, 8(14). Available at: https://doi.org/10.3310/hta8140.

Turner-Bowker, D.M. *et al.* (2016) ‘The use of patient-reported outcomes in advanced breast cancer clinical trials: a review of the published literature’, *Current Medical Research and Opinion*, pp. 1709–1717. Available at: https://doi.org/10.1080/03007995.2016.1205005.

Vanderhout, S. *et al.* (2022) ‘Patient-reported outcomes and target effect sizes in pragmatic randomized trials in ClinicalTrials.gov: A cross-sectional analysis’, *PLOS Medicine*, 19(2), p. e1003896. Available at: https://doi.org/10.1371/JOURNAL.PMED.1003896.

Velanovich, V. (2007) ‘Behavior and analysis of 36-item short-form health survey data for surgical quality-of-life research’, *Archives of Surgery*, 142(5), pp. 473–477. Available at: https://doi.org/10.1001/archsurg.142.5.473.

Velasquez, M. and Hester, P. (2013) ‘An analysis of multi-criteria decision making methods’, *International Journal of Operations Research*, 10(2), pp. 56–66.

Vickers, A.J. and Altman, D.G. (2001) ‘Statistics Notes: Analysing controlled trials with baseline and follow up measurements’, *BMJ*, 323(7321), pp. 1123–1124. Available at: https://doi.org/10.1136/bmj.323.7321.1123.

Walters, S., Young, T. and Kwon, J. (2018) ‘Comparison of statistical methods for the analysis of patient-reported outcomes in RCTS’, *Quality of life research*, 27(Suppl 1, pp. S59–S60.

Walters, S.J. (2009) *Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation, Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation*. Chichester, UK. Available at: https://doi.org/10.1002/9780470840481.

Walters, S.J. *et al.* (2017) ‘Recruitment and retention of participants in randomised controlled trials: A review of trials funded and published by the United Kingdom Health Technology Assessment Programme’, *BMJ Open*, 7(3), p. 15276. Available at: https://doi.org/10.1136/bmjopen-2016-015276.

Walters, S.J. and Campbell, M.J. (2004) ‘The use of bootstrap methods for analysing health-related quality of life outcomes (particularly the SF-36)’, *Health and Quality of Life Outcomes*, 2(70). Available at: https://doi.org/10.1186/1477-7525-2-70.

Walters, S.J. and Campbell, M.J. (2005) ‘The use of bootstrap methods for estimating sample size and analysing health-related quality of life outcomes’, *Statistics in Medicine*, 24(7), pp. 1075–1102. Available at: https://doi.org/10.1002/sim.1984.

Walters, S.J., Campbell, M.J. and Lall, R. (2001) ‘Design and analysis of trials with quality of life as an outcome: a practical guide’, *Journal of Biopharmaceutical Statistics*, 11(3), pp. 155–176. Available at: https://doi.org/10.1081/BIP-100107655.

Wang, C. and Tu, D. (2020) ‘A bootstrap semiparametric homogeneity test for the distributions of multigroup proportional data, with applications to analysis of quality of life outcomes in clinical trials’, *Statistics in Medicine*, 39(12), pp. 1715–1731. Available at: https://doi.org/10.1002/sim.8507.

Ware, J.E. (2000) 'SF-36 Health Survey update', *Spine*, 25(24), pp. 3130–3139. Available at: https://doi.org/10.1097/00007632-200012150-00008.

Ware, J.E., Kosinski, M. and Gandek, B. (1993) *The SF-36 Health Survey: Manual and Interpretation Guide*. Boston: The Health Institute, New England Medical Center.

Ware, J.E. and Sherbourne, C.D. (1992) 'The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection.', *Medical care*, 30(6), pp. 473–83. Available at: https://doi.org/10.1097/00005650-199206000-00002.

Waterhouse, J. *et al.* (2010) 'A randomised 2 × 2 trial of community versus hospital pulmonary rehabilitation, followed by telephone or conventional follow-up', *Health Technology Assessment*, 14(6). Available at: https://doi.org/10.3310/hta14060.

Watson, A.J.M. *et al.* (2017) 'A pragmatic multicentre randomised controlled trial comparing stapled haemorrhoidopexy with traditional excisional surgery for haemorrhoidal disease: The eTHoS study', *Health Technology Assessment*, 21(70), pp. 1–223. Available at: https://doi.org/10.3310/hta21700.

van der Weijst, L. *et al.* (2017) 'Systematic literature review of health-related quality of life in locally-advanced non-small cell lung cancer: Has it yet become state-of-the-art?', *Critical Reviews in Oncology/Hematology*, 119, pp. 40–49. Available at: https://doi.org/10.1016/j.critrevonc.2017.09.014.

Weindling, A.M. *et al.* (2007) 'Additional therapy for young children with spastic cerebral palsy: A randomised controlled trial', *Health Technology Assessment*, 11(16). Available at: https://doi.org/10.3310/hta11160.

Weldring, T. and Smith, S.M.S. (2013) 'Article Commentary: Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs)', *Health Services Insights*. SAGE Publications Ltd, p. 61. Available at: https://doi.org/10.4137/HSI.S11093.

Wiggins, M. *et al.* (2004) 'The social support and family health study: A randomised controlled trial and economic evaluation of two alternative forms of postnatal support for mothers living in disadvantaged inner-city areas', *Health Technology Assessment*, 8(32). Available at: https://doi.org/10.3310/hta8320.

Wilhelm, M.O. (2008) 'Practical Considerations for Choosing Between Tobit and SCLS or CLAD Estimators for Censored Regression Models with an Application to Charitable Giving*', *Oxford Bulletin of Economics and Statistics*, 70(4), pp. 559–582. Available at: https://doi.org/10.1111/j.1468-0084.2008.00506.x.

Williams, J.G. *et al.* (2016) 'Comparison Of iNfliximab and ciclosporin in STeroid Resistant Ulcerative Colitis: Pragmatic randomised trial and economic evaluation (CONSTRUCT)', *Health Technology Assessment*, 20(44), pp. 1–322. Available at: https://doi.org/10.3310/hta20440.

Williams, M.A. *et al.* (2015) 'Strengthening and stretching for rheumatoid arthritis of the hand (SARAH). A randomised controlled trial and economic evaluation', *Health Technology Assessment*, 19(19), p. 221. Available at: https://doi.org/10.3310/hta19190.

Williams, N.H. *et al.* (2017) 'Subcutaneous injection of Adalimumab Trial compared with Control (SCIATiC): A randomised controlled trial of adalimumab injection compared with placebo for patients receiving physiotherapy treatment for sciatica', *Health Technology Assessment*, 21(60), pp. 1–179. Available at: https://doi.org/10.3310/hta21600.

Wirth, R., Houts, C. and Deal, L. (2016) 'Rasch Modeling With Small Samples: A Review Of The Literature', *Value in Health*, 19(3), p. A109. Available at: https://doi.org/10.1016/j.jval.2016.03.1841.

Wirth, R.J. and Houts, C. (2018) 'Multidimensional PRO measures: consequences for treatment effect detection and proposed solutions', *Quality of Life Research*, 27(Suppl 1(27), pp. S16–S16. Available at:

https://link.springer.com/article/10.1007/s11136-018-1991-2.

Yensy, N.A. (2021) ‘The comparison of the ordinal logistic model with the classical regression model’, in *Journal of Physics: Conference Series*. IOP Publishing, p. 12033. Available at: https://doi.org/10.1088/1742-6596/1731/1/012033.

Zheng, X., Qin, G. and Tu, D. (2017) ‘A generalized partially linear mean-covariance regression model for longitudinal proportional data, with applications to the analysis of quality of life data from cancer clinical trials’, *Statistics in Medicine*, 36(12), pp. 1884–1894. Available at: https://doi.org/10.1002/sim.7240.

Zigmond, A.S. and Snaith, R.P. (1983) ‘The Hospital Anxiety and Depression Scale’, *Acta Psychiatrica Scandinavica*, 67(6), pp. 361–370. Available at: https://doi.org/10.1111/j.1600-0447.1983.tb09716.x.

Zou, K.H., Carlsson, M.O. and Quinn, S.A. (2010) ‘Beta-mapping and beta-regression for changes of ordinal-rating measurements on Likert scales: A comparison of the change scores among multiple treatment groups’, *Statistics in Medicine*, 29(24), pp. 2486–2500. Available at: https://doi.org/10.1002/sim.4012.

Appendix A Supplementary for the review of statistical methods for the analysis of PROs in Chapter 2

A.1 Detailed inclusion and exclusion criteria for study selection

Table A.1 A detailed inclusion and exclusion criteria

Selection Criteria	Inclusion	Exclusion
Study type	Reviews on statistical methods for the analysis of PROs; Studies compared different methods for the analysis of PROs; Recommendations on what methods to use for the analysis of PROs; Studies developed statistical methods for the analysis of PROs.	Trials; Studies that looked for association and correlation; Studies that developed PROs or tested the feasibility, validity and reliability of PROs; Methods or reviews that developed or summarised the statistical methods for cost-effectiveness analysis; Reviews that only reported the clinical effectiveness of PROs but did not summarise the applied statistical methods; Reviews that summarised different PROs but not statistical methods; Protocols; Pilot studies.
Analysing methods	Various statistical methods for the analysis of the between group difference in PROs.	Studies that analysed PROs as explanatory variables; Methods for factor analysis; Methods for mapping a PRO to another.
Measures	PROs; QoL.	Studies that not focused on PROs.

HTA, health technology assessment; PROs, patient-reported outcomes; QoL, quality-of-life.

A.2 Studies excluded from the secondary screening

Table A.2 Studies excluded from the secondary screening with reasons (N = 29)

Author (Year of publication)	Exclusion Reason
Ahn and Ahn (2020)	Dealt with missingness
Altun and Turkan (2016)	Developed or validated a PRO
Anota *et al.* (2017)	Conference abstract
Brombin and Di Serio (2016)	Introduced a method for multivariate analysis
Cocks, Tharmanathan and Smith (2013)	Conference abstract
Conigliani, Manca and Tancredi (2015)	Mapped a PRO to EQ-5D
Fu *et al.* (2012)	Introduced a Bayesian approach to analyse device preference studies. It cannot be applied to the analysis of PRO data
Ge, Peng and Tu (2020)	Introduced a threshold linear mixed model for the identification of treatment-sensitive subsets
Gilet *et al.* (2011)	Conference abstract
Hinds *et al.* (2018)	Focused on statistical methods for developing or validating a PRO
Hong *et al.* (2013)	Introduced a Bayesian method to make indirect treatment comparison
Kaciroti *et al.* (2006)	Dealt with missingness
Lee and Daniels (2013)	Dealt with missingness
Lee, Daniels and Sargent (2010)	Dealt with missingness
Lv *et al.* (2019)	Dealt with missingness
Majewska *et al.* (2019)	Conference abstract
Marinacci *et al.* (2001)	Used statistical methods to fit PRO data as endpoint in a trial
Mishra and Ghosh (2009)	Dealt with censored data due to follow-up losses and study termination
Moerkerke *et al.* (2005)	Introduced a method for multivariate analysis
O'Kelly *et al.* (2019)	Conference abstract
Quintana *et al.* (2019)	Developed a Bayesian latent variable repeated measures model to determine disease progression.
Ribaudo and Thompson (2002)	Extended a hierarchical logistic regression to the multivariate analysis
Sajobi *et al.* (2018)	Reviewed statistical model- and design-based methods have been developed to test for response shift in longitudinal PROs
van der Weijst *et al.* (2017)	Conducted a systematic review on trials using PROs as endpoints. The statistical methods were not specifically summarised.
Steen, Curran and Molenberghs (2001)	Fit different models to a dataset, mainly solve the problem of missingness
Velanovich (2007)	Introduced a 'top-box' method to interpret PRO data
Walters, Young and Kwon (2018)	Conference abstract
Wirth, Houts and Deal (2016)	Conference abstract
Wirth and Houts (2018)	Conference abstract

Appendix B Quantitative MCDA for filtering different statistical methods in Chapter 5

B.1 Scoring and weighting system

We adapted the essential/highly desirable properties based on expert opinions from the SISAQOL Consortium (Coens *et al.*, 2020) to establish the criteria for the qualitative MCDA. A total of 18 experts with multiple roles as statistician, researcher, trials methodologist etc. composed the SISAQOL statistical methods working group, and 16 of them involved in the establishment of the statistical criteria. The essential/highly desirable properties are whether the method can compare two treatment arms, adjust for baseline score, be clinically relevant, allow for confounding factors, handle missing data, and handle clustered data.

Weights attached to included statistical properties are shown in Table B.1. Essential/highly desirable statistical properties composed the base-case criteria set (Set A) for primary analysis; and all 19 statistical properties that were listed for discussion were adapted for sensitivity analysis (Set B). In the base-case criteria set, the criteria 'Be clinically relevant' criterion was divided into two sub-criteria – 'within-individual clinical relevance' and 'within-group and between group clinical relevance' according to the SISAQOL guidance (Coens *et al.*, 2020) .The weight of each criterion is calculated by the proportion of total votes that criterion represents, and the two sub-criteria of 'Be clinically relevant' took the average weight of it. For example, 16 experts rated criteria P1 (compare two treatment groups) as an important statistical property, accounting for 0.21 of the total votes (16/77) in the Set A, and for 0.10 of the total votes (16/153) in Set B.

The scoring system which assessed each method according to its statistical properties was developed based on both the coding scheme of the SISAQOL guidance (Coens *et al.*, 2020) and our understanding of identified methods. In the established MCDA, the scoring for each criterion ranged from 0 to 100 (worst to best), and equal intervals were taken for different levels of performance. For example, criteria P7 (Handle cluster data with repeated measurements) was defined to have three values: 100, where both the repeated assessments of each individual and the order of measurements over time are taken into account; 50, where either of the two components is taken into account; and 0, where neither of the two components is taken into account.

To test the robustness of the established MCDA, sensitivity analyses were conducted. First, expand the base-case criteria set by adding another 13 statistical properties to the criteria (Set B in Table B.1); and second, change the weights of essential/highly desired statistical properties by applying another four sets to the MCDA (Set C-F in Table B.1).

Table B.1 Statistical properties proposed by the SISAQOL group with weights

Statistical properties	N	Criteria sets with weights					
		Set A	Set B	Set C	Set D	Set E	Set F
Essential/highly desirable statistical properties							
P1 Compare 2 treatment arms	16	0.21*	0.10	0.14	0.40	0.17	0.16
P2 Adjust for baseline score	14	0.18	0.09	0.14	0.04	0.00	0.09
Be clinically relevant	13						
P3 Within-individual		0.08	0.04	0.14	0.04	0.17	0.04
P4 Within and between group		0.08	0.04	0.14	0.40	0.17	0.08
P5 Allow for confounding factors	12	0.16	0.08	0.14	0.04	0.17	0.08
P6 Handle missing data (Part 1: with least restrictions)	11	0.14	0.07	0.14	0.04	0.17	0.07
P7 Handle clustered data (Part 1: over time)	11	0.14	0.07	0.14	0.04	0.17	0.07
Sum	*77*	*1.00*		*1.00*	*1.00*	*1.00*	
Other statistical properties that did not meet the essential/highly desirable criteria							
P8 Compare more than 2 treatment arms	9		0.06				0.06
P9 Handle unbalanced designs (Part 2: due to the dependency of assessment time)	9		0.06				0.06
P10 Calculate sample size	8		0.05				0.00
P11 Handle unbalanced designs (Part 1: due to practical reasons)	7		0.05				0.05
P12 Robustness	7		0.05				0.05
P13 Ability to maintain the ITT population	6		0.04				0.00
P14 Handle multiplicity	6		0.04				0.04
P15 Allow for time-varying covariates	5		0.03				0.03
P16 Handle clustered data (Part 2: within group)	5		0.03				0.03
P17 Handle a bounded scale	5		0.03				0.03
P18 Handle clustered data (Part 3: between group)	4		0.03				0.03
P19 Handle unbalanced designs (Part 3: due to non-adherence)	3		0.02				0.02
P20 Handle missing data (Part 2: with uncertainty estimates)	2		0.01				0.01
Sum	*153*		*1.00*				*1.00*

N, the number of votes for each statistical property by the expert panel; SISAQOL, Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data.

*The weight of P1 compare 2 treatment arms is calculated as the number of votes for P1 (16) divided by the total number of values for desired statistical properties (77).

B.2 Calculation methods

The evaluation process in the MCDA was independently conducted by YQ, SW and RJ who have backgrounds in mathematics, statistics, economics, and health-related science; and inconsistencies in scoring were discussed until reaching a conclusion.

Multiple approaches can be used to aggregate values and weights for the MCDA. The simple linear additive model (SLAM) was adapted to calculate the score of each statistical method, as the SLAM can compensate among criteria and it is simple to be performed without the use of complex computer programs (Velasquez and Hester, 2013; Diaby and Goeree, 2014).

The score of a statistical method $x(x = 1, 2, \ldots n)$ can be calculated by aggregating the multiplication of the weight attached to each statistical property, W_p $(0 < W_p < 1, p = 1, 2, \ldots m)$, and the score of each statistical method on each property, $S_{x,p}(p = 1, 2, \ldots m;\ x = 1, 2, \ldots n)$, as shown in Table B.2.

$$Score\ x = \sum_{p=1}^{m} S_{x,p} \times W_p \tag{B.1}$$

$$W = \sum_{p=1}^{m} W_p = 1 \tag{B.2}$$

Scores for each statistical method range from 0 to 100, and statistical methods with higher scores are considered to have more desired statistical properties than others. The scores of statistical methods can then be ranked, compared and selected.

Table B.2 The structure of scoring and weighting options (simple linear additive model)

Options (x)	**Criteria (p)**			**Total**
	Statistical property 1	**Statistical property 2**	**…**	
Statistical method 1	$S_{1,1}$	$S_{1,2}$	…	$Score$ 1
Statistical method 2	$S_{2,1}$	$S_{2,2}$	…	$Score$ 2
…	…	…	…	
Weights (sum = 1)	W_1	W_2	…	

B.3 Performance matrix

Table B.3 shows the performance matrices of different statistical methods when considering base-case set of weighted criteria (Set A). In general, multivariable methods are scored higher than univariable methods. Linear regression and Tobit regression are scored the same for sharing similar model theories and assumptions. The ANOVA and ANCOVA are scored less than linear regression since they are purely hypothesis tests and do not generate estimates or CIs for the treatment effect (which is regarded as good statistical practice), even if ANOVA and ANCOVA can be reformulated as multiple linear regression models. The scores of beta regression, median regression and CLAD are equal but smaller than linear regression because their estimands of treatment effect cannot be explained as mean or odds ratios without transformation.

Statistical methods that are developed for analysing ordinal data, i.e., ordinal regression (ordered logit & ordered probit) and binomial regression (beta-binomial and binomial-logit-Normal), have the same score. Again, unlike linear models, coefficients estimated by these methods cannot be explained and interpreted directly on the original scale or measurement or scoring system for the PRO, and the coefficients need to be transformed to odds ratios to estimate the treatment effect, which increases the difficulty in interpreting the results.

The sensitivity analysis shows that the ranks of statistical methods remain stable, with exceptions of *t*-test (Table B.4). The *t*-test stands out when giving heavy weights (Set 2) to ‘Compare 2 treatment arms’ (P1) and ‘Clinically relevant’ (P4), but *t*-test scored zero for the rest of the criteria set; and when considering additional statistical properties (Set B), CLAD and beta regression ranked higher than other methods for cross-sectional data, as both methods can account for boundaries and have few assumptions on data distribution.

Table B.3 Performance matrix with scoring for base-case criteria set (Set A)

Statistical methods	Statistical properties							Score
	P1	P2	P3	P4	P5	P6	P7	(Set A)
Multivariable methods for uncorrelated responses								
Tobit regression	100	100	0	100	100	0	0	62.99
Linear regression	100	100	0	100	100	0	0	62.99
Beta regression (logit link)	100	100	0	50	100	0	0	58.77
CLAD regression	100	100	0	50	100	0	0	58.77
Median regression	100	100	0	50	100	0	0	58.77
Ordered logit model	100	100	0	50	100	0	0	58.77
Beta-binomial regression	100	100	0	50	100	0	0	58.77
Binomial-logit-Normal regression	100	100	0	50	100	0	0	58.77
Ordered probit model	100	100	0	50	100	0	0	58.77
ANOVA or ANCOVA	100	100	0	0	100	0	0	54.55
Univariable methods								
t-test	100	0	0	100	0	0	0	29.22
Sign test	100	0	100	0	0	0	0	29.22
Kruskal-Wallis test	100	0	0	0	0	0	0	20.78
Mann-Whitney U test	100	0	0	0	0	0	0	20.78
Wilcoxon signed rank test	100	0	0	0	0	0	0	20.78

ANCOVA, analysis of covariance; ANOVA, analysis of variance; CLAD, censored least absolute deviations. P1, compare two treatment arms; P2, adjust for baseline score; P3, be clinically relevant (within-individual); P4, be clinically relevant (within-group and between group); P5, allow for confounding factors; P6, handle missing data; and P7, handle clustered data.

Table B.4 Scoring by criteria sets for sensitivity analysis

Statistical methods	**Criteria sets**				
	Set B	**Set C**	**Set D**	**Set E**	**Set F**
Multivariable methods for uncorrelated responses					
Tobit regression	46.41	57.14	88.00	50.00	55.56
Linear regression	40.85	57.14	88.00	50.00	50.00
Beta regression (logit link)	46.57	50.00	68.00	41.67	53.76
CLAD regression	46.57	50.00	68.00	41.67	53.76
Median regression	43.30	50.00	68.00	41.67	50.49
Ordered logit model	43.30	50.00	68.00	41.67	50.49
Beta-binomial regression	41.01	50.00	68.00	41.67	48.20
Binomial-logit-Normal regression	41.01	50.00	68.00	41.67	48.20
Ordered probit model	41.01	50.00	68.00	41.67	48.20
ANOVA or ANCOVA	38.89	42.86	48.00	33.33	44.12
Univariable methods					
t-test	22.22	28.57	80.00	33.33	26.14
Sign test	19.28	28.57	44.00	33.33	24.51
Kruskal-Wallis test	18.63	14.29	40.00	16.67	23.86
Mann-Whitney U test	17.97	14.29	40.00	16.67	17.97
Wilcoxon signed rank test	12.75	14.29	40.00	16.67	17.97

ANCOVA, analysis of covariance; ANOVA, analysis of variance; CLAD, censored least absolute deviations.

Appendix C Supplementary for the empirical analysis in Chapter 7

C.1 Recoding techniques for empirical analysis

For the scores of SF-36 bodily pain (BP) and social functioning (SF) dimension, two different types of scoring are seen from our nine trial data due to the application of non-standardised scoring strategies when collecting and calculating the SF-36v1 dimension scores. Appropriate recoding techniques are developed and applied to accommodate these two types of scoring, as shown in Table C.1.

Table C.1 Recoding techniques for two scoring patterns for BP and SF scores in SF-36v1 and SF-36v2

Two scoring patterns for BP					
SF-36v1 ($n = 10$)			SF-36v2 ($n = 10$)		
Scoring values	**Recoding techniques**	**Recoded values**	**Scoring values**	**Recoding techniques**	**Recoded values**
0	[0, 5.55]	0	0	[0, 5]	0
11.1	(5.55, 16.65]	1	12	(5, 16]	1
22.2	(16.65, 27.75]	2	22	(16, 26]	2
33.3	(27.75, 38.85]	3	31	(26, 36]	3
44.4	(38.85, 49.95]	4	32		
55.6	(49.95, 61.05]	5	41	(36, 47]	4
66.7	(61.05, 72.15]	6	42		
77.8	(72.15, 83.25]	7	51	(47, 57]	5
88.9	(83.25, 94.35]	8	61	(57, 67]	6
100	(94.35, 100]	9	62		
			72	(67, 77]	7
			84	(77, 92]	8
			100	(92, 100]	9
Two scoring patterns for SF					
SF-36v1 ($n = 10$)			SF-36v2 ($n = 9$)		
Scoring values	**Recoding techniques**	**Recoded values**	**Scoring values**	**Recoding techniques**	**Recoded values**
0	[0, 5.55]	0	0	[0, 6.25]	0
11.1	(5.55, 16.65]	1	12.5	(6.25, 18.75]	1
22.2	(16.65, 27.75]	2	25.0	(18.75, 31.25]	2
33.3	(27.75, 38.85]	3	37.5	(31.25, 43.75]	3
44.4	(38.85, 49.95]	4	50.0	(43.75, 56.25]	4
55.6	(49.95, 61.05]	5	62.5	(56.25, 68.75]	5
66.7	(61.05, 72.15]	6	75.0	(68.75, 81.25]	6
77.8	(72.15, 83.25]	7	87.5	(81.25, 93.75]	7
88.9	(83.25, 94.35]	8	100	(93.75, 100]	8
100	(94.35, 100]	9			

n denotes the number of possible levels. Note that for both BP and SF dimensions, the type 1 scoring is the non-standardised scoring in our dataset with SF-36v1, and the type 2 scoring is the standardised scoring in our dataset with SF-36v2.

Recoding techniques for other six dimensions in both SF-36 versions are provided in Table C.2.

Table C.2 Recoding techniques for RE, RP, VT, GH, PF, and MH dimensions in SF-36v1 and SF-36v2

(a) Scoring strategies for SF-36v1

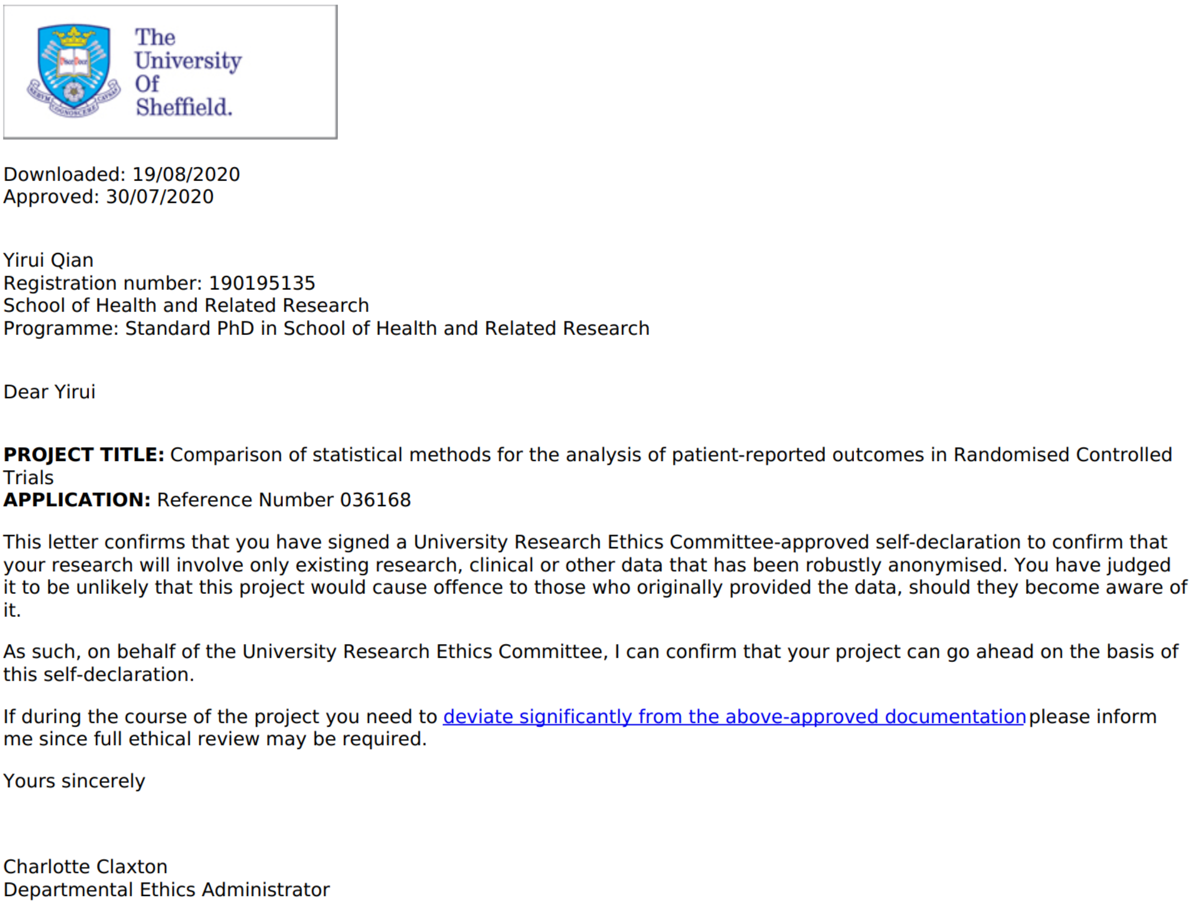
Dimensions	SF-36v1		
	Scoring values	Recoding techniques	Recoded values
RE	0(33.3)100	16.65(33.3)83.35	0(1)3
RP	0(25)100	12.5(25)87.5	0(1)4
VT	0(5)100	2.5(5)97.5	0(1)20
GH	0(5)100	2.5(5)97.5	0(1)20
PH	0(5)100	2.5(5)97.5	0(1)20
MH	0(4)100	2(4)98	0(1)25

(b) Scoring strategies for SF-36v2

Dimensions	SF-36v2		
	Scoring values	Recoding techniques	Recoded values
RE	0(8.3)100	4.15(8.3)95.45	0(1)12
RP	0(6.25)100	3.125(6.25)96.875	0(1)16
VT	0(6.25)100	3.125(6.25)96.875	0(1)16
GH	0(5)100	2.5(5)97.5	0(1)20
PH	0(5)100	2.5(5)97.5	0(1)20
MH	0(5)100	2.5(5)97.5	0(1)20

GH, general health; MH, mental health; PF, physical functioning; RE, role limitation – emotional; RP, role limitation – physical; VT, vitality. Explanation of the values using SF-36v1 RE as an example: 0(33.3)100 represents the scoring values ranging from 0 to 100, with 33.3 as intervals, i.e. 0, 33.3, 66.6, and 100; 16.65(33.3)83.35 means the recoding cut-points starting at 16.65, with 33.3 as intervals, i.e. 16.65, 49.95, and 83.35; 0(1)3 denotes the recoded values ranging from 0 to 3, with 1 as interval, i.e. 0, 1, 2, and 3.

C.2 Ethics Approval Letter



The
University
Of
Sheffield.

Downloaded: 19/08/2020
Approved: 30/07/2020

Yirui Qian
Registration number: 190195135
School of Health and Related Research
Programme: Standard PhD in School of Health and Related Research

Dear Yirui

PROJECT TITLE: Comparison of statistical methods for the analysis of patient-reported outcomes in Randomised Controlled Trials
APPLICATION: Reference Number 036168

This letter confirms that you have signed a University Research Ethics Committee-approved self-declaration to confirm that your research will involve only existing research, clinical or other data that has been robustly anonymised. You have judged it to be unlikely that this project would cause offence to those who originally provided the data, should they become aware of it.

As such, on behalf of the University Research Ethics Committee, I can confirm that your project can go ahead on the basis of this self-declaration.

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since full ethical review may be required.

Yours sincerely

Charlotte Claxton
Departmental Ethics Administrator

C.3 Stata codes for the recoding and regression analysis

We hereby present the Stata codes for recoding SF-36 dimension scores and conducting regression analysis, using SF-36v2 MH scores at 6-month post-randomisation from the Lifestyle Matters (LM) trial as an example.

C.3.1 Variable Specification

The variables are defined as follows:

- `mh6` is the score of SF-36 mental health score at 6-month follow-up in the LM trial.
- `omh6` is the recoded ordinal score of `mh6` to fit ordinal regression and the binomial regression;
- `fmh6` is the recoded score of `mh6` on a [0,1] scale to fit fractional logistic regression;
- `bmh6` is the recoded score of `mh6` on an (0,1) scale to fit beta regression;
- `SZ` is the sample size of patients;
- `mh0` is the SF-36 mental health score at baseline;
- `omh0` is the recoded ordinal score of `mh0` to fit ordinal regression and the binomial regression;
- `fmh0` is the recoded score of `mh0` on a [0,1] scale to fit fractional logistic regression;
- `bmh0` is the recoded score of `mh0` on an (0,1) scale to fit beta regression;
- `group` is a binary variable for treatment (treatment group = 1, control group = 0);
- `N` is the number of possible observable values of a SF-36 dimension score. In our example, N equals 21 for the SF-36 version 2 mental health dimension.

C.3.2 Recoding techniques

** Recoding to an ordinal scale

```
gen omh0 = irecode(mh0, 2.5, 7.5, 12.5, 17.5, 22.5, 27.5, 32.5,
37.5, 42.5, 47.5, 52.5, 57.5, 62.5, 67.5, 72.5, 77.5, 82.5, 87.5,
92.5, 97.5)

gen omh6 = irecode(mh6, 2.5, 7.5, 12.5, 17.5, 22.5, 27.5, 32.5,
37.5, 42.5, 47.5, 52.5, 57.5, 62.5, 67.5, 72.5, 77.5, 82.5, 87.5,
92.5, 97.5)
```

** Recoding to a [0, 1] scale

```
gen fmh6 = mh6/100

gen fmh0 = mh0/100
```

** Recoding to a (0, 1) scale

```
egen SZ6 = count(mh6)

gen bmh6 = (fmh6*(SZ6-1)+0.5)/SZ6
```

```
egen SZ0 = count(mh0)
```

```
gen bmh0 = (fmh0*(SZ0-1)+0.5)/SZ0
```

C.3.3 Regression analysis

** Multiple linear regression

```
regress mh6 group mh0
```

** Tobit regression

```
tobit mh6 group mh0, ll(0) ul(100)
```

** CLAD regression [package ‘sg153’ is required]

```
clad mh6 group mh0, rep(1000) ul(100)
```

** Median regression

```
qreg mh6 group mh0
```

** Ordered logit model

```
ologit omh6 group omh0
```

** Ordered probit model

```
oprobit omh6 group omh0
```

** Beta-binomial regression [package ‘betabin’ is required]

```
betabin omh6 group omh0, n(N) link(logit)
```

** Binomial-logit-Normal regression

```
glm omh6 group omh0, link(logit) family(binomial N)
```

** Fractional logistic regression

```
fracreg logit fmh6 group pmh0
```

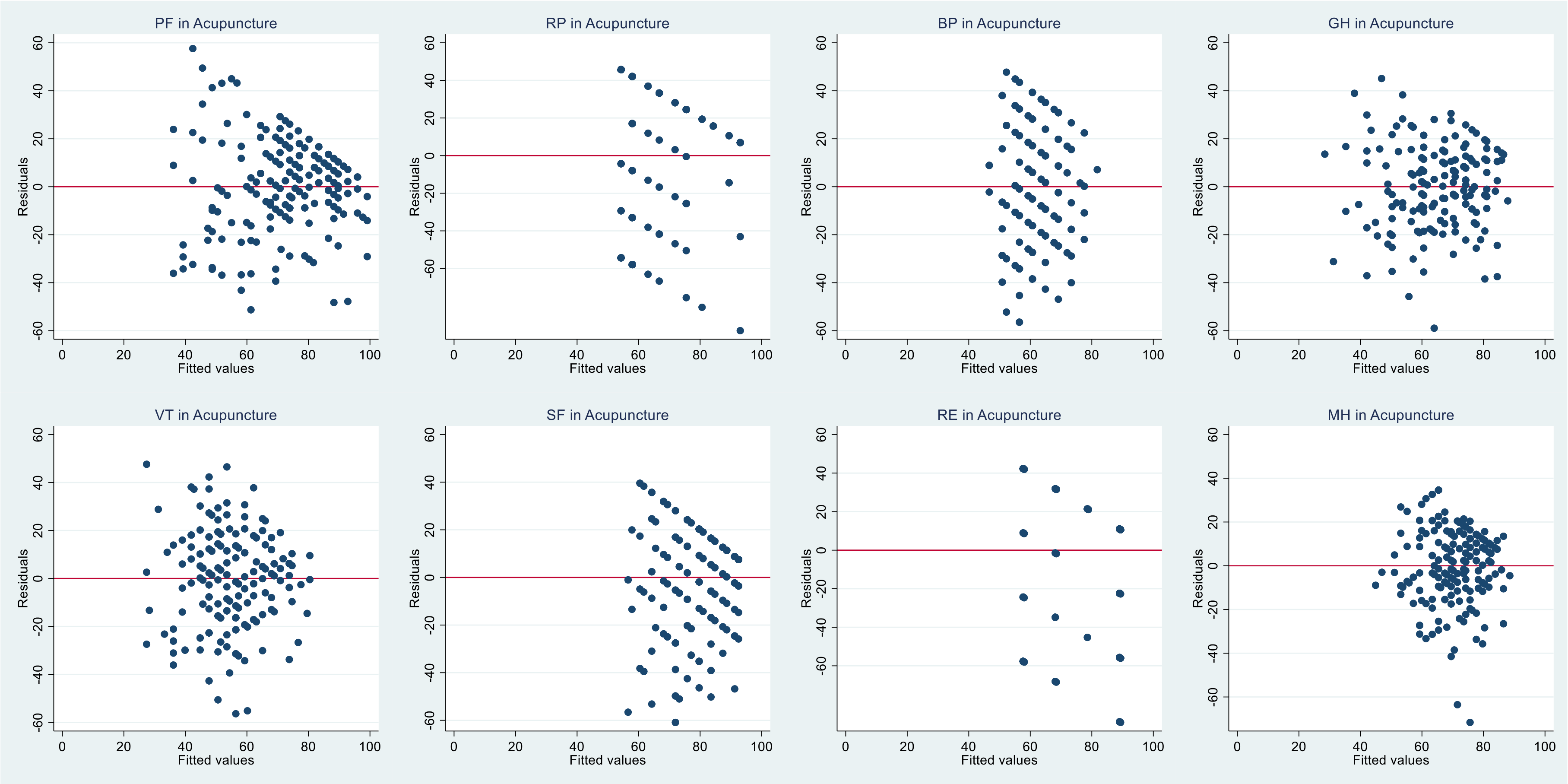
** Beta regression

```
betareg bmh6 group bmh0
```

C.4 Post estimation plots MLR and Tobit

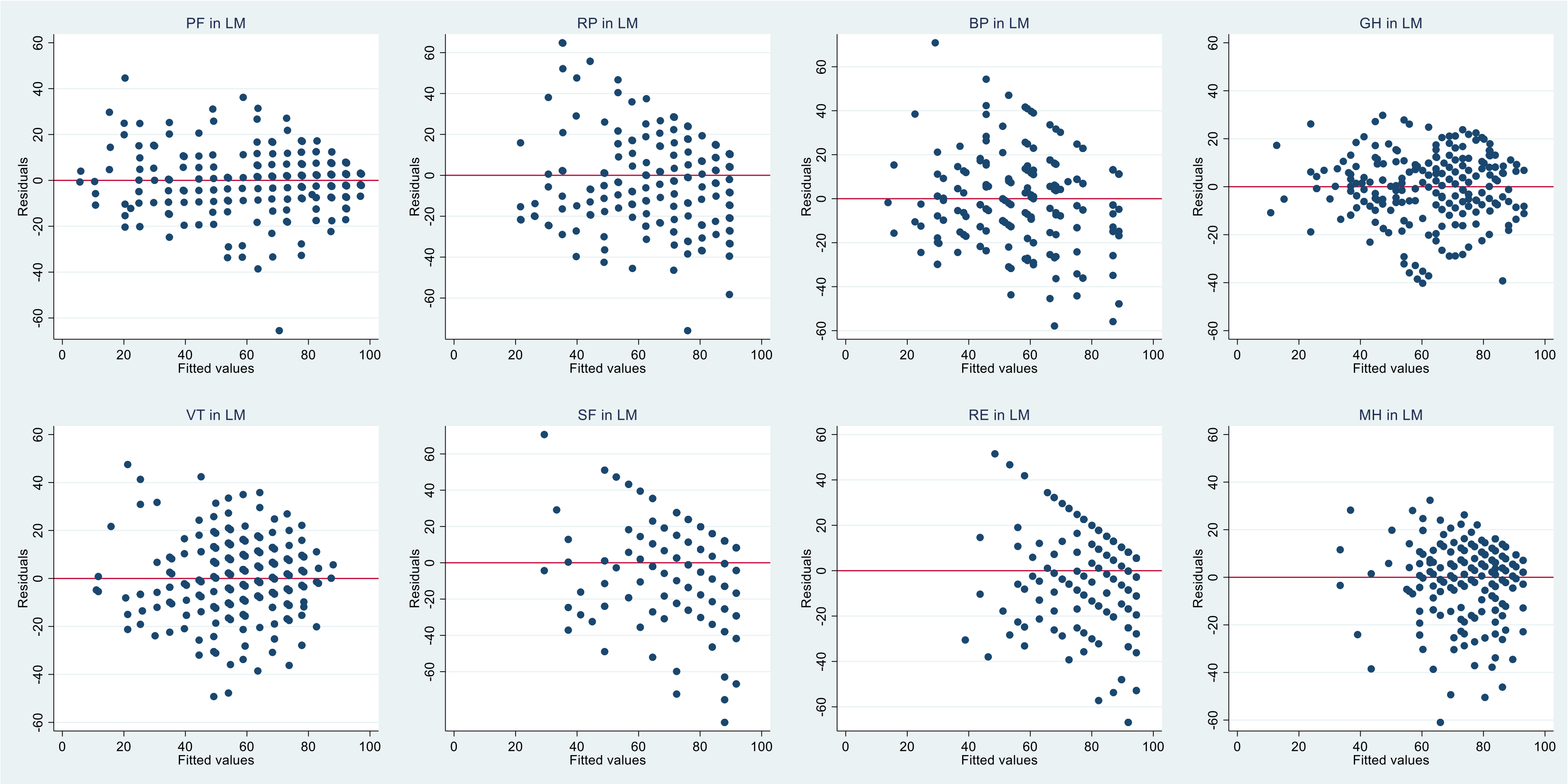
A series of post estimation plots were generated to test model assumptions where possible, including homoscedasticity and Normality of residuals for post-estimation of MLR and Tobit, shown in Figure C.1, Figure C.2, Figure C.3, and Figure C.4.

(a) Acupuncture



Continued on the next page

(b) LM



Continued on the next page

(c) PLINY

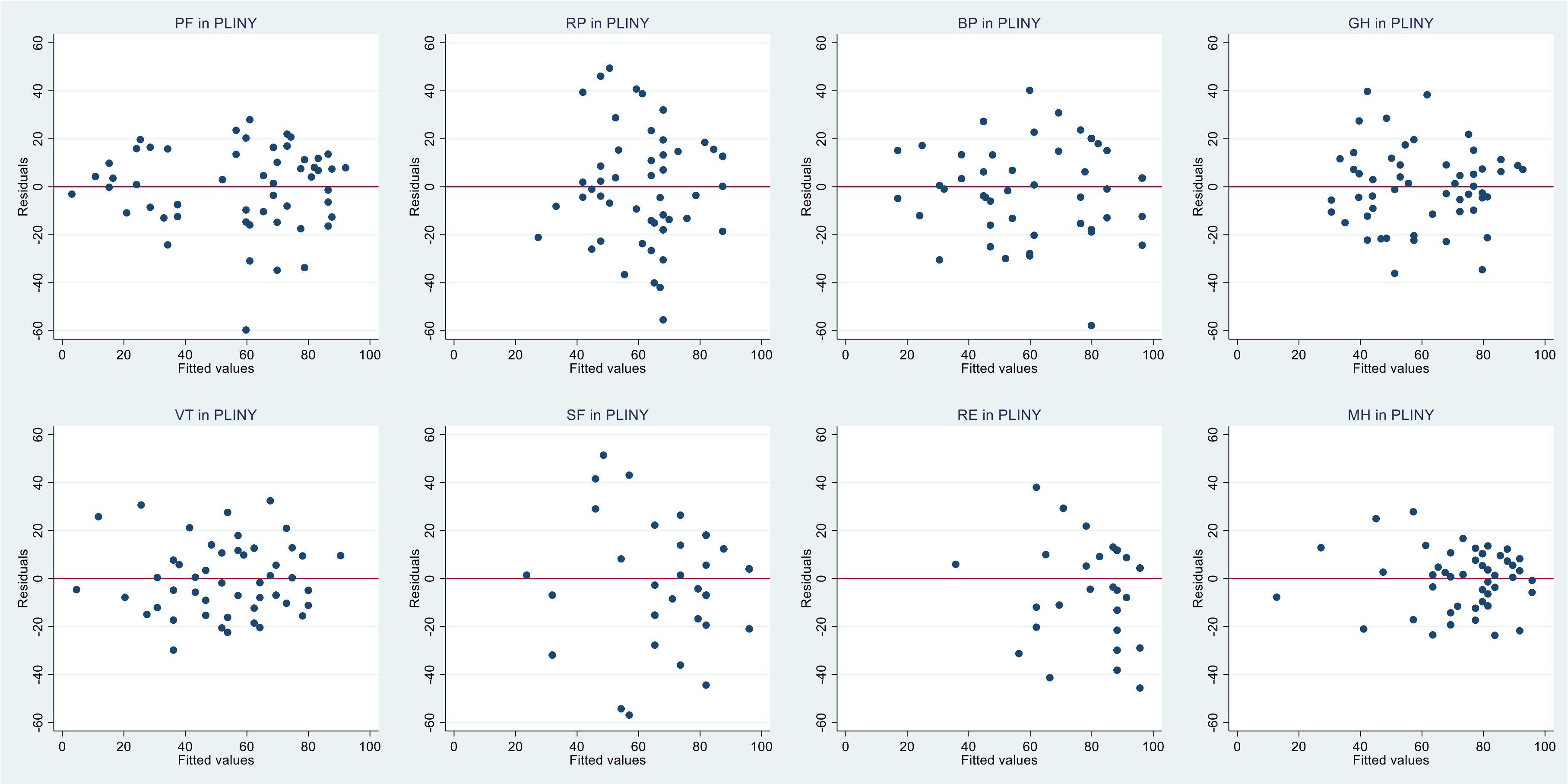
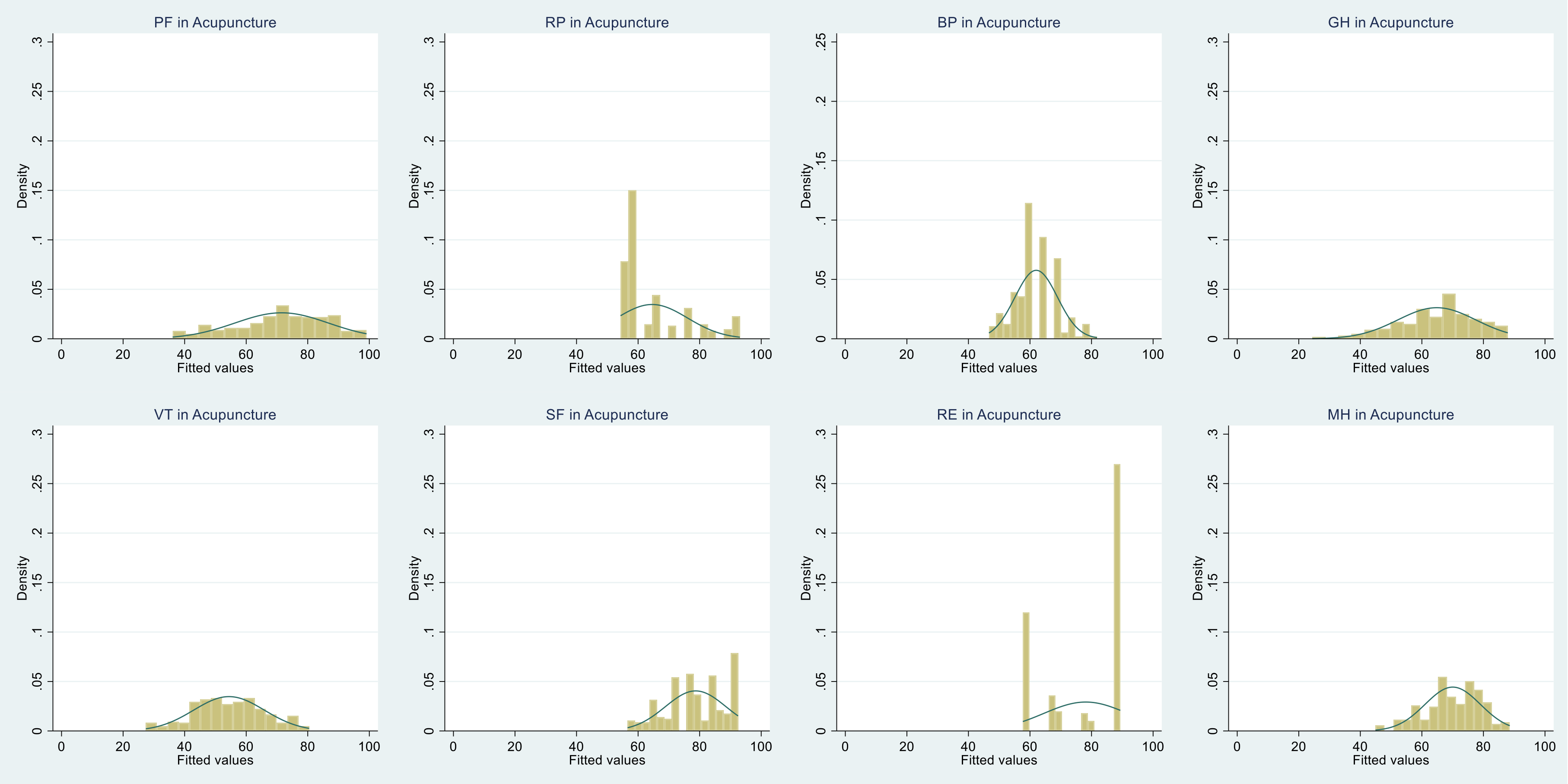


Figure C.1 Residual plot against fitted values after multiple linear regression estimation of SF-36 eight dimension scores

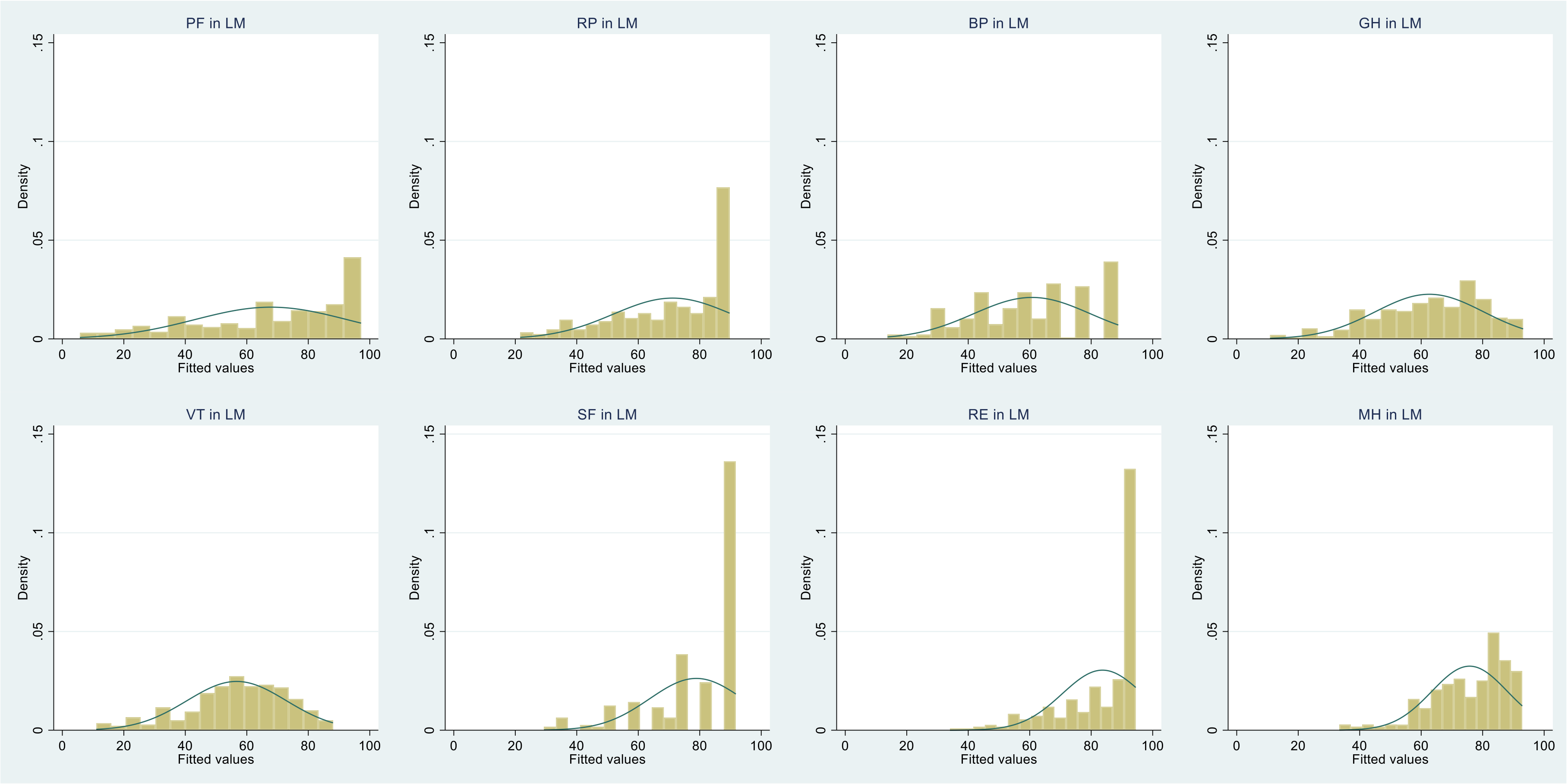
BP, bodily pain; GH, general health; MH, mental health; PF, physical functioning; RE, role limitation – emotional; RP, role limitation – physical; SF, social functioning; VT, vitality.

(a) Acupuncture



Continued on the next page

(b) LM



Continued on the next page

(c) PLINY

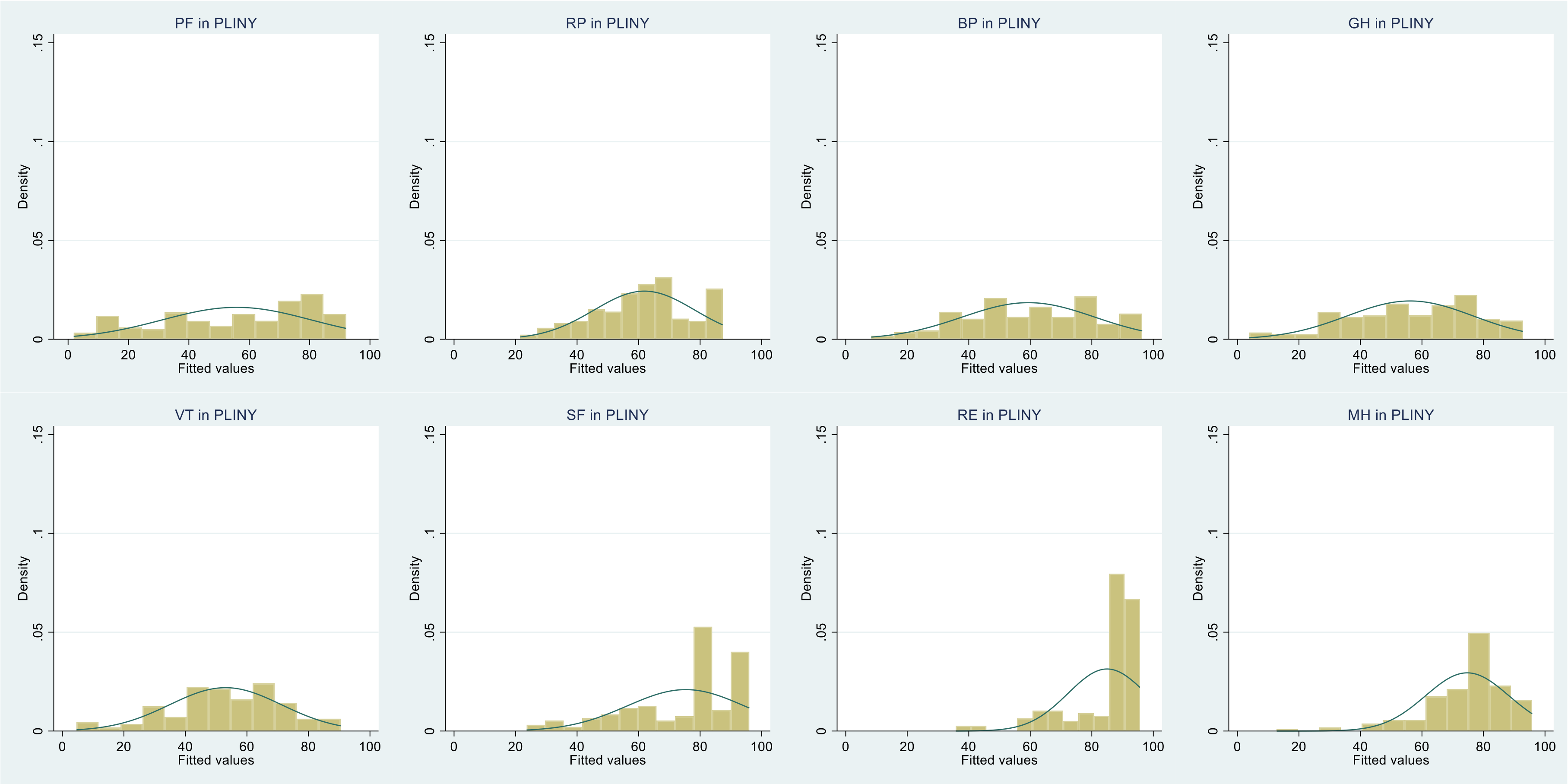
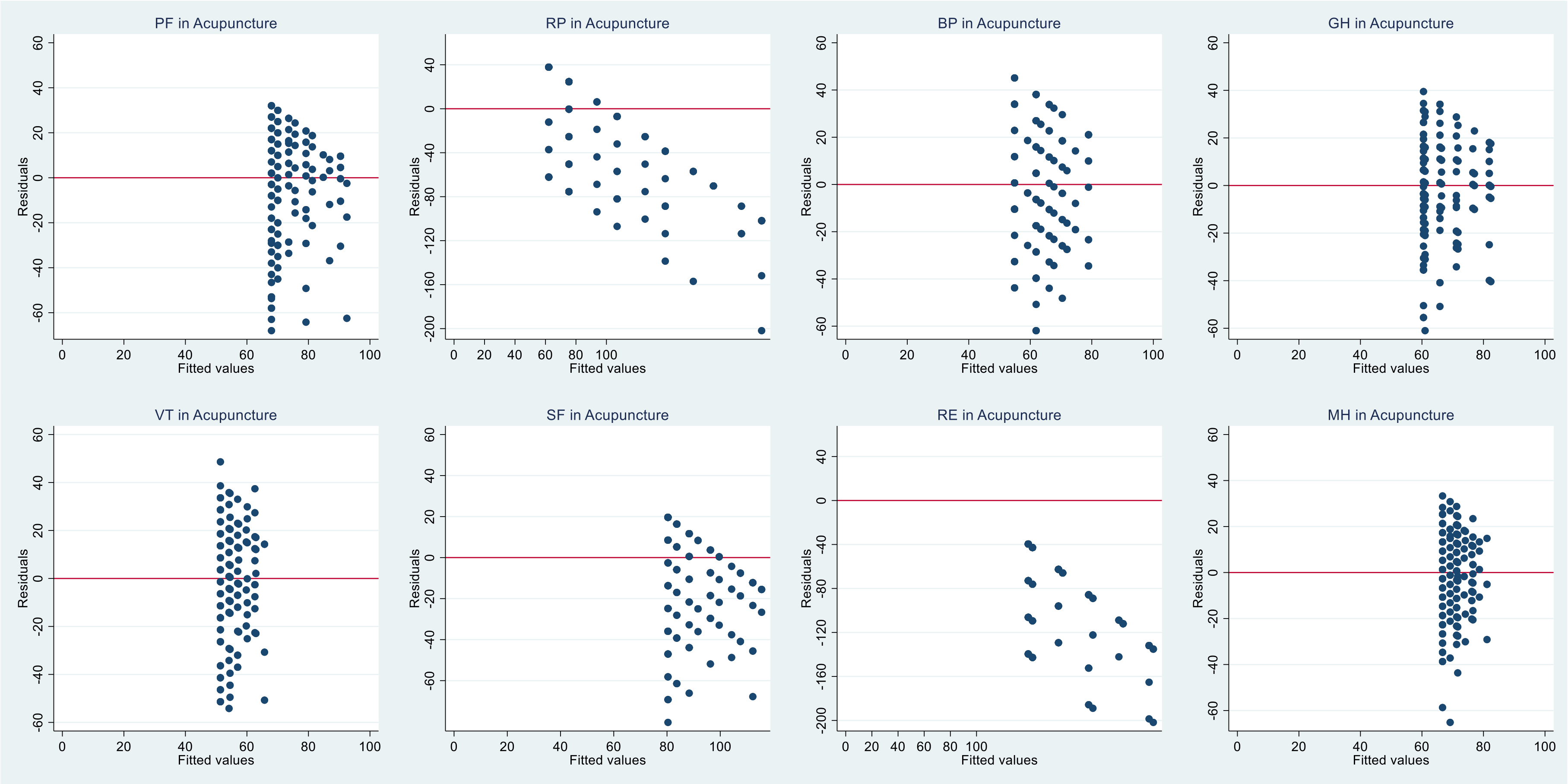


Figure C.2 Histogram of residuals after multiple linear regression estimation of SF-36 eight dimension scores

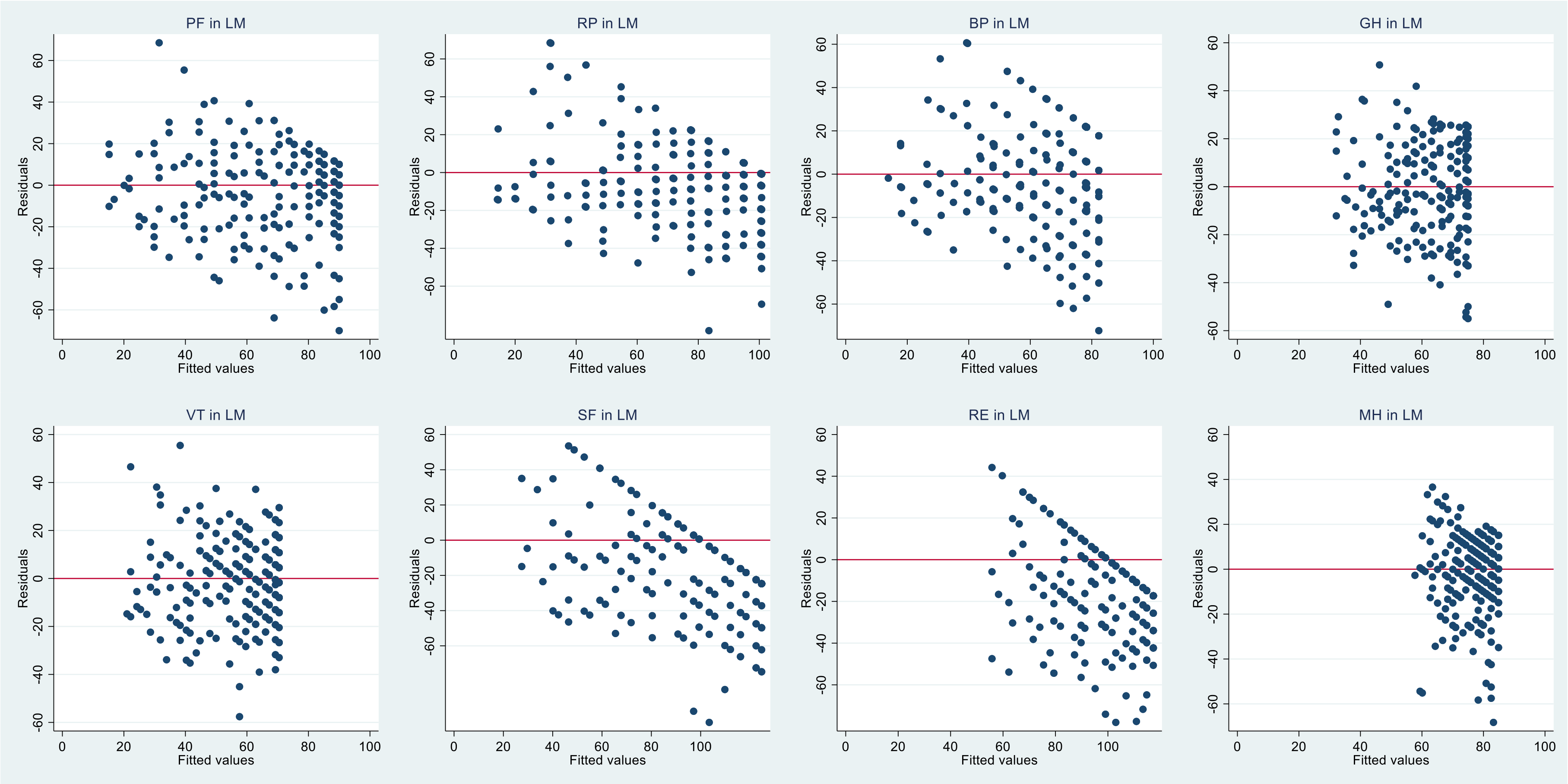
BP, bodily pain; GH, general health; MH, mental health; PF, physical functioning; RE, role limitation – emotional; RP, role limitation – physical; SF, social functioning; VT, vitality.

(a) Acupuncture



Continued on the next page

(b) LM



Continued on the next page

(c) PLINY

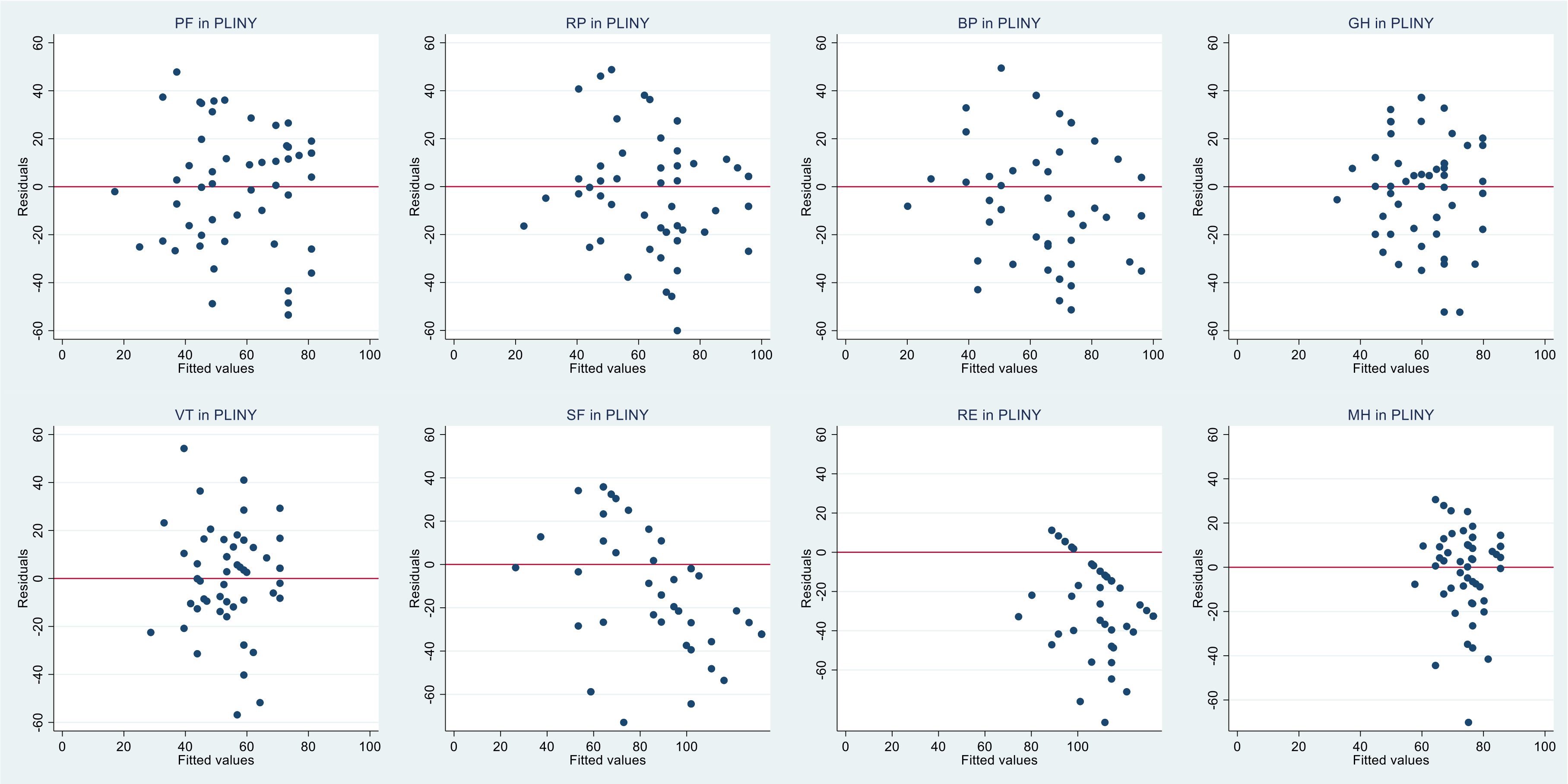
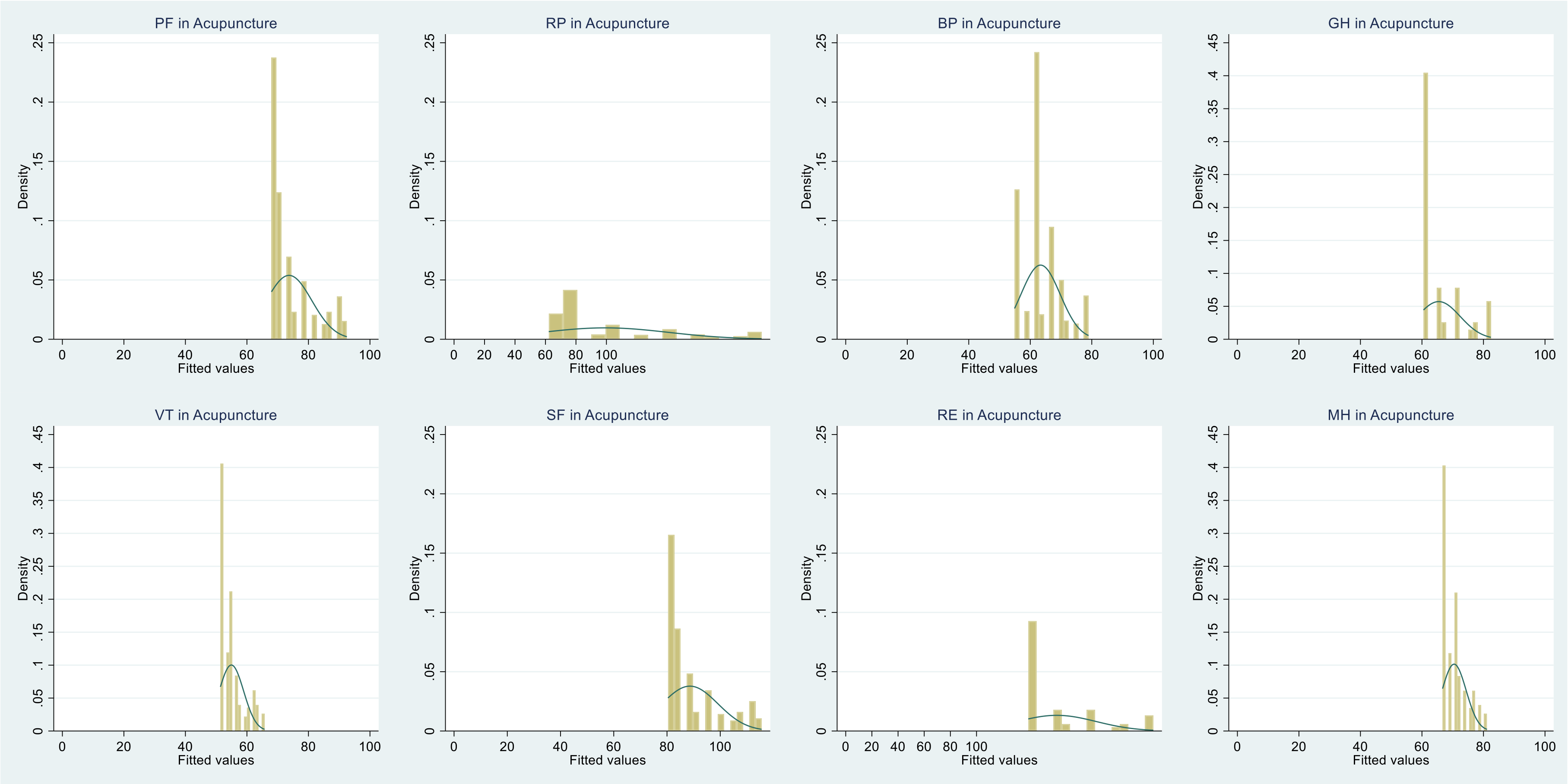


Figure C.3 Residual plot against fitted values after Tobit regression estimation of SF-36 eight dimension scores

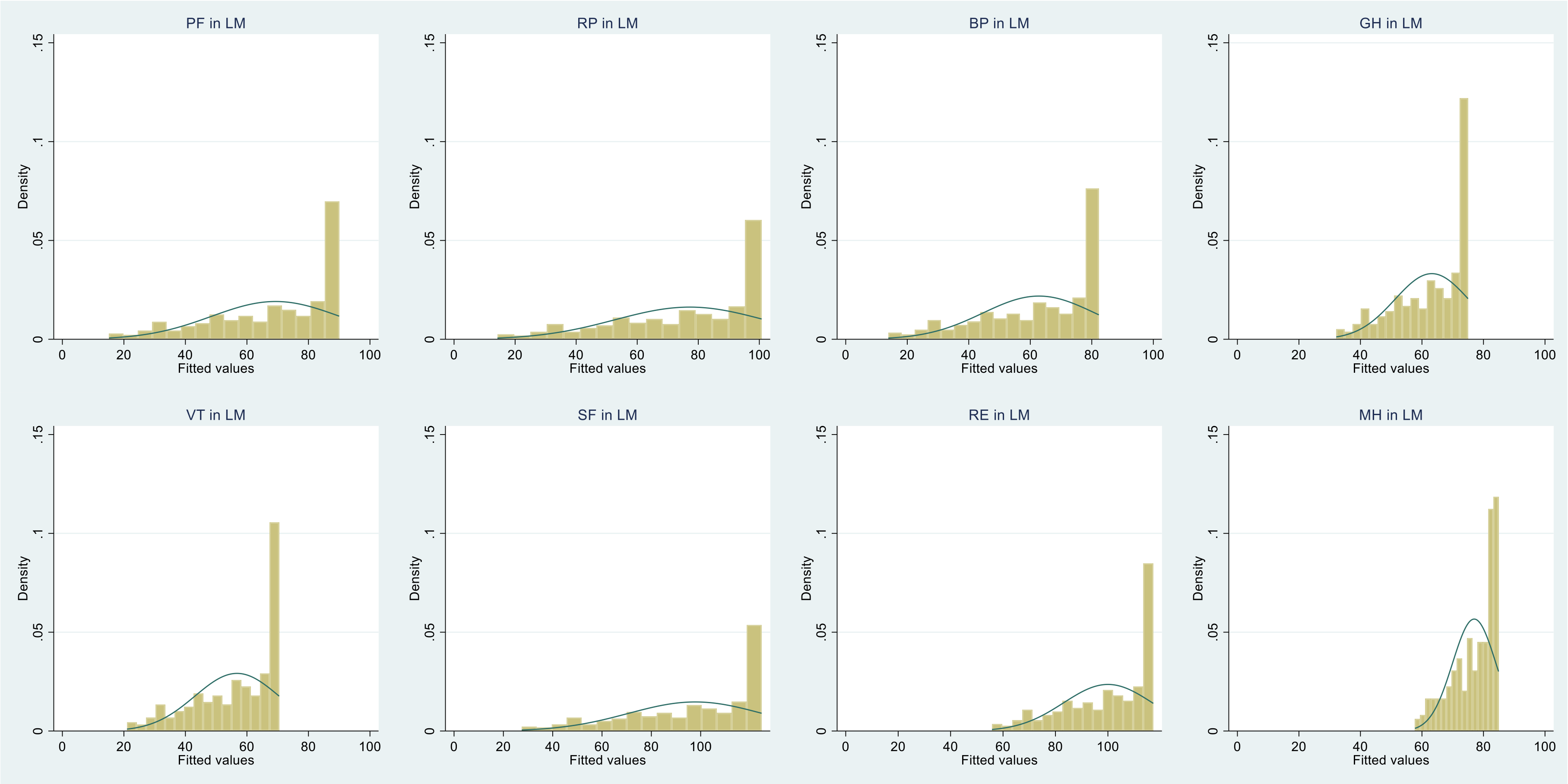
BP, bodily pain; GH, general health; MH, mental health; PF, physical functioning; RE, role limitation – emotional; RP, role limitation – physical; SF, social functioning; VT, vitality.

(a) Acupuncture



Continued on the next page

(b) LM



Continued on the next page

(c) PLINY

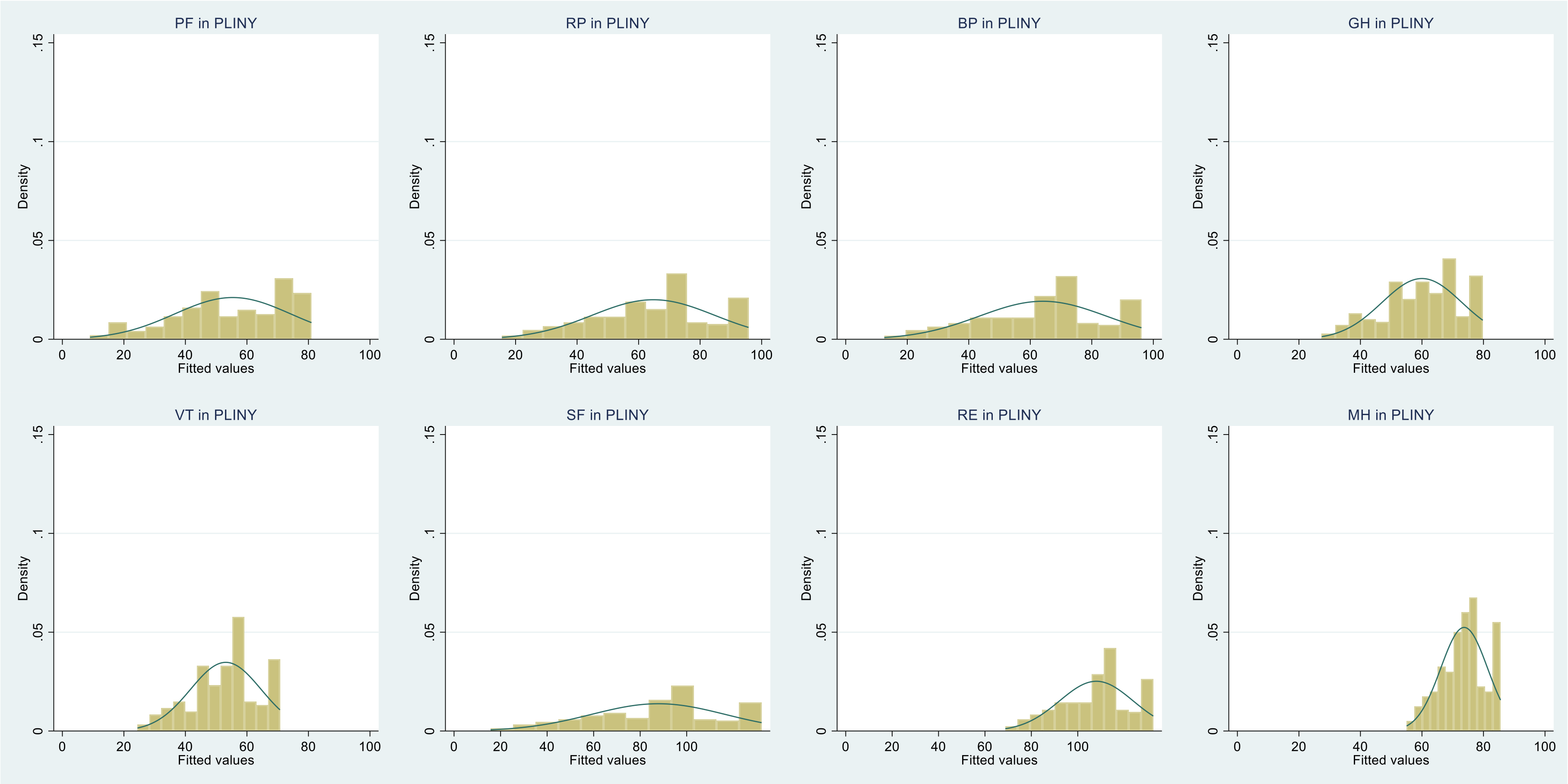
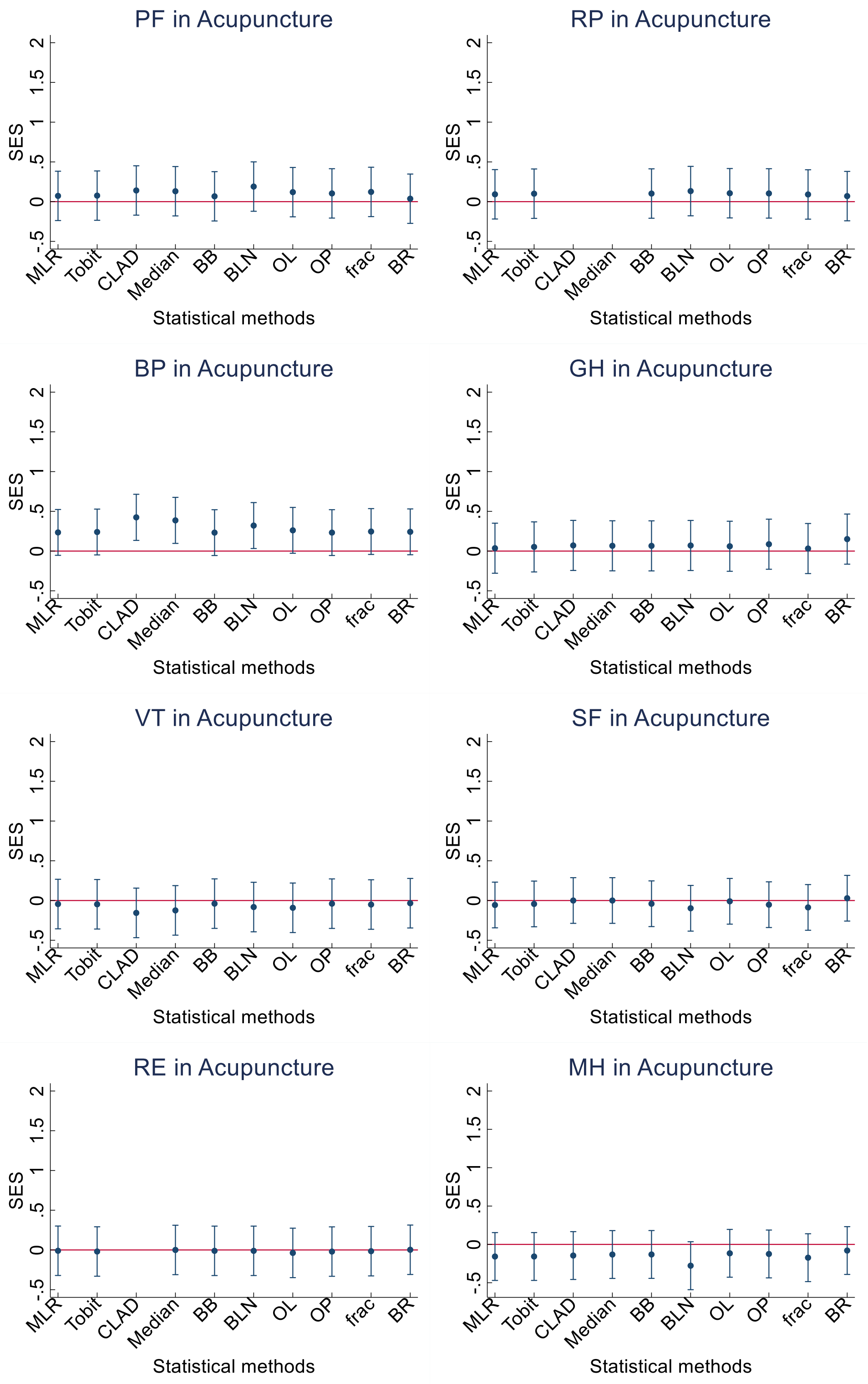


Figure C.4 Histogram of residuals after Tobit regression estimation of SF-36 eight dimension scores

BP, bodily pain; GH, general health; MH, mental health; PF, physical functioning; RE, role limitation – emotional; RP, role limitation – physical; SF, social functioning; VT, vitality.

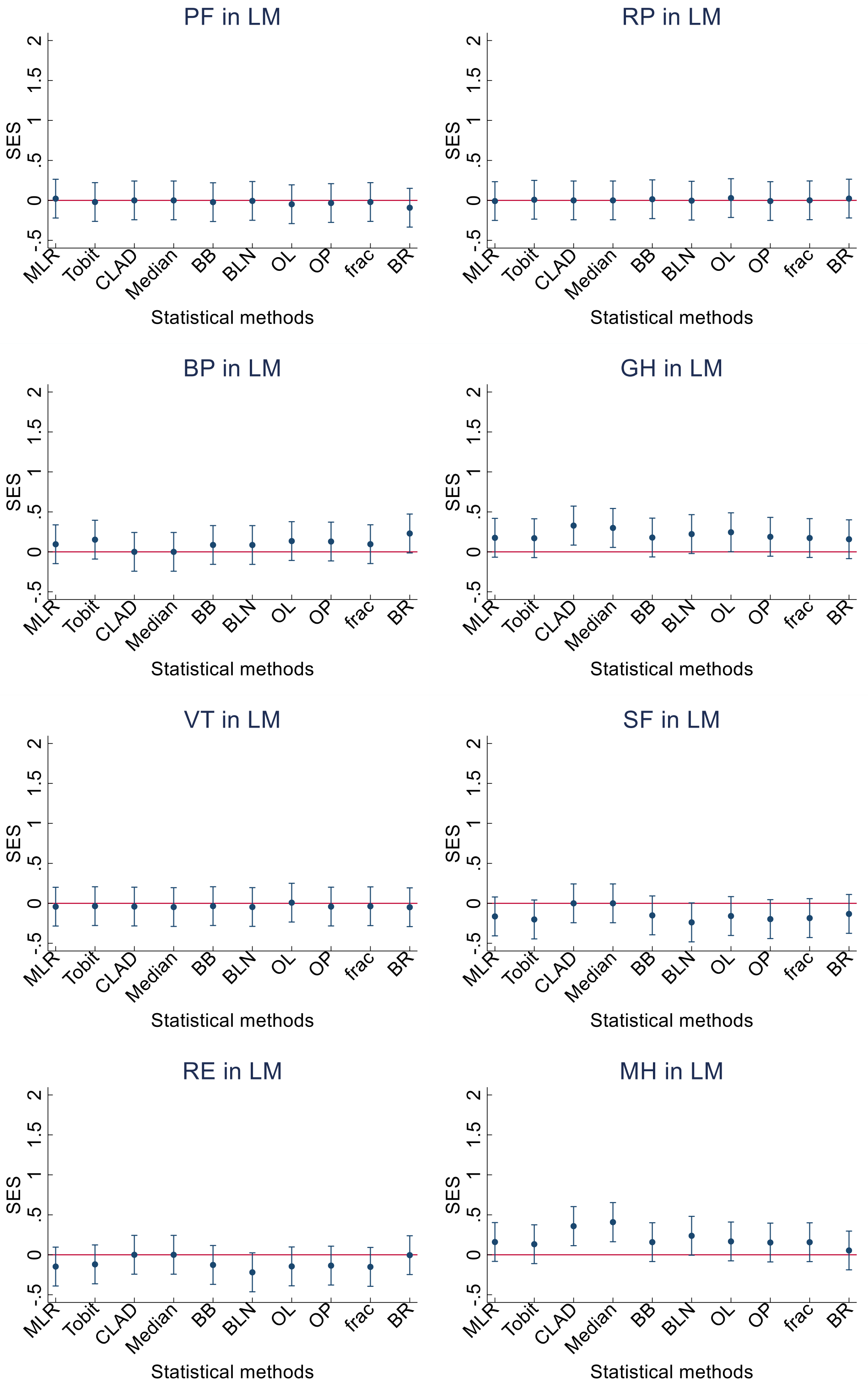
C.5 Effect size plots for Acupuncture, LM, and PLINY

(a) Acupuncture



Continued on the next page

(b) LM



Continued on the next page

(c) PLINY

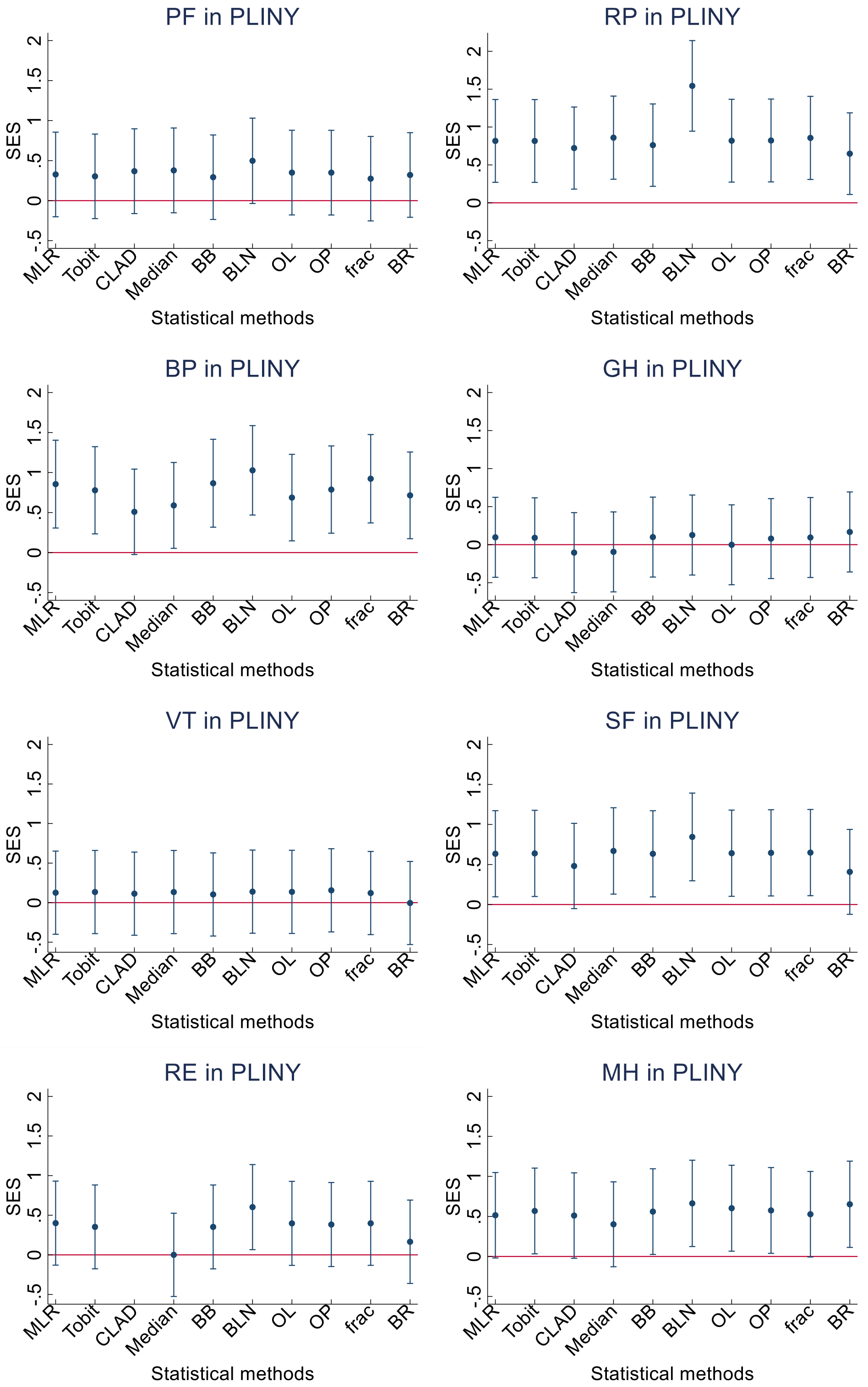


Figure C.5 SES with associated 95% CIs of treatment estimates by ten different statistical methods for SF-36 eight dimension scores using three RCTs

The bars on two sides of the vertical line for each method represents the 95% CIs for SES. The red horizontal represents the SES for no different between two treatment arms (i.e. y = 0). BP, bodily pain; GH, general health; MH, mental health; PF, physical functioning; RE, role limitation – emotional; RP, role limitation – physical; SES, standardised effect size; SF, social functioning; VT, vitality.

Appendix D Supplementary for the simulation analysis in Chapter 9

D.1 Additional tables and figures for exploratory analysis

Detailed information on the predefined parameter values to generate the simulated dataset, and the observed parameter values generated from the simulated dataset under each DGM and each number of observations (i.e. sample sizes) for three levels is shown in Table D.1. The average estimated treatment differences of 5,000 estimates from each statistical method under two estimand frameworks for five DGMs and six sample sizes considering three levels are shown in Table D.2. The detailed information on the number and percentage of missing values under the five DGMs for each level is shown in Table D.3. The scatterplots of estimated SESs from different statistical methods in comparison with MLR are presented in Figure D.1.

Table D.1 Comparison of predefined parameter values and observed parameter values (Ntotal = 2,700,000)

n_{obs}	DGM	Predefined means			Observed means								
					Level 4			Level 10			Level 26		
		Control	Treat	Group Difference	Control	Treat	Group Difference	Control	Treat	Group Difference	Control	Treat	Group Difference
100	1	50	50	0	49.93	49.98	0.05	49.98	50.05	0.07	49.93	49.98	0.05
200					49.95	50.04	0.08	50.01	50.06	0.05	49.96	50.00	0.05
400					49.99	50.01	0.01	50.03	50.05	0.01	49.98	50.00	0.02
800					50.01	50.01	-0.01	50.06	50.05	-0.01	50.01	50.00	-0.02
1200					50.01	49.99	-0.02	50.04	50.04	0.00	50.00	50.00	0.00
1600					50.02	50.01	-0.01	50.06	50.05	0.00	50.00	50.01	0.00
100	2	50	54.4	4.4	49.93	54.37	4.44	49.98	54.34	4.36	49.93	54.28	4.35
200					49.95	54.37	4.42	50.01	54.36	4.35	49.96	54.30	4.34
400					49.99	54.35	4.36	50.03	54.36	4.33	49.98	54.29	4.31
800					50.01	54.35	4.34	50.06	54.36	4.29	50.01	54.29	4.28
1200					50.01	54.34	4.34	50.04	54.35	4.31	50.00	54.29	4.30
1600					50.02	54.36	4.34	50.06	54.37	4.31	50.00	54.30	4.30
100	3	50	61	11	49.93	60.85	10.92	49.98	60.77	10.79	49.93	60.68	10.74
200					49.95	60.82	10.86	50.01	60.78	10.77	49.96	60.69	10.73
400					49.99	60.82	10.82	50.03	60.76	10.73	49.98	60.68	10.70
800					50.01	60.83	10.82	50.06	60.76	10.70	50.01	60.68	10.67
1200					50.01	60.84	10.83	50.04	60.76	10.72	50.00	60.68	10.69
1600					50.02	60.83	10.81	50.06	60.77	10.71	50.00	60.69	10.69
100	4	50	67.8	17.8	49.93	67.18	17.25	49.98	67.01	17.04	49.93	66.92	16.99
200					49.95	67.17	17.22	50.01	67.03	17.03	49.96	66.93	16.97
400					49.99	67.18	17.18	50.03	67.02	16.99	49.98	66.92	16.94
800					50.01	67.18	17.17	50.06	67.02	16.96	50.01	66.92	16.91
1200					50.01	67.19	17.19	50.04	67.02	16.98	50.00	66.92	16.92
1600					50.02	67.20	17.18	50.06	67.03	16.97	50.00	66.93	16.93
100	5	50	72	22	49.93	71.32	21.39	49.98	71.07	21.09	49.93	70.95	21.01
200					49.95	71.32	21.37	50.01	71.06	21.06	49.96	70.95	20.99
400					49.99	71.31	21.32	50.03	71.05	21.02	49.98	70.95	20.97
800					50.01	71.30	21.29	50.06	71.05	20.99	50.01	70.95	20.93
1200					50.01	71.31	21.30	50.04	71.06	21.01	50.00	70.94	20.95
1600					50.02	71.32	21.30	50.06	71.06	21.01	50.00	70.96	20.95

Table D.2 Estimated treatment difference in by the six statistical methods for each level and each DGM

(a) Level 4

n_{obs}	DGM	Scale-based estimand framework						SES estimand framework					
		Mean or median			**logORs**			**SES**					
		MLR	**Tobit**	**Median**	**Frac**	**OL**	**BB**	**MLR**	**Tobit**	**Median**	**Frac**	**OL**	**BB**
100	1	0.050	0.041	0.155	0.002	0.006	0.000	0.002	0.002	0.005	0.002	0.003	0.000
	2	4.443	5.083	13.037	0.179	0.358	0.145	0.188	0.188	0.341	0.188	0.184	0.141
	3	10.925	12.730	16.593	0.446	0.886	0.351	0.465	0.462	0.577	0.461	0.444	0.344
	4	17.252	20.706	16.623	0.723	1.422	0.543	0.741	0.726	0.723	0.724	0.677	0.532
	5	21.388	26.370	16.633	0.919	1.792	0.665	0.928	0.896	0.598	0.891	0.812	0.651
200	1	0.082	0.099	0.573	0.003	0.006	0.004	0.003	0.004	0.012	0.003	0.003	0.004
	2	4.419	5.078	15.245	0.178	0.349	0.143	0.186	0.186	0.318	0.186	0.182	0.139
	3	10.863	12.664	16.597	0.443	0.867	0.347	0.460	0.456	0.571	0.456	0.440	0.340
	4	17.220	20.655	16.597	0.720	1.399	0.532	0.736	0.720	0.656	0.719	0.676	0.522
	5	21.367	26.318	16.597	0.915	1.764	0.657	0.923	0.890	0.581	0.887	0.814	0.644
400	1	0.013	0.018	-0.133	0.001	0.001	0.004	0.001	0.001	-0.002	0.001	0.000	0.004
	2	4.359	4.997	15.704	0.175	0.343	0.139	0.184	0.183	0.265	0.183	0.179	0.135
	3	10.822	12.600	16.084	0.440	0.860	0.345	0.458	0.453	0.501	0.454	0.439	0.339
	4	17.183	20.600	16.084	0.717	1.387	0.527	0.734	0.717	0.509	0.717	0.675	0.518
	5	21.317	26.247	16.084	0.912	1.748	0.657	0.919	0.885	0.502	0.883	0.813	0.645
800	1	-0.007	-0.011	0.320	0.000	0.000	-0.002	0.000	0.000	0.004	0.000	0.000	-0.002
	2	4.339	4.966	16.297	0.174	0.342	0.134	0.183	0.182	0.241	0.182	0.179	0.131
	3	10.817	12.582	16.304	0.440	0.858	0.339	0.457	0.452	0.410	0.454	0.439	0.333
	4	17.166	20.555	16.304	0.716	1.383	0.529	0.733	0.715	0.410	0.716	0.676	0.520
	5	21.288	26.178	16.304	0.910	1.741	0.652	0.917	0.883	0.410	0.882	0.814	0.640
1200	1	-0.016	-0.016	0.147	-0.001	-0.002	-0.001	-0.001	-0.001	0.002	-0.001	-0.001	-0.001
	2	4.337	4.970	16.077	0.174	0.340	0.138	0.182	0.182	0.230	0.182	0.179	0.135
	3	10.831	12.604	16.077	0.440	0.857	0.335	0.458	0.453	0.353	0.454	0.439	0.329
	4	17.187	20.583	16.077	0.717	1.383	0.528	0.733	0.715	0.353	0.716	0.676	0.519
	5	21.299	26.190	16.077	0.910	1.740	0.653	0.918	0.883	0.353	0.882	0.814	0.641
1600	1	-0.008	-0.012	0.346	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000
	2	4.343	4.969	15.831	0.174	0.342	0.136	0.183	0.182	0.229	0.183	0.179	0.133
	3	10.811	12.575	15.831	0.440	0.856	0.333	0.457	0.452	0.316	0.453	0.439	0.327
	4	17.182	20.573	15.831	0.717	1.382	0.528	0.733	0.715	0.316	0.716	0.676	0.519
	5	21.304	26.196	15.831	0.911	1.739	0.651	0.918	0.883	0.316	0.882	0.814	0.640

Continued on the next page

(b) Level 10

n_{obs}	DGM	Scale-based estimand framework						SES estimand framework					
		Mean or median			logORs			SES					
		MLR	Tobit	Median	Frac	OL	BB	MLR	Tobit	Median	Frac	OL	BB
100	1	0.069	0.059	0.164	0.003	0.008	0.002	0.003	0.003	0.004	0.003	0.004	0.003
	2	4.362	4.549	4.889	0.176	0.353	0.158	0.202	0.203	0.173	0.202	0.197	0.203
	3	10.790	11.359	11.102	0.440	0.874	0.391	0.502	0.503	0.394	0.498	0.479	0.502
	4	17.037	18.208	17.429	0.713	1.395	0.621	0.803	0.799	0.622	0.786	0.737	0.796
	5	21.091	22.868	21.874	0.904	1.746	0.776	1.006	0.995	0.779	0.971	0.891	0.989
200	1	0.050	0.051	0.202	0.002	0.004	0.002	0.002	0.002	0.009	0.002	0.002	0.002
	2	4.353	4.553	5.268	0.175	0.345	0.158	0.200	0.201	0.195	0.200	0.195	0.201
	3	10.770	11.351	11.117	0.439	0.860	0.390	0.498	0.498	0.462	0.495	0.476	0.497
	4	17.025	18.207	17.149	0.711	1.375	0.620	0.798	0.792	0.642	0.781	0.734	0.789
	5	21.058	22.830	21.842	0.900	1.721	0.774	1.000	0.985	0.906	0.965	0.889	0.980
400	1	0.013	0.013	-0.020	0.001	0.001	0.000	0.001	0.001	0.000	0.001	0.001	0.001
	2	4.326	4.521	5.333	0.174	0.342	0.157	0.199	0.199	0.198	0.198	0.194	0.199
	3	10.727	11.296	10.727	0.436	0.853	0.388	0.496	0.495	0.487	0.492	0.474	0.494
	4	16.991	18.155	16.652	0.709	1.366	0.618	0.795	0.788	0.621	0.778	0.732	0.785
	5	21.020	22.776	21.441	0.898	1.710	0.772	0.996	0.980	0.972	0.961	0.887	0.975
800	1	-0.015	-0.017	0.116	-0.001	-0.001	-0.001	-0.001	-0.001	0.004	-0.001	0.000	-0.001
	2	4.292	4.484	5.488	0.172	0.339	0.155	0.197	0.197	0.208	0.197	0.193	0.197
	3	10.699	11.265	10.867	0.435	0.850	0.387	0.494	0.493	0.408	0.490	0.473	0.492
	4	16.957	18.115	16.592	0.707	1.361	0.617	0.793	0.785	0.628	0.776	0.731	0.782
	5	20.991	22.744	21.514	0.896	1.704	0.770	0.995	0.977	0.808	0.959	0.887	0.972
1200	1	-0.001	-0.002	0.031	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000
	2	4.308	4.499	5.394	0.173	0.340	0.156	0.198	0.198	0.217	0.197	0.193	0.198
	3	10.717	11.282	10.700	0.436	0.850	0.387	0.495	0.493	0.351	0.491	0.473	0.492
	4	16.976	18.137	16.494	0.708	1.361	0.617	0.794	0.785	0.675	0.776	0.732	0.782
	5	21.011	22.766	21.298	0.897	1.703	0.771	0.995	0.977	0.699	0.960	0.887	0.972
1600	1	-0.003	-0.004	0.085	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000
	2	4.312	4.506	5.313	0.173	0.340	0.156	0.198	0.198	0.226	0.198	0.193	0.198
	3	10.713	11.278	10.659	0.435	0.849	0.387	0.494	0.493	0.318	0.490	0.473	0.492
	4	16.975	18.133	16.413	0.708	1.360	0.617	0.794	0.785	0.702	0.776	0.732	0.782
	5	21.006	22.760	21.186	0.896	1.702	0.771	0.995	0.976	0.631	0.959	0.887	0.972

Continued on the next page

(c) Level 26

n_{obs}	DGM	Scale-based estimand framework						SES estimand framework					
		Mean or median			logORs			SES					
		MLR	Tobit	Median	Frac	OL	BB	MLR	Tobit	Median	Frac	OL	BB
100	1	0.048	0.042	0.066	0.002	0.006	0.001	0.002	0.002	0.002	0.002	0.003	0.001
	2	4.348	4.477	4.461	0.175	0.352	0.170	0.203	0.205	0.157	0.203	0.199	0.205
	3	10.742	11.132	11.040	0.438	0.871	0.423	0.506	0.509	0.389	0.502	0.483	0.507
	4	16.989	17.820	17.543	0.710	1.392	0.679	0.810	0.811	0.619	0.794	0.744	0.802
	5	21.013	22.294	21.895	0.900	1.741	0.852	1.015	1.010	0.772	0.980	0.900	0.995
200	1	0.047	0.044	0.099	0.002	0.003	0.002	0.002	0.002	0.003	0.002	0.002	0.002
	2	4.344	4.475	4.481	0.175	0.345	0.171	0.202	0.203	0.162	0.202	0.197	0.204
	3	10.733	11.131	11.042	0.437	0.858	0.423	0.503	0.504	0.398	0.499	0.480	0.503
	4	16.973	17.807	17.579	0.708	1.373	0.678	0.806	0.803	0.634	0.789	0.741	0.796
	5	20.993	22.274	21.921	0.897	1.718	0.851	1.010	1.000	0.791	0.975	0.897	0.987
400	1	0.020	0.021	-0.041	0.001	0.002	0.001	0.001	0.001	-0.001	0.001	0.001	0.001
	2	4.314	4.448	4.370	0.173	0.342	0.169	0.201	0.201	0.157	0.200	0.195	0.202
	3	10.705	11.098	10.946	0.435	0.852	0.422	0.501	0.501	0.394	0.497	0.478	0.500
	4	16.944	17.768	17.551	0.706	1.364	0.676	0.803	0.799	0.632	0.786	0.739	0.792
	5	20.969	22.236	21.920	0.895	1.707	0.849	1.007	0.995	0.789	0.971	0.896	0.982
800	1	-0.017	-0.018	0.046	-0.001	-0.001	-0.001	-0.001	-0.001	0.002	-0.001	0.000	-0.001
	2	4.278	4.409	4.492	0.172	0.339	0.168	0.199	0.199	0.163	0.199	0.194	0.200
	3	10.669	11.061	10.849	0.434	0.848	0.420	0.499	0.498	0.395	0.495	0.477	0.497
	4	16.908	17.730	17.682	0.704	1.359	0.675	0.801	0.796	0.649	0.784	0.738	0.789
	5	20.932	22.195	22.006	0.893	1.701	0.847	1.004	0.992	0.807	0.969	0.895	0.979
1200	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	2	4.295	4.425	4.534	0.172	0.339	0.168	0.199	0.200	0.161	0.199	0.194	0.200
	3	10.685	11.075	10.666	0.434	0.848	0.420	0.499	0.499	0.383	0.495	0.477	0.498
	4	16.924	17.745	17.717	0.705	1.358	0.675	0.801	0.796	0.654	0.784	0.738	0.789
	5	20.947	22.213	21.949	0.893	1.700	0.848	1.005	0.992	0.812	0.969	0.895	0.979
1600	1	0.004	0.005	0.026	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000
	2	4.299	4.431	4.624	0.172	0.340	0.168	0.200	0.200	0.162	0.199	0.195	0.200
	3	10.689	11.079	10.570	0.434	0.848	0.421	0.499	0.499	0.373	0.495	0.478	0.498
	4	16.928	17.748	17.790	0.705	1.358	0.675	0.801	0.796	0.648	0.784	0.739	0.789
	5	20.953	22.216	21.967	0.893	1.700	0.848	1.005	0.992	0.802	0.970	0.896	0.979

Table D.3 A Number and percentage of missing values due to non-convergence under the five DGMs for each level

(a) Level 4

DGM	n_{obs}	MLR		Tobit		Median		Frac		OL		BB		Total	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%
1	100	0	0.0	0	0.0	65	1.3	0	0.0	0	0.0	2,634	52.7	2,699	9.0
	200	0	0.0	0	0.0	1	0.0	0	0.0	0	0.0	3,041	60.8	3,042	10.1
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	3,436	68.7	3,436	11.5
	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	3,856	77.1	3,856	12.9
	1200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4,152	83.0	4,152	13.8
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4,277	85.5	4,277	14.3
	Total	*0*	*0.0*	*0*	*0.0*	*66*	*0.2*	*0*	*0.0*	*0*	*0.0*	*21,396*	*71.3*	*21,462*	*11.9*
2	100	0	0.0	0	0.0	37	0.7	0	0.0	0	0.0	2,922	58.4	2,959	9.9
	200	0	0.0	0	0.0	1	0.0	0	0.0	0	0.0	3,138	62.8	3,139	10.5
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	3,471	69.4	3,471	11.6
	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	3,886	77.7	3,886	13.0
	1200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4,135	82.7	4,135	13.8
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4,229	84.6	4,229	14.1
	Total	*0*	*0.0*	*0*	*0.0*	*38*	*0.1*	*0*	*0.0*	*0*	*0.0*	*21,781*	*72.6*	*21,819*	*12.1*
3	100	0	0.0	0	0.0	26	0.5	0	0.0	0	0.0	2,960	59.2	2,986	10.0
	200	0	0.0	0	0.0	8	0.2	0	0.0	0	0.0	3,338	66.8	3,346	11.2
	400	0	0.0	0	0.0	3	0.1	0	0.0	0	0.0	3,704	74.1	3,707	12.4
	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	3,967	79.3	3,967	13.2
	1200	0	0.0	0	0.0	1	0.0	0	0.0	0	0.0	4,142	82.8	4,143	13.8
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4,164	83.3	4,164	13.9
	Total	*0*	*0.0*	*0*	*0.0*	*38*	*0.1*	*0*	*0.0*	*0*	*0.0*	*22,275*	*74.3*	*22,313*	*12.4*
4	100	0	0.0	0	0.0	28	0.6	0	0.0	0	0.0	2,979	59.6	3,007	10.0
	200	0	0.0	0	0.0	8	0.2	0	0.0	0	0.0	3,340	66.8	3,348	11.2
	400	0	0.0	0	0.0	3	0.1	0	0.0	0	0.0	3,705	74.1	3,708	12.4
	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	3,953	79.1	3,953	13.2
	1200	0	0.0	0	0.0	1	0.0	0	0.0	0	0.0	4,144	82.9	4,145	13.8
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4,287	85.7	4,287	14.3
	Total	*0*	*0.0*	*0*	*0.0*	*40*	*0.1*	*0*	*0.0*	*0*	*0.0*	*22,408*	*74.7*	*22,448*	*12.5*
5	100	0	0.0	0	0.0	22	0.4	0	0.0	0	0.0	3,115	62.3	3,137	10.5
	200	0	0.0	0	0.0	5	0.1	0	0.0	0	0.0	3,562	71.2	3,567	11.9
	400	0	0.0	0	0.0	3	0.1	0	0.0	0	0.0	3,859	77.2	3,862	12.9
	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4,109	82.2	4,109	13.7
	1200	0	0.0	0	0.0	1	0.0	0	0.0	0	0.0	4,207	84.1	4,208	14.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4,290	85.8	4,290	14.3
	Total	*0*	*0.0*	*0*	*0.0*	*31*	*0.1*	*0*	*0.0*	*0*	*0.0*	*23,142*	*77.1*	*23,173*	*12.9*

Continued on the next page

(b) Level 10

DGM	n_{obs}	MLR		Tobit		Median		Frac		OL		BB		Total	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%
	100	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	2	0.0	2	0.0
	200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
1	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*2*	*0.0*	*2*	*0.0*
	100	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	0.0	1	0.0
	200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
2	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*1*	*0.0*	*1*	*0.0*
	100	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4	0.1	4	0.0
	200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
3	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*4*	*0.0*	*4*	*0.0*
	100	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	0.0	1	0.0
	200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
4	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1200	0	0.0	0	0.0	1	0.0	0	0.0	0	0.0	0	0.0	1	0.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	*0*	*0.0*	*0*	*0.0*	*1*	*0.0*	*0*	*0.0*	*0*	*0.0*	*1*	*0.0*	*2*	*0.0*
	100	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	0.0	1	0.0
	200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
5	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*1*	*0.0*	*1*	*0.0*

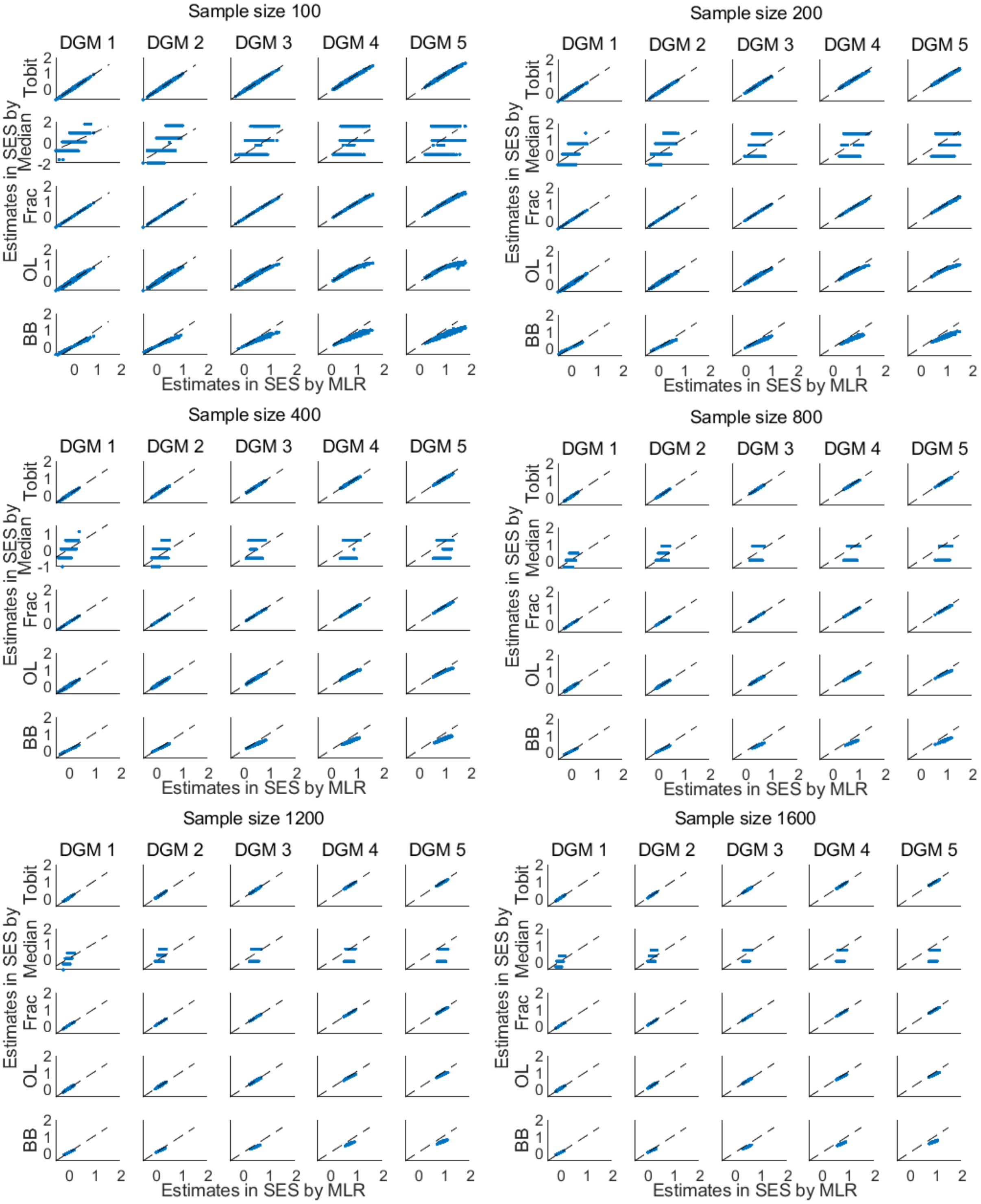
Continued on the next page

(c) Level 26

DGM	n_{obs}	MLR		Tobit		Median		Frac		OL		BB		Total	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%
	100	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
1	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*
	100	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
2	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*
	100	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
3	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	100	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
4	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1200	0	0.0	0	0.0	1	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	*0*	*0.0*	*0*	*0.0*	*1*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*
	100	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	400	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
5	800	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1200	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	1600	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
	Total	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*	*0*	*0.0*

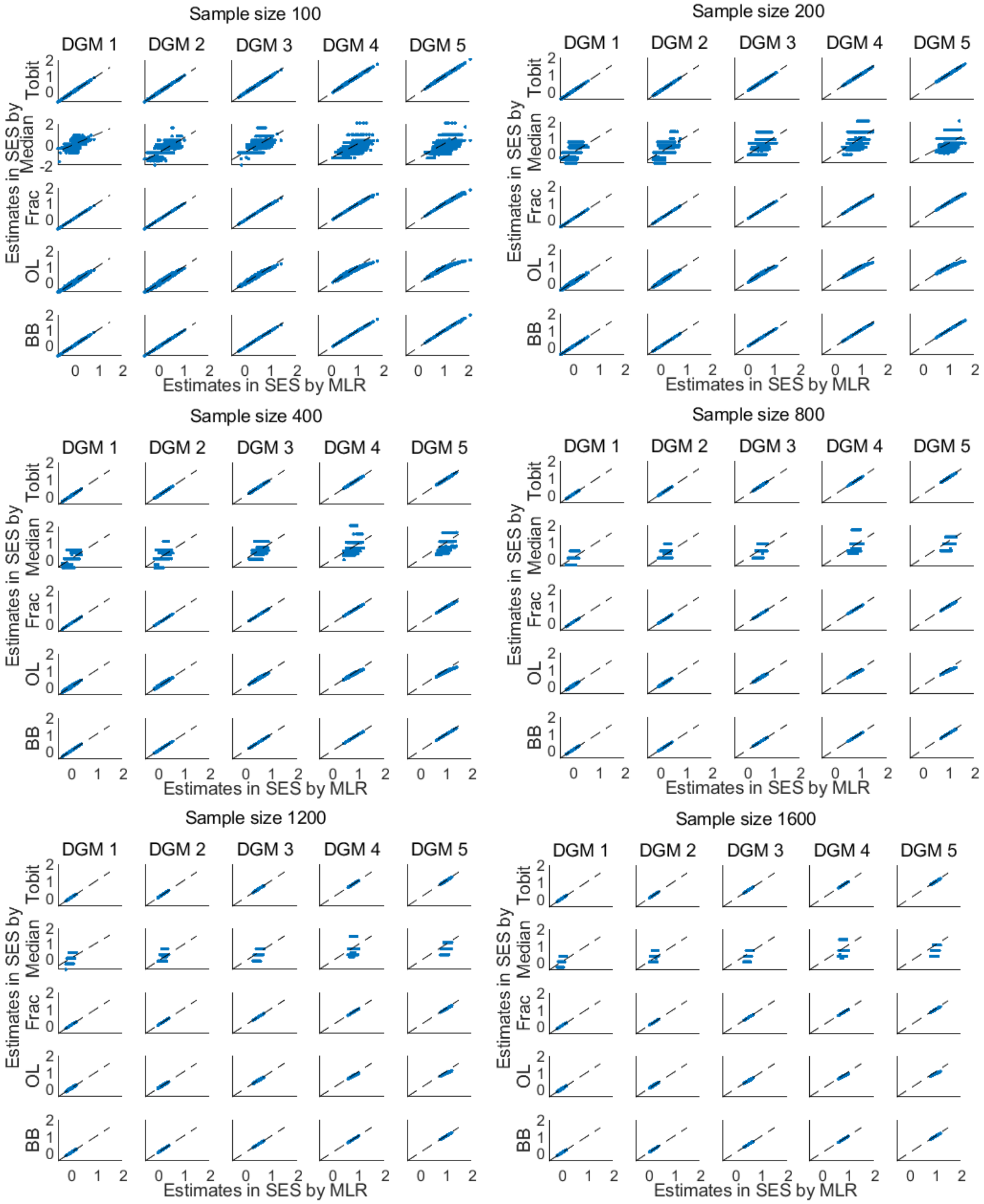
Each cell contains up to a maximum of 5,000 estimates from 5,000 simulations.

(a) Level 4



Continued on the next page

(b) Level 10



Continued on the next page

(c) Level 26

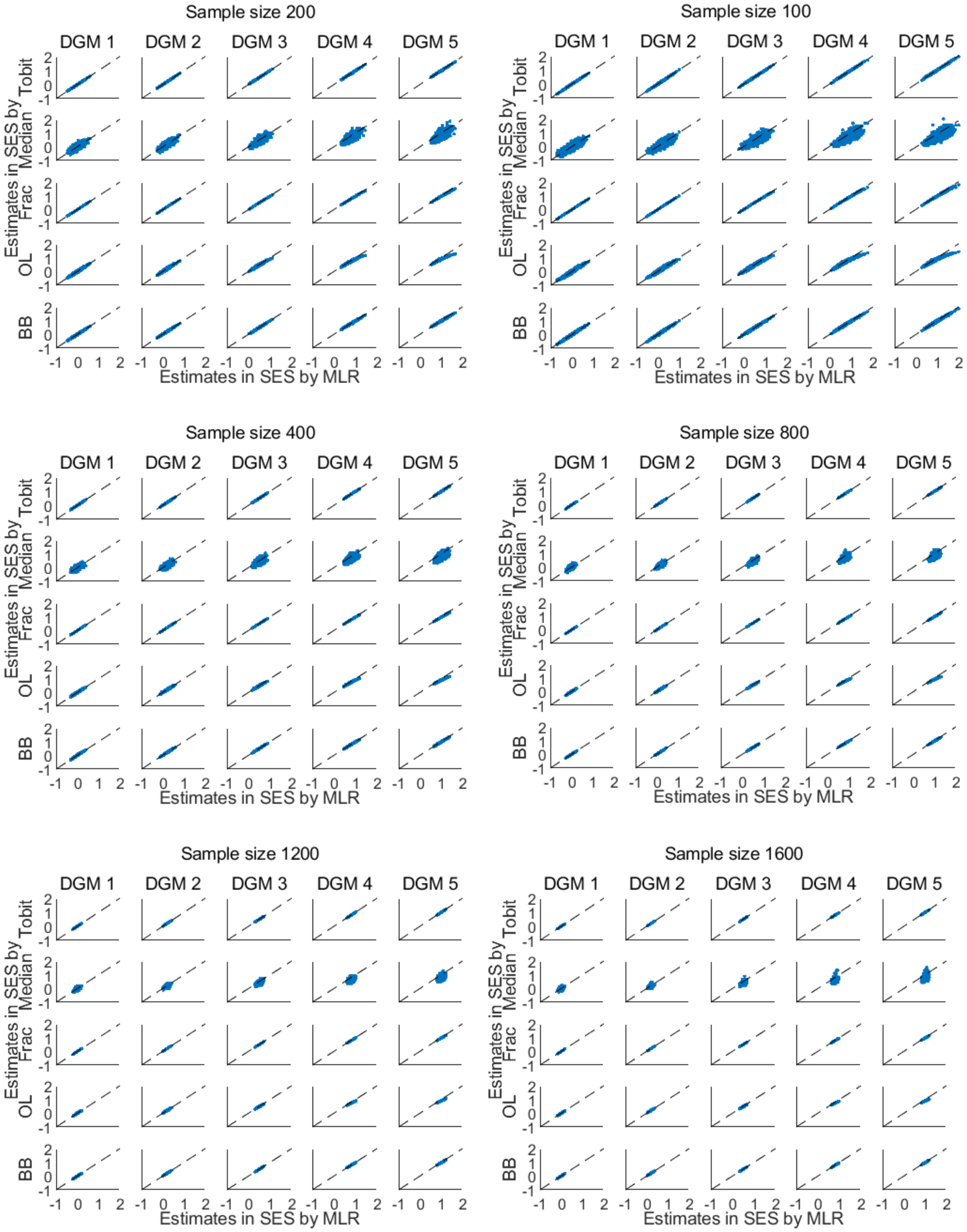


Figure D.1 Scatterplots of SES using MLR as baseline

BB, beta-binominal regression. DGM, data-generating mechanism; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model; Tobit, Tobit regression.

D.2 Stata codes for the simulation analysis

Stata codes to simulate datasets and to analyse the simulated datasets are presented under this section. We first generate datasets using random-number generator under the Normal distribution, using six different numbers of repetitions (i.e. 100, 200, 400, 800, 1,200, and 1,600). The generated continuous data are rescored to pre-specified categorical values using the recoding techniques. Then, we analyse the simulated dataset with the six statistical methods (i.e. MLR, Tobit, Median, Frac, OL, and BB), and record their number of missing values, estimated treatment differences, associated standard errors, and the observed treatment differences. The dimension score with 10 possible ordinal categorical values (`level = 10`) is used as an example.

D.2.1 Main codes

```
set seed 1

tempname postDGM

tempname postseed

postfile `postseed' str2000 s1 str2000 s2 str1100 s3 using
seedfileL10_fixedsd.dta, replace

postfile `postDGM' int(i) byte(dgm) str7(method) float(level SZ n1
n2 mean1 mean2 Dmean theta se) using
DGMscorewloopNulllevel10_fixedsd, replace

local a = 1 //sample size controller

local m = 50 //mean value of the Normal distribution

local sd = 22 //standard deviation of the Normal distribution

local d1 = 4.4 //treatment difference (small effect)

local d2 = 11 //treatment difference (median effect)

local d3 = 17.6 //treatment difference (large effect)

local d4 = 22 //treatment difference (very large effect)

local count = 0 //number of iteration tracer

local level = 10 //number of possible ordinal categorical values


while `a' <= 6 {

     forvalues i = 1/5000 { //i is the number of simulation

          clear

          include "Loop4SampleSize.do"

          post `postseed' (substr(c(rngstate),1,2000))
(substr(c(rngstate),2001,2000)) (substr(c(rngstate),4001,.))
```

```
           set obs `SZ'
           //DGM 1: null
           gen rannum = uniform()
           egen group = cut(rannum), group(2)
           gen score1 = rnormal(`m', `sd')
           //DGM 2-5: alternative
           gen score2 = score1
           replace score2 = score1 + `d1' if group == 1
           gen score3 = score1
           replace score3 = score1 + `d2' if group == 1
           gen score4 = score1
           replace score4 = score1 + `d3' if group == 1
           gen score5 = score1
           replace score5 = score1 + `d4' if group == 1

           local x = 1
           local y = 5 //number of score distributions
           include "Loop4Regression.do"
           local count = `count' + 1
           di `count'
      }
      local a = `a' + 1
}
postclose `postDGM'
postclose `postseed'
```

D.2.2 Codes for specify different sample sizes ("Loop4SampleSize.do")

Loop for sample size

```
if `a' == 1 {
      local SZ = 100
      }
      else if `a' == 2 {
            local SZ = 200
            }
      else if `a' == 3 {
            local SZ = 400
            }
      else if `a' == 4 {
            local SZ = 800
            }
      else if `a' == 5 {
            local SZ = 1200
            }
      else {
            local SZ = 1600
      }
```

D.2.3 Codes for analysing simulated dataset ("Loop4Regression.do")

```
while `x' <= `y' {
      include "Loop4Recoding.do"
      su oscore`x' if group == 0
      local n1 = r(N)
      local mean1 = r(mean)
      su oscore`x' if group == 1
      local n2 = r(N)
      local mean2 = r(mean)
      local Dmean = `mean2' - `mean1'
      //MLR
      regress oscore`x' group
      post `postDGM' (`i') (`x') ("MLR") (`level') (`SZ') (`n1')
(`n2') (`mean1') (`mean2') (`Dmean') (r(table)[1,1]) (r(table)[2,1])
      //Tobit
      tobit oscore`x' group, ll(0) ul(100)
      post `postDGM' (`i') (`x') ("Tobit") (`level') (`SZ') (`n1')
(`n2') (`mean1') (`mean2') (`Dmean') (r(table)[1,1]) (r(table)[2,1])
      //Median
      capture qreg oscore`x' group
      if _rc == 1 {
            exit 1
      }
      else if _rc == 0 {
            local theta = r(table)[1,1]
            local se_theta = r(table)[2,1]
      }
      else {
            local theta = .
            local se_theta = .
      }
      post `postDGM' (`i') (`x') ("Median") (`level') (`SZ') (`n1')
(`n2') (`mean1') (`mean2') (`Dmean') (`theta') (`se_theta')
      //Frac
      fracreg logit pscore`x' group

      post `postDGM' (`i') (`x') ("Frac") (`level') (`SZ') (`n1')
(`n2') (`mean1') (`mean2') (`Dmean') (r(table)[1,1]) (r(table)[2,1])

      //OL

      ologit oscore`x'_c group

      post `postDGM' (`i') (`x') ("OL") (`level') (`SZ') (`n1')
(`n2') (`mean1') (`mean2') (`Dmean') (r(table)[1,1]) (r(table)[2,1])

      //BB

      capture betabin oscore`x'_c group, n(`level') link(logit) nolog

      if _rc == 1 {

            exit 1

      }

      else if _rc == 0 {

            local theta = r(table)[1,1]

            local se_theta = r(table)[2,1]

      }

      else {

            local theta = .

            local se_theta = .

      }

      post `postDGM' (`i') (`x') ("BB") (`level') (`SZ') (`n1')
(`n2') (`mean1') (`mean2') (`Dmean') (`theta') (`se_theta')

      local x = `x' + 1

}
```

D.2.4 Codes for recoding continuous data to ordinal categorical data ("Loop4Recoding.do")

```
if `level' == 10 {

      recode score`x' (min/5.55=0) (5.55/16.65=11.1)
(16.65/27.75=22.2) (27.75/38.85=33.3) (38.85/49.95=44.4)
(49.95/61.05=55.6) (61.05/72.15=66.7) (72.15/83.25=77.80)
(83.25/94.35=88.9) (94.35/max=100.0), gen(oscore`x')

      generate oscore`x'_c = irecode(score`x', 5.55, 16.65, 27.75,
38.85, 49.95, 61.05, 72.15, 83.25, 94.35)

      gen pscore`x' = round(oscore`x'/100,0.001)
```

```
        }

        else if `level' == 4 {

                recode score`x' (min/16.65=0) (16.65/49.95=33.3)
(49.95/83.25=66.6) (83.25/max=100.0), gen(oscore`x')

                generate oscore`x'_c = irecode(score`x', 16.65, 49.95,
83.25)

                gen pscore`x' = round(oscore`x'/100,0.001)

        }

        else if `level' == 26 {

                recode score`x' (min/2=0) (2/6=4) (6/10=8) (10/14=12)
(14/18=16) (18/22=20) (22/26=24) (26/30=28) (30/34=32) (34/38=36)
(38/42=40) (42/46=44) (46/50=48) (50/54=52) (54/58=56) (58/62=60)
(62/66=64) (66/70=68) (70/74=72) (74/78=76) (78/82=80) (82/86=84)
(86/90=88) (90/94=92) (94/98=96) (98/max=100), gen(oscore`x')

                generate oscore`x'_c = irecode(score`x', 2, 6, 10, 14,
18, 22, 26, 30, 34, 38, 42, 46, 50, 54, 58, 62, 66, 70, 74, 78, 82,
86, 90, 94, 98)

                gen pscore`x' = round(oscore`x'/100,0.001)

        }

        else {

                local a = 999

        }
```

Notes:

1. The seed and stream for each simulated data is recorded using a seed file. This can help retrieve the failure by rerunning a certain iteration using the recorded seed (in the coding example, it is called `seedfileL10_fixedsd.dta`). As we aim to use the same set of seed and stream for different levels, the recorded seeds from these simulations can be recorded and they are expected to be identical across different levels.
2. Command `capture` is used in Stata to continue the loop despite non-convergence or other errors are shown. It is used for Median and BB in our codes, as they are the methods which have shown missing values previously.