Improving the specification of the target difference in the sample size calculation of a randomised trial of treatments for osteoarthritis

Bethan Copsey

Pembroke College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

For Ellie,

who never doubted that I could do this.

Personal Acknowledgements

Thank you to:

Mum and Dad for always picking up for phone in good times and bad, and to Ellie and James for making time for me, giving me lifts, and making me gin & tonics.

The rest of my amazing family for their never-ending love and support (even if they have no idea what I work on): Grandma Phyllis; Sheila, Dave and John; Sarah and Paul; Tom, Eliza and Finnick; Tess and Shay; Roly and Helen.

Jenny Leech for taking the time to send long WhatsApp messages and wonderful care packages to cheer me up. I am so grateful to have you in my life.

Dymphna Murphy and Katy Bryer for the drinks, dinners and many visits to Rick's caf. You each remind me of my strength and I always feel better about myself after seeing you.

Alex Edge for the roast dinners, cups of tea, movies and countryside walks. Our time together has created some of my happiest memories.

Melissa Jansen van Rensburg, Katriel Cohn-Gordon and Michael Peyton Jones for Friday night meatballs dinners, board games and baking sessions. I feel privileged to have been welcomed into your house and lives. A special thank you to Melissa for our Nashville and crochet sessions, your excellent book recommendations and always helpful advice.

The many other housemates who have put up with me throughout this process including Tamla Wharvell, Dylan Barrett, David Quartey and Rachel Johnston.

My Warwick friends, Julian Bhardwaj, Solene van der Wielen, Hayley Clissold, and Josh Wyatt, for sticking by me and reminding me how much I've grown.

My 'home home friends' from Cheshire for always being there for me and reminding me that I am loved, even if we only see each other a few times a year, especially Tay and Mike Wolffe, Caz and Dave Warderley, Alice Marston, Joel and Emma Clarke, Tom Towers, and Jess Clark.

The Oxford drunken knitwits for giving me an escape every Wednesday, the weekend retreats and letting me be part of such a friendly, amazing community, with special mentions to Caroline Phipps, Tom Jackson, and Jess Doondeea, as well as Lisa Vokes, Janey Messina, Sue Aungier, and Amy Barnes.

My primary supervisor, Jonathan Cook, for all of your help in the past few years, listening to my stress waffle and teaching me to celebrate even the smallest achievements. I wouldn't have been able to do this without you.

My co-supervisors, James Buchanan, Ray Fitzpatrick, Sallie Lamb and Sue Dutton, for their positivity and encouragement.

All of the RRIO team for making me feel valued. Special thank you to Amanda Hall, Helen Richmond, Jacqueline Thompson, Sue Davolls, Beth Fordham, Pippa Nicolson, Hopin Lee, and Gill Jones for your coffees, walks in the park, and phone calls.

Corran Roberts and Paula Dhiman for the coffees, lunches and gossip, and for laughing at my funny stories.

Everyone else at CSM and OCTRU for making me feel welcome and supporting me from application to completion of this DPhil, including Virginia Chiocchia, Emmanuel Ogundimu, Ayo Odutayo, Sally Hopewell, Steve Gerry, Jen de Beyer and Angela MacCarthy.

Several other NDORMS staff and students who have helped me in very different but all incredibly important ways including Pete Salmond, Ed Burn, Klara Berensci, and Afsie Sabokbar.

David Newport who coached pilates at the Botnar and Jen Middleton at Bodyburn Oxford for motivating me to exercise and de-stress. Both of your classes have helped to improve my productivity and positivity during the PhD process and reminded me what is important in life.

Technical Acknowledgements

This project was funded by a doctoral studentship from the EPSRC (Engineering and Physical Sciences Research Council) and Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences via the Medical Sciences Division of the University of Oxford (grant number MR/K501256/1).

During the time I have completed this DPhil, I have received financial recompense for employment from other areas in the University of Oxford, including Pembroke College, the IT Learning Centre, National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care Oxford at Oxford Health NHS Foundation Trust, and the National Institute for Health Research, Health Technology Assessment (HTA) Programme. The views expressed in this thesis are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health, or the EPSRC.

I would like to acknowledge the work of Jacqueline Thompson, Karan Vadher and Usama Ali, who conducted eligibility screening and data extraction for the systematic review in Chapter 2.

I wish to acknowledge Dr Simon Abram for his helpful advice and comments during the early conception of the discrete choice experiment in Chapter 4. I also wish to acknowledge the project management team and panel members at ResearchNow for providing the sample of participants for the discrete choice experiment in Chapter 4. Ethical approval for the discrete choice experiment in Chapter 4 was obtained from the University of Oxford Central University Research Ethics Committee (reference R55785/RE002).

Data used in the preparation of Chapter 5 were obtained from the Osteoarthritis Initiative (OAI) database, which was downloaded on 18 January 2018 and is available for public access at http://www.oai.ucsf.edu/. The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. Chapter 5 of this thesis was prepared using an OAI public use data set and

does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.

I would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out the work in Chapter 6 (http://dx.doi.org/10.5281/zenodo.22558).

Finally, I acknowledge English language editing by Jennifer A de Beyer of the Centre for Statistics in Medicine, University of Oxford.

Abstract

Thesis title: Improving the specification of the target difference in the sample size calculation of a randomised trial of treatments for osteoarthritis

Candidate name: Bethan Copsey (Pembroke College)

Thesis submitted for the degree of: Doctor of Philosophy (in Trinity 2019)

The sample size of a clinical trial is the number of participants the trial aims to recruit. Sample size is a critical aspect of clinical trial design and has ethical and financial implications. The sample size depends on the target difference, the difference in outcome that the trial is powered to detect. This thesis aims to improve methods for specifying the target difference in randomised trials of osteoarthritis.

I conducted a systematic review of sample size calculations in hip and knee osteoarthritis trials published in 2016. It found that most sample size calculations were poorly reported and could not be reproduced. The target difference in the sample size calculation was commonly justified by a published minimum clinically important difference (MCID).

Several versions of the WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) were commonly used in hip and knee osteoarthritis trials. It was often unclear which version was used, hindering interpretation of trial results.

I conducted a discrete choice experiment examining patient preferences when choosing between osteoarthritis medications. Duration of treatment effect was shown to be important to participants, viewed with similar importance to the amount of symptom relief provided and risks of the treatment.

I analysed a cohort of people with osteoarthritis and showed that MCID estimates for the WOMAC varied across different follow-up time points. However, there was no visual trend in the change in MCID estimates over time. Longitudinal methods were feasible to calculate MCID estimates, but did not improve precision.

A simulation study that I conducted found that the pattern of the treatment effect (its duration and consistency) affected the optimal statistical method of analysis for a randomised trial using the WOMAC as the primary outcome.

Future research is needed to examine whether the findings are generalisable to different datasets, outcome measures and health conditions.

Contents

List of Figures

List of abbreviations

ANCOVA: Analysis of Covariance

ARC: Advanced Research Computing

AUC: Area Under the Curve

CES-D: Center for Epidemiological Studies - Depression

CI: Confidence Interval

CONSORT: Consolidated Standards of Reporting Trials

DCE: Discrete Choice Experiment

DELTA: Difference Elicitation in Trials

EMA: European Medicines Agency

EmpSE: Empirical Standard Error

FDA: United States Food and Drug Administration

GEE: Generalised Estimating Equation

GP: General Practitioner

GROC: Global Rating of Change

HOOS: Hip disability and Osteoarthritis Outcome Score

ICC: Intra-class Correlation Coefficient

IQR: Interquartile Range

JSN: Joint Space Narrowing

KOOS: Knee injury and Osteoarthritis Outcome Score

LCI: Lower Confidence Interval

MAR: Missing At Random

MCAR: Missing Completely At Random

MCIC: Minimum Clinically Important Change

MCID: Minimum Clinically Important Difference

MCSE: Monto Carlo Standard Error

MNAR: Missing Not At Random

ModSE: Average Model-based Standard Error

MRS: Marginal Rate of Substitution

NIM: Non-inferiority Margin

NNT: Number Needed to Treat

NPRS: Numerical Pain Rating Scale

NR: Not Reported

NRS: Numerical Rating Scale

NSAID: Non-Steroidal Anti-Inflammatory Drugs

OA: Osteoarthritis

OAI: Osteoarthritis Initiative

OARSI: Osteoarthritis Research Society International

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

REML: Restricted Maximum Likelihood

ROC: Receiver Operating Characteristic

SD: Standard Deviation

SES: Standardised Effect Size

SF-12: 12-item Short Form Health Survey

UCI: Upper Confidence Interval

UK: United Kingdom

VAS: Visual Analogue Scale

WOMAC: Western Ontario and McMaster Universities Osteoarthritis Index

WTR: Willingness to Risk

Chapter 1

Introduction

1.1 The importance of an appropriate sample size

The target sample size of a clinical trial is the number of participants the trial aims to recruit. The sample size is a critical aspect of trial design. The use of an inappropriate sample size has ethical and financial implications.

A sample size that is smaller than appropriate could lead to inconclusive or potentially misleading trial results. Trials with insufficiently large sample sizes are more likely to produce a statistically insignificant result, missing a true clinically meaningful treatment effect [2]. An overly large sample size will have sufficient power to detect a clinically meaningful effect and has the potential to change clinical practice. However, it will be more costly and consume more resources than necessary [3].

In both cases, whether the sample size is too small or too large, a trial will consume research-based resources and funding that were not required to improve the scientific

integrity of the trial and some or all of the participants will be recruited unnecessarily. These limited resources could have been allocated to other trials designed to appropriately answer the research question. For overly large trials, a trial could have included fewer participants and reduced the length of the trial. Some of the 'extra' participants may also have unnecessarily received an inferior (and potentially harmful) treatment. An overly large trial could also be misleading in detecting a statistically significant treatment difference that is not clinically worthwhile.

Trials with an inappropriate sample size could even harm the evidence base. An overly large trial will usually have longer recruitment and follow-up periods. This will delay the publication and dissemination of the trial results. If a treatment was found to be more effective than conventional care, the delayed dissemination could then delay the uptake of the more effective treatment and its implementation into clinical practice. A trial with too few participants could produce an insignificant result when a true clinically important treatment effect exists. This could prevent or delay further trials being conducted on the same treatment if potential researchers wrongly believe that the overly small study provides evidence of an absence of treatment effect [4]; this can cause contention and confusion [5]. In addition, researchers may choose not to focus on the area considered while the small study is on-going to avoid duplication of effort. In this way, a small underpowered study can actually diminish the evidence base if it delays the initiation of larger confirmatory trials.

Therefore, it is undesirable to have a sample size that is either too small or too large as it: (i) wastes research funding and resources, (ii) unnecessarily subjects trial participants to potentially useless or harmful treatments, and (iii) delays the implementation of effective treatments into clinical practice.

1.2 Sample size calculation

The sample size of a randomised trial can be calculated using various methods [6], for example:

- Using the conventional Neyman-Pearson approach (described in this section below)
- Precision approach: Calculating the sample size, assuming known parameters (such as standard deviation for continuous outcomes) to achieve the specified precision of the effect estimate (e.g., width of confidence interval).
- Bayesian methods: One example of a Bayesian approach to sample size calculation assumes a joint prior distribution for the treatment difference and other unknown parameters (e.g., the standard deviation of a continuous treatment outcome). Simulations are used to specify a sample size that maximises the average power across this distribution. This could include adaptive trial designs.
- Value of information: This approach determines an optimal sample size based on the 'value' of additional information gained compared to the cost of recruiting additional participants. If the trial finds the 'wrong' result and an intervention is incorrectly implemented (or not implemented), this would result in lost health gain. This loss in health gain would be more likely if the trial is smaller. This aims to find the sample size that maximises the difference between the expected loss in health gain and the estimated cost of the trial.

By far the most common method used to calculate the sample size of a randomised controlled trial is the conventional Neyman-Pearson approach, which is the focus for this thesis. Equation 1.1 shows the formula based on the Neyman-Pearson approach for a two-arm trial with a continuous primary outcome, assuming that the outcome

is normally distributed, there is equal allocation (1:1) between treatment arms and that the standard deviation of the outcome is known and equal in the two treatment arms.

$$\frac{2\sigma^2(z_{\alpha/2} + z_{\beta})^2}{(\mu_1 - \mu_2)^2} \tag{1.1}$$

The components necessary to conduct the calculation are the:

1. Target difference $(\mu_1 - \mu_2)$
2. Standard deviation (σ)
3. Significance level or type I error (α)
4. Power $(1 - \beta)$, where β is the type II error

Although trials commonly collect and analyse outcomes at multiple assessment time points, the sample size calculation is usually based on a treatment comparison at a single time point. This implicitly assumes that the target difference in the outcome is constant across all assessment time points, such that the study power remains high for treatment comparisons at any time point in the follow-up period. Alternatively, in practice, trialists may power their study for the main time point and consider the treatment effect at other time points to be less important (especially when the effect is likely to be smaller).

1.2.1 Specification of target differences

Although the required sample size is dependent on the outcome type, statistical parameters and planned analysis, it is typically most sensitive to the target difference, the difference in outcome that the trial is powered to detect [7]. The choice of target

difference can have a dramatic impact on the required sample size. Using equation (1.1) and keeping the type I and type II errors constant, halving the target difference that you are aiming to detect increases the target sample size by approximately four times as many participants.

Specifying the target difference is arguably one of the most important and most difficult decisions when designing a clinical trial [3]. In theory, we would like a trial to be able to detect a treatment effect no matter how small. However, in practice, this would not be feasible because it would require an infinitely large sample size. However, when using a large target difference, it is more likely that the trial will 'miss' a small treatment effect that would be relevant in clinical practice. Therefore, trialists attempt to determine the target difference that is the smallest clinically meaningful difference between the two treatments. Pocock stated that a trial should be designed to identify "the smallest difference that is of such clinical value that it would be very undesirable to fail to detect it" [8].

Cook *et al.* reviewed methods for specifying the target difference in randomised trials and identified the seven methods described below [9].

- Anchor: Comparing the difference in the primary outcome to an important difference in a gold-standard measure.
- Distribution: Using solely the statistical properties of the primary outcome measure.
- Standardised effect size: Relating the distributional variation to classifications of standardised effect size based on Cohen's d.
- Health economic: Selecting the target difference that corresponds to the cost-effectiveness threshold (e.g., cost/benefit ratio) based on the associated treatment costs.

- Opinion seeking: Surveying the opinions of key stakeholders, for example, using a DELPHI-style consensus-based study.
- Pilot study: Using a pilot study for the trial to inform the target difference for a full-scale trial.
- Review of evidence base: Reviewing the treatment effects found in the literature from previous randomised trials, meta-analyses or observational studies.

1.3 Anchor-based methodology

Anchor-based methods are the most commonly used to specify target differences [9]. The anchor-based approach specifies the target difference by calculating the difference in the primary outcome of the trial that corresponds to an important difference in an anchor measure.

Anchor-based methods determine the smallest difference that is considered clinically meaningful. This corresponds directly to the concept of a minimum clinically important difference (MCID). Jaeschke defines the MCID as the "smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" [10].

1.3.1 MCID terminology and methodology

There is currently no consensus on the optimal method to calculate an MCID estimate [11, 12]. The anchor-based method is favoured because it incorporates patient opinion and quantitative data [13]. The anchor-based method calculates the mean change in the primary outcome among a group of participants who have been deemed to have

a small but important change. This group of participants are selected based on the change in the 'anchor' measure. The anchor measure is often a patient-reported measure of the perceived treatment benefit, but could be a clinician-reported outcome or an objective biological measure. For example, a study using a global rating of change (GROC) as the anchor measure may examine the change within, and between, participants who describe their change from baseline as 'slightly improved' [13].

As well as dispute around the methodology for MCID calculation, there is also debate surrounding the appropriate terminology to use to define an important difference. The literature includes various definitions of what constitutes an MCID and these definitions are not always consistent, which can cause confusion. The term 'minimum clinically important difference' could be seen as a misnomer as it may be based on patient opinion and does not necessarily account for the opinion of a clinician. Therefore, some researchers prefer to use the term 'minimum important difference' (MID) [14, 15]. The same terms are often applied interchangeably to refer to within-person differences and between-group differences. In practice, it is important to differentiate between a within-person and between-group difference because a difference that is important when comparing two interventions may not be important if it is experienced as a change over time for a single person.

The definitions used in this thesis are given below. Although I am interested in patient opinion, I have used the abbreviation MCID as this is more commonly used than 'MID', likely due to its use in Jaeschke's seminal paper [10].

MCID (Minimum Important Difference): The MCID is defined as the smallest between-group difference that is considered worthwhile. This is used to assess the importance of a difference between groups of people, such as comparing treatment arms, at a single time point.

MCIC (Minimum Important Change): The MCIC is defined as the smallest within-person change that is considered worthwhile. This is used to assess the importance of a change over time of an individual person.

1.3.2 Use of MCID estimates

The usefulness of an MCID (and MCIC) as a concept is widely supported [12]. If an intervention produces a statistically significant difference (between groups or for a single participant), the difference may be very small and therefore not produce a noticeable effect on the participant's quality of life. The MCID provides a minimal value that the between-group difference should exceed to enact change in the management or treatment regime of a group of people. It facilitates the evaluation of treatment interventions by allowing us to assess whether the group of people achieved a worthwhile difference in their condition.

As discussed above, MCID estimates can be used in trial design to specify the target difference in the sample size calculation. As well as being applied in trial design, MCID estimates are used in the interpretation of clinical trial results. Trials can compare the proportion of participants who exceed an MCIC between treatment arms [16, 17, 18, 19, 20]. The proportion of participants reaching the MCIC could also be examined in observational cohort or case-control studies [21, 22, 23, 24, 25]. Some studies use a participant exceeding the MCIC as a predictor of treatment success [26, 27, 28, 29, 30, 31]. MCID estimates can also be used in the interpretation of between-group differences in randomised trials and meta-analyses, to determine whether the between-group mean difference is clinically meaningful [32, 33, 34, 35]. The use of MCID estimates to interpret the pooled between-group differences in meta-analyses has been recommended to incorporate a measure of clinical relevance into the results of systematic reviews [36].

1.3.3 Variability of MCID estimates

Some argue that the usefulness of an MCID is limited by its variability [12]. MCID estimates have been shown to vary based on the calculation methods used and clinical factors, such as the population and intervention being considered. Studies have found that MCID estimates differ based on baseline disease severity [37, 38, 39, 40]. Others have hypothesised that MCID estimates may vary by other participant characteristics, such as age, sex, and ethnicity, but few have found strong associations [38, 41]. Whether a change in outcome is seen as clinically important may also depend on psychological factors, such as patient expectation, optimism or mental well-being [42, 43, 44]. The MCID estimate can also depend on co-morbidities, especially if the intervention effects are specific to only one condition. An improvement in one condition could be obscured by symptoms such as pain and fatigue due to other conditions [45]. The intervention being considered could also influence the MCID estimate. It may be that a larger difference is required to be seen as worthwhile for interventions that are more burdensome, such as surgical treatments [46]. Researchers may need to pay careful attention to these clinical and methodological factors when selecting which MCID estimate from the literature to apply.

Maltenfort *et al.* suggested that MCID values may vary based on "the timetable of recovery" [47]. A small number of studies have examined variability in MCID estimates over time. Some studies have found MCID estimates differ by follow-up time point [46, 48, 49]. However, other studies have found MCID values to be similar across follow-up time points [41, 50, 51].

Existing methods, particularly those used in economics, demonstrate that time may be an important factor in the evaluation of treatment effectiveness. For example, quality-adjusted life years have been used to summarise outcomes while allowing effectiveness to vary over the participant's lifetime. Similarly, economists often use

discounting to reduce the weighting of health states that occur further into the future [52]. This provides evidence that people may value treatment benefits differently depending on when the treatment benefits occur.

1.4 Rationale for thesis

Studies commonly calculate the MCID for a disease-specific patient-reported outcome at a single time point. For long-term conditions, the length of time that the benefit from a treatment lasts is important. It is unclear how MCID values might vary over time and how changes in the effect over time should be viewed. For example, a person could find a sustained improvement below the MCID to be more meaningful than a short-term difference that meets the MCID at a single time point but fades rapidly.

The time point at which an important difference is assessed may also have implications for the analysis and subsequent interpretation of trials and how the scale of benefit is appraised in trials in long-term conditions. If MCID estimates do indeed vary over time, it is possible that MCID values are being applied inappropriately in trial design and interpretation.

The research presented in this thesis examined the variability of MCID estimates over time to indicate whether the assessment time point should be incorporated into the target difference specification when calculating the sample size for a trial. It also explored the use of longitudinal methods to incorporate time into the calculation of MCID estimates.

The findings of this research will help trialists to understand whether the assessment time point should be considered when specifying the target difference in a sample size calculation and will compare different methods that could be used to account for this. A more appropriately chosen sample size will help to ensure that trials are

well-designed and will help to reduce waste in research and improve patient health outcomes. The results will also raise broader issues related to how we estimate the scale of benefit in trials in long-term conditions.

1.4.1 Simple example to demonstrate the potential impact of different target differences

Tashjian *et al.* found an MCID of 1.37cm on the 0-10cm visual analogue scale (VAS) for pain among rotator cuff repair patients [53]. We can hypothetically assume the MCID at an alternative time point to be 1.2cm or 1.54cm. If we assume a standard deviation of 2.5 to be constant across all time points, 90% power and 5% two-sided level of significance:

1. Original estimate: To detect a target difference of 1.37 will require a sample size of 142 participants
2. Estimate using a lower MCID: To detect a target difference of 1.2 will require a sample size of 186 participants.
3. Estimate using a higher MCID: To detect a target difference of 1.54 will require a sample size of 114 participants.

The reduction of the target difference to 1.2cm increases the sample size by 44 participants, over 30% more than the original estimate. Similarly, in the opposite direction, to detect a target difference of 1.54, 28 fewer participants will be required, 20% less than the original estimate.

These variations have only adjusted the target difference by 0.17cm, which is less than 2% of the full measurement scale. However, these differences could substantially affect the sample size and thus the conduct of the trial, in terms of the number of centres, length of the recruitment period and overall costings. A review by Speich

et al. found the median costs per patient recruited to a randomised trial was $409 (ranging from $41 to $6990) [54].

The chosen MCID value could also have implications for the interpretation of trial results. It could affect whether between-group differences are viewed as clinically important, if they fall within the range of possible MCID values. If only the MCID estimate of 1.37 points was available, a between-group difference of 1.3 points might not be seen as clinically significant. However, if the MCID of 1.2 points had been calculated, e.g., using data from a shorter follow-up period, the between-group difference of 1.3 points would have exceeded the MCID, so would have been seen as clinically meaningful.

Small variations in the MCID value used can have important implications when using the MCID estimate to calculate the sample size and interpreting the results of future trials. Therefore, if MCID values vary by assessment time point, it is important to account for these time-varying effects and apply the appropriate MCID for the corresponding follow-up period.

1.5 Clinical application: Osteoarthritis

1.5.1 Osteoarthritis

The research presented in this thesis focuses on osteoarthritis, which is highly prevalent and can substantially affect people's quality of life. There are over 25 million people living with osteoarthritis in the US alone [55]. Osteoarthritis is associated with significant healthcare resource use and economic costs [56]. It has also been shown to have a high economic burden [57, 58] and is one of the leading causes of disability, contributing 17.1 million years lived with disability in 2010 globally [59].

In the UK, osteoarthritis is reported to be the most common cause of disability [60] and, almost half of people aged 75 and over have sought treatment for osteoarthritis. As the prevalence of osteoarthritis increases with age, rising life expectancy and population growth mean that the number of people living with osteoarthritis is expected to increase rapidly over the next decade [61, 62].

1.5.2 Research into osteoarthritis treatments

Current treatments for osteoarthritis include surgical interventions (primarily joint replacement), pharmacological therapies (such as non-steroidal anti-inflammatory drugs (NSAIDs) or hyaluronic acid) and other conservative treatments (including education and exercise) [63, 64, 65]. However, currently available treatments largely treat the symptoms of osteoarthritis, particularly pain and physical function, rather than the underlying disease pathology [62]. According to the ACR Clinical Classification Criteria, diagnosis of osteoarthritis is dependent on the specific joint being examined and includes physical symptoms with or without radiographic imaging or laboratory findings [66, 67]. However, there is no single definition for what osteoarthritis actually is, and the mechanisms for osteoarthritis remain unclear [68, 69].

A recent update of the Osteoarthritis Research Society International (OARSI) guidelines found uncertainty regarding the appropriateness of many treatments for knee osteoarthritis either generally or for specific phenotypes (for example, those with co-morbidities or multiple joint osteoarthritis) [70]. As more is learned about the pathology of osteoarthritis, further research could explore tailoring treatments based on the patient's pathology [65, 68]. Although short-term disease markers and radiological outcomes may be preferred to assess the efficacy of novel treatments [68], large, confirmatory trials can be used to ensure this translates into improvements in the quality of life of people living with osteoarthritis.

1.5.3 Justification of application to osteoarthritis

There is a need for clinical research on osteoarthritis due to the high prevalence, societal impact and uncertainty surrounding the optimal treatment regimens. Osteoarthritis is therefore an appropriate disease area to consider to explore the methodology of clinical trial design.

This thesis explores the incorporation of time effects into trial design using a continuous, patient-reported primary outcome. The research is therefore relevant to osteoarthritis and other long-term conditions with low mortality rates. For example, long-term monitoring of osteoarthritis is recommended in clinical practice and randomised trials [60, 71]. The research would be less relevant for other clinical areas. For high mortality conditions, such as cancer, survival is of primary importance and therefore trial design considerations will be based on survival outcomes, as opposed to patient-reported outcomes. There is also limited applicability to acute conditions where the patient is expected to recover quickly and thus long-term follow-up may not be the main focus. Osteoarthritis is a suitable clinical application because trials in osteoarthritis commonly measure patient-reported outcomes using repeated assessment over time.

People with osteoarthritis tend to live many years before having joint replacement surgery; for example, Gademan *et al.* found that only 10% of people with early osteoarthritis symptoms received an arthroplasty after 9 years [72]. Due to concerns about the longevity of joint replacements, many people are encouraged to live with symptoms for as long as possible before proceeding to surgery [72, 73]. Some people living with osteoarthritis may never be considered suitable for surgery, due to their age or personal preference [74]. Hence, long-term effects of treatments on osteoarthritis symptoms are important.

1.6 Thesis aims and objectives

Trialists commonly determine the MCID at only one time point, which does not account for changes in effect over time and hence ignores the chronic nature of osteoarthritis. Increased understanding is needed about how the magnitude of difference that participants consider to be important varies over time. Calculations of the sample size and underlying target difference in a primary outcome need to take any time-dependent effects into account. The existence of these effects also has implications for analysing trial results and interpreting trial findings.

It is unclear how investigators currently specify target differences for disease-specific patient-reported outcomes and which methods are appropriate for calculating target differences of treatment effects with potentially long durations, such as in osteoarthritis trials.

This thesis aims to explore whether it is necessary to incorporate time into sample size calculations for trials in long-term conditions, such as osteoarthritis, and, if so, how trialists should approach this.

1.7 Structure of thesis

Chapter 1, this chapter, introduces the concepts of sample size calculation and important differences, presents the aims of the thesis and justifies the application to osteoarthritis.

Chapter 2 reviews current practice for sample size calculations in hip and knee osteoarthritis trials. The systematic review examines the methodology, reporting and reproducibility of the sample size calculations.

Chapter 3 reviews the use, analysis and reporting of the WOMAC Index, a commonly used outcome measure for hip and knee osteoarthritis trials. It uses a cohort of hip and knee osteoarthritis trials identified from the review in Chapter 2.

Chapter 4 presents the design and results of a discrete choice experiment to evaluate how duration of treatment benefit affects treatment preferences in people living with osteoarthritis.

Chapter 5 assesses the stability of important differences over time and explores the use of longitudinal methods to calculate important differences using secondary analysis of an observational cohort study.

Chapter 6 uses simulations to compare the statistical properties of different longitudinal methods to analyse trial results.

Chapter 7 summarises the over-arching results and discusses the potential implications and limitations of the findings.

Chapter 2

Current practice regarding sample size calculation in randomised trials of hip and knee osteoarthritis: a systematic review

Prior publication:

The methods and results for this chapter were published as manuscripts. Conferences abstracts for presentations on this chapter have also been published (see Appendix F.1 for details).

Copsey B, Dutton S, Fitzpatrick R, Lamb SE, Cook JA: Current practice in methodology and reporting of the sample size calculation in randomised trials of hip and knee osteoarthritis: a protocol for a systematic review. Trials 2017, 18(1):466.

Copsey B, Thompson JY, Vadher K, Ali U, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA: Sample size calculations are poorly conducted and reported in many randomized trials of hip and knee osteoarthritis: results of a systematic review. Journal of Clinical Epidemiology 2018, 104:52-61.

2.1 Introduction

2.1.1 Importance of sample size calculation

A key component of clinical trial design is the sample size calculation. As discussed in Chapter 1, the choice of sample size is important for ethical, practical and financial reasons. An overly large sample size is undesirable as it increases the costs of the trial and likely delays dissemination of the findings [3]. Too large a sample size is also unethical as it may result in additional participants receiving a treatment when there is already sufficient evidence to show it is inferior to an alternative [75]. A sample size that is too small may lead to the trial not having sufficient precision and a clinically important treatment effect, should it exist, is more likely to be 'missed' [2, 76].

Altman *et al.* emphasised the importance of justifying the sample size when reporting the results of a trial [77]. An *a priori* sample size calculation encourages trialists to explain why the number of participants recruited is smaller than the target sample size and indicates the primary outcome of the trial. It also indicates that the trial was properly designed [77]. The justification of the sample size of a trial is also key to avoiding research waste.

Williamson *et al.* discuss the need to assess the scientific validity of a trial "Will the research project, as designed, answer the question being asked?". Justification of the sample size is an important component of this [78]. When a sample size justification is adequately reported, it allows the reader to understand what the trial was designed to achieve. The sample size for phase III randomised trials is commonly justified using a sample size calculation, usually done in the same way. Typically, the primary outcome and the difference between treatments that the trial was statistically designed to detect (the target difference) is specified, along with the assumptions made, such

as the anticipated magnitude of variability and effect in the control arm [79]. If well justified, the target difference can also inform the interpretation of the trial's findings, clarifying the presence (or absence) of a meaningful difference. As a consequence, appropriate calculation and reporting of the sample size calculation can help to avoid research waste, preventing the conduct of trials that are likely to produce inconclusive and potentially misleading results.

2.1.2 Prior literature

Reviews have previously examined the reporting of sample size calculations. A study of peer reviewer comments found that sample size calculations were often not performed or completed post-hoc [80]. Where sample size calculations have been reported, many studies have found that they were reported inadequately and based on inaccurate assumptions in both journal publications and unpublished protocols [79, 81, 82]. For example, inadequate reporting of a sample size calculation could mean that key components of the sample size calculation have been omitted, such as the standard deviation being omitted from the sample size calculation for a continuous primary outcome. Poor reporting of sample size calculations has also been found for specific study designs including cluster-randomised and cross-over trials, even after specific guidelines were produced aiming to improve this [83, 84, 85]. Other reviews have examined trials of specific interventions [86, 87, 88].

Discrepancies in the assumptions for parameters in sample size calculations can affect power [89, 90]. For instance, studies can be underpowered if the standard deviation is underestimated [91]. This can be the case if the assumed standard deviation is estimated using the results of a small pilot study in a more homogeneous population [89].

2.1.3 Potential influence of trial characteristics

Prior research has found that methodological quality and reporting are associated with study-specific features, including funding source, type of intervention, type of comparison treatment and the number of study centres. Studies have highlighted the complexities of surgical trials and have highlighted their poor methodological quality and reporting [92, 93, 94]. Trials with an active control arm may also have methodological differences to placebo-controlled trials, e.g., trials with an active control arm may use a smaller target difference and thus require a larger sample size [95, 96]. Studies have also suggested that multi-centre trials may have better methodological quality than single-centre trials [92, 97, 98]. Previous reviews have shown that industry-funded studies may differ in terms of transparency and outcome reporting [99, 100, 101].

2.1.4 Rationale for choice of condition

Several reviews have examined the reporting of sample size calculations in a handful of specific conditions, including musculoskeletal conditions. In back pain, Froud *et al.* found that sample size calculations were reported in only 41% of trials and that only one-third were powered to detect a standardised mean difference ≤ 0.5 [102]. Similarly, a review of rheumatology trials published in 2001-2002 found that trials were often underpowered and sample size calculations were poorly reported [103]. A review of rehabilitation trials found that reporting of sample size calculations had increased over time (from 3.4% in 1998 to 57.3% in 2008) [104]. In a more recent review of rehabilitation trials, Castellini *et al.* found that the completeness of reporting of sample size calculations had also increased over time; however, among the 80 randomised trials that reported a power calculation, only 13 (16.3%) completely

described the sample size calculation [105].

To review all trials of all musculoskeletal conditions would produce a highly heterogeneous sample. For instance, the treatments available for back pain differ greatly from those for hand osteoarthritis. Instead, this review focuses on a single condition, where the outcome measures and interventions used are more homogeneous, in order for the results to be useful in practice. This review focuses solely on osteoarthritis of the hip and knee. As discussed in Section 1.5.1, osteoarthritis is highly prevalent and the prevalence is expected to increase in the near future [55, 61, 62]. As osteoarthritis is such a common condition, the choice of sample size for osteoarthritis trials should be less affected by recruitment difficulties that may be present for more rare conditions [106].

Currently, there is no available cure for osteoarthritis (Section 1.5.2). As osteoarthritis is a leading cause of disability [59] and has a high economic burden [57, 58], many trials are conducted in osteoarthritis. These trials often use a patient-reported outcome measure as a primary outcome [107].

The scope of this review was refined to include only hip and knee osteoarthritis because the choice of intervention and outcome measures used are often specific to the affected area of the body. The hip and knee are the most common sites to be affected by osteoarthritis [108].

No previous systematic reviews of the sample size calculations in hip and knee osteoarthritis trials has been done. In general, few reviews have attempted to replicate the sample size calculation of published trials [79, 81, 87, 109].

2.1.5 Objectives

The primary objective of this review is to summarise current practice regarding sample size calculation for trials of hip and knee osteoarthritis, including the sample size, target difference, and justification of the chosen inputs.

The secondary objectives are to assess the reporting and replicability of the sample size calculation.

2.2 Methods

2.2.1 Identification of studies

Studies were identified by searching electronic databases (Medline, Cochrane Central Register of Controlled Trials (CENTRAL), CINAHL, EMBASE, PsycINFO, PEDro and AMED). An example search strategy is given in Appendix A.1. Searches were restricted to articles published between 1 January 2016 and 31 December 2016. The initial search was performed on 15 December 2016 and updated on 31 March 2017 to allow for a time lag of at least 3 months between publication and database indexing.

For the search strategy, terms related to trial design were selected for individual databases to find a combination of terms with high precision whilst also retaining high sensitivity [110, 111, 112, 113]. The strategies used were found to be comprehensive and detected almost all randomised trials, without identifying a large proportion of articles that were not reporting the results of randomised trials. In all databases, where possible, search results were limited to only include human studies.

Search terms related to osteoarthritis were compiled based on Cochrane reviews in this area [114, 115, 116, 117, 118, 119, 120]. Terms were used if they related to osteoarthritis. More general terms, such as musculoskeletal pain, were not included. No restrictions were used based on the affected body part, e.g., terms specifically related to hip or knee.

2.2.2 Study selection

Eligibility was determined based on the criteria outlined below. All of the trials that met the eligibility criteria were included.

2.2.2.1 Inclusion criteria

Design:

- Included trials were randomised controlled trials of two treatment arms.
- Superiority, non-inferiority and equivalence trials were eligible.

Population:

- Hip and/or knee osteoarthritis (OA)
- Trials of, for example, total knee arthroplasty were eligible only if it was clearly stated in the article that all participants had hip and/or knee osteoarthritis.

Intervention:

- Any intervention and any comparison treatments were included.

Outcomes:

- Inclusion was not restricted based on the trial outcomes. Trials with binary, continuous or time-to-event primary outcomes were eligible.

Article:

- Primary report of a trial
- Published in 2016

2.2.2.2 Exclusion criteria

Design:

- Articles on non-randomised studies were excluded, including case-control and cross-sectional studies.
- Quasi-randomised studies, or studies not stating that the allocation was randomised, were excluded.
- Factorial design and cross-over trials were excluded.
- Pilot studies were excluded, as were 'feasibility', 'proof of concept' and 'exploratory' studies.
- Studies that intended to use their results to inform future definitive phase III trial were not included.

Population:

- Trials with mixed populations were excluded, e.g., including both rheumatoid arthritis and osteoarthritis.

Intervention:

- Studies that did not evaluate treatments (e.g., compare different screening methods) were excluded.
- Studies on the prevention of osteoarthritis were excluded.

Article:

- Non-English language articles were excluded.
- Articles were excluded if they were conference abstracts or study protocols.

- Separate publications for secondary analyses were excluded, e.g., long-term follow up analyses.

2.2.2.3 Conduct of study selection

Search results from each of the databases were combined and duplicates removed (using referencing software and by hand). Abstracts and full texts were screened independently by two reviewers according to the eligibility criteria in Sections 2.2.2.1 and 2.2.2.2. I screened 100% of the articles. The second screening was split between three other authors (Usama Ali, Karan Vadher and Jacqueline Thompson). Disagreements regarding study eligibility were resolved by discussion between the two reviewers who assessed the paper and, where necessary, a third reviewer was involved to achieve a majority decision.

2.2.3 Data extraction and management

Data were extracted using a standardised form. Where a study protocol was cited or attached as an appendix, the details on the sample size calculation were also extracted from the protocol. Where there were conflicts between the information in the protocol and the main publication, information from the main publication was used.

The following information was extracted from each article when reported:

- Article: Country
- Design: Study design (e.g., superiority, non-inferiority)
- Population: Condition (including how osteoarthritis was defined), setting, eligibility criteria (e.g., disease severity).
- Treatment: Intervention, comparison.

- Outcome: Primary outcome measure(s)
- Sample size details: Statistical approach (conventional, other), target sample size, method for calculation, values used and justification (e.g., effect size, target difference, standard deviation, adjustment for loss to follow-up, sidedness of test, significance level, power), whether sample size could be replicated, whether sample size re-estimation was planned (e.g., using interim analysis), and whether a sensitivity analysis was conducted to examine the impact of assumptions on sample size. Post-hoc sample size calculations were not considered.
- Follow-up: Number of participants randomised, number lost to follow up, and whether compliance was measured.

The data extraction form was piloted before conducting the main data extraction to ensure all relevant information was collected and to improve clarity. Data extraction was performed independently by a second reviewer (Usama Ali, Karan Vadher or Jacqueline Thompson) on a sample of 20% of the included studies to check accuracy.

2.2.4 Data synthesis

Data were summarised across the studies, including the general characteristics of the included studies, using appropriate summary statistics (e.g., n and %, median and interquartile range).

The proportion of studies that justified the sample size and target difference and that reported each component of the sample size calculation was calculated.

2.2.5 Sample size replication

Core values for the sample size calculation were defined as the power, level of significance, sidedness of test (one-tailed or two-tailed), level of attrition and:

i For continuous outcomes for superiority trials: Target difference as standardised effect size or mean difference and standard deviation.

ii For continuous outcomes for non-inferiority trials: Non-inferiority margin, mean difference and standard deviation.

iii For binary outcomes for superiority trials: Any two of the anticipated between-group risk difference, effect in the intervention group and effect in the control group.

Using the reported values, I attempted to re-calculate the target sample size. Unless otherwise stated, it was assumed that 80% power and 5% two-sided significance level were used, that there was no attrition and that the trial hypothesis was superiority. For non-inferiority trials, the anticipated mean difference was assumed to be 0 if this was not reported.

When comparing the replicated and reported target sample sizes, the ratio was calculated as:

$$\frac{\text{replicated value} - \text{reported value}}{\text{reported value}} \tag{2.1}$$

The number of studies with a ratio above 1.1 and 1.3 or below 0.9 and 0.7 is presented (i.e., where the replicated value was over 10% or 30% greater or less than the reported sample size). A calculation was determined to be replicated if the replicated calculation produced a value within 10% as rounding errors and the use of different statistical software could produce differences of this proportion for equivalent calculations, especially for smaller sample sizes.

Attempts to replicate the sample size calculation were primarily conducted using Stata IC 14 [121], following the extension of Equation 1.1 using the Student's t-distribution. For trials where the sample size could not be replicated (replicate value not within 10% of the reported target sample size), the replication calculation was conducted by a second statistician and the replication was attempted based on the z-test and using the SampSize application [122]. Where the article referred to the use of particular software to calculate the sample size and the sample size could not be replicated, the calculation was repeated using the alternative software, where available.

2.2.6 Accuracy of components

The components of the sample size were compared to the corresponding values in the study results. The components examined were the pooled standard deviation (for continuous outcomes) and treatment effect in the control arm (for binary outcomes). The corresponding value in the study results for the primary time point was used. Where a primary time point was not specified, the value for the final follow-up time point was used. As for the replicated sample size, the number of studies where the ratio was above 1.1 and 1.3 or below 0.9 and 0.7 is presented.

2.2.7 Subgroup analysis

Subgroup analysis was conducted to explore the associations between study-level characteristics and key aspects of the sample size calculation: (i) observed sample size (number of participants randomised), (ii) whether the sample size calculation was reported, (iii) whether all core components of the sample size calculation were reported and (iv) whether the sample size calculation could be replicated.

For formal subgroup comparisons, the following categories were used:

- Intervention types: Surgical vs non-surgical

  A trial was classified as 'surgical' if any study arm received a surgical intervention as part of their treatment in the study. This included trials comparing different surgical interventions and trials comparing different treatments as part of pre-operative or post-operative care. Trials were not classified as surgical if the outcome measurement occurred only before surgery.

- Centres: Single vs multi-centre
- Funding: Industry-funded (all or in part) vs no industry funding
- Comparator: Placebo/waitlist vs active control

The (i) target sample size was compared between groups using the median difference and 95% confidence interval estimated using Hodges-Lehmann [123, 124].

Absolute risk differences with 95% confidence intervals are reported for (ii) reporting of a sample size calculation, (iii) reporting of core sample size components, and (iv) the replicated sample size being >10% larger than the reported sample size.

A two-sided significance level of 0.05 was used. As the sub-group analysis was exploratory, no adjustment was made for multiple testing.

2.3 Results

2.3.1 Flow of studies

From the databases searches, 116 of 2955 articles were eligible for this review. The most common reasons for exclusion at the full text eligibility stage were not being a randomised trial (observational study or review) and not being a primary report of a full trial (pilot study, secondary analysis or conference abstract) (Figure 2.1). The list of included studies is presented in Appendix A.1.

Of the 116 included trials, 78 (67%) reported a power calculation (Figure 2.1).

Figure 2.1: PRISMA flow diagram

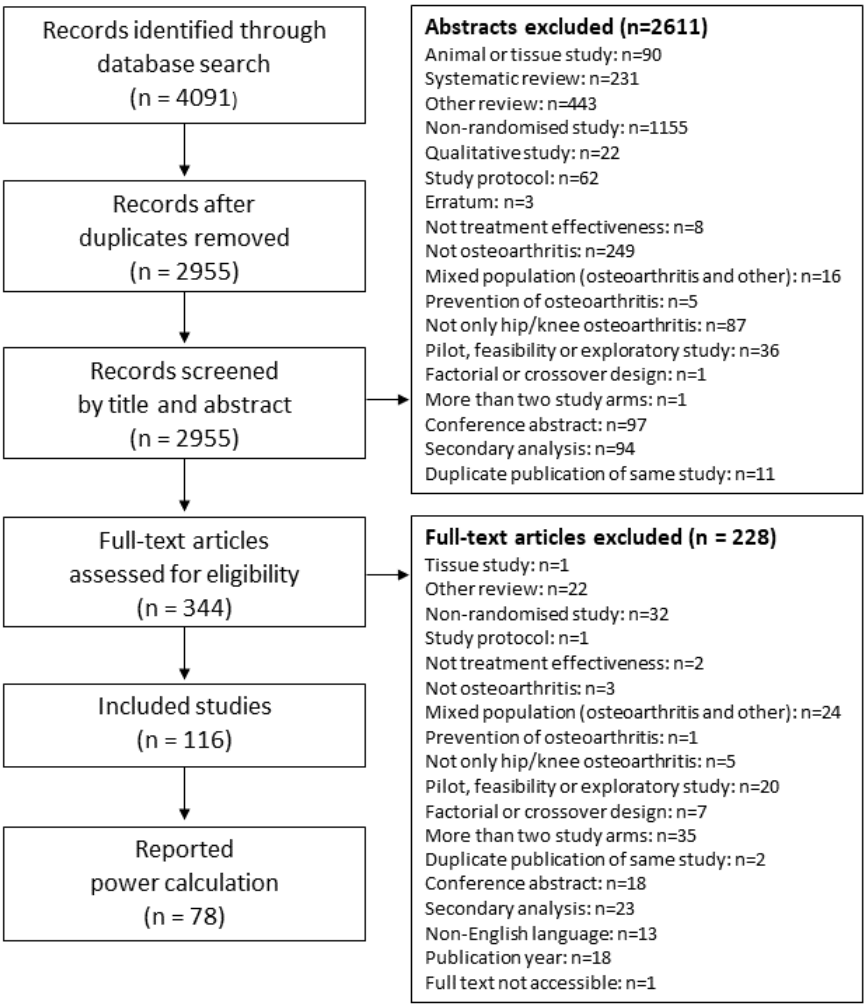


Figure reproduced from Copsey *et al.* (2018) [125], published by CC BY-NC-ND.

2.3.2 Characteristics of study sample

2.3.2.1 Characteristics of study design

The characteristics of the included studies are summarised in Tables 2.1 and 2.2. The majority were parallel-group, superiority, single-centre trials of knee osteoarthritis funded by non-industry sources. Only one trial was cluster-randomised (1%, n=1/116) and all others were participant-randomised. The allocation ratio was 1:1 in almost all trials (98%, n=114/116). A wide variety of interventions were evaluated; around 18% (21/116) were surgical interventions, 28% (33/116) drugs and 19% (22/116) exercise. Although 18% assessed a surgical intervention, participants received surgical treatment in 28% of trials (n=32/116); for example, if the trial examined different treatments as part of post-surgical care. Most studies compared the intervention to an active treatment (65%, n=75/116).

Only 13 trials (13/116, 11%) referred to or provided access to the trial protocol, and all of these trials reported a power calculation. The number of participants randomised ranged from 20 to 633 (median 73). The follow-up period ranged from immediately after treatment to 7 years (median 4 months from baseline).

Table 2.1: Study characteristics: Design (n=116)

	Reported power calculation					
	Yes		**No**		**Total**	
Total	N=78		N=38		N=116	
	n	%	n	%	n	%
Study design:						
Parallel	77	99%	38	100%	115	99%
Cluster	1	1%	0	0%	1	1%
Study hypothesis:						
Superiority	51	65%	18	47%	69	59%
Non-inferiority	8	10%	1	3%	9	8%
Multiple	1	1%	0	0%	1	1%
Unclear	18	23%	19	50%	37	32%
Study centres:						
Single centre	55	71%	28	74%	83	72%
Multi-centre	16	21%	3	8%	19	16%
Unclear	7	9%	7	18%	14	12%
Funding source:						
Industry	13	17%	2	5%	15	13%
Non-industry	31	40%	9	24%	40	34%
Combination	4	5%	3	8%	7	6%
No funding	7	9%	3	8%	10	9%
Not reported	23	29%	21	55%	44	38%
Follow up assessments:						
1	17	22%	15	41%	32	28%
2	24	31%	10	27%	34	30%
3 or more	37	47%	12	32%	49	42%
Patient-reported outcome measure:						
Yes	75	96%	31	82%	106	91%
No	3	4%	7	18%	10	9%
Number randomised:						
Median (IQR)	86.5 (55 - 150)		59 (40 - 76)		73 (50 - 120)	
Range	26 - 633		20 - 140		20 - 633	
n	78		37		115	
Follow-up period (months):						
Median (IQR)	5.5 (1.5 - 12)		3 (1.3 - 6)		4 (1.5 - 12)	
Range	0 - 84		0.2 - 27		0 - 84	
n	78		38		116	

Table 2.2: Study characteristics: Population (n=116)

	Reported power calculation					
	Yes		No		Total	
Total	N=78		N=38		N=116	
	n	%	n	%	n	%
Population						
Knee osteoarthritis	69	88%	32	84%	101	87%
Hip osteoarthritis	7	9%	4	11%	11	9%
Hip or knee osteoarthritis	2	3%	2	5%	4	3%
Intervention						
Drug	21	27%	12	32%	33	28%
Surgery	17	22%	4	11%	21	18%
Exercise	19	24%	3	8%	22	19%
Other	21	27%	19	50%	40	34%
Comparator						
Active treatment	48	62%	27	71%	75	65%
Usual care	12	15%	5	13%	17	15%
Placebo or sham	18	23%	6	16%	24	21%
Definition of osteoarthritis						
Radiological and symptomatic	27	35%	5	13%	32	28%
Radiological only	7	9%	6	16%	13	11%
Symptomatic only	11	14%	5	13%	16	14%
Unclear	33	42%	22	58%	55	47%
Restrict by previous surgery						
No previous surgery	41	53%	18	47%	59	51%
Not previous surgery in specified time period	15	19%	3	8%	18	16%
Post-surgery only	1	1%	0	0%	1	1%
No	21	27%	17	45%	38	33%
Restrict by radiological findings						
By Kellgren Lawrence classification	33	42%	20	53%	53	46%
Other restriction	9	12%	2	5%	11	9%
No	36	46%	16	42%	52	45%
Restrict by symptoms						
Pain with stiffness and/or swelling	3	4%	3	8%	6	5%
Pain only	31	40%	11	29%	42	36%
No	44	56%	24	63%	68	59%
Restrict by primary osteoarthritis						
Exclude secondary osteoarthritis	20	26%	7	18%	27	23%
No	58	74%	31	82%	89	77%

2.3.2.2 Characteristics of study population

The majority of trials included only participants with knee osteoarthritis (87%, n=101/116). Studies varied in how osteoarthritis was defined and this was often not reported. When reported, the majority of studies used both symptomatic and radiological criteria (28%, n=32/116). Several studies referred to the ACR criteria for diagnosing osteoarthritis (41%, n=47/116) but often did not report whether they used radiographic imaging criteria or only clinical criteria [66, 67]. To restrict eligibility based on disease severity, several studies used pain symptoms (41%, n=48/116) or the Kellgren-Lawrence classification system (a system classifying knee osteoarthritis based on the presence of osteophytes, joint space narrowing and bone deformity) (46%, n=53/116) [126].

Most studies restricted inclusion by age (76%, n=88/116). Several studies excluded participants based on prior surgery (66%, n=77/116). Some studies excluded secondary osteoarthritis (23%, n=27/116), and few studies restricted eligibility based on gender (6%, n=7/116) or post-traumatic disease (15%, n=17/116).

2.3.2.3 Comparison based on reporting of power calculation

Of the 116 included trials, 78 (67%) reported a power calculation. Studies which did not report a power calculation were similar in the majority of characteristics to those where a power calculation was reported. However, studies reporting a power calculation were on average more likely to have a larger sample size, multiple follow-up assessments, refer to the study protocol and report the trial funding source (Table 2.1).

Among the 38 trials that did not report a power calculation, 1 trial reported a post-hoc power calculation and 6 trials reported that a power calculation was conducted

but did not provide any details on this calculation (e.g., stated that the sample size was calculated based on 90% power but did not report any other information regarding the target difference). Of the remaining 31 trials that did not report any reference to a power calculation, 4 trials reported that the sample size was based on the pre-defined recruitment period and 27 trials did not state anything regarding their choice of sample size. In the 31 studies that did not report any reference to a power calculation, 58% (18/31) mentioned the small sample size as a limitation of the trial and an additional 10% (3/31) stated that future trials with larger sample sizes were necessary. The lack of a power calculation was mentioned as a limitation in 5 studies (16%, 5/31).

The results that follow (Section 2.3.3 onwards) are based on the 78 studies that reported details of a power calculation.

2.3.3 Methods of sample size calculation

All of the included studies used conventional power calculation approaches (Neyman-Pearson or statistical hypothesis testing) [127, 128]. None used alternative techniques such as Bayesian approaches or simulations [129, 130, 131].

Most studies had a continuous primary outcome (97%, n=76/78) (Table 2.3). For the primary outcome, only one trial used a binary outcome and no trial used a time-to-event measure. The type of primary outcome was unclear in one trial: the effect size was reported as a percentage, but it was unclear whether this related to a risk difference in a binary outcome or a percentage change in a continuous outcome. The one cluster-randomised trial reported the adjustment for clustering.

Studies most commonly powered their study on a single primary outcome (91%, n=71/78). For three studies (4%, n=3/78), the number of outcomes the study was powered on was unclear as only the standardised effect size used was reported and the primary outcome was not specified.

Only one trial reported planning to re-estimate their sample size during their trial if attrition was higher than expected, but this was not found to be necessary. Unplanned re-estimation of the sample size was conducted in three trials (4%, n=3/78) due to poor recruitment, low attrition or post-hoc analysis.

Sensitivity analysis on the sample size calculation was conducted in nine trials (12%, n=9/78) to assess the impact of adjusting the assumptions. This was most commonly done to assess the power of the study for secondary outcomes (n=4). Sensitivity analysis was also conducted to power for subgroup analysis (n=1) and examine the assumptions on the rate of attrition (n=1), normality (n=1), target difference (n=1) and power required (n=1).

2.3.4 Reporting of sample size calculation

Less than a quarter of studies reported all core components of the sample size calculation (21%, n=16/78). Many individual components of the sample size calculation were well-reported, with 96% of trials reporting both the power and significance level (n=75/78). However, the sidedness of the test (one-tailed or two-tailed) was less well-reported with only 41% of studies reporting the value (n=32/78).

The majority of studies used 5% level of significance (88%, n=69/78) or 2.5% for one-tailed tests (8%, n=6/78). The statistical power was at least 80% for all studies, with most using 80% power (71%, n=55/78). It was most common to use 80% power and 5% level of significance (65%, n=51/78).

Table 2.3: Reporting of components (n=78)

	n	%
Type of primary outcome		
Continuous	76	97%
Binary	1	1%
Unclear [a]	1	1%
Number of primary outcomes		
1	71	91%
2	3	4%
3	1	1%
Primary outcome not specified	3	4%
Alpha		
One-sided	8	10%
0.025	6	8%
0.05	2	3%
Two-sided	23	29%
0.05	21	27%
0.10	1	1%
Not reported	1	1%
Multiple (planned one-sided and two-sided)	1	1%
0.05	1	1%
Sidedness of test not reported or unclear	46	59%
0.05	45	58%
Not reported	1	1%
Power		
0.80	55	71%
0.85	3	4%
0.90	11	14%
0.95	4	5%
Other or multiple	3	4%
Not reported	2	3%
Attrition		
<5%	2	3%
5-9%	2	3%
10-14%	15	19%
15-19%	12	15%
20-24%	20	26%
≥25%	6	8%
Not reported	20	26%
Unclear	1	1%

[a] In 1 study, the type of primary outcome was unclear [132]. It reported a target difference of 20% but it was unclear if this was a between-group risk difference or difference in a continuous measure.

The level of attrition assumed was not reported in a quarter of trials (27%, n=21/78). Where reported, the level of attrition assumed in the sample size calculation was most commonly 10-24% (60%, n=47/78), but ranged from 2% to 44%.

For superiority trials powered on a continuous outcome, the mean difference was reported by almost all trials (90%, n=61/68) and most reported the standard deviation (66%, n=45/68) (Table 2.4). The standardised effect size was only reported in 12 trials (18%, n=12/68) but could be calculated from reported information for an additional 42 trials (62%, n=42/68). The standardised effect size could most often not be calculated because the standard deviation was missing. Few studies reported the assumed effect in the control arm (6%, n=4/68).

Where reported, the standardised effect size that the study was powered to detect was large for the majority of trials (median 0.75, IQR 0.50 to 0.86). Only one study reported a target difference of 0.2 or less; however it is likely that this target difference was incorrectly reported as the replicated sample size for this study was much greater than the reported sample size [133].

All of the non-inferiority trials reported the non-inferiority margin and most reported the standard deviation (75%, n=6/8). Around half reported the anticipated mean difference between treatment arms (38%, n=3/8) and the effect in the control arm (50%, n=4/8).

Only one cluster-randomised trial was included. It reported the intra-class correlation coefficient (ICC) assumed in the power calculation to adjust for clustering. The cluster size was reported in the methods but not in the 'Sample size calculation' section. A parallel-group study also reported the ICC used to adjust for clustering as the intervention was delivered in a group format, although the cluster size was only reported in the trial results and not in the methods section.

For the single trial with a binary outcome, the risk difference between the two arms was reported but the anticipated effect in the intervention or control group was not reported.

Of the 13 trials that referred to a study protocol, 11 protocols were published in a journal and 2 were accessible as online appendices. The sample size calculation was reported consistently in the study protocol and results publication for the majority of trials (77%, n=10/13). For 7 trials, the reported sample size calculation in the protocol was identical to the publication results. In 3 trials, there were no discrepancies but the protocol provided additional details, such as justification for the anticipated level of attrition. In the remaining 3 trials, there were discrepancies between the reported sample size calculation in the protocol and the results publication. One trial reported a lower assumed attrition rate but a higher sample size in the publication. Another trial did not report a power calculation in the protocol and planned to recruit 140 participants, however the power calculation in the results publication gave a target sample size of 42 participants and 46 participants were randomised. For 1 trial, it was clear from the protocol that the reported sample size calculation in the publication was a revised calculation that was adjusted during the recruitment period; however, this was not apparent from the results publication alone and reasons for the revisions were not reported in the protocol.

Table 2.4: Reporting of components for continuous outcomes (n=76)

Superiority trials (N=68)	n	%
Mean difference		
Mean difference reported	59	87%
Range of values reported	2	3%
Not reported	7	10%
Standard deviation		
Reported	45	66%
Not reported	23	34%
Standardised effect size		
Reported	12	18%
Calculated from reported information	42	62%
Insufficient information	14	21%
Effect in control arm		
Reported	4	6%
Not reported	64	94%

Non-inferiority trials (N=8) [a]	n	%
Non-inferiority margin		
Reported	8	100%
Mean difference		
Mean difference reported	3	38%
Not reported	5	63%
Standard deviation		
Reported	6	75%
Not reported	2	25%
Standardised effect size		
n/a - non-inferiority trial	8	100%
Effect in control arm		
Reported	4	50%
Not reported	4	50%

[a] Note: One study with multiple hypotheses categorised as non-inferiority. Studies where hypothesis is unclear were assumed to be superiority

2.3.5 Justification of components of sample size calculation

In trials with continuous outcomes, the justification for the mean difference was provided in 40 trials (53%, n=40/76). The mean difference was most commonly based on the treatment difference found in a published trial (29%, n=22/76) or a published MCID (18%, n=14/76) (Table 2.5).

Justification for the standard deviation used was reported in half of trials with continuous outcomes (46%, n=35/76) with most referring to a published trial (33%, n=25/76). Although rarely reported, the anticipated effect in the control arm was also most often based on a published trial (7%, n=5/76).

Very few trials provided a justification for the anticipated level of attrition (3%, n=2/76).

Table 2.5: Justification of components for continuous outcomes (n=76)

	n	%
Justification for mean difference		
Combination (MCID and other)	1	1%
MCID	13	17%
Published trial	22	29%
Published observational study	1	1%
Uncited pilot study	3	4%
No justification	24	32%
n/a - not reported	12	16%
Justification for non-inferiority margin		
Combination (MCID and other)	1	1%
MCID	3	4%
Published trial	3	4%
No justification	1	1%
n/a - not a non-inferiority trial	68	89%
Justification for standard deviation		
Published trial	25	33%
Published observational study	2	3%
Systematic review	1	1%
Uncited pilot study	7	9%
No justification	16	21%
n/a - not reported	25	33%
Justification for attrition		
Published trial	1	1%
Uncited pilot study	1	1%
No justification	54	71%
n/a - not reported	20	26%

2.3.6 Replicability of sample size calculation

Of the 78 trials that reported a power calculation, the sample size was only replicated in 41 trials (53%, n=41/78) (Table 2.6 and Figure 2.2). A quarter of trials did not provide sufficient information for the sample size calculation to be replicated (28%, n=22/78). The replicated calculation produced a sample size over 10% larger than the reported value in 12% of studies (n=9/78) and over 30% larger in 6% of studies (n=5/78) (Table 2.6 and Figure 2.3).

In trials that reported all core components (no assumptions had to be made in calculating the sample size), the sample size could not be replicated in 13% of trials (13%, n=2/16).

Figure 2.2: Replicability of sample size calculation

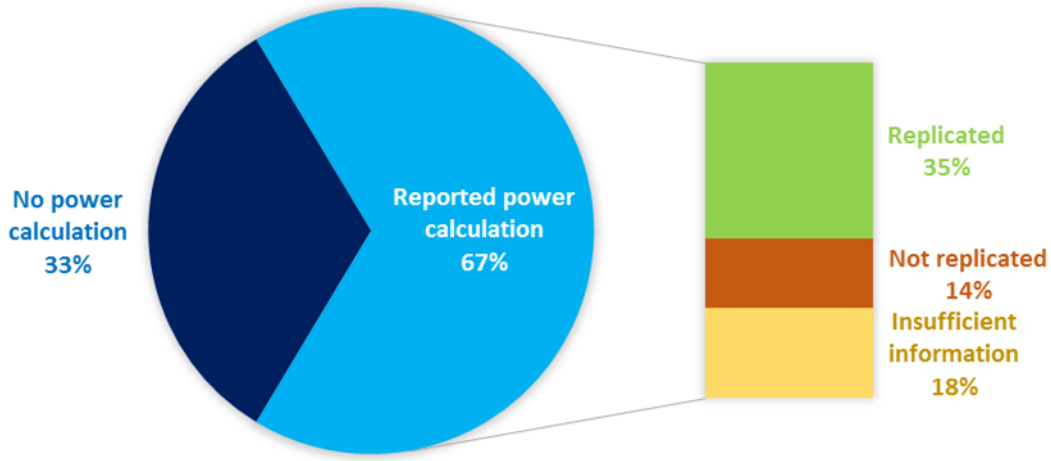


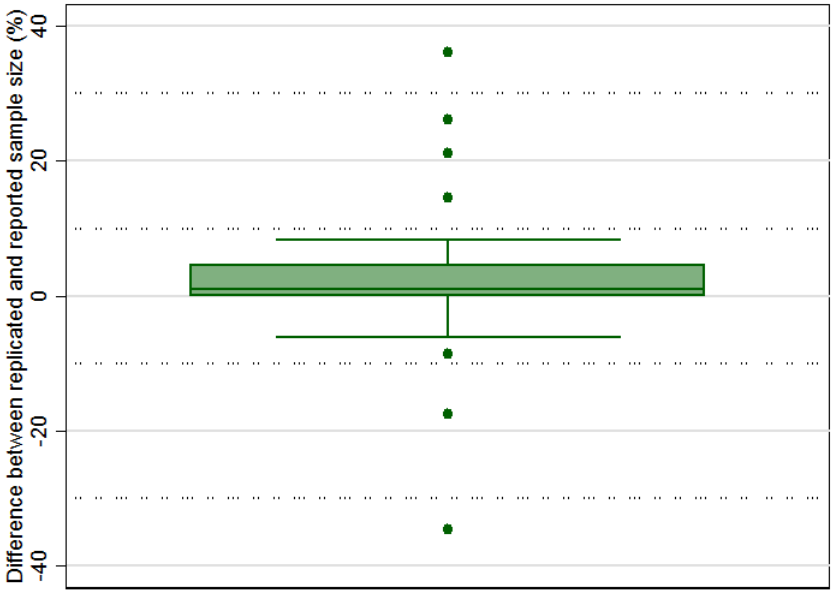
Table 2.6: Replicability of sample size calculation (n=78)

	Report all core components					
Replicated vs reported n	**Yes**		**No**		**Total**	
Total	N=16		N=62		N=78	
	n	%	n	%	n	%
Replicated >30% larger	0	0%	6	10%	6	8%
Replicated value 10-30% larger	0	0%	3	5%	3	4%
Within 10%	14	88%	27	44%	41	53%
Replicated value 10-30% smaller	0	0%	1	2%	1	1%
Replicated >30% smaller	2	13%	3	5%	5	6%
Insufficient information	0	0%	22	35%	22	28%

Four trials that could be replicated required additional assumptions to interpret the reported information. In two of these, the reported target difference had to be translated into a different scale to replicate the sample size calculation. For example, the sample size could be replicated using a target difference of 2 points on a 0-10 scale when it was reported as 20% [134, 135]. In two studies, the sample size calculation could be replicated if the unusually high attrition rate was not accounted for, such that both the reported value and replicated value before accounting for attrition were compared [136, 137]. For instance, one study was forced to inflate the sample size due to national regulation on the minimum number of participants required. However, the replicated value assuming no inflation was compared to the sample size reported as required by the power calculation if the regulation had not been in place [136].

For two trials (3%, n=2/78), the reported sample size calculation was inappropriate for the study design. One study used a sample size calculation for a paired sample and the other trial used a survey-based approach to calculate their sample size [138, 139]. For both of these studies, the replicated sample size calculation assumed the use of a two-sample independent t-test.

Figure 2.3: Comparison of replicated and reported sample sizes (as a percentage of reported sample size)



Note: Total number of studies: n=56. Excluded from Figure 2.3 are 5 studies where the difference was over 50% and 4 studies where the difference was below -50%.
Figure reproduced from Copsey *et al.* (2018) [125], published by CC BY-NC-ND.

The absolute difference between the replicated and reported sample size in terms of the number of participants was small for most studies (median difference 1 participant, IQR 0 to 5). There were five studies where the difference between the replicated and reported sample size was greater than 50 participants (9%, n=5/56). Within these five trials, there were four trials where the sample size was underestimated (replicated value was over 30% larger than the reported value) and one trial where the sample size was overestimated (replicated value was at least 30% smaller than the reported value).

2.3.7 Accuracy of components

When the standard deviation assumed in the sample size calculation could be compared to the value in the trial results, the estimate used in the sample size calculation was accurate (within 10%) in only a quarter of studies where they could be compared (31%, n=9/29) (Table 2.7). In six studies, the follow-up standard deviation was over 30% larger than the value assumed in the sample size calculation leading to a reduction in power.

As few studies reported the assumed treatment effect in the control arm, the accuracy of this value was not assessed.

Table 2.7: Accuracy of the standard deviation (n=78)

Accuracy of standard deviation	n	%
Follow-up SD >30% larger	6	8%
Follow-up SD 10-30% larger	3	4%
Within 10%	9	12%
Follow-up SD 10-30% smaller	5	6%
Follow-up SD >30% smaller	6	8%
Assumed and follow-up SD could not be compared	47	60%
n/a - not continuous outcome	2	3%

2.3.8 Subgroup analysis

Exploratory subgroup analysis did not detect any significant differences in reporting based on study intervention or type of comparator (Table 2.8). Multi-centre trials and industry-funded trials had significantly larger sample sizes (in terms of number of randomised participants).

Table 2.8: Subgroup analysis: Reporting of power calculation

	Power calculation reported:		
	Yes	No	Risk difference (95% CI)
Surgical	78% (25/32)	22% (7/32)	15.0% (-2.6% to 32.7%)
Non-surgical	63% (53/84)	37% (31/84)	
Single centre	66% (55/83)	34% (28/83)	-17.9% (-37.2% to 1.3%)
Multi-centre	84% (16/19)	16% (3/19)	
Industry funding	77% (17/22)	23% (5/22)	12.4% (-7.6% to 32.4%)
No industry funding	65% (61/94)	35% (33/94)	
Active control	64% (48/75)	36% (27/75)	-9.2% (-26.5% to 8.2%)
Placebo or usual care	73% (30/41)	27% (11/41)	

	Report all core components:		
	Yes	No	Risk difference (95% CI)
Surgical	16% (4/25)	84% (21/25)	-6.6% (-24.9% to 11.6%)
Non-surgical	23% (12/53)	77% (41/53)	
Single centre	20% (11/55)	80% (44/55)	-5.0% (-28.7% to 18.7%)
Multi-centre	25% (4/16)	75% (12/16)	
Industry funding	12% (2/17)	88% (15/17)	-11.1% (-29.7% to 7.4%)
No industry funding	23% (14/61)	77% (47/61)	
Active control	17% (8/48)	83% (40/48)	-10.0% (-29.0% to 9.0%)
Placebo or usual care	27% (8/30)	73% (22/30)	

	Replicated sample size calculation:		
	Yes	No	Risk difference (95% CI)
Surgical	48% (12/25)	52% (13/25)	-6.7% (-30.4% to 17.0%)
Non-surgical	55% (29/53)	45% (24/53)	
Single centre	47% (26/55)	53% (29/55)	-21.5% (-47.7% to 4.8%)
Multi-centre	69% (11/16)	31% (5/16)	
Industry funding	53% (9/17)	47% (8/17)	0.5% (-26.4% to 27.3%)
No industry funding	52% (32/61)	48% (29/61)	
Active control	52% (25/48)	48% (23/48)	-1.3% (-21.5% to 24.0%)
Placebo or usual care	53% (16/30)	47% (14/30)	

Table 2.8: Subgroup analysis: Reporting of power calculation

	Sample size (number randomised):	
	Median sample size (IQR)	Median difference (95% CI)
Surgical	75 (60-100)	5 (-11 to 19)
Non-surgical	73 (43-124)	
Single centre	63 (46-100)	-76 (-14 to -163)
Multi-centre	169 (74-260)	
Industry funding	122 (62-169)	36 (8 to 80)
No industry funding	65 (46-100)	
Active control	76 (54-110)	9 (-8 to 23)
Placebo or usual care	64 (40-122)	

2.3.9 Case studies

This section provides case studies to demonstrate sample size calculations that have been reported well or poorly. This relates to providing sufficient information on all relevant components of the sample size calculation for the calculation to be replicated [6].

2.3.9.1 Examples of good reporting of sample size calculations

The quote below is from the publication by Beselga *et al.* [140]. The sample size calculation for this trial was well-reported because it includes the target difference, assumed standard deviation and level of attrition, as well as the level of significance, sidedness of the test and power. It also justifies the target difference and standard deviation with a citation where appropriate.

> "The sample size was calculated using Ene 3.0 software (GlaxoSmithKline, Autonomic University of Barcelona, Spain). The calculations were based on detecting differences of 2.0 units in the NPRS, considered as the minimum clinical important difference (MCID) (Farrar *et al.*, 2001), assuming a standard deviation of 1.7 (based on pilot data), an α of 0.05, β of 90%, and a 2-tailed t-test. The estimated sample size was calculated to be at least 17 subjects in each group. The sample was increased to 20 subjects in each group to allow a drop out of 15%."

Transparency and accuracy in a sample size calculation do not necessarily imply that the sample size was calculated using appropriate inputs. It could be argued that a target standardised effect size greater than 1 is difficult to achieve. A sample size calculation being well-reported does not imply that the components have been chosen appropriately. Good reporting allows the reader to judge for themselves whether they feel the components are appropriate.

An example of a more complicated sample size calculation that is reported to a good standard is given by Allen *et al.* for a cluster-randomised trial [141].

> "We based our sample size of 300 patient participants on detection of a moderate effect size of approximately 0.30 for the difference in mean WOMAC scores between groups, with 80% power and a type I error rate of 0.05. This translates to a 4.2-point difference at 12 months, which is equivalent to an improvement of approximately 11% from the anticipated mean baseline score; this allowed sufficient power to detect a clinically relevant difference (12% to 18%, based on prior relevant literature) (37-39). We used a 2-sample t test sample size calculation for the between-group difference at 12 months multiplied by a factor of $(1 - \rho)^2$, where

> ρ represents the Pearson correlation between baseline and follow-up outcome measures (0.60) (40). This sample size was then adjusted to reflect provider clustering using an intraclass correlation coefficient of 0.02 (41) and was inflated to compensate for potential attrition (12%). On the basis of our pilot work, we assumed a mean baseline WOMAC score of 38 (SD, 14)."

They provided additional information including the correlation between baseline and follow-up measurements and the intraclass correlation as their trial was cluster-randomised and used ANCOVA for the statistical analysis. They also translated the target difference into the change from baseline and the standardised effect size, which makes it easier to interpret. However, it would be more informative if the assumed number of clusters was reported.

2.3.9.2 Example of ambiguous reporting of a sample size calculation

Bryk *et al.* reported the key components of the sample size calculation, except from the assumed level of attrition. However, the target difference was reported to be a '20% improvement' [134].

> "Sample size was calculated assuming 80% power to detect a 20% improvement in pain (Numerical Pain Rating Scale - NPRS), with a standard deviation of 2 points and a significance level of 5%. The required sample was 17 patients per group."

This is ambiguous because it could be interpreted in multiple ways. Although it could represent a 20% improvement in the risk of a binary outcome, this seems unlikely as the outcome is pain measured using a numerical rating scale. The outcome could be a 20% improvement from the baseline score, but the mean score at baseline is not

reported. In fact, the 20% difference relates to the score range. The sample size can be replicated if the target difference is taken to be 2 points (i.e., a 2-point difference on 0-10 scale or 20-point difference on a 0-100 scale). This also causes confusion as the reporting of the target difference suggests a 0-100 scale was used, whereas the standard deviation is based on a 0-10 scale. Thus, although all of the key components were reported, they were not reported in a clear and transparent way, which makes it more difficult to interpret.

2.3.9.3 Example of poor reporting of a sample size calculation

Schinsky *et al.* reported that a sample size calculation was performed but did not provide any details of the treatment effect that the trial was powered to detect [142].

> "Power analysis demonstrated that this study contained a sufficient sample size to prove non-inferiority for each of the power analysis parameters analyzed with an appropriate equivalence margin, 80% power, and a significance level of 0.05."

2.4 Discussion

2.4.1 Summary of findings

This systematic review summarises current practice in the methodology and reporting of sample size calculations of randomised trials of hip and knee osteoarthritis. Two-thirds of the trials reported a sample size calculation and most of the remaining one-third made no reference to their choice of sample size. Almost all sample size justifications were based on a conventional power calculation approach using a continuous outcome. However, the sample size calculation was fully described in very few studies, with most not reporting the anticipated level of attrition or assumed standard deviation.

In most instances where the values were reported, the target mean difference and assumed standard deviation used were justified. The target difference was most commonly based on the findings of a previous trial and/or an estimate of the MCID. The standard deviation was most commonly based on the results of previously published trials. It was rare for studies to justify the assumed level of attrition. The standard deviation assumed in the power calculation was also largely inaccurate (too small or too large) when compared with the follow-up results of the trial.

The reported sample size could often not be replicated, especially when the sample size calculation was not fully described. Trials were only powered to detect moderate-to-large standardised differences in study outcomes (according to Cohen's d), suggesting that clinically relevant treatment effects could have been missed due to insufficient participants being included in the studies [143].

2.4.2 Comparison with related literature

The proportion of recently published trials reporting a sample size calculation has been found to be similar in other clinical or methodological areas, with estimates between 50% and 70% [83, 84, 102]. Several reviews in other areas also had similar results, indicating that around one-quarter of sample size calculations included all core components [83, 104, 105]. However, these findings were inconsistent with the review by Charles *et al.*, who found the proportion of studies reporting a power calculation to be much higher at 95%, with a higher proportion reporting all core components. The higher quality of reporting found by Charles *et al.* may have been because eligibility was restricted to journals with a high impact factor [81].

Among those trials that reported a power calculation, the findings on the reporting of individual components (power, level of significance and target difference) were generally consistent with the results of other reviews [81, 104, 109]. However, this review found some differences compared with the findings of Rutterford *et al.*, who identified lower levels of reporting for several components, including the assumed standard deviation, level of attrition and justification for the target treatment difference. These differences may have been due to improved reporting over time or differences in common practice between clinical areas. In terms of the accuracy of components, Vickers found that the majority of studies underestimated the standard deviation, whereas this review found an equal balance between overestimation and underestimation [90]. However, both reviews highlighted that the standard deviation used in the sample size calculation was often inaccurate.

The target difference was most commonly justified based on previous trials. This aligns with the results of the DELTA survey and DELTA^2 review suggesting that there was a high level of awareness of this method for justification and that its use was highly recommended by trialists [9, 82]. However, this review disagreed with the

results of the DELTA survey in the use of pilot studies to inform the target difference. The DELTA survey indicated that there was high awareness and usage of pilot studies to justify a target difference, yet very few studies in this review justified their target difference based on pilot study data [9]. The reason for this is unclear, but it may be due to arguments that pilot studies should not be used to inform sample size estimation because the highly selective and small samples used often produce overly optimistic and imprecise results [144]. A more recent review of target difference elicitation found that it was rare for HTA trials to use pilot data to inform the target difference used in the sample size calculation, which is consistent with the findings of this review [82].

The replicability of the sample size calculations was similar to the findings by Clark *et al.* and studies in other clinical areas such as dentistry [79, 109]. Conversely, the review by Charles *et al.* and a study of reporting in anaesthesia journals found fewer discrepancies between the replicated and reported sample sizes [81, 104]. Considering only those trials that reported all core components of the sample size calculation, the proportion of trials where the replicated sample size was consistent with the reported value found in this review was similar to the results found by Charles *et al.* Therefore, the decreased replicability of the sample size calculations found compared to other reviews is likely due to the lower quality of reporting of the sample size calculation in osteoarthritis studies [81, 104]. This suggests that, in osteoarthritis trials, the issue is that the components of the sample size calculation are less frequently reported, as opposed to being reported inaccurately.

2.4.3 Strengths and limitations

The main strengths of this review are the systematic search strategy and restricted eligibility criteria. The review assessed an up-to-date sample of trials that should em-

ulate current practice in trial methodology. Although it included only 116 trials, this is a reasonable number considering that these trials were all published in 1 year. The focus on published trials also meant that reporting was assessed from the perspective of the information available to a reader who is interested in the results of the trial.

Focusing only on trials of hip and knee osteoarthritis is also a key strength as this produced a more homogeneous sample in terms of the outcome measures used and the population from which the trials recruited. Increased variability would be more likely between trials of different conditions because differences in the effectiveness of currently used treatments, the rarity of the condition and the outcome measures used would affect the target difference the trial is designed to detect and the anticipated difficulty in recruitment. The use of a more restricted sample also allowed a more in-depth assessment of trial design and methodology. For instance, I could examine the justification of the components used in the sample size calculation and compare these components with the corresponding values in the results.

However, this review has several limitations. Although the overall sample was 116 trials, some of the subgroups were quite small. As this limited the power of subgroup analyses, especially when subgroup factors are imbalanced, the findings of these should be interpreted cautiously. A larger sample would have enabled confirmatory subgroup analysis to establish whether specific study characteristics were associated with the quality of reporting.

The review only included trials published in 2016 and therefore the results cannot establish whether there have been changes in reporting over time. Although all trials were published in the same year, there will be variation in the time the studies were designed, for example, due to longer follow-up assessment or recruitment period, or even the time between submission and publication in a journal. This could explain differences in the study methods.

Assessment of the sample size calculation and related assumptions relied only on information reported in the articles. There was no attempt to contact the investigators for additional information. Poor reporting does not necessarily mean poor methodology [145]. For example, a trial team could have used previously published studies to inform the choice of target difference but not reported this information. However, reliance on reported information could also lead to over-estimation of methodological quality. There is potential for modification to the sample size calculation prior to publication, as seen in recent findings on outcome switching [146]. If this is the case, the reported information may not accurately reflect the a priori sample size calculation when the study was planned. It may be that post-hoc power calculations are conducted but reported as if it was calculated *a priori*. In this review, study protocols were only examined if referred to in the trial article. However, a more accurate assessment of the methodology used in study design may be seen by examining uncited trial protocols or ethics applications as in the review by Clark *et al.* [79, 96, 147]. The findings are also limited in that we did not contact authors of the included studies to establish the reasons that the sample size calculation was not replicated, for example, whether the sample size calculation was conducted incorrectly or the input values (e.g. standard deviation) was reported incorrectly.

There may have been relevant trials that were not identified by the searches due to variations in terminology and subject headings [148, 149]. Studies that were indexed into databases after March 2017 would also not have been identified. This may have made the sample less representative as publications in lower impact journals are likely to be indexed later [150]. As the review focused on reporting, only published articles were included. The number of underpowered trials with small sample sizes may have been underestimated as trials with non-significant results are less likely to be published [151, 152]. The results of this review may not be applicable to other clinical areas, particularly where dichotomous or time-to-event primary outcomes are

common.

2.4.4 Implications

There are clear areas for improvement in the reporting of sample size calculations in trials of hip and knee osteoarthritis. There is the potential for deficiencies in the reporting of sample size calculations more generally in randomised trials of other disease areas. Whilst there were examples of good practice in the literature reviewed, it is concerning that a third of the reviewed trials provided no justification for their choice of sample size. This is also surprising as it would be expected that research ethics committees, funding bodies and peer reviewers would ask for some justification of the choice of sample size. In trials that did not report a power calculation, the small number of participants or the lack of sample size calculation was often stated as a limitation of the trial, indicating that the missing power calculations were not due to a lack of awareness. This suggests that the sample size calculations were not performed during the trial design stage, rather than omitted in the reporting of the trial.

Although trialists may argue that the restrictive word count of journal articles prevents the description of detailed methodology, this may become less of a hindrance as publication of trial protocols becomes common practice and it is increasingly easy to publish additional information in supplementary materials available as online appendices [153]. Publication of trial protocols should be encouraged as standard, whether through journal articles, as an appendix to trial publications or in free online document repositories (e.g., figshare or arXiv) [154, 155].

When a power calculation was reported, there was often insufficient information to allow the calculation to be re-produced. In particular, there is potential for improvement in the reporting of the predicted level of attrition, sidedness of the test,

standard deviation (or control group level), and justification for the values used in the calculation. The absence of this information prevents verification of the sample size calculation and makes interpretation of the trial results more ambiguous. It could be unclear whether the target difference was not achieved due to a lack of treatment effect or because underestimation of the standard deviation led to the trial being underpowered.

There is a role for statistical peer reviewers to ensure that details of sample size calculations are adequately reported [78, 156]. The CONSORT statement is now recommended by many journals and recommends stating how the sample size was determined [157]. If sample size calculations are not reported adequately, journals and peer reviewers could recommend the use of the DELTA^2 checklist for more detail on what should be included in the sample size calculation [6].

The results highlighted the problem of inaccuracy in the components of the sample size calculation. While all studies calculate the sample size to achieve at least 80% power, studies may not achieve sufficient power as the standard deviation is often underestimated. To enable identification of this issue, trialists should pre-specify the target difference in terms of the between-group mean difference in the original scale of the primary outcome measure and corresponding standard deviation, rather than specifying only the standardised effect size. If there is uncertainty in the predicted standard deviation, sensitivity analysis should be conducted before the trial to indicate how adjusting the estimate of the standard deviation would affect the power of the study.

For some trials, the value produced by attempting to replicate the sample size calculation was very different to the sample size reported in the trial publication. This may have been driven by misleading or inaccurate information in the articles, in which case trials should be more explicit in the methodology used (e.g., accounting

for repeated measures). Alternatively, it could have been due to fundamental error when undertaking the calculation; this was the case for at least a few trials where the calculation was clearly inappropriate for the design. This is a more problematic issue as it should ideally have been checked by the trial team, funding panel (if grant-based funding was used), ethics committee and journal peer reviewers.

The overarching implication is that poor reporting and lack of replicability of sample size calculations is contributing to increased research waste [158, 159]. The issue of research waste applies to sample size calculations in three main areas: lack of transparency in the study design, conduct of underpowered trials and difficulties in the interpretation of trial results. Regarding transparency, the reporting of a power calculation ensures that trialists specify their primary outcome and the treatment effect that they believe to be clinically meaningful [78, 160]. An *a priori* power calculation during the trial design stage can prevent underpowered trials from being carried out if they are likely to be uninformative, allowing the resources that would have been allocated to that trial to be used in trials that have been designed with the ability to answer the proposed research question. A well-reported sample size calculation allows the reader to interpret the trial results alongside the initial aims of the trialists, so the reader can decide for themselves the likelihood of a false result and the clinical relevance of the findings. Although reporting guidelines have attempted to improve reporting quality, there is a clear need for peer review by statisticians and methodologists of trial protocols, funding applications, ethical approval and trial results publications [159]. Trial teams should be encouraged to involve members with formal training in statistics and research design early on in their trials [159].

2.4.5 Future research

Future research could explore the reasons for the lack of replicability of sample size calculations, for example, by contacting trial teams for further information on the methods used to calculate the sample size and whether the components were accurately reported. Future work could also explore other factors that may be associated with high quality reporting of the sample size calculation, such as whether the article was peer-reviewed by a statistician [161, 162, 163]. It could also explore whether the reporting of a power calculation is indicative of other methodological quality, for example, by examining the association with the study's risk of bias assessment [164] or level of pragmatism [165].

Future studies could examine the values of components used in sample size calculations in more detail. This review has not considered the appropriateness of the target differences used in the sample size calculations. The target difference could be compared to the MCIDs given in the literature for the corresponding outcome measure. It may be that studies are adequately powered based on the reported target difference but that clinically important differences could be missed if the reported target difference is larger than the MCID.

Another area where additional research would be beneficial is improving methods to predict the variability in study outcomes to more accurately predict the standard deviation used in the power calculation. For example, components of sample size calculations estimated using pilot studies could be compared to actual trial results (e.g., standard deviation) to assess the level of accuracy [91, 166]. It could be interesting to explore reasons for inaccuracy of the parameters, for example, differences in the study populations when the parameters are estimated based on the results of previous trials.

Future work in later chapters of this thesis build on this review. Chapter 3 reviews randomised trials using a single outcome measure, the WOMAC Index, and includes a summary of the use of the WOMAC in sample size calculations. Focusing on a single outcome measure allows more in-depth assessment of the components included in the sample size calculation, including whether the sample size calculation aligns with the statistical methods used to analyse trial results. Target differences and their justification can be compared more easily when the same scale is used across trials.

This review also highlights that sample size calculations in osteoarthritis trials do not account for the assessment time point or follow-up duration in the sample size calculation. Chapters 4 and 5 explore whether the duration of treatment effect is important in terms of the patient perspective and clinical importance, providing insight into whether follow-up duration should be considered in the sample size calculation and target difference specification. Chapter 6 uses simulations to examine the statistical properties of different methods for analysing randomised trials, including the implications for the statistical power.

2.4.6 Conclusion and recommendations

It was common for sample size calculations in trials of hip and knee osteoarthritis to be reported inadequately. Even when the sample size calculation was reported, the calculation frequently could not be replicated. This raises concerns about whether the sample size calculation was performed correctly and whether the trial was appropriately designed to achieve its primary objective. It also makes it difficult to establish how likely it is that a meaningful difference between the treatments exists. Clear and accurate reporting of a sample size calculation (or sample size justification) should be mandated by journal editors and peer reviewers for grant applications, study protocols and results publications.

Chapter 3

A systematic review of the collection, analysis and reporting of the WOMAC in hip and knee osteoarthritis trials

Prior publication:

The results for this chapter were published as a manuscript. Conferences abstracts for presentations on this chapter have also been published (see Appendix F.1 for details).

Copsey B, Thompson JY, Vadher K, Ali U, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA: Problems persist in reporting of methods and results for the WOMAC measure in hip and knee osteoarthritis trials. Quality of Life Research 2019, 28(2):335-343.

3.1 Introduction

Chapter 2 reported the findings of a literature review that highlighted the poor reporting of sample size calculations in randomised trials in osteoarthritis populations. However, this review included a heterogeneous sample of trials that used a variety of outcome measure. The Western Ontario and McMaster Universities Arthritis Index (WOMAC) is a disease-specific outcome measure designed for use in people with osteoarthritis [167]. The WOMAC is a commonly used outcome in osteoarthritis trials. This chapter builds on the previous review by examining the use of the WOMAC. Interpretation of the results of clinical trials can be hindered if outcome measures are used inconsistently across trials [168, 169]. This chapter examines the use of the WOMAC in sample size calculations and other aspects, including data collection, statistical analysis and reporting of trial results.

The WOMAC is one of the most widely-used outcome measures for hip/knee osteoarthritis [170, 171, 172] and has been used in clinical trials to evaluate the efficacy of surgical and pharmacological treatments [173, 174, 175, 176]. The FDA (United States Food and Drug Administration) and EMA (European Medicines Agency) have recommended the use of the WOMAC as an efficacy endpoint in medical research [176, 177, 178, 179].

The measurement properties of the WOMAC have been extensively reviewed [180]. Although it is limited for some psychometric aspects [181, 182], many studies have recommended the WOMAC as the superior outcome measure for knee and hip osteoarthritis in terms of reliability, validity, responsiveness and interpretability [183, 184]. A recent review found the WOMAC to be the "best-performing lower limb measure for hip/knee patients" undergoing arthroplasty and "the second most promising measure" when considering more specific populations of only hip replacement or only

knee replacement [185]. However, there are different versions of the WOMAC outcome measure and the WOMAC outcome score can be measured and calculated in different ways [186].

The WOMAC is made up of 24 items over 3 subscales (5 for pain, 2 for stiffness and 17 for function) and, for each item, participants report their difficulty with an activity, for example, climbing stairs. There are variations in how the items are rated. For example, items can be rated on a Likert scale (0=no difficulty to 4=extremely difficult) or using a 0-100mm visual analogue scale (VAS) or an 11-point numerical rating scale (NRS). The Likert scale version of the WOMAC questionnaire is given in Appendix B.

A user guide for the WOMAC is not freely available and requires the user to submit a request to www.womac.org [187]. The user guide includes information on how the WOMAC was derived, calculation of scores, and specific clinimetric and statistical issues. To obtain the user guide, users (researchers or clinicians) are required to submit a request to the developer of the WOMAC via the website www.womac.org, including their personal details and information on the intended use of the WOMAC measure.

The seminal paper did not provide the wording for the individual questions used in the questionnaire. The Likert scale and VAS have been found to be highly correlated [188] but these different approaches can cause confusion because the possible range of scores can be different for the two scales [189]. There are also variations in how scores are combined; for example, some studies calculate the overall score for the WOMAC scale as the sum of the individual items scores, while others calculate the average (mean) of the item scores. In a trial that reports that the WOMAC was used as an outcome measure, if the version of the WOMAC used is not reported, readers may assume the score for the pain subscale ranges from 0 to 20 points (using the sum of

Likert scale items), when in fact it could range from 0 to 10 points (using the average of a 0-10 NRS). This could lead to overestimation of the clinical significance of the results.

The effect on the total score range of the different options for scoring and combining individual items of the WOMAC are shown in Table 3.1.

Table 3.1: Score range for the WOMAC measure using different versions and methods to combine 24 individual item scores

	Combine 24 item scores using:		
Version used for item scores:	**Total**	**Average**	**Percentage**
Likert scale (item 0-4)	0-96	0-4	0-100
Numerical rating scale (item 0-10)	0-240	0-10	0-100
Visual analogue scale (item 0-100)	0-2400	0-100	0-100

Patient-reported quality of life measures are increasingly being used, if not as the primary outcome, then as one of the key outcomes collected within randomised trials across a range of conditions [190, 191, 192]. The WOMAC is a useful outcome measure to focus on when examining the use of patient-reported outcomes in randomised trials, as it is one of the most commonly used outcome measures in osteoarthritis [193].

A previous review by Woolacott *et al.* examined the use and reporting of the WOMAC and found poor reporting on the WOMAC Index total scale and pain subscale in trials of physical therapies [189]. The review included studies published until 2010. Since then, there have been considerable efforts to improve the quality of reporting in clinical trials [157], which may have encouraged trialists to more clearly describe the outcome measurement using the WOMAC.

As well as examining more recent use of the WOMAC, the review presented in this chapter adds to the existing literature by synthesising information on the WOMAC that has not been assessed previously, including its use in study design, statistical analysis and how the results of the WOMAC score were reported and interpreted.

3.1.1 Aims and objectives

Using a sample of published osteoarthritis trials, this review examines the use of WOMAC as a case study to explore issues in quality of life outcome measurement in randomised trials of any intervention (primarily around their analysis and input into study design).

The primary objective is to examine the statistical analysis methods used for the WOMAC outcome.

The secondary objectives are to i) review the data collection methods, ii) review the reporting of the study results of the WOMAC and, iii) where the WOMAC is used as a primary outcome, to examine the target difference and its justification. The final of these was carried out in preparation for the analysis presented in Chapter 5.

3.2 Methods

3.2.1 Identification of studies

Studies were identified from a cohort of trials included in a prior systematic review of sample size calculations for osteoarthritis trials (Chapter 2). In the previous review, several databases (Medline, Cochrane Central Register of Controlled Trials (CENTRAL), CINAHL, EMBASE, AMED, PsycINFO and PEDro) were searched for articles that reported the results of clinical trials of osteoarthritis. An example search strategy is given in Appendix A.1.

3.2.2 Eligibility criteria of previous review

The information below summarises the eligibility criteria of the original review. This is described in more detail in Chapter 2.

- Design: Randomised controlled trial of two or more treatments.
- Condition: Hip and/or knee osteoarthritis
- Exclusions: Study protocols, factorial, multi-arm, crossover, non-randomised, feasibility or pilot studies.

3.2.3 Additional eligibility criteria

Outcomes: Studies were included if they used the WOMAC Index or any one of the WOMAC subscales - pain, stiffness or physical function - as an outcome. Studies that only measured the WOMAC at baseline were excluded.

3.2.4 Selection of studies

Eligibility screening was conducted independently by pairs of reviewers.

3.2.5 Data Extraction and Management

Study characteristics were previously extracted for the earlier review in Chapter 2, including study design, population, primary outcome, sample size, and follow-up assessment. Data extraction for the previous review was conducted independently by a second reviewer for 20% of the included studies.

Where the WOMAC was the primary outcome, I had previously extracted the target difference and anticipated variability in the sample size calculation, including any justification for those values for the review in Chapter 2.

I extracted additional information using a standardised form, with no second author check.

The additional information extracted was:

- Study characteristics including eligibility criteria based on the WOMAC, follow-up assessment time points and baseline demographics (e.g., WOMAC mean score).
- Data collection methods: Version of pain score used (VAS, NRS, or Likert scale), how subscales were combined (using average, sum or other method), score range.
- Statistical analysis methods used to evaluate the WOMAC for the main study results, including whether results were analysed as post-treatment or change scores, statistical technique (e.g., mixed effects regression, t-test), dichotomi-

sation using WOMAC score (e.g., >20% improvement), covariate adjustment, handling of missing data (e.g., intention-to-treat or complete case).

- Statistical results reported: Within-group (mean score, mean change, percentage change, median), variability (standard deviation, confidence interval, range), between-group comparison (mean difference, p-value, standardised mean difference), binary outcomes (proportion, odds ratio, NNT (number needed to treat)).
- Clinical interpretation: Any reference to the clinical importance of the difference in the WOMAC (in design or discussion of results, e.g., MCID (minimum clinically important difference)).

3.2.6 Data Synthesis

Characteristics of the included studies were summarised using the number and proportion within each category, or the median and interquartile range for continuous outcomes.

For the data collection methods, analysis techniques and reporting of results, the number and proportion of studies in each category were reported. Data were summarised within subgroups based on the scale (or subscale) used.

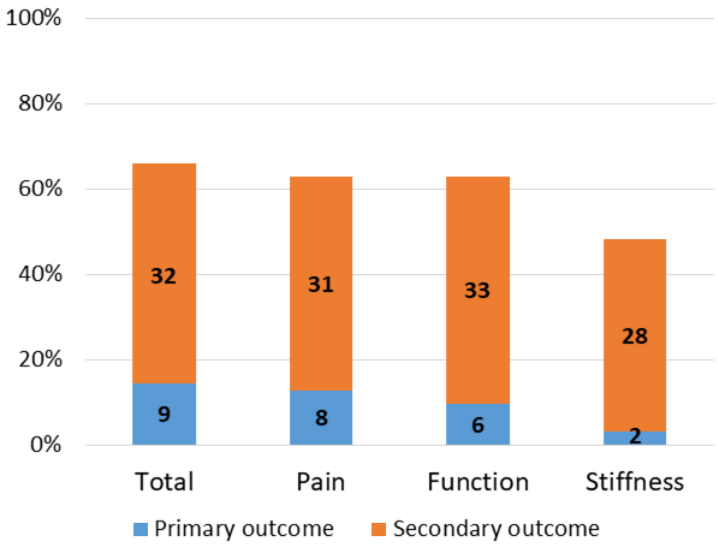
3.3 Results

3.3.1 Characteristics of included studies

The identification of studies is presented in a flow diagram in Chapter 2 (Figure 2.1). The original cohort of trials included 116 studies that were two-arm randomised trials of hip and/or knee osteoarthritis. Of the 116 trials included in the previous review, 62 reported using the WOMAC outcome measure and were included in this review. The characteristics of these 62 trials are summarised in Table 3.2. The majority of the included trials were single-centre superiority trials of drug or exercise interventions for knee osteoarthritis. Only one trial was cluster-randomised (2%, n=1/62).

The WOMAC total score was reported in 41 trials (66%, n=41/62), and it was used as the primary outcome in 22% of these trials (n=9/41) (Figure 3.1). The pain and function subscales were reported in 39 trials (63%, n=39/62) (Figure 3.1). However, only 30 trials reported the stiffness subscale (48%, n=30/62).

Figure 3.1: Reporting of WOMAC total and subscales as a trial outcome measure



Total number of studies: n=62.
Number of trials shown in graph bars, percentage of trials shown on y-axis. Figure adapted from Copsey *et al.* (2019) [194], published by CC BY 4.0.

Of the 41 trials reporting the WOMAC total score, less than half reported the results for all three individual subscales (pain, stiffness and function) (37%, n=15/41). Of the 62 included trials, 15 reported the total score and all WOMAC subscales, 15 reported all WOMAC subscales without the total score and 22 reported only the total score. The remaining 10 trials that did not report all subscales mostly reported the pain and function subscales only, excluding the stiffness subscale (without or without the total score) (Table 3.3).

Only 4 of the 62 trials restricted eligibility of the trial participants using the WOMAC measure. For example, 1 trial only included participants >300 on a 0-500 scale for the WOMAC pain subscale. An alternative version of the WOMAC was used in 12 trials; 8 used a translated version, 1 used a shortened version, 2 combined only two of the three subscales and 1 used one used metric scaling to weight the items. For the individual subscales, 5 trials used a translated version.

Table 3.2: Characteristics of included studies (n=62)

	n	%
Study hypothesis:		
Superiority	40	65%
Non-inferiority	3	5%
Multiple	1	2%
Unclear	18	29%
Study centres:		
Single centre	46	74%
Multi-centre	9	15%
Unclear	7	11%
Funding source:		
Industry	8	13%
Non-industry	25	40%
Combination	3	5%
No funding	4	6%
Not reported	22	35%
Population:		
Knee osteoarthritis	57	92%
Hip osteoarthritis	4	6%
Hip or knee osteoarthritis	1	2%
Intervention:		
Drug	19	31%
Surgery	4	6%
Exercise	14	23%
Other[a]	25	40%
Comparator:		
Active treatment	41	66%
Usual care	8	13%
Placebo or sham	13	21%
Number randomised:		
Median (IQR)	75 (50 - 148)	
Range	20 - 606	
n	61	
Follow-up period (months):		
Median (IQR)	4.5 (1.5 - 6)	
Range	0 - 36	
n	62	

[a] Other interventions included laser therapy, kinesio taping, acupuncture and ultrasound.

Table 3.3: Reporting of WOMAC total and subscales (n=62)

	n	%
Reported WOMAC total score:	41	66%
Total and all 3 subscales	15	24%
Total and 2 subscales (pain and function)	4[a]	6%
Total only	22	35%
Did not report WOMAC total score:	21	34%
3 individual subscales	15	24%
2 subscales (pain and function)	4	6%
1 subscale	2[b]	3%

[a] For 1 study, the total score used was the sum of the pain and function subscales only. For 3 studies, the stiffness subscale was used to calculate the total score but results on the stiffness subscale were not reported separately.
[b] 1 study reported the pain subscale only and 1 study reported the function subscale only.

None of the trials made the WOMAC questionnaire available or provided details of the wording of individual items as part of the results publication. Of the 10 trials which referred to a published protocol, 1 trial provided the WOMAC questionnaire used as an appendix to the study protocol [195].

The Knee injury and Osteoarthritis Outcome Score (KOOS) and Hip disability and Osteoarthritis Outcome Score (HOOS) are both longer joint-specific measures based on the WOMAC. One advantage of using the KOOS or HOOS is that the WOMAC score can be calculated using a subset of the included items [196, 197]. However, none of the 62 trials that reported the WOMAC used the KOOS or HOOS outcome measures.

Where reported, trial publications stated that the WOMAC was completed by the participant (34%, n=21/62) or by the participant with a clinician assessor (34%, n=21/62). No trial reported that the WOMAC was collected by a proxy for the participant. However, 1 trial noted that the questionnaire was read by the investigator for illiterate participants [198].

3.3.2 Measuring the WOMAC

The majority of the studies used a 5-point Likert scale to measure the items of the WOMAC measure, with fewer studies reporting the 0-10 numerical rating scale or 0-100 visual analogue scale (Figure 3.2). The item range was unclear for a quarter of studies reporting the WOMAC total score (Table 3.4). The most common range for the total WOMAC score was 0-96 (41%, n=17/41). A range of 0-96 corresponds to summing the score for a 0-4 Likert scale for all 24 items. However, the total score range was as small as 0-10 (averaging item scores using a 0-10 scale) and as large as 0-2400 (summing item scores using a 0-100 scale).

Figure 3.2: An overview of measurement of the WOMAC

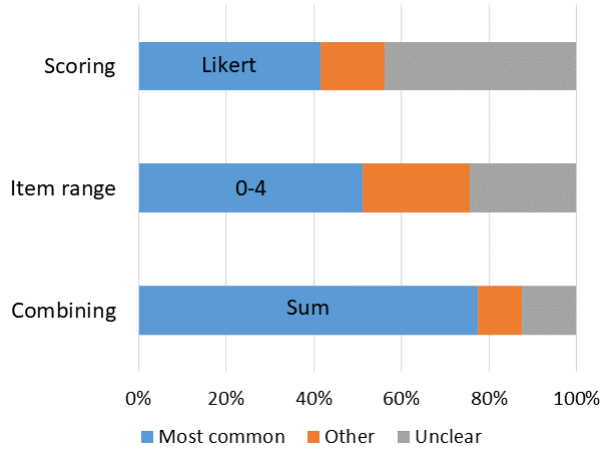


Table 3.4: Measurement of the WOMAC

	Total		**Pain**		**Function**		**Stiffness**	
Denominator	n = 41		n = 39		n = 39		n = 30	
	n	%	n	%	n	%	n	%
Scoring version:								
Likert scale	17	41%	25	64%	24	62%	18	60%
NRS	3	7%	3	8%	3	8%	1	3%
VAS	3	7%	4	10%	4	10%	4	13%
Unclear	18	44%	8	21%	8	21%	7	23%
Item range:								
0-4	21	51%	25	64%	25	64%	19	63%
0-10	7	17%	7	18%	7	18%	5	17%
0-100	2	5%	5	13%	5	13%	5	17%
1-4	1	2%	0	0%	0	0%	0	0%
Unclear	10	24%	2	5%	8	21%	1	3%
Maximum of scale:								
Median (IQR)	96 (96-100)		20 (20-50)		68 (68-150)		8 (8-20)	
Range	10-2400		10-500		10-1700		8-200	
Method to combine subscales:								
Sum	31	78%	n/a		n/a		n/a	
Average	3	8%	n/a		n/a		n/a	
Sum and convert to percentage	1	3%	n/a		n/a		n/a	
Unclear	5	13%	n/a		n/a		n/a	

The details of measuring the WOMAC score were often not reported and required assumptions based on other reported information. For the WOMAC total score, the version used for data collection (e.g., Likert scale, VAS or NRS) was only reported by 56% of trials (n=23/41). The range used for a single item of the WOMAC score was reported by half of trials. The item range was assumed based on reported information for a quarter of trials, most commonly assumed from the mean score in the trial results. Although the range of a single item suggested the use of a particular version (e.g., 0-4 for a Likert scale), the item range was not used to make assumptions about

the version used. For example, an item score ranging from 0-10 could be collected using a Likert scale, numerical rating scale or visual analogue scale.

The range for the combined score and method used to combine items for the WOMAC total scale were not reported in most studies. Although the range of the combined total score could often be assumed from the mean score or range of a single item, it was still unclear for 20% of trials (n=8/41). The level of reporting was higher for the individual subscales, with around 75% of trials reporting the version used and the range for a single item. Although 50% of trials reported the combined range for each subscale, it could be assumed for the remaining trials based on the mean score or the item range. It was more difficult to make assumptions on the combined range of the total score because it was difficult to differentiate between a 0-96 and 0-100 range using only the reported mean values in the trial results.

For the study results, four of the included studies categorised participants based on their WOMAC score (two using the WOMAC total score and two using the WOMAC pain subscale). Dichotomisation was based on their post-treatment score in one study [199] and based on their percentage change score in three studies [200, 201, 202]. Two studies used a combination of the WOMAC post-treatment score, WOMAC change score and participant global assessment score to define a treatment responder based on the OARSI responder criteria [201, 203, 204]. One study categorised participants into subgroups based on their baseline WOMAC score [205].

3.3.3 Analysing the WOMAC

The methods of analysis were similar between the different subscales. The majority of studies used a t-test, repeated-measures ANOVA or mixed effects model (Table 3.5). The proportion of trials using covariate adjustment ranged from 27% for the WOMAC total score to 46% for the pain subscale. For the trials that adjusted for baseline covariates, the most common covariates to adjust for were baseline score (13 trials), BMI (7 trials), sex (6 trials) and age (5 trials). Only two-thirds of trials reported the methods used to handle missing outcome data, with around a quarter using complete case analysis and a quarter using imputation.

Table 3.5: Analysis of the WOMAC

	Total		Pain		Function		Stiffness	
Denominator	n = 41		n = 39		n = 39		n = 30	
	n	%	n	%	n	%	n	%
Statistical analysis method:								
t-test	13	33%	11	28%	11	28%	10	33%
Repeated measures ANOVA	9	23%	7	18%	8	21%	6	20%
Mixed effects	8	20%	10	26%	10	26%	7	23%
ANCOVA	3	8%	5	13%	4	10%	4	13%
Mann-Whitney U-test	6	15%	4	10%	4	10%	3	10%
Other	1	3%	2	5%	2	5%	0	0%
Adjusted for covariates:								
Yes	11	27%	18	46%	17	44%	12	40%
No	26	63%	21	54%	22	56%	18	60%
Unclear	4	10%	0	0%	0	0%	0	0%
Method to handle missing data:								
Complete case	10	24%	10	26%	10	26%	7	23%
Multiple imputation	3	7%	5	13%	5	13%	4	13%
Single imputation (e.g., LVCF)	3	7%	5	13%	5	13%	3	10%
Mixed model without imputation	5	12%	5	13%	5	13%	3	10%
No missing data	4	10%	2	5%	2	5%	2	7%
Unclear	16	39%	12	31%	12	31%	11	37%

A quarter of studies measured the WOMAC at only a single follow-up time point, i.e., once at baseline and once immediately after the end of treatment (26%, n=16/62). Nearly half of trials measured the WOMAC at three or more follow-up time points post-randomisation (42%, n=26/62). The longest follow-up assessment period was 3 years.

Of the trials with more than one follow-up assessment, 26 trials used a longitudinal analysis method (42%, n=26/62), 18 conducted multiple analyses each using data from a single time point (39%, n=18/62) and 2 trials only conducted between-group analysis at the final assessment time point (4%, n=2/62).

3.3.4 Interpreting results from the WOMAC

The majority of studies reported the within-group mean post-treatment score (88%, n=36/41 for WOMAC total) (Table 3.6). Some reported the mean change and post-treatment score (20%, n=8/41 for WOMAC total). Most studies also reported the corresponding standard deviation (88%, n=36/41 for WOMAC total). The level of reporting of the mean difference and standard deviation were similar for the individual WOMAC subscales. For the WOMAC total score, only one-third of studies reported the mean difference between the treatment arms (32%, n=13/41 for WOMAC total). The proportion of studies reporting the between-group difference was higher for the WOMAC subscales but still remained below half. Where the between-group difference was not reported, trials often reported only the p-value for the between-group analysis.

Table 3.6: Reporting of the WOMAC results

	Total		**Pain**		**Function**		**Stiffness**	
Denominator	n = 41		n = 39		n = 39		n = 30	
	n	%	n	%	n	%	n	%
Group summary score:								
Mean post-treatment score	28	68%	18	46%	20	51%	15	50%
Mean change score	2	5%	3	8%	3	8%	3	10%
Mean post-treatment and change scores	8	20%	13	33%	12	31%	7	23%
Median post-treatment score	1	2%	2	5%	2	5%	2	5%
Median change score	0	0%	0	0%	0	0%	0	0%
Median post-treatment and change scores	0	0%	0	0%	0	0%	2	5%
Multiple reported	1	3%	2	5%	1	3%	1	3%
None reported	1	3%	1	3%	1	3%	0	0%
Within-group variation:								
Standard deviation	26	61%	22	56%	22	56%	17	57%
Standard error	0	0%	1	3%	1	3%	0	0%
95% confidence interval	5	12%	6	15%	6	15%	4	13%
Range	1	2%	0	0%	0	0%	1	3%
Standard deviation and 95% confidence interval	2	5%	4	10%	4	10%	4	13%
Other reported	0	0%	1	3%	1	3%	1	3%
Multiple reported	1	2%	1	3%	1	3%	1	3%
None reported	7	17%	4	10%	4	10%	2	7%
Between-group score:								
Mean difference	13	32%	18	46%	18	46%	12	40%
None reported	28	68%	21	54%	21	54%	18	60%
Between-group variation:								
95% confidence interval and p-value	10	24%	12	31%	12	31%	10	33%
95% confidence interval	3	7%	5	13%	5	13%	2	7%
p-value	20	49%	19	49%	18	46%	16	53%
p-value interval (e.g $p<0.05$)	6	15%	3	8%	3	8%	2	7%
None reported	2	5%	0	0%	1	3%	0	0%

All four studies that dichotomised participants based on the WOMAC score reported the proportion of participants achieving the dichotomised score. Of these, one study also reported the corresponding odds ratio and two studies reported the corresponding p-value. Two studies dichotomised participants based on the percentage change in the WOMAC (or its subscales): Allen *et al.* used cut-offs of 12% and 18% improvement in the WOMAC total score and Jin *et al.* used cut-offs of 20% and 50% improvements in the WOMAC pain subscale [201, 206]. Losina *et al.* dichotomised participants based on whether they achieved the MCID of 10 points in the WOMAC function score [199]. Yuan *et al.* categorised participants based on whether the WOMAC total score decreased by $\geq \frac{2}{3}$, $\geq \frac{1}{3}$ or $< \frac{1}{3}$ [202].

One additional study used dichotomisation of the WOMAC as part of a wider definition of clinical response. If participants had at least 50% improvement and an absolute change of at least 20 points on the WOMAC pain or function subscales, Wang *et al.* defined this group as 'clinically improved' [203].

3.3.5 Sample size calculations using the WOMAC

Among the 41 studies that reported a power calculation, 17 studies estimated the sample size based on the WOMAC scale or one of its subscales. Among these 17 studies, 7 studies were powered on the WOMAC total score, 6 studies on the pain subscale and 4 studies on the function subscale. The target sample size ranged from 40 to 376 for the WOMAC total score and 60 to 300 for the WOMAC function subscale. The sample size was generally higher for the WOMAC pain subscale, ranging from 200 to 560. The information from the sample size calculations is presented in Table 3.7.

The components of the sample size calculation were difficult to compare due to the different scale ranges across trials, as discussed in section 3.3.2. There was variation in

the values used for the target difference and assumed standard deviation, even within studies that used the same scale range. For example, the target difference ranged from 2 to 15 for trials where the sample size was estimated based on the WOMAC total score using a 0-96 scale. Where a justification was provided, the standard deviation assumed in the sample size calculation was justified based on a published trial (n=7), an uncited pilot study (n=1) or both (n=1). The target difference was justified based on a published estimate for an MCID (n=5) or a published trial (n=4).

The commonly referenced MCID estimates were those published by Angst *et al.* in 2001 and 2002 and by Tubach *et al.* in 2005 [39, 193, 207]. None of the MCID estimates were calculated based on the same scale range as the scale range used in the trial. For example, Hinman *et al.* used the 0-68 function subscale but cited the MCID publication by Tubach *et al.* published in 2005 based on the 0-100 function subscale [39, 208]. Allen *et al.* used the 0-96 total scale but cited the MCID publication from Angst *et al.* based on the 0-10 total scale. However, Allen *et al.* justified the target difference by translating it into the percentage change from baseline [193, 200, 206, 207].

Table 3.7: Sample size calculation information for trials powered on WOMAC

Trial	Range	Target difference	NIM	Assumed standard deviation	Target SES	Target sample size	Justifications: Target difference	Standard deviation	Alpha	Power
Total scale										
Allen (a)	0-96	4.2	n/a	14	0.3 [r]	300	MCID	Pilot	0.05	0.80
Allen (b)	0-96	2	n/a	NR	0.27 [r]	376	MCID	n/a	0.05	0.80
Bagnato	0-240	10	n/a	NR	NR	66	NR	n/a	0.05	0.80
Bisicchia	Unclear	6	n/a	12	0.5 [c]	150	NR	NR	0.05	0.80
Calatayud	0-96	10	n/a	11	0.91 [c]	40	NR	Trial	0.05	0.80
Rahimzadeh	0-96	8.5	n/a	5.2	1.63 [c]	42	Trial	Trial	0.05	0.95
Simental -Mendia	0-96	15	n/a	20	0.75 [c]	65	Trial	Trial	0.05	0.95
Pain subscale										
Wadsworth	0-20	1.3	n/a	4.5	0.29 [c]	260	Trial	Trial	0.10	0.80
Jin	0-500	20	n/a	NR	NR	400	Trial	n/a	0.05	0.80
Conrozier	0-20	NR	1.35	NR	n/a	208	n/a	n/a	0.05	0.80
Wang (b)	0-500	NR	20	100	n/a	180	n/a	Trial	0.05	0.80
Hochberg	0-500	NR	8	26	n/a	560	n/a	Trial	0.025	0.90
Hill	0-50	NR	n/a	NR	0.4 [r]	200	n/a	n/a	0.05	0.80
Function subscale										
Hinman	0-68	6	n/a	10.5	0.57 [c]	164	MCID	Pilot and trial	NR	0.90
Losina	0-100	10	n/a	NR	0.5 [r]	300	MCID & SES	n/a	0.05	0.8
Monaghan	0-68	10.4	n/a	13.6	0.76 [r]	60	MCID	Trial	0.05	0.80
de Rooij	0-68	NR	n/a	NR	0.4 [r]	154	NR	n/a	0.05	0.80

SES: standardised effect size, NIM: non-inferiority margin, NR: not reported. [c] SES calculated from reported values [r] SES reported.

3.4 Discussion

3.4.1 Summary of findings

The WOMAC measure is widely used in trials of hip and knee osteoarthritis, with around half of trials using the WOMAC or at least one of its corresponding subscales. This review found that poor reporting on the measurement of the WOMAC scale remains a problem. There was also high variability in the version of the WOMAC scale that was used and how the individual items or subscales were combined. This high variability makes it difficult to elicit information on the range of the scale when it is not reported. In the majority of trials, the scoring range for the WOMAC total score was 0-96. However, for some trials, the range was as small as 0-10 or as large as 0-2400. The problems of poor reporting and lack of clarity on how the scale was measured apply to both the WOMAC total score and the individual pain, function and stiffness subscales.

It was common for trials to use a t-test, repeated measures ANOVA or mixed effects model to analyse the WOMAC score. In the analysis of the results, most trials did not use baseline covariate adjustment and one-third of trials did not report how missing data were handled. The majority of trials reported the mean score and standard deviation (or equivalent) within each treatment arm. However, only one-third of trials reported the between-group difference for the WOMAC score (total or subscale), hindering the interpretation of trial results.

3.4.2 Comparison with existing literature

In recent years, there has been increasing awareness of the need for high-quality reporting of the methods and results of randomised trials. However, despite examining

an up-to-date sample of trials, this review found little evidence of improvement in the consistency or reporting of how the WOMAC was measured.

The poor reporting and variability in measurement of the WOMAC align with the findings of Woolacott *et al.* [189]. Both reviews found that the majority of studies used the 5-point (0-4) Likert scale. Woolacott and colleagues found that around half of trials reported the type of scale and the score range for the pain subscale. They also found that the range and type of scale could not be assumed based on the reported data for 20% and 10% of trials respectively. These results are similar to those found in this review. The literature also shows that lack of information on how missing data are handled is a common problem, not restricted to the WOMAC outcome or osteoarthritis trials [209, 210, 211, 212].

A key difference in this review compared to the review by Woolacott *et al.* was that most trials reporting the WOMAC pain subscale also reported the function and stiffness subscales, whereas Woolacott *et al.* found that the function subscale was rarely reported. This seems counter-intuitive when the review by Woolacott *et al.* was restricted to physical therapy interventions. This difference may be due to changes in common practice over time. However, both reviews showed that many trials reported the results using the WOMAC total score without reporting the results of the component subscale scores.

Some trials used published MCID estimates to justify their choice of target difference. However, in all cases, the MCID was calculated based on a different version of the WOMAC scale to the version used in the randomised trial. A recent review of published MCID estimates in degenerative knee disease outcomes found only three studies that produced 'credible' MCID estimates, none of which were used in the sample size calculations of the included trials [213].

3.4.3 Strengths and limitations

This review examines the use and reporting of the WOMAC up-to-date sample of trials identified using a systematic search strategy. This is the first review to examine the WOMAC subscales of function and stiffness, as well as the pain subscale and total score. Eligibility for inclusion in this review was not restricted by intervention, thus making the results relevant to trials of interventions including physical therapies, surgery and pharmacological treatments.

The key limitation of this review is that the included studies were restricted to publication during 1 year (January - December 2016). The results do not provide information on the trends in the use and reporting of the WOMAC over time. The use of the WOMAC is likely to differ in observational studies or multi-arm trials, compared to two-arm randomised trials; however, this was outside the scope of this review.

3.4.4 Implications

The results of this review indicate that the problems of poor reporting of information on the measurement of the WOMAC scale highlighted by Woolacott *et al.* persist in osteoarthritis trials [189]. The less restrictive eligibility criteria used in this review also demonstrate that the issues of poor reporting and unclear measurement methods apply to the WOMAC subscales and trials evaluating various types of interventions.

The key implication of these results is that the lack of clarity in how the WOMAC was measured hinders the interpretation of trial results. When the range of the scale is ambiguous, it is difficult to interpret both the treatment effect and level of disease severity. For example, a 20-point change on a 0-96 scale would be seen as much more important than the same difference on a 0-2400 scale. Poor reporting of the effect estimates from between-group analyses could also hinder the interpretation of

study results. The inconsistency in the version of the WOMAC scale used and how the items are combined makes it difficult to compare the results of different trials. Although some research has suggested that the different versions of the WOMAC produce correlated results, there could still be discrepancies due to the methods used [188].

Variation in how the WOMAC is scored also makes it more difficult to assess the clinical importance of treatment effects when the results between trials cannot be directly compared. MCIDs can be used to provide a uniform method of assessing whether a treatment effect is considered worthwhile. WOMAC measure, for instance, assessing the psychometric properties. This research, as well as methogological research on the psychometric properties of the measure, would need to be repeated for all versions of the WOMAC measure. For example, when assessing validity and responsiveness, the sensitivity of a 0-4 Likert scale, 0-10 numerical rating scale and 0-100 visual analogue scale are likely to be very different. Until further research has been conducted on the psychometric properties of each version, researchers should choose the 0-4 point Likert scale version as it appears to be the most commonly used version of the WOMAC in osteoarthritis trials.

Although the reporting of individual subscales has improved since the review by Woolacott *et al.*, there are still several studies that report the WOMAC total score without reporting the individual subscales for pain, function and stiffness. It would be judicious to report the results for all three domains, as this would allow readers to see which domains are generating the treatment effects. An intervention may produce different effects across the three domains, which may not be clear when the subscales are combined. For example, an intervention may significantly improve function but increase pain compared to the control. Although there would be no difference in the WOMAC total score, the differences in individual domains would be

useful information for choosing between the two treatments. Presenting the results of the individual subscales would be useful when the changes in some domains are more desirable than others due to patient preference or the hypothesised treatment mechanisms [214]. Reporting the results of individual subscales as well as the total score would also facilitate comparison of the trial results with other trials where only individual subscales are reported without the total score.

There is a clear need for improved reporting of how the WOMAC is measured to facilitate the interpretation of study results and comparison of results across trials.

3.4.5 Future research

Future research should look further into the most appropriate version of the WOMAC measure to use for particular situations. Studies have found that the different versions are correlated, however it is unclear whether the psychometric properties are similar [188]. For example, there are likely to be differences in the sensitivity of a 0-4 Likert scale, 0-10 numerical rating scale and 0-100 visual analogue scale formats. Future studies should examine the validity, responsiveness and usability of measures in specific settings. Along with existing evidence, this could then be used to produce recommendations on how to measure and analyse the WOMAC scale and its subscales. This could help to improve the consistency in how the WOMAC is scored.

Future studies could assess whether 'translating' the scores measured using one version of the WOMAC into another version produces comparable results. Researchers could then explore whether it is appropriate to standardise the results of different versions of the WOMAC onto the same scale (e.g., converting all versions to a 0-100 scale, equivalent to the percentage of the maximum possible score). This would facilitate the comparison and synthesis of trial results.

Future clinical trials in knee osteoarthritis should use the 5-point Likert scale version of the WOMAC and calculate scores by summing individual items, especially when trialists feel it is important to compare their results with previous studies. Trialists who decide to measure the WOMAC using lesser-used versions should provide sufficient justification for doing so.

Future reviews could examine the characteristics of the sample used to calculate the MCID of the WOMAC and compare this to the characteristics of randomised trials where the MCID estimates have been applied. This was not completed in this review due to the small number of studies that justified the target difference using an MCID estimate. Future studies could also examine the generalisability of MCID estimates for the different versions of the WOMAC measure, for example, to see whether it is appropriate to translate an MCID calculated for the Likert scale version into the VAS version. This would provide more information on the applicability of MCID estimates in sample size calculations.

As the 0-4 point Likert scale was found to be the most commonly used in this review, this version was used in the work in subsequent chapters (secondary analysis in Chapter 5, discrete choice experiment in Chapter 4 and simulation study in Chapter 6). This review also found that several trials measured individual subscales of the WOMAC. Chapter 4 builds on this by examining how people living with osteoarthritis value the different domains of the WOMAC measure and explores whether the WOMAC adequately captures what is important to people with osteoarthritis.

Several trials included in this review used published MCID estimates for the WOMAC to inform the specification of the target difference in their sample size calculation. Chapter 5 explores whether it is appropriate to use MCID estimates to inform the design of trials using different follow-up time points. Chapter 5 estimates the MCID for the WOMAC using secondary analysis of a cohort dataset at different follow-up

time points to assess whether the MCID estimate is consistent over time. This review found that different methods were used to analyse the WOMAC outcome data. Chapter 6 compares the statistical properties of some of these methods using simulated data for the WOMAC. The data collected on the mean and standard deviation of the treatment effect estimates and baseline outcome scores informed the data generating mechanism using in the simulation study in Chapter 6.

3.4.6 Conclusion

This review found that the WOMAC measure was widely used in randomised trials of hip and knee osteoarthritis. However, there was large variation in the version of the scale used and how the data were analysed. There was poor reporting of the study methodology and results for the WOMAC score and its subscales. The interpretation of findings and comparisons with other studies was hindered by limitations, such as the ranges of the outcome scales being unclear, not reporting an effect size estimate, or not exploring the effects of assumptions on missing data mechanisms.

The version of the WOMAC used, how the item scores were combined and the range of the scale (and subscales) should be reported in the methods of the published papers. The between-group difference and corresponding confidence interval should be reported in the results when between-group analyses are conducted. When the WOMAC total score is analysed, the results for the individual subscales (pain, function and stiffness) should also be reported. This would allow readers to assess whether the treatment effect on the total score is mainly due to a change in a single domain.

Future research should examine which version of the WOMAC measure has optimal properties in different situations. Increasing consistency between trials and clear reporting of the measurement of the WOMAC would facilitate clinical interpretation and comparison of trial findings.

Chapter 4

How important is duration of treatment effect to people living with osteoarthritis? A discrete choice experiment

Prior publication:

The results for this chapter were published as a manuscript. Conferences abstracts for presentations on this chapter have also been published (see Appendix F.1 for details).

Copsey B, Buchanan J, Fitzpatrick R, Lamb SE, Dutton SJ, Cook JA: Duration of treatment effect should be considered in the design and interpretation of clinical trials: Results of a discrete choice experiment. Medical Decision Making 2019, 39(4):461-473.

4.1 Introduction

This chapter reports a stated preference study exploring whether the duration of the treatment effect is an important aspect for people with osteoarthritis and therefore whether it should be accounted for in clinical trial design. Chapter 2 showed that randomised trials did not account for the longitudinal nature of data collection in their sample size calculations. If the duration of the treatment effect is important to people with osteoarthritis, this could justify accounting for duration in the design and analysis of osteoarthritis trials. The stated preference study considered the duration of treatment effect on three different domains of osteoarthritis symptoms: pain, stiffness and function. These three domains align with the three subscales of the WOMAC Index, the patient-reported composite outcome measure reviewed in Chapter 3. This stated preference study builds on Chapter 3 to explore the differences in how people with osteoarthritis value symptom relief in each of the three areas, using the most common version of the WOMAC.

4.1.1 Motivation

There has been increasing attention on ensuring that both clinical practice and medical research are patient-focused [215, 216]. In clinical practice, there has been a shift towards shared decision-making, where the person is informed about their treatment choices and involved in decisions about their care [217, 218, 219, 220]. In medical research, patients and the public are increasingly involved in prioritising research questions, designing and selecting patient-reported outcome measures, and disseminating findings to a wider audience [221, 222, 223, 224].

When considering the impact on time in clinical trial design and interpretation, it is important to explore how the duration of treatment effect affects preferences for

treatment from a patient or participant perspective. Lasting symptom relief is an attractive characteristic for medications [225]. Randomised trials often find that that short-term treatment effects are not sustained over time and do not always translate into long-term effects [226, 227, 228]. Providers are aware that it is desirable for medication effects to be quick, strong and long-lasting [229]. For chronic conditions, long-term follow-up assessment is advocated to evaluate whether treatment effects are sustained over time [71, 230, 231, 232]. Clinicians have suggested that it is more likely that medications will be adopted if they have been tested in trials of longer follow-up duration [233]. However, it is unclear how people choose between medications that could provide a stronger or longer treatment effect. A person could be willing to take a medication that provides reduced symptom relief if the medication is likely to remain effective for a longer time. A persons's preference for a particular treatment could be influenced by the level of treatment effect, the duration of effect and potential risks and side effects, as well as other treatment characteristics.

Understanding the factors that are important to patients when making treatment decisions can help clinicians to know which issues should be discussed with patients to inform their shared decision-making. This information could also be used to inform decision support aids and provide evidence for researchers, funding bodies, guideline developers and commissioners to inform decisions on prioritising research directions and implementing treatments into clinical practice [234, 235]. If duration of treatment effect is found to be important, this could affect the design and interpretation of clinical trials, for example, the selection of trial endpoints, the incorporation of duration into the analysis of treatment effectiveness and assessment of the risk-benefit of different interventions [236].

This study focuses on the importance of the duration of treatment effect in the context of medications for osteoarthritis. This is a suitable case example because osteoarthri-

tis is a long-term condition and medications can relieve symptoms of osteoarthritis for varying lengths of time. Medications for osteoarthritis, such as NSAIDs (non-steroidal anti-inflammatory drugs), can also provide different benefits and risks, in terms of relieving different symptoms and being associated with an increased risk of serious events. The motivation for focusing on osteoarthritis is described in more detail in Chapter 1.

Background outline

This chapter describes a stated preference study (a discrete choice experiment) designed to examine whether the duration of treatment effect is important to people with osteoarthritis when choosing between different medications. The results of the study can be used to inform decisions on the methods used to specify the target difference in sample size calculations. This Introduction describes revealed and stated preference studies (Section 4.1.2), what a discrete choice experiment is (Section 4.1.3), and why a discrete choice experiment is an appropriate study design (Section 4.1.4). It concludes with a summary of the literature examining the importance of time in patient preference studies in any health condition and specifically in osteoarthritis (Section 4.1.5).

4.1.2 Revealed and stated preference studies

People's preferences can be explored using both revealed and stated preference studies. In revealed preference studies, participants make choices in real-life scenarios, and these choices can be used to quantify their preferences. For example, data on prescriptions can highlight which medications are preferred and explore factors that influence this choice. However, in clinical practice, the clinician often drives the decision of which medication someone should be prescribed. Prescribing practices will likely not

provide robust information about what the person prefers. Another disadvantage of revealed preference studies is that certain medications may not be suitable for specific groups of people, for instance, due to contraindications with other medications they are taking.

This study used a stated preference study, in which participants stated which choice they would make in a hypothetical scenario. Although participants are not forced to act on their choices, stated preference studies have the advantage that the medication options available and how these options are described can be restricted in a controlled setting. As stated preference studies are based on hypothetical scenarios, the same participant can state which option they would prefer in multiple scenarios, allowing researchers to gain more information from each participant and reduce potential confounding.

4.1.3 What is a discrete choice experiment?

A discrete choice experiment (or DCE) is a type of stated preference study in which respondents select their preferred choice from multiple options in hypothetical scenarios. Each option is described in terms of several 'attributes' or characteristics. The results of a discrete choice experiment can be used to quantify the relative importance of different factors in the decisions made by respondents.

4.1.4 Reasons for using a discrete choice experiment

There are several alternative methodologies that could have been used instead of a discrete choice experiment. These alternative study designs include contingent valuation (willingness-to-pay), standard gamble methods or time-trade-off. However, standard gamble methods have been criticised due to their complexity, as participants

must consider probabilities of different health states within the same scenario [237]. Willingness-to-pay has been shown to relate more to participant's financial situation than their health state [238].

A key advantage of a discrete choice experiment is that the relative valuations of multiple attributes can be assessed simultaneously, whereas in a time-trade-off study, only two attributes can be included. It has also been suggested that discrete choice experiments are less cognitively demanding and may encourage more qualitative decision-making [239, 240, 241, 242]. Discrete choice experiments have also been found to be less prone to anchoring effects and could reduce the risk of 'gamification'. For example, in some time trade-off experiments, instead of considering the different available options, participants may select responses to finish the survey more quickly. [243]. A discrete choice experiment was used in this study because it allows more than two attributes to be examined without increasing cognitive demand or introducing bias due to participant characteristics.

4.1.5 Existing literature

Preference studies on time in any health condition

Patient preference studies in other conditions have examined the importance of time in different ways, for example, exploring trade-offs between quality of life and life expectancy [244, 245]. Previous discrete choice experiments have included duration of illness in benefit-harm trade-offs [246, 247]. However, this attribute describes the length of time until the illness is cured. There is no cure for osteoarthritis and therefore the duration of symptom reduction is more relevant.

Other studies have used discrete choice experiments to estimate time preferences in terms of discounting, exploring how the valuation of a health state changes when

the onset of worsening health is postponed to a future time point [52]. A discrete choice experiment was used to explore preferences on duration of treatment effect in people with psoriasis [248]. The results suggested that people with psoriasis had a strong negative preference for treatments with short duration of beneficial effects and a meta-analysis found that they would accept a much higher risk of adverse events to increase duration of effect compared to reducing symptoms [249]. However, this is only one example for a single condition. Overall, there is limited existing literature and no consensus on the strength of these preferences for duration relative to other aspects of treatment.

Preference studies on time in osteoarthritis

In osteoarthritis, preference studies have considered the impact of time to the onset of action of medication [250, 251], waiting time for surgery [252, 253], and time in hospital [254, 255, 256, 257]. Preference studies have examined how the length of time in a particular health state affects treatment choices [258, 259]. However, the level of treatment effect has been broadly defined in these studies, for instance comparing mild and severe disease states or comparing a successful operation to one requiring revision.

The majority of discrete choice experiments conducted in people living with osteoarthritis consider treatment effects of fixed duration. For example, Hauber *et al.* found that decreased risk of adverse events, pain levels and functional problems were important factors in osteoarthritis medication choice. However, they only considered the treatment effect 1 hour after taking medication and did not consider whether this effect may decline over time [214]. Few studies have examined the importance of the duration of treatment effect from a course of medication. Posnett *et al.* considered the duration of effect for oral medications and injections and found that duration

of pain relief from injections was one of the most important factors for treatment preferences [251]. However, the duration of pain relief was based on one dose of oral medication (8 or 12 hours).

It is unclear how the duration of treatment effect affects preferences for treatment in people with osteoarthritis.

4.1.6 Objectives

This discrete choice experiment aimed to evaluate how duration of treatment effect affects treatment preferences in people with osteoarthritis and quantify the relative importance of the duration of effect, level of treatment effect and potential risks of the treatment. This aligns with the main objectives of this thesis by indicating whether participants value treatment effects lasting different lengths of time differently. This could provide justification for accounting for the time point of assessment in randomised trials of treatments for osteoarthritis when specifying the target difference in sample size calculations and analysing the results.

4.2 Development and Methods

The discrete choice experiemnt was planned following the practice guidelines by Bridges *et al.* [260].

4.2.1 Attribute identification and selection

A literature search was performed to identify factors that are relevant to people with osteoarthritis when choosing between different treatments. Searches were performed in MEDLINE, EMBASE and EconLit from inception to 16 August 2017. The search strategies combined terms on osteoarthritis and methods for stated preference studies (Appendix C.1).

The search identified 868 articles (324 from MEDLINE, 544 from EMBASE, 0 from EconLit). From these articles, 157 relevant preference studies were identified, of which 61 were stated preference studies. The potential attributes extracted from these studies covered three main areas:

1. Treatment effectiveness: These included the overall treatment effect, pain, function, and speed and length of symptom relief.

2. Safety risks and side effects: These included risk of addiction, risk of serious events (such as heart attack, stomach bleeding or renal failure), adverse effects (such as nausea or diarrhoea), and surgical complications (such as risk of infection).

3. Treatment administration and cost: These varied depending on the treatments being considered. For medications, the factors included cost, treatment administration (pill, cream or injection), and convenience of fitting the treatment into lifestyle (treatment schedule and need for a prescription).

The results of the literature search were used to create a long list of potential attributes that could be included in the discrete choice experiment, excluding those relating specifically to non-pharmacological treatments. As discrete choice experiments can include only a limited number of attributes, patient input was used to identify which of the potential attributes were most important. A group of 10 people with osteoarthritis were recruited using an online advert placed on the PAIRS (Patients Active in Research) Thames Valley website. These patient representatives conducted rating and ranking exercises on the importance of the potential attributes (results presented in Appendix C.2).

Attributes were selected for the discrete choice experiment based on the results of the patient input and existing literature on preferences, avoiding the selection of attributes with overlapping constructs. Attributes were chosen from all three domains and are described in Sections 4.2.1.1-4.2.1.3.

4.2.1.1 Treatment effectiveness

The effect on pain, stiffness and ability to perform daily activities were included as they were consistently rated of high importance. This aligns with the three domains in the WOMAC outcome measure, which is commonly used in randomised trials of osteoarthritis [167, 261]. The WOMAC outcome measure is described in more detail in Chapter 3.

The effect on quality of sleep and social activities were rated as somewhat important by the patient group. However, these factors were considered to be correlated with other included attributes: sleep quality with pain at night, and social life with ability to perform daily activities. Therefore, sleep and social activities were not included.

The effect on work activities and anxiety level were excluded as they were rated as not very important by the patient group.

4.2.1.2 Safety risks and side effects

Side effects and risk of serious adverse events were both considered to be important by the patient group, however this was variable across the respondents.

Increased risk of cardiovascular events is known to be associated with NSAID use and was seen as important by respondents [262]. A composite outcome of cardiovascular events was not used as the prevalence and perceived importance of the risk may differ across cardiovascular events, including stroke, heart attack and coronary heart disease. A prior discrete choice experiment found that the importance of risk of heart attack and risk of stroke were very similar [214]. The risk of heart attack is known to vary between different NSAIDs and doses [263]. Therefore, risk of heart attack (also known as myocardial infarction) was included as an attribute.

Although patient input suggested that stomach ulcers were less important than cardiovascular events, gastrointestinal toxicity is considered to be an important risk by clinicians when deciding whether to treat using NSAIDs. Gastrointestinal events have been described as "the most serious health hazards of NSAIDs" [264, 265, 266]. To ensure clinical relevance, the risk of peptic ulcer bleeding was included as an attribute.

For risk of serious events, the ratings of importance were similar across individual serious events. Renal and hepatic problems were not included as these are less prevalent among osteoarthritis populations and NSAID users than cardiovascular events and are also measured less frequently in osteoarthritis trials [267].

Common side effects were rated as slightly less important than risk of serious adverse events. Side effects (e.g., dyspepsia) are more prevalent in NSAID users than more serious events (e.g., heart failure). However, more minor side effects may be less important to people with osteoarthritis as they are mild and reversible after stopping treatment [268]. Therefore, as only a limited number of attributes can be included

in each choice profile, side effects, other than risk of heart attack and risk of stomach ulcer bleeding, were assumed to be constant across all choices.

4.2.1.3 Treatment administration and cost

Duration of symptom relief was included as it was rated by the patient group as very important. Speed of symptom relief was considered somewhat important, but this was likely correlated with duration of symptom relief. Therefore, speed of symptom relief was assumed to be constant across all choices.

Cost (to the individual and the NHS), frequency of treatment administration and requirement for a prescription were excluded because these factors were ranked of low importance by the patient group.

4.2.1.4 Included attributes

The final six attributes included in the discrete choice experiment were:

1. Pain,
2. Stiffness,
3. Difficulty with daily activities (function),
4. Duration of symptom relief,
5. Risk of heart attack, and
6. Risk of stomach ulcer bleed.

These attributes included the key benefits and risks associated with NSAID treatments, which have been shown to vary between NSAIDs and dose levels used [1, 263, 269, 270].

4.2.2 Level selection

The attributes and levels are presented in Figure 4.1 and Table 4.1.

Figure 4.1: Attributes included in the discrete choice experiment (adapted from van Walsem 2015) [1].

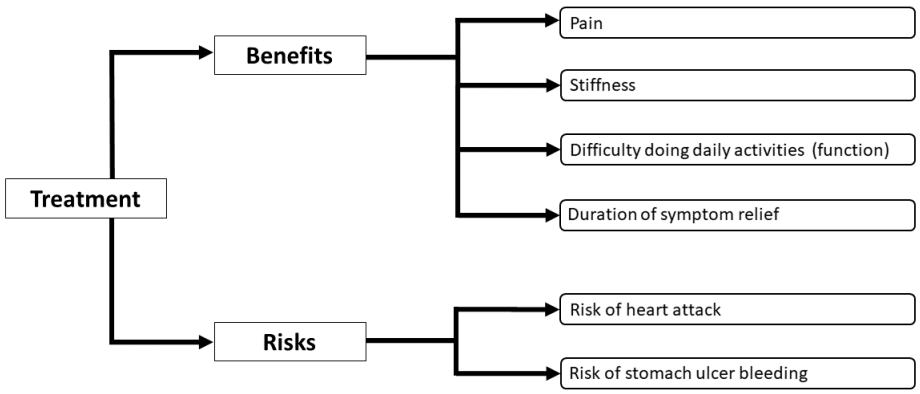


Figure reproduced from the author-accepted manuscript Copsey *et al.* (2019) [271], published with permission.

Table 4.1: Summary of attributes and levels

Attribute	**Levels**
Pain:	None (0 out of 100mm) Mild (25 out of 100mm) Moderate (50 out of 100mm) Severe (75 out of 100mm)
Difficulty with daily activities (function):	None (0 out of 100mm) Mild (25 out of 100mm) Moderate (50 out of 100mm) Severe (75 out of 100mm)
Stiffness:	None (0 out of 100mm) Mild (25 out of 100mm) Moderate (50 out of 100mm) Severe (75 out of 100mm)
Length of symptom relief:	1 month 3 months 6 months 12 months
Risk of heart attack in the next year:	0% (0 out of 100 people) 0.5% (0.5 out of 100 people i.e. 1 out of 200 people) 1% (1 out of 100 people) 2% (2 out of 100 people)
Risk of stomach bleed in the next year:	0% (0 out of 100 people) 1% (1 out of 100 people) 2% (2 out of 100 people) 4% (4 out of 100 people)

4.2.2.1 Treatment effectiveness

The three attributes related to treatment effect corresponded to the three domains in the WOMAC scale: pain, stiffness and function. For pain, stiffness and function, the levels used were none, mild, moderate or severe. These levels matched the set of responses in the Likert scale version of the WOMAC. The corresponding value on a 0-100mm visual analogue scale was also shown with additional graphical representation to demonstrate that the differences between adjacent levels were equivalent. The 'extreme' level was not used as it was assumed that the hypothetical person in the scenario had 'severe' ratings at baseline and the treatment should not result in worsening on any of these domains. In addition, participants rated as 'extreme' would likely require surgical treatment.

4.2.2.2 Risk of serious adverse events

The risk of a heart attack in 1 year is estimated to be around 0.2% for the general (non-NSAID user) population, translating to 2 cases per 1000 person years [263, 268]. The existing literature suggests that the risk of heart attack among NSAID users ranges from 1 to 13 cases per 1000 person years (or 0.1-1.3% over 1 year), with most studies reporting a 0.4% risk over 1 year [265, 268, 272, 273, 274]. The maximum risk level of heart attack was thus set at 2% (or 2 cases per 100 person years), with intermediate levels set at 0.5% and 1%.

Studies have suggested that the risk of gastrointestinal complications among NSAID users is between 0.5-4% during a one-year period [266, 273, 275, 276, 277, 278, 279, 280]. The risks vary depending on the NSAID used, dose and drug exposure time and the use of preventative treatments (e.g., misoprostol or proton pump inhibitors), as well as characteristics of the person, such as age and history of gastrointestinal events

[275, 278, 279, 280, 281]. To cover the range of potential clinically relevant values, the risk levels included were an absolute risk of 0%, 1%, 2% and 4%. Absolute risk levels were used instead of increased risk compared with the general population to make interpretation less cognitively demanding for participants.

4.2.2.3 Duration of symptom relief

This attribute captured the improvement in the person's outcomes gained from taking a course of medication or continuously taking the medication, rather than the duration of symptom relief gained from a single dose. The levels of duration for this attribute reflected commonly assessed time points in studies of NSAIDs for osteoarthritis [269, 282, 283, 284]. As he majority of trials included a 3-month and 6-month outcome assessment, these were included as intermediate levels. Levels of duration of 1 month and 12 months were also used to ensure the results would be applicable to the length of follow-up assessment in the majority of randomised trials of NSAID treatments.

4.2.3 Experimental design

Each choice task involved comparing two medications. An example choice task is shown in Figure 4.2. An example including the text presented at the start of the choice task is presented in Appendix C.2. An opt-out option, for example, to be able to choose to take neither medication, would have allowed estimation of uptake of different medications among the osteoarthritis population. However, an opt-out option was not included because this would increase the cognitive demand of the task and medication uptake was not being considered. Unlabelled profiles (where medication names are not given) were used to prevent the participant from introducing prior knowledge or previous experiences into the choice task, as was found by Rochon *et al.* [285].

Figure 4.2: Example choice task

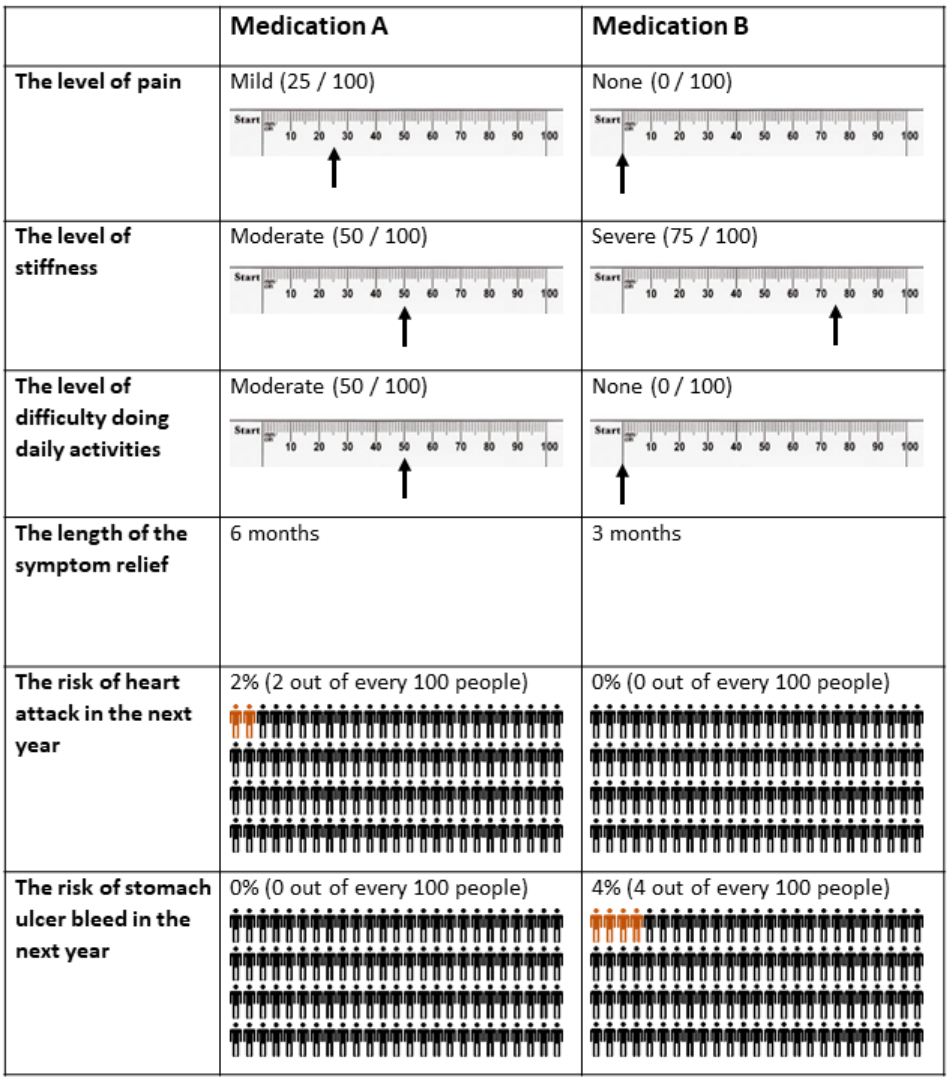


Figure reproduced from the author-accepted manuscript Copsey *et al.* (2019) [271], published with permission.

Detailed descriptions of the attributes and levels and an example question were presented before the choice tasks to improve respondent comprehension. The practice choice task included a dominated alternative, where one medication was the same or better than the other medication for all attributes. The hypothetical scenario that the participant should imagine was repeated at the beginning of each choice task as a reminder and to re-iterate that participants should imagine being in this situation, rather than considering their current condition.

The questionnaire was piloted using a sample of patient representatives to ensure that the language used was clear and lay-person friendly, as well as checking that the questionnaire was not too cognitively burdensome to complete.

Including all possible combinations of different levels for all attributes (a full factorial design) is often infeasible because it requires an infeasibly large number of choice tasks. Therefore, a fractional factorial design was used, where participants completed a subset of all possible pairs of medication options. The number of choice tasks to be completed per participant was limited to 17 (16 tasks with 1 duplicate to check consistency). All participants completed the same set of choice tasks. Completion of 16 choice tasks has been found to be feasible, including for older populations [236, 260, 286].

The 16 choice tasks were selected to produce an experimental design with a high level of d-efficiency. A design with a high level of d-efficiency minimises the level of imprecision around the estimates of the model parameters [287]. Optimising on d-efficiency allows the selection of the smallest possible design (reducing the number of choice tasks) that allows precise estimation of all required parameters. The final design of 16 choice tasks was able to achieve high levels of d-efficiency, whilst limiting cognitive burden and avoiding errors from respondent fatigue or boredom [288, 289].

The experiment used a fractional factorial design consisting of 16 choice tasks on 6 attributes with 4 levels. The experimental design was developed using Ngene software to produce a maximally efficient design [290]. Treatment effect attributes were included as effects-coded variables. Risk attributes were entered as continuous variables, assuming linearity based on evidence from prior discrete choice experiments [291].

Fixed prior values of zero were used to generate the experimental design for the pilot questionnaire. For the main design, coefficients from analysis of the pilot data were used as priors. A model-averaging approach was used, based on (i) a model where all coefficients were used as the prior values and (ii) a model where zero priors were used for the function and duration attributes as the coefficients followed a logical direction.

The same process was used to select a final experimental design for both the pilot and main survey. From the 10 most efficient designs, the designs containing dominant choice tasks (tasks where one choice was equivalent or superior to the alternative choice for all attributes), without balanced levels (the levels of an attribute were presented an unequal number of times) and level overlap (attributes having the same level in both alternatives in a single choice task) were excluded. Of the remaining designs, the design with the highest d-efficiency was selected. The results from the pilot survey that were used to determine the priors were not included in the main analysis.

Ngene code is presented in Appendix C.4 and the experimental designs for the pilot and main questionnaires are presented in Appendix C.5.

4.2.4 Data collection

The main survey sample was recruited from a chronic pain panel via a market research company, ResearchNow. Participants were eligible to complete this survey if they were UK residents at least 18 years of age, had a self-reported diagnosis of hip and/or knee osteoarthritis and provided informed consent.

The survey was delivered in an online format using LimeSurvey software [292]. Participants had previously consented to be part of a market research panel and be contacted by ResearchNow regarding surveys. Participants received an invitation from ResearchNow with an online link to the survey. Each participant received points from ResearchNow for completing the survey that could be redeemed for rewards, such as online shopping vouchers.

Participants provided additional demographic information on age, gender, work status and co-morbidities. Information on their condition was also assessed, including disease duration and prior treatments. The WOMAC outcome measure was used to assess disease severity using the pain, stiffness and function subscales (Likert scale version 3.1) [261]. Additional items assessed the participant's mood, level of optimism and risk-taking attitude.

Participants were also asked to rate the overall difficulty of completing the choice tasks to enable assessment of the cognitive demand of comparing the profiles.

Ethical approval was obtained from the University of Oxford Central University Research Ethics Committee [reference R55785/RE002].

4.2.5 Statistical analysis

The target sample size was 300 participants. Historically, formulas to calculate the sample size of a discrete choice experiment have been rules of thumb based on the number of attribute levels. Reviews have found that around 40% of discrete choice experiments recruited 100-300 participants [293, 294]. Orme recommends a sample size of 300 as a rule of thumb for main effects analyses [295]. Subgroup analysis was considered exploratory.

Descriptive statistics were used to summarise the sample of participants, including age, gender, prior treatment and disease severity. Continuous variables were summarised using the mean and standard deviation. Categorical variables were described using the frequency and proportion of responses.

The responses to the choice questions were assessed using three methods:

1. Consistency check: A duplicate choice task was presented to assess whether participants selected the same response when presented with the same choice task later in the survey.

2. Rationality check: A choice task with a dominated alternative was included, such that one medication option was at least as good or 'better' than the other option on all attributes.

3. Dominance check: Choices made by each respondent were compared to alternative choice profiles that assumed that respondents focused only on a single attribute. The similarity between the respondents' choices and the alternative choice profiles was examined for each of the individual attributes to explore whether participants were considering all of the presented attributes. Dominance was considered to be indicated if participants chose the medication with the 'better' level for the single attribute in at least 14 of the 16 choice tasks.

Effects-coding was used for the attributes of pain, function, stiffness and duration of effect. These attributes were not expected to have a linear effect [214]. In effects coding, the coefficient for the reference category (severe or 12 months) is calculated by negating the sum of the coefficients for the other levels. The coefficient for the reference term is thus not encompassed in the intercept term (unlike conventional dummy coding), allowing the intercept in this study to be meaningfully interpreted because it only represented the preference of choosing the medication option shown on the left over the option shown on the right, regardless of the levels of the attributes [296]. Effects-coding also allowed estimation of interaction effects between attributes as the interaction was based on the 'central level' for one variable, as opposed to at the reference level.

There were no missing data because it was mandatory for participants to complete all items of the online survey before submitting their responses.

4.2.5.1 Regression analysis

Two forms of logistic regression were used to model the responses to the choice tasks because the choice response was a dichotomous variable: each medication was either selected or not.

Attributes of pain, stiffness, function and duration of treatment effect were included as effects-coded variables. The base levels were 'severe' for the WOMAC domain attributes and '1 month' for duration. Risk attributes for heart attack and stomach ulcer bleeding were included as continuous variables, assuming a linear form.

The magnitude, direction and level of significance of the attribute-level coefficients are reported. For each attribute, statistical significance was defined as a two-sided p-value below 0.05. The relative importance of the attributes is presented graphically, showing the effects-coded coefficients with the corresponding 95% confidence interval.

Conditional logistic regression analysis

Conditional logistic regression was used because it allowed the model to account for the 'matching' within each choice task, in that each medication A profile was compared to a single medication B profile. It conditioned on the fact that either medication A or medication B must be selected; participants could not select neither or both medications. A positive coefficient for an attribute in this regression represented an improvement in utility, with a higher coefficient indicating a greater probability of choosing the medication. This did not account for the repeated-task nature of the experiment, where the same participant responded to multiple choice tasks. The main analysis used a conditional logistic regression model using the clogit command in Stata [297].

Alternative-specific conditional regression was not used because no variables were case-specific and alternatives were unlabelled. However, the alternative-specific conditional regression following McFadden's choice model produced very similar results to the standard conditional regression [298].

Mixed effects regression analysis

As a secondary analysis, a mixed effects model was used to account for preference heterogeneity. The simulation used 400 random Halton draws. Mixed logit models must be evaluated numerically as they do not have a closed form. Halton sequences are superior to random draws as they reduce the run time and reduce simulation error in mixed logit analyses [299].

Initially, all variables were included as random effects and those with non-significant standard deviations were included as fixed effects variables in a stepwise procedure. In the mixed effects analysis, the terms where the standard deviation of the coeffi-

cient was statistically significant were included as random effects in the final model because a significant standard deviation is indicative of preference heterogeneity. In the final model, fixed effects variables were stiffness (none and moderate), pain (moderate), function (mild and moderate), and duration (3 and 6 months). Random effects variables were pain (none and mild), duration (12 months), stiffness (mild), risk of heart attack and risk of stomach ulcer bleeding. All random-effects coefficients were assumed to be normally distributed. In the final mixed effects model, random coefficients were specified to be potentially correlated. Density functions for the random effects coefficients were produced using the conditional expectation of the coefficient for each respondent with 500 random Halton draws [300].

Interpreting the model coefficients

The formula for the probability of choosing medication A is:

$$\frac{exp(\sum \beta_{iA})}{exp(\sum_i \beta_{iA}) + exp(\sum_i \beta_{iB})} \tag{4.1}$$

where β_{iA} is the coefficient in the model for the level of medication A for the i^{th} attribute.

Focusing on the difference in the sums of the coefficients, Equation 4.1 can be re-arranged to:

$$\frac{exp(\sum \beta_{iA})}{exp(\sum_i \beta_{iA}) + exp(\sum_i \beta_{iB})} = \frac{1}{exp(\sum_i \beta_{iA} - \beta_{iB}) + 1} \tag{4.2}$$

An example of calculating the probability of choosing a medication is presented in Appendix C.6.

Marginal rates of substitution and the willingness-to-risk were calculated to indicate the trade-off between different attributes.

The marginal rate of substitution (MRS) between attributes x and y is:

$$MRS_{x,y} = \frac{coeff(x)}{coeff(y)} \tag{4.3}$$

When the MRS is equal to 1, the two attributes have the same coefficient and therefore have an equal effect on medication choice. When the MRS is equal to 2, the coefficient for attribute x is twice the coefficient for attribute y and attribute x has twice the effect on medication choice. If the attributes are measured on the same scale, have an MRS of 2 and medication A is higher than medication B in attribute x by value V, medication B will be:

- preferred to medication A if medication B is higher in attribute y by more than $2V$,
- equivalent to medication A if medication B is higher in attribute y by exactly $2V$,
- not preferred to medication A if medication B is higher in attribute y by less than $2V$.

The risk attributes were entered as continuous parameters and thus the coefficients could be interpreted in terms of the level of risk participants would be willing to accept. The rate at which respondents were willing to trade off between the attributes was calculated relative to the risk attributes. For example, the increase in risk of heart attack the respondent was willing to trade off to reduce their pain level from severe to none, rather than severe to mild, was calculated.

The willingness-to-risk is:

$$WTR_{i-j} = \frac{coeff(x_i) - coeff(x_j)}{coeff(risk)} \tag{4.4}$$

where $coeff(x_i)$ is the coefficient for the i^{th} level of attribute x. If medication A has level i for attribute x and medication B has level j for attribute x, then:

- medication A will be preferred if the risk for medication B is higher by more than WTR,
- medications A and B will be equivalent if the risk for medication B is higher by exactly WTR,
- medication B will be preferred if the risk for medication B is higher by less than WTR.

4.2.5.2 Sensitivity analyses

Sensitivity analyses were performed to test the robustness of the results to different model assumptions [301, 302, 303]. Sensitivity analyses were conducted excluding:

- Respondents who failed the rationality check (selected a dominated alternative),
- Respondents who failed the consistency check (selected different medications when presented with the same choice task on different occasions),
- Respondents who found it difficult to choose between medications during the choice tasks (selecting 'quite difficult' or 'very difficult'), and
- Respondents who found it difficult to imagine a hypothetical scenario when completing the choice tasks (selecting 'quite difficult' or 'very difficult').

4.2.5.3 Exploratory and subgroup analyses

Subgroup analyses were performed using interaction terms to explore differences in the results due to age, gender, baseline disease severity or previous joint replacement surgery. However, these additional analyses were considered exploratory due to the anticipated low level of precision. Interaction terms were included between baseline covariates and all attributes and levels for age, baseline score and gender.

Additional exploratory subgroup analyses for specific attributes were:

- Prior stomach bleed combined with risk of stomach bleed
- Prior heart attack combined with risk of heart attack
- Risk-taking attitude combined with risk of stomach bleed and risk of heart attack
- Time since diagnosis combined with duration

Age and the total WOMAC score at baseline were assumed to be continuous linear terms in the interactions. Prior stomach bleed and prior heart attack were included as dummy variables (classified as having the prior event if the respondent replied 'yes' and not having the event if they replied 'no' or 'don't know'). Risk-taking attitude was included as two dummy variables, one for a response of 'risk loving' and one for a response of 'risk averse' (the base case response was 'risk neutral' or 'don't know'). Time since diagnosis was included as two dummy variables, one for 3-5 years and one for 6 or more years since diagnosis (the base case response was less than 3 years since diagnosis or 'don't know').

Interactions for each covariate were tested in separate models. In the mixed effects model, interaction terms were initially assumed to be random effects and replaced as fixed effects terms if the standard deviation for the coefficient was non-significant.

Following this, significant interaction terms were combined in a single model and non-significant interactions were sequentially removed.

There were insufficient degrees of freedom to fully test interactions between the duration and treatment effect attributes. Exploratory analyses were conducted excluding non-significant attributes and including interaction terms between pain and duration.

All analyses of interaction effects should be interpreted with caution, due to potential design inefficiency and lack of power.

4.3 Results

4.3.1 Respondent characteristics

Of the 547 potential participants who clicked the link to enter the survey, 342 were eligible and consented to participate. Of the 342 eligible and consenting responders, 300 participants completed the survey and were included in the analysis (88%, n=300/342).

The demographics of the 300 respondents are described in Table 4.2. The average age of respondents was 60 years old. Most participants had osteoarthritis in their knees, hips and hands and had been treated previously with exercise, physiotherapy and medication treatments. Around a quarter of participants had undergone joint replacement surgery (n=83, 28%). The majority of participants had been diagnosed with osteoarthritis for at least 3 years and also suffered from high blood pressure. Respondents were evenly balanced in terms of their sex (male/female) and work status (working/retired). The majority of participants had moderate levels of symptoms on the WOMAC pain, function and stiffness subscales and had good mental health status, with high levels of optimism generally.

Table 4.2: Demographics of the discrete choice experiment respondents

	Mean (SD) or n	**Range or %**
Age (years)	60.5 (SD 13.3)	23-92
Sex:		
Male	136	45.3%
Female	164	54.7%
Work status:		
Working full-time	76	25.3%
Working part-time	38	12.7%
Homemaker / caregiver	14	4.7%
Retired	141	47.0%
Unemployed	17	5.7%
Other	14	4.7%
Osteoarthritis sites:		
Shoulder	111	37.0%
Elbow	65	21.7%
Hip	186	62.0%
Hand or wrist	158	52.7%
Knee	241	80.3%
Foot or ankle	97	32.3%
Number of osteoarthritis sites:		
Single site	83	27.7%
Multiple sites	217	72.3%
Prior treatment:		
Medication (tablet)	229	76.3%
Medication (cream or gel)	156	52.0%
Injection	96	32.0%
Physiotherapy	155	51.7%
Exercise	152	50.7%
Joint replacement surgery	83	27.7%
Prior injury of knee or hip before osteoarthritis	95	31.7%
Medications:		
Warfarin	53	17.7%
Glucocorticoid	87	29.0%

Table 4.2: Demographics

	Mean (SD) or n	Range or %
General health:		
Excellent	24	8.0%
Very good	59	19.7%
Good	97	32.3%
Fair	74	24.7%
Poor	46	15.3%
Time since osteoarthritis diagnosis:		
Less than a year	14	4.7%
1 - 2 years	39	13.0%
3 - 5 years	79	26.3%
6 - 10 years	70	23.3%
Over 10 years	93	31.0%
Don't know	5	1.7%
WOMAC scores:		
WOMAC pain	8.1 (SD 4.7)	0-20
WOMAC stiffness	3.5 (SD 1.9)	0-8
WOMAC function	27.4 (SD 16.6)	0-68
WOMAC total	39.0 (SD 22.3)	0-96
Co-morbidities:		
Stomach bleed	42	14.0%
Peptic ulcer disease	45	15.0%
Stroke	43	14.3%
Heart attack	40	13.3%
High blood pressure	162	54.0%
Other heart problems	66	22.0%
Time during the past four weeks felt calm and peaceful:		
None of the time	21	7.0%
A little of the time	73	24.3%
Some of the time	72	24.0%
Good bit of the time	50	16.7%
Most of the time	66	22.0%
All of the time	18	6.0%

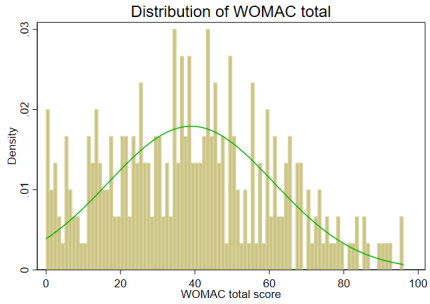
Table 4.2: Demographics

	Mean (SD) or n	Range or %
Time during the past four weeks felt downhearted and blue:		
None of the time	60	20.0%
A little of the time	79	26.3%
Some of the time	59	19.7%
Good bit of the time	43	14.3%
Most of the time	33	11.0%
All of the time	26	8.7%
Level of optimism:		
Very optimistic	64	21.3%
Quite optimistic	121	40.3%
Neither optimistic, nor pessimistic	66	22.0%
Quite pessimistic	25	8.3%
Very pessimistic	17	5.7%
Don't know	7	2.3%
Risk-taking attitude:		
Risk loving	42	14.0%
Risk neutral	109	36.3%
Risk averse	146	48.7%
Don't know	3	1.0%

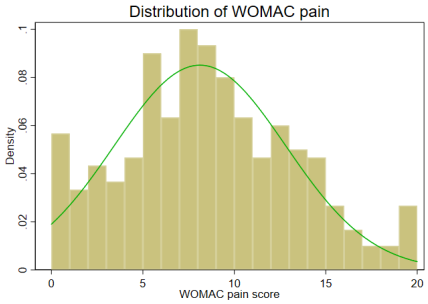
SD: Standard deviation.

The continuous demographic variables were approximately normally distributed. The histograms for the distribution of the WOMAC subscales, WOMAC total score and age are presented in Figure 4.3.

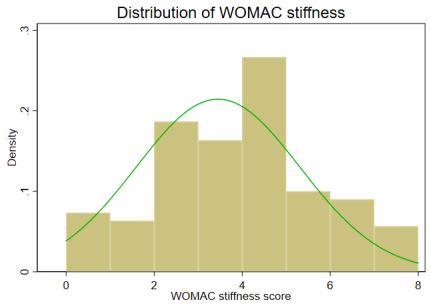
Figure 4.3: Histograms of WOMAC scores and age



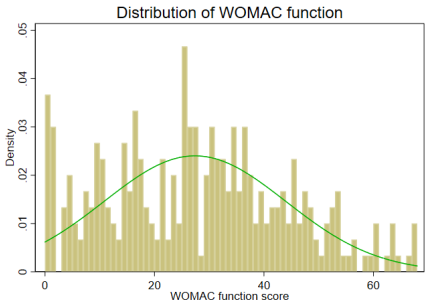
(a) WOMAC total score



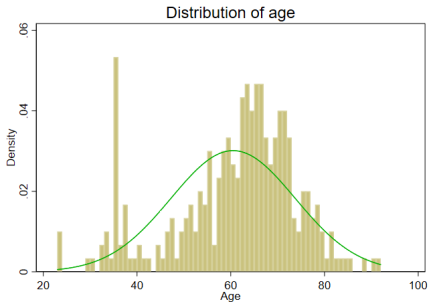
(b) WOMAC pain subscale



(c) WOMAC stiffness subscale



(d) WOMAC function subscale



(e) Age

4.3.2 Assessments of difficulty, rationality and consistency

Most participants found it 'quite easy' or 'quite difficult' to choose between the two medications presented in the choice tasks (Table 4.3). Very few respondents found it 'very difficult' to choose between the two medications (n=10/300, 3%). When performing the choice tasks, the majority of participants reported that it was easy to consider an imaginary patient scenario, rather than their own situation.

Table 4.3: Difficulty performing choice tasks

	n	%
Level of difficulty choosing between two medications:		
Very easy	42	14.0%
Quite easy	98	32.7%
Neither easy, nor difficult	50	16.7%
Quite difficult	100	33.3%
Very difficult	10	3.3%
Level of difficulty considering an imaginary patient scenario:		
Very easy	69	23.0%
Quite easy	94	31.3%
Neither easy, nor difficult	66	22.0%
Quite difficult	59	19.7%
Very difficult	12	4.0%

When presented with a duplicate task later in the survey (tasks 4 and 17 were identical), around a fifth of respondents selected a different medication (n=55/300, 18%). A small proportion of respondents failed the rationality check by selecting a dominated alternative during a practice choice task (n=20/300, 7%). Participants who failed the rationality check were more likely to fail the consistency check (n=9/20, 45% compared to n=46/280, 16%). There was no association between the difficulty ratings and failing the rationality or consistency check.

Most participants spent 10-20 minutes on the survey, with 10-40 seconds spent on each choice task. The median time to complete a single choice task was 26 seconds (IQR 17 to 35), ranging from 3 seconds to 5 minutes. The median time to complete the full survey was 18 minutes (IQR 14 to 26), ranging from 3 minutes to 2.75 hours. Of the 300 participants, 250 spent less than 30 minutes on the survey, 35 spent 30-45 minutes and only 15 took more than 45 minutes to complete the survey. Participants who took longer than 45 minutes usually had a large gap between two questions, indicating that they took a break from the survey. For example, one participant finished the survey 1 hour and 40 minutes after starting but spent 1 hour and 10 minutes on a single choice task. In this case, it is likely that the participant was not working on the survey during this period.

4.3.3 Results of the conditional logistic regression model

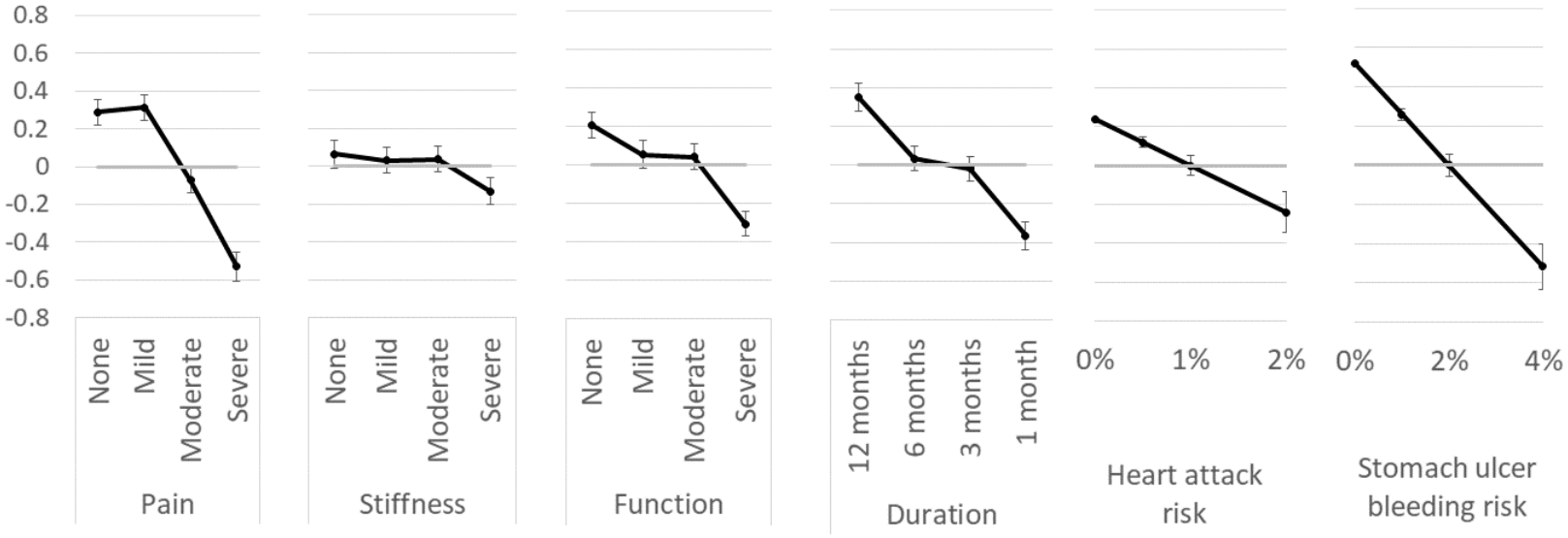
The coefficients of the conditional logistic regression model demonstrated the importance of the different attributes and levels. As mentioned in Section 4.2.5.1, a positive coefficient for an attribute in this regression represented an improvement in utility, with a higher coefficient indicating a greater probability of choosing the medication. The coefficients for this model are shown in Table 4.4 and presented graphically in Figure 4.4.

Table 4.4: Coefficients from fixed effects conditional logit model

		Coefficient	**95% LCI**	**95% UCI***
Pain	None	0.28865	0.22186	0.35545
	Mild	0.31266	0.24365	0.38167
	Moderate	-0.07161	-0.13938	-0.00383
	Severe	-0.52971	-0.60751	-0.45190
Stiffness	None	0.06354	-0.00895	0.13603
	Mild	0.03014	-0.03643	0.09670
	Moderate	0.03809	-0.03091	0.10710
	Severe	-0.13177	-0.20066	-0.06288
Function	None	0.20856	0.14317	0.27393
	Mild	0.05509	-0.01518	0.125362
	Moderate	0.04310	-0.02426	0.11045
	Severe	-0.30675	-0.37189	-0.24160
Duration	12 months	0.35157	0.27998	0.42316
	6 months	0.03287	-0.03123	0.09698
	3 months	-0.01916	-0.08358	0.04527
	1 month	-0.36529	-0.43880	-0.29178
Heart attack risk	0%	0	-	-
	0.5%	-0.11942	-0.14547	-0.09337
	1%	-0.23884	-0.29094	-0.18675
	2%	-0.47769	-0.58188	-0.37350
Stomach bleed risk	0%	0	-	-
	1%	-0.25894	-0.28792	-0.22996
	2%	-0.51788	-0.57584	-0.45992
	4%	-1.03576	-1.15168	-0.91983

* LCI: Lower confidence interval, UCI: Upper confidence interval.

Figure 4.4: Coefficients for fixed effects model



The coefficients displayed a logical direction for most between-level differences, with lower risk, fewer severe symptoms and longer duration of effect being seen as more favourable. Illogical directions in the coefficients were seen between 'none' and 'mild' levels for pain and 'mild' and 'moderate' levels for stiffness. However, the difference in the coefficients between adjacent levels was small in both cases.

The overall 'height' of each attribute in Figure 4.4 indicates that stiffness was not considered important. However, the results suggested that the levels of all other attributes were associated with medication choice. Pain symptoms, duration of overall effect and risk of stomach ulcer bleeding were seen as the most important attributes, with physical function and risk of heart attack seen as important to a lesser extent.

For duration, there were important differences between 6 months and 12 months, and between 1 month and 3 months. However, there was a plateau between the two intermediate levels, suggesting that participants did not differentiate between effects lasting 3 and 6 months. There was a similar pattern for function, where the importance of 'mild' and 'moderate' levels were not significantly different.

For pain, there were large differences in importance between mild, moderate and severe levels. However, 'no pain' and 'mild pain' had similar levels of importance.

A marginal rate of substitution close to 1 suggested that participants would trade-off a 1% increase in the risk of heart attack for a 1% reduction in the risk of stomach ulcer bleeding.

The difference between the coefficients for duration for 12 months and 1 month was 0.72 (0.35 to -0.37, Table 4.4). The marginal rate of substitution relative to risk of heart attack was 3 (0.72 / 0.24, Appendix C.7). Participants would thus trade-off an increase in the risk of heart attack of 3% for an increase in the duration of treatment effect from 1 month to 12 months. The values were similar relative to the risk of stomach ulcer bleeding.

None of the pre-planned sensitivity analyses produced any changes in the predicted medication choice for the different tasks or any large differences in the magnitude of the coefficients. The model was able to predict the actual medication chosen for 64% of the choice tasks (n=3059/4800). For all tasks except 1 (task 15 of 16), the medication predicted by the model was the medication most commonly chosen by the respondents (Table 4.5). The area under the curve (AUC) was calculated to be 0.68 (95% CI 0.67 to 0.70).

Table 4.5: Predictive ability of fixed effects model by choice task

Choice task	**Predicted medication**	**Probability of selecting predicted medication (from model)**	**Proportion selecting predicted medication (in sample)**
1	Medication A	65.3%	68.3%
2	Medication A	56.7%	62.0%
3	Medication B	87.0%	84.3%
4	Medication B	62.1%	57.7%
5	Medication A	82.7%	85.3%
6	Medication A	61.4%	66.0%
7	Medication B	70.2%	66.7%
8	Medication B	60.0%	55.7%
9	Medication A	65.4%	69.3%
10	Medication B	55.6%	50.3%
11	Medication A	56.1%	61.0%
12	Medication A	62.3%	65.7%
13	Medication B	64.3%	60.7%
14	Medication B	55.6%	52.3%
15	Medication B	52.7%	49.3%
16	Medication A	62.0%	65.0%

The effect on the probability of choosing a medication of a one-level improvement in each attribute is shown in Appendix C.8.

The number of participants who presented with dominant preferences (14 or more choices matching the dominant alternative for one attribute) is shown in Table 4.6. Very few participants selected based on only a single attribute. The attributes where several participants selected the dominant alternatives for the majority of the choice tasks were pain and risk of stomach ulcer bleeding.

Table 4.6: Participants exhibiting dominant preferences

Attribute	**Proportion selecting dominant alternative:**	
	In 14 or more tasks	**In all 16 tasks**
Pain	26 (8.7%)	4 (1.3%)
Stiffness	0 (0%)	0 (0%)
Function	7 (2.3%)	0 (0%)
Duration	5 (1.7%)	2 (0.7%)
Risk of heart attack	2 (0.7%)	1 (0.3%)
Risk of stomach ulcer bleeding	22 (7.3%)	2 (0.7%)

4.3.4 Results of mixed effects model

The fixed-effects coefficients and mean values for the random-effects coefficients indicated the same relative preferences between the attributes and levels. Although the probabilities of selecting each medication differed between the participants, the predicted medication choice did not vary between the fixed-effects and mixed-effects models. The similarity between the two models can be seen visually by comparing Figures 4.4 and 4.5. Using a mixed effects model, the AUC remained at 0.68 (95% CI 0.67 to 0.70), showing no improvement in the predictive ability of the fixed effects model.

However, the mixed effects model added information because it indicated the presence of preference heterogeneity for certain attributes and levels (Table 4.7). The standard deviation was significant for pain (none and mild), stiffness (mild), function (none), duration (12 months), risk of heart attack and risk of stomach ulcer bleeding. With the exception of mild stiffness, where the main effect was non-significant, all other levels where preference heterogeneity was present were the most extreme levels, with coefficients greater than 0.45. The standard deviation was larger than or close to the mean value for the coefficient, indicating high levels of heterogeneity. This suggests that heterogeneity in participant preferences was related to the relative importance of large improvements in symptoms, improvements lasting for a long period of time and associated risks of treatment.

Figure 4.5: Coefficients for mixed effects model

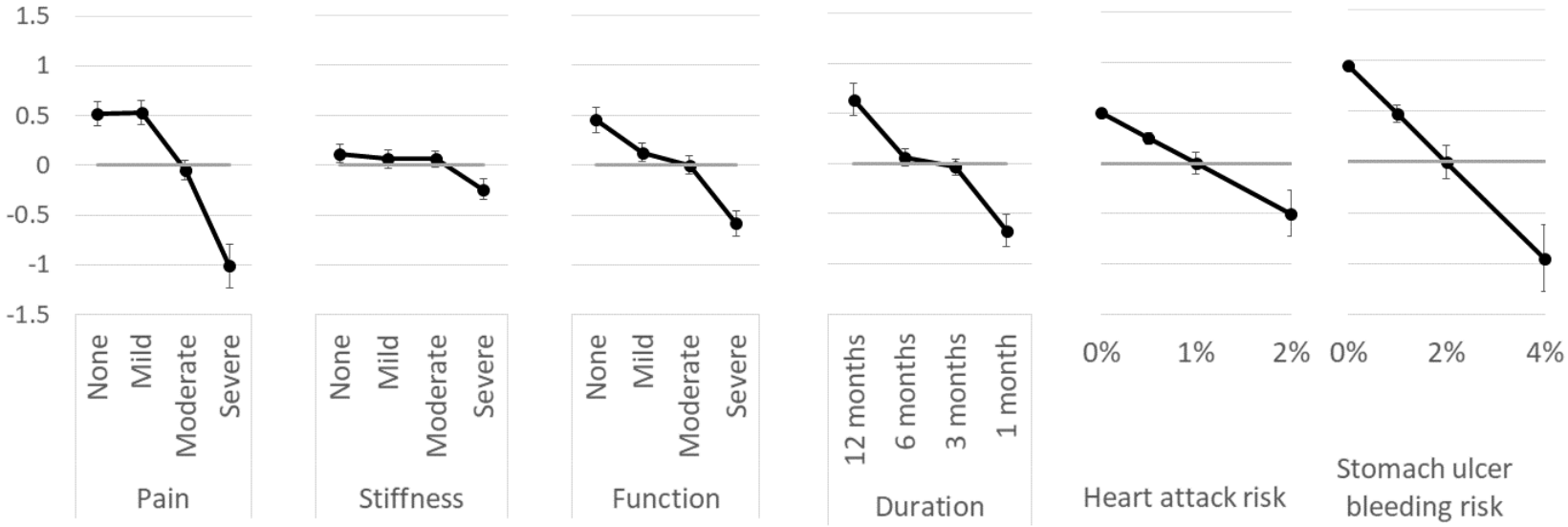


Figure reproduced from the author-accepted manuscript Copsey *et al.* (2019) [271], published with permission.

The results of the mixed effects model can be more easily interpreted by calculating the willingness-to-risk. Table 4.8 demonstrates the maximum increase in the risk of heart attack that a participant was willing to accept for a given increase in treatment effect. All other attributes being constant, medications A and B were seen as equivalent if medication A reduced pain to moderate and medication B reduced pain to mild but medication B had 1.2% higher risk of heart attack. As a second example, if a person's current medication reduced their stiffness to 'mild', then they would switch to a different medication that offered the same treatment effects on pain and function but would reduce their stiffness to 'none' only if the new medication had at most 0.1% higher risk of heart attack.

Comparing benefits and risks associated with the treatment, participants were willing to accept an increase in the risk of heart attack of 2.6% to increase the duration of the treatment effect from 1 month to 12 months. To reduce symptoms from 'severe' to 'none', respondents would be willing to accept an increased risk of heart attack of 2.1% for improvements in function, and 3.1% for improvements in pain symptoms (Table 4.8).

Table 4.7: Coefficients from mixed effects logit model

		Coefficient	95% LCI	95% UCI*
Mean:				
Pain	None	0.52020	0.39721	0.64319
	Mild	0.53673	0.41536	0.65811
	Moderate	-0.04781	-0.14337	0.04775
	Severe	-1.00912	-1.22513	-0.79311
Stiffness	None	0.11672	0.02148	0.21196
	Mild	0.06088	-0.03603	0.15779
	Moderate	0.06177	-0.02271	0.14625
	Severe	-0.23937	-0.34103	-0.13770
Function	None	0.45272	0.32141	0.58403
	Mild	0.12813	0.03410	0.22215
	Moderate	0.00092	-0.08692	0.08877
	Severe	-0.58177	-0.70805	-0.45549
Duration	12 months	0.63846	0.48101	0.79591
	6 months	0.06214	-0.01866	0.14294
	3 months	-0.03104	-0.11067	0.04858
	1 month	-0.66956	-0.82973	-0.50939
Risk of heart attack		-0.49789	-0.61133	-0.38444
Risk of stomach ulcer bleeding		-0.47407	-0.55524	-0.39289
SD:				
Pain	None	0.69390	0.55868	0.82912
	Mild	0.56137	0.41953	0.70322
Stiffness	Mild	0.32988	0.21653	0.44323
Function	None	0.69724	0.56001	0.83447
Duration	12 months	0.56718	0.41081	0.72355
Risk of heart attack		0.54769	0.42734	0.66804
Risk of stomach ulcer bleeding		0.47850	0.40304	0.55397

* LCI: Lower confidence interval, UCI: Upper confidence interval.

Table 4.8: Trade-offs between treatment effectiveness and risk of heart attack [#]

Improvement in treatment effect:		**Maximum risk increase willing to accept (%)**		
		Willingness to risk	LCI	UCI
Pain	Severe to moderate	1.9	1.4	2.5
	Moderate to mild	1.2	0.8	1.6
	Mild to none	0.0	-0.3	0.2
	Total: severe to none	**3.1**	**2.4**	**3.7**
Stiffness	Severe to moderate	0.6	0.3	0.9
	Moderate to mild	0.0	-0.3	0.3
	Mild to none	0.1	-0.2	0.4
	Total: severe to none	**0.7**	**0.4**	**1.0**
Function	Severe to moderate	1.2	0.8	1.6
	Moderate to mild	0.3	0.0	0.6
	Mild to none	0.7	0.3	1.0
	Total: severe to none	**2.1**	**1.6**	**2.6**
Duration	1 month to 3 months	1.3	0.9	1.7
	3 months to 6 months	0.2	-0.1	0.4
	6 months to 12 months	1.2	0.8	1.5
	Total: 1 month to 12 months	**2.6**	**2.0**	**3.2**

[*] LCI: Lower confidence interval, UCI: Upper confidence interval.
[#] Coefficients were translated into willingness to risk heart attack because willingness to risk stomach ulcer bleeding varied with disease severity.

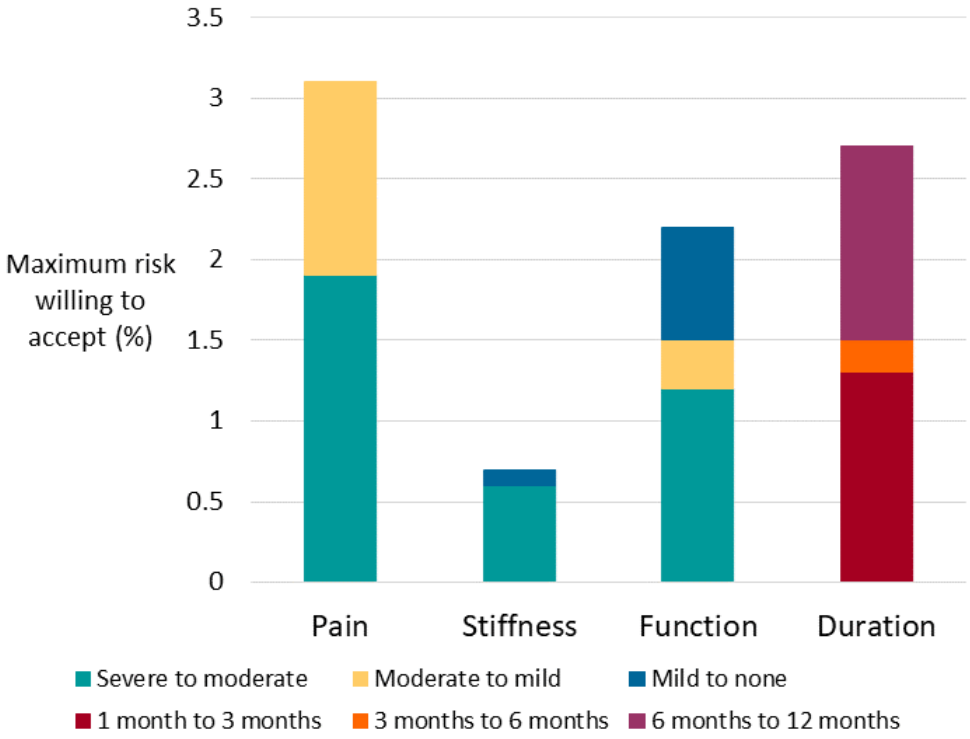
The overall importance of the treatment effect attributes can be demonstrated by assessing the trade-off in the risk of serious events that the respondents were willing to make to improve the treatment effect from the worst to the best levels (Figure 4.6). Pain was the most important attribute overall. Participants would be willing to increase the risk of heart attack by 3.1% to increase the medication's effectiveness in reducing their pain from severe to none. Participants were willing to increase the risk of heart attack by only 0.7% to increase the medication's effectiveness in reducing their stiffness from severe to none.

Assume that two medications are identical apart from their effect on stiffness and risk of heart attack. If medication A has no effect on stiffness (i.e., stiffness remains severe) and medication B reduces stiffness from 'severe' to 'none', then participants will prefer medication B over medication A if the risk of heart attack is at most 0.7% higher for medication B compared to medication A.

The trade-off in terms of the risk of heart attack was similar to the trade-off in terms of the risk of stomach ulcer bleeding because of the similarity in the coefficients of the two risk attributes.

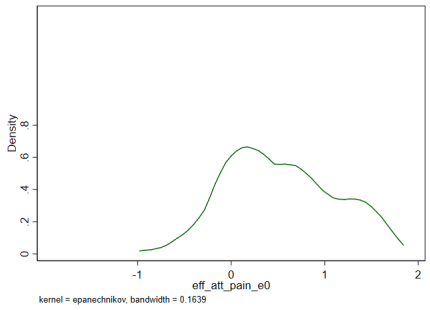
Although very few respondents exhibited dominant preferences, dominant preferences were more common for the pain and ulcer bleeding attributes (Table 4.6). This supports the finding that the effect on pain was highly important.

Figure 4.6: Importance of treatment effect and duration attributes: willingness to trade-off improvement from worst to best level in terms of risk of heart attack

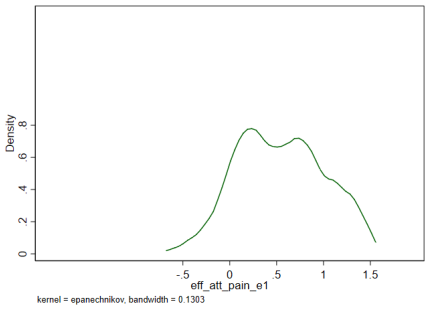


The density functions of the random effect coefficients are shown in Figure 4.7. The stiffness coefficient has approximately symmetric distribution, indicating that, while there was some heterogeneity, it was centred on the average coefficient (Figure 4.7(d)). The pain (mild) coefficient indicates heterogeneity in preferences, with two peaks in the distribution of the coefficients (Figure 4.7(b)). The pain (none) and function (none) coefficients have flatter distributions, suggesting some evidence of heterogeneity (Figures 4.7(a) and 4.7(c)). The risk attributes and duration coefficients demonstrate heterogeneity by the somewhat skewed distribution (Figures 4.7(e), 4.7(f) and 4.7(g)). This suggests that the majority of participants placed high importance on increasing duration of effect to 12 months with a minority group placing very low importance on this difference. Similarly, for the risk attributes, the majority of participants considered risks to be important, with a 1% change in risk having a coefficient of between 0 and -0.5. However, there were also many participants who considered risks to be extremely important, with coefficients up to 1.5.

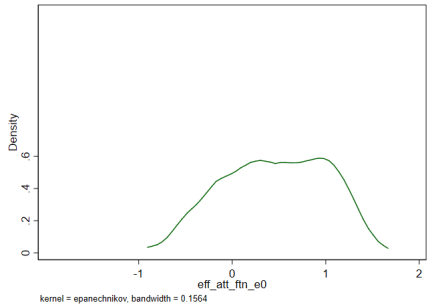
Figure 4.7: Density functions of random effects coefficients



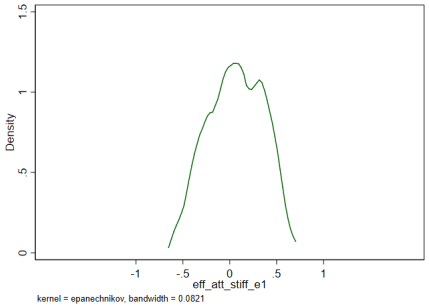
(a) Pain (none)



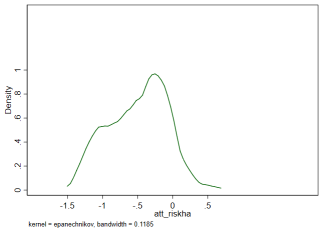
(b) Pain (mild)



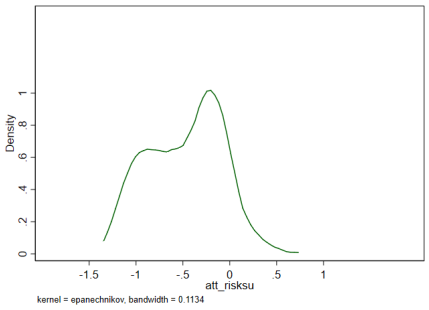
(c) Function (none)



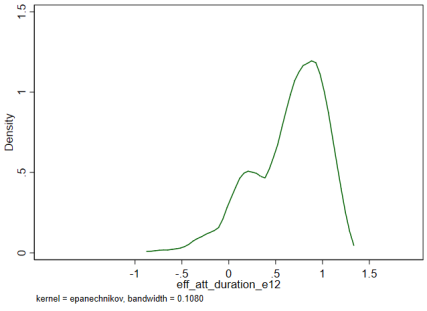
(d) Stiffness (mild)



(e) Risk of heart attack



(f) Risk of stomach ulcer bleeding



(g) Duration (12 months)

4.3.5 Exploratory subgroup analysis: Interactions with respondent characteristics

The results of the subgroup analysis presented below are based on including interaction terms in the mixed effects model. The results of including interaction terms in the conditional logistic regression model are presented in Appendix C.9.

Subgroup analysis of the WOMAC score using single interaction terms suggested that there were three potentially significant interactions: WOMAC score and (i) stomach bleed risk, (ii) heart attack risk and (iii) function (moderate). The only other significant interaction term identified in the single interaction models was prior joint replacement and risk of heart attack. When these terms were combined into a single model, the only remaining significant interaction term was WOMAC score and stomach bleed risk. The results of the final model including the significant interaction term are presented in Table 4.9. This model indicated that respondents with more severe disease placed less importance on the risk of stomach ulcer bleeding.

The model with interaction terms predicted different medication choices from the main model for 3% of choice tasks (n=153/4800). The discrepancies mostly affected respondents with very low or very high WOMAC scores and related almost exclusively to 2 of the 16 choice tasks (151 of the 153 tasks). Both choice tasks (10 and 15) had predicted probabilities close to 0.5 in the main mixed effects model, and there were large differences between the two medications in the risk of stomach ulcer bleeding (0% or 1% compared to 4% risk).

Figure 4.8: Trade-off between maximal improvement in attributes in terms of increased risk of heart attack: increasing importance of stomach bleed risk for respondents with less severe disease symptoms

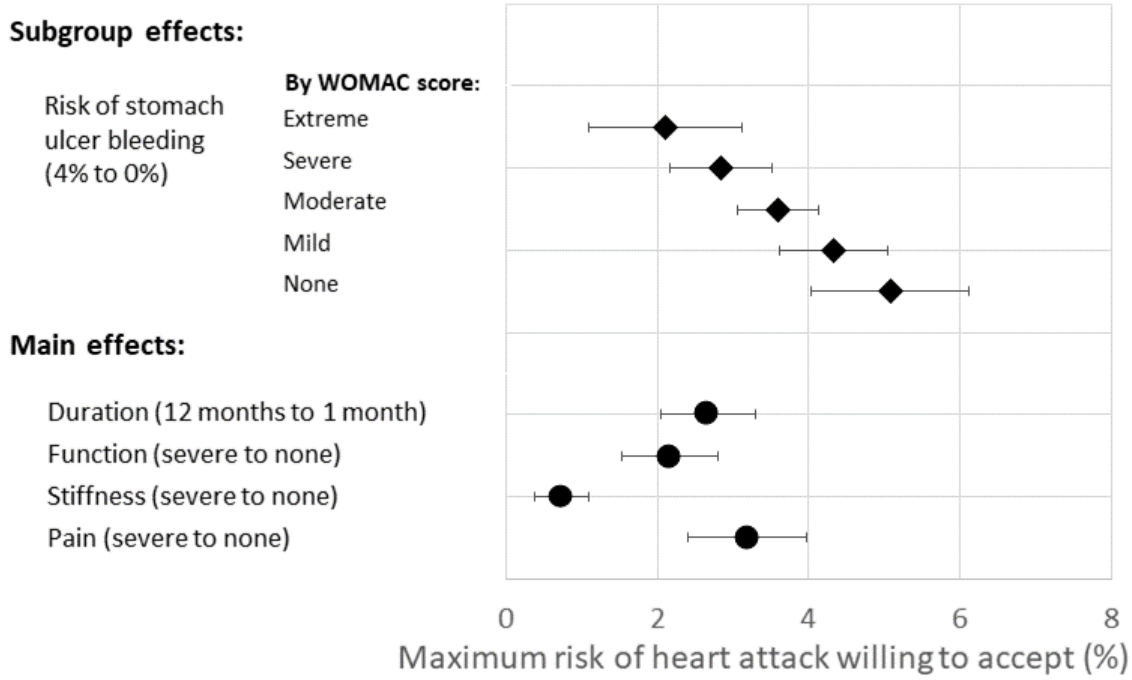


Figure 4.8 shows the difference in the importance of different attributes, relative to the risk of heart attack. To reduce the risk of stomach ulcer bleeding from 4% to 0%, respondents with an average WOMAC score (39 out of 96) would accept a 4% increase in the risk of heart attack, compared to 5% for the minimum WOMAC score of 0 out of 96 and 2% for the maximum WOMAC score of 96 out of 96. For participants with the minimum WOMAC score (0 out of 96), the risk of stomach ulcer bleeding was the most important attribute, whereas for participants with a moderate WOMAC score (48 out of 96), the risk of stomach ulcer bleeding was of similar importance to pain.

The difference in the importance of risk of stomach bleeding can be also shown in terms of participants' willingness-to-risk. A participant with a severe WOMAC score (72 out of 96) would be willing to change from medication A giving moderate pain to medication B that gives mild pain if the increase in the risk of heart attack is no more than 1.7% higher than for medication A. Alternatively, a participant with a mild WOMAC score (24 out of 96) would be willing to accept medication B if the increase in the risk of stomach ulcer bleeding is no more than 1.1% higher than for medication A. Because participants with more severe disease placed less importance on the risk of stomach ulcer bleeding, they would accept a higher increased risk for the same improvement in pain.

Including an interaction between the WOMAC score and risk of stomach ulcer bleeding gave similar results to the main mixed effects analysis, indicating heterogeneity in the importance of the risk attributes (Section 4.3.4 and Figure 4.7). However, there remained a high level of preference heterogeneity that was not explained by the subgroup analyses. In addition, including the interaction term in the mixed effects model did not improve the predictive ability of the model; the AUC and corresponding confidence interval did not change.

Table 4.9: Coefficients from mixed logit model with interaction terms

		Coefficient	95% LCI	95% UCI*
Main effects:				
Pain	None	0.53137	0.40737	0.65537
	Mild	0.53712	0.41435	0.65989
	Moderate	-0.05553	-0.15307	0.04202
	Severe	-1.01297	-1.23452	-0.79141
Stiffness	None	0.11199	0.01681	0.20717
	Mild	0.06952	-0.02990	0.16894
	Moderate	0.06319	-0.02166	0.14804
	Severe	-0.24470	-0.34825	-0.14115
Function	None	0.45589	0.32423	0.58755
	Mild	0.13141	0.03726	0.22555
	Moderate	0.00052	-0.08793	0.08898
	Severe	-0.58782	-0.71452	-0.46111
Duration	12 months	0.62458	0.46308	0.78608
	6 months	0.06569	-0.01521	0.14659
	3 months	-0.02506	-0.10479	0.05468
	1 month	-0.66521	-0.82969	-0.50074
Heart attack risk		-0.48324	-0.59213	-0.37434
Stomach bleed risk		-0.46728	-0.54444	-0.39011
Interactions:				
WOMAC score x Stomach bleed risk		0.00375	0.00167	0.00583

* LCI: Lower confidence interval, UCI: Upper confidence interval.

4.3.6 Trade-off between pain and duration

The trade-off between duration and treatment effect was more complex to determine because both were included as effects-coded variable. However, marginal rates of substitution provided some insight into the potential trade-off.

As an example, consider two medications with equal utility scores: Medication A results in mild pain and has a treatment effect lasting 1 month. Medication B results in moderate pain and has a treatment effect lasting 3 months.

Table 4.8 showed the trade-off that participants were willing to make that increased the risk of heart attack. showing that several improvements in the treatment effect or extensions in the duration of treatment effect were equivalent. A reduction in the risk of heart attack by approximately 1.2% was equivalent to any of the following:

- Reducing effectiveness on reducing pain from moderate to mild,
- Increasing the duration of treatment effect from 1 month to 3 months, or
- Increasing the duration of treatment effect from 6 months to 12 months.

Similarly, for functional difficulties, reducing functional problems from severe to moderate was equivalent to increasing the duration of treatment effect from 1 month to 3 months. Reducing functional problems from severe to mild was equivalent to increasing the duration of effect from 1 month to 6 months. Extending the duration of treatment effect from 1 month to 12 months was always more important, whatever the improvement in the effect on functional problems. So if the effect of medication A is sustained for 12 months, the effect of medication B is sustained for 1 month and the two medications are otherwise equivalent, then the results suggest that participants would prefer medication A regardless of the effect on function.

Further examples of comparing medications with different levels for pain and duration are shown in Table 4.10.

Table 4.10: Comparison of medications with different durations of effect and effects on pain

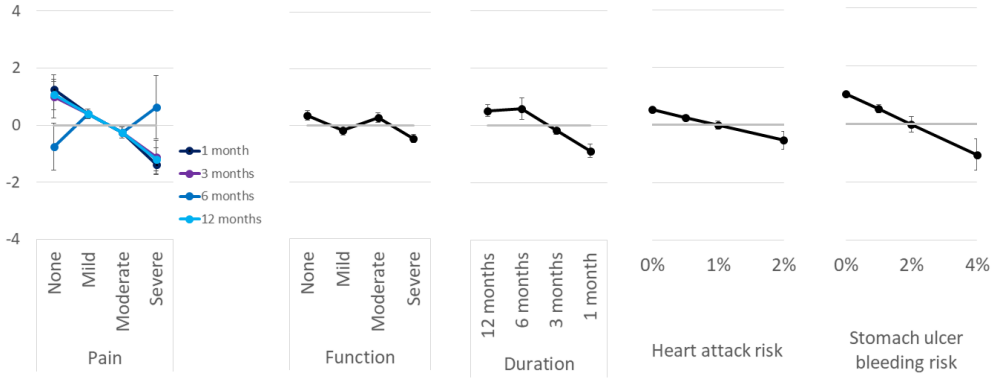
Medication A		**Medication B**		**Equivalent risk change for pain**	**Equivalent risk change for duration**	**Which difference is more important?**	**Which medication is more preferred?**
Pain	**Duration**	**Pain**	**Duration**				
Severe	12 months	Moderate	1 month	2.0%	2.7%	Duration	Medication A
Severe	12 months	Moderate	3+ months	2.0%	$\leq$ 1.4%	Pain	Medication B
Severe	12 months	Mild or none	1+ months	3.2%	$\leq$ 2.7%	Pain	Medication B
Severe	6 months	Moderate	1+ months	2.0%	$\leq$ 1.5%	Pain	Medication B
Severe	6 months	Mild or none	1+ months	3.2%	$\leq$ 1.5%	Pain	Medication B
Severe	3 months	Moderate	1 month	2.0%	1.3%	Pain	Medication B
Severe	3 months	Mild or none	1 month	3.2%	1.3%	Pain	Medication B
Moderate	12 months	Mild or none	6 months	1.2%	1.2%	Same	Equivalent
Moderate	12 months	Mild or none	$\leq$ 3 months	1.2%	$\geq$ 1.4%	Duration	Medication A
Moderate	6 months	Mild or none	3 months	1.2%	0.2%	Pain	Medication B
Moderate	6 months	Mild or none	1 month	1.2%	1.5%	Duration	Medication A
Moderate	3 months	Mild or none	1 month	1.2%	1.3%	Same	Equivalent

4.3.7 Interaction models between pain and duration

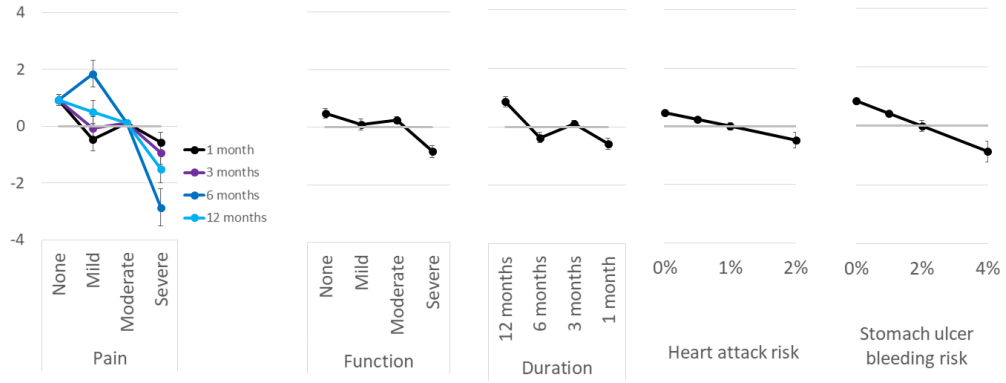
Exploratory models including interactions between pain and duration indicated that the importance of reductions in pain levels could differ based on the duration of the treatment effect. The coefficients for models with interactions between pain and duration are presented in Appendix C.10. To have sufficient degrees of freedom to include the interaction terms, the stiffness attribute was excluded because it did not significantly affect medication choice in the main model.

The interactions between pain and duration in Figure 4.9 suggest that reductions in pain were more important for treatment effects lasting 6 or 12 months, compared to effects lasting 1 or 3 months. This can be seen mostly clearly in the interactions between pain (mild) and duration (Figure 4.9(b)). However, the direction of some coefficients became illogical when interaction terms were included, for example, a reduction to no pain was seen as less important than a reduction to mild pain at 6 months duration in Figure 4.9(a). This could be confounding between the interaction terms and other attributes as the experimental design was not optimised for analysis of interactions between attributes. The results of interactions between pain and duration should be interpreted with caution and future studies would be needed to make confirmatory statements about the interaction effects between attributes.

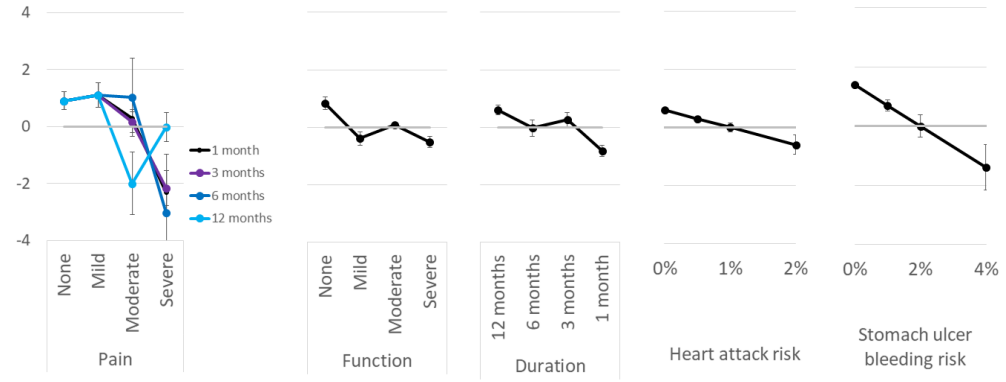
Figure 4.9: Coefficients of mixed effects models with interactions between pain and duration (excluding stiffness)



(a) Between duration and pain (none)



(b) Between duration and pain (mild)



(c) Between duration and pain (moderate)

4.4 Discussion

4.4.1 Summary of findings

This study aimed to examine whether duration of treatment effect is an important factor for people when choosing between different osteoarthritis medications, and to quantify the level of importance of duration of treatment effect relative to treatment benefits and risks. The results suggest that improvements in pain and function, duration of treatment effect and risks of serious events are important to people with osteoarthritis when choosing between medications.

The overall importance of duration of treatment effect was similar to the importance of improvements in pain levels and function and risks of serious cardiovascular and gastrointestinal events. The results suggest that participants would be willing to accept a medication that was less effective in relieving their pain symptoms if the effect of the medication was longer-lasting. For example, participants viewed a medication reducing pain from 'severe' to 'moderate' for 3 months as equivalent to a medication reducing pain from 'severe' to 'mild' for 1 month. Therefore, the duration of treatment effect should be incorporated into risk-benefit assessments of pharmacological treatments for osteoarthritis. However, there was considerable preference heterogeneity, suggesting that risk-benefit trade-offs will vary between individuals.

Improvements in pain were highly important unless pain levels were mild. Improvements in physical function were important to a lesser extent, however stiffness was considered relatively unimportant. The importance of moving between different levels of the WOMAC domains was not consistent across the levels. For example, improving pain levels to 'mild' compared to 'moderate' was important, but a reduction to 'no pain' was seen as equivalent to 'mild pain'.

Risks of serious events were also considered to be important. An absolute risk reduction of 1% was equivalent in importance to the risk of either a heart attack or stomach ulcer bleed. Respondents thus focused on the incidence risk of the serious events and did not distinguish between serious cardiovascular events or serious gastrointestinal events. This could have been due to lack of knowledge and awareness of the varying impact of the different events.

There was preference heterogeneity in the importance of the risk of serious events. Subgroup analyses indicated that participants with more severe disease placed less importance on the risk of stomach ulcer bleeding. Respondents with more severe disease will have worse symptoms and therefore may be more willing to accept risk to improve their current health state. Additionally, the reduced importance of the risk of stomach bleeding to those with more severe disease may be because respondents with more severe disease were more aware of the side-effects associated with NSAID medications and were therefore aware that acid-reducing and ulcer-preventative treatments, such as proton-pump inhibitors, can be taken to mitigate the increased risk of stomach ulcer bleeding. However, the subgroup analyses should be interpreted with caution due to low power and increased type I error due to multiple testing.

4.4.2 Comparison with literature

4.4.2.1 Representativeness of the study sample

The age and gender profile of the study sample was very similar to the UK Clinical Practice Research Datalink (CPRD) sample of initial management of osteoarthritis and late-stage osteoarthritis requiring joint replacement [304, 305, 306, 307]. Incidence of osteoarthritis is known to be higher in older age groups and females [308, 309]. Observational data and clinical trials had a similar age to the study sample (mean age

around 60 years old) but often had a slightly larger gender imbalance than in the study sample (60-70% female) [108, 269, 281, 282, 283, 284, 304, 305, 306, 307, 308, 310, 311, 312, 313, 314, 315, 316, 317]. The study sample was also representative in terms of the sites of osteoarthritis occurrence, with many participants having osteoarthritis at multiple sites. Knee osteoarthritis was most common, followed by hip osteoarthritis, then hand osteoarthritis [305, 312, 318, 319].

The proportion of women and disease severity (WOMAC scores) varied between clinical trials of NSAIDs for osteoarthritis [269, 282, 316]. However, the characteristics of the study sample were within the region of trials included in a recent review of oral NSAIDs for osteoarthritis (WOMAC pain 50-60 out of 100) [269]. Other studies included samples with similar characteristics, including disease duration [273, 284, 308, 320, 321, 322], hypertension [265, 274, 312, 314], prior knee injury or surgery [323], and comorbidities, including previous gastrointestinal and cardiovascular problems [265, 310, 312, 313, 314, 318].

4.4.2.2 Comparison of results with other preference studies

The heterogeneity in preferences for osteoarthritis medication found in this study has been demonstrated in other studies [214, 324, 325].

Regarding symptomatic improvement, the results of this study align with the literature in that the treatment effects on pain and physical function were important, whereas stiffness was relatively unimportant [214, 238, 291]. One key difference between this study and the existing literature was that other studies suggested the importance of moving from 'mild' to 'none' on the pain scale was more important than moving from 'moderate' to 'mild' levels [214, 291]. This difference in findings could be due to variations in participant characteristics, how participants interpreted the levels of the WOMAC domains or the consideration of long-term effects.

Other studies have also found that risk of adverse events was important to people with osteoarthritis [250, 324, 325, 326, 327, 328]. The results of this study differed from Hauber *et al.* in the importance of the risk of different events [214]. Hauber *et al.* found that the risk of cardiovascular events (heart attack and stroke) was viewed as more important than the risk of a bleeding ulcer. This could be due to differences in how the events were described or the different ranges considered for the incidence levels of the events. Hauber *et al.* described a bleeding ulcer requiring an operation, as opposed to focusing on the risk of mortality. Hauber *et al.* also included risks of 0-2.5% for bleeding ulcers and 0-3% for heart attack, compared to 0-4% and 0-2% respectively in this study, which could have affected the results. Risks of 2% for heart attack and 4% for stomach bleed are within the maximum acceptable risk levels for people with osteoarthritis [329].

The studies by Hauber *et al.* and Arden *et al.* considered the treatment effect 1 hour after taking the medication, assuming that the effect did not degrade over time. Our study found that the duration of treatment effect was important. This aligns with the results of other studies. For example, Cordero-Ampuero *et al.* found lasting symptom relief was important to people [225]. Posnett *et al.* also found that duration of relief was more important than the amount of pain reduced by injections [251].

4.4.3 Examples of use and interpretation of results

The results of the study can be applied to real-world medications to infer whether one medication would be preferred to another.

4.4.3.1 Trade-off between risk and duration: Example of piroxicam

Louthrenoo *et al.* compared piroxicam to diacerein for the treatment of painful knee osteoarthritis. Both produced similar effects for the first months [330]. However, after 3 months (on treatment discontinuation), the effects persisted for diacerein until 6 months but symptoms returned for the piroxicam group. A carry-over effect for diacerein was also found when compared to diclofenac [331]. Both found that the risk of adverse events was lower for diacerein. Hence, diacerein would be the preferred medication in terms of risks and treatment effects.

The lack of carry-over effect for NSAIDs means they must be taken for a longer time for the effect to be sustained. However, the risk of serious adverse events will also increase with the exposure period. Therefore, the results of this study could be used to assess whether people with osteoarthritis would be willing to continue taking an NSAID to produce a longer-lasting effect, given the increased risk of adverse events.

For piroxicam, pain, function and stiffness were reduced from moderate (around 50/100) to mild (around 20/100) levels for 3 months [330]. Assuming that the symptom reduction would be sustained if the participant continued to take this medication for longer than 3 months, the marginal rate of substitution in terms of the risk of heart attack is -0.08 for 3 months, 0.14 for 6 months and 1.47 for 12 months for the fixed effects model (Appendix C.7). Therefore, to sustain the effect for 6 months (i.e., an additional 3 months), the participant would be willing to increase the risk of heart attack by 0.22%. To sustain the effect for 12 months (i.e., an additional 9 months), the participant would be willing to increase the risk of heart attack by 1.55%. The risk of a heart attack is around 1-2% per year when taking an NSAID [272, 313]. Assuming that the risk of a heart attack increases linearly with the exposure period, then once the medication has been taken for 3 months, the additional risk from taking the medication for a further 3 months would not be worth the increased length of

treatment effect. However, participants would be willing to expose themselves to a higher risk of a heart attack if taking the medication for 12 months sustained the treatment effect.

4.4.3.2 Trade-off between risk and treatment effect: Example of rofecoxib

Rofecoxib, a drug marketed to treat osteoarthritis and acute pain, was discontinued due to the significantly increased risk of cardiovascular events (0.4% compared to 0.1%, as found in the VIGOR study) [332, 333, 334]. Other studies have estimated the risk of heart attack while taking rofecoxib to be around 1.5% for 1 year [335, 336].

Our results suggest that some people may be willing to tolerate a high risk of heart attack if the medication provides sufficient and sustained symptom reduction. For example, participants would be willing to increase the risk of heart attack by at least 1.5% if the medication reduces pain from moderate to mild, reduces functional problems from severe to mild, or increases duration of treatment effect from 3 months to 12 months. The willingness-to-risk could also be satisfied for a medication with a combination of smaller improvements, such as reducing functional problems from mild to none, reducing stiffness from severe to mild and reducing risk of stomach bleeding by 0.5%. However, even given these preferences, clinicians may still be unwilling to subject patients to a higher risk of cardiovascular events and may be cautious when basing medication choices on stated preferences only.

4.4.4 Strengths

4.4.4.1 Strengths of study methodology

The use of a discrete choice experiment overcame methodological issues with alternative study designs, such as time trade-off studies. In a discrete choice experiment, more than two attributes can be considered in the same choice task. In this discrete choice experiment, the same number of questions were completed regardless of the responses to those questions. The problem of 'satisficing' was thus avoided, where participants respond in a particular way to reduce the time to complete the task.

The choice of attributes and levels were informed by clinical relevance, existing literature on preferences and patient representative feedback. This should increase the applicability of the results to a wide range of osteoarthritis medications and ensure that the included attributes are important to most people living with osteoarthritis.

Using a condition-specific sample, as opposed to sampling from the general population, increased the relevance of the results to the clinical situation. People living with osteoarthritis should be able to more easily cast their mind backwards or forwards to the relevant point in their own condition. Those living with the condition are also likely to be more motivated to participate and to fully consider the attributes and background information.

Designing the choice tasks to be based on a hypothetical scenario increased the homogeneity of the treatment choices made by different respondents. This reduced the possibility that treatment preferences were due to the characteristics of the particular respondent and increased the power of the study. Using unlabelled profiles for the medications also reduced the number of unobservable assumptions made by participants during the choice tasks.

The use of a mixed logit model allows preference heterogeneity to be incorporated. Alternatively, a probit model could have been used. Probit and logit models have been shown to produce similar results in most circumstances [337]. However, a probit model requires assumptions on the correlation between errors across choices, and the coefficients are more difficult to interpret than for logistic regression as they are not on the log-odds scale [338]. A practical drawback of a probit model is its high computational complexity. It cannot be solved analytically as it does not have a closed form expression and the computational complexity is more of a barrier when the structure of the covariate matrix is unrestricted. In addition, the experimental design software (Ngene) is not able to optimise design efficiency for probit models.

4.4.4.2 Strengths of study conduct

This is one of the first discrete choice experiments to examine the importance of the duration of treatment effect in decision-making. This study has demonstrated that duration can be included in preference studies, along with effectiveness and risks of treatment, without causing prohibitive increases in cognitive difficulty and participant burden. It is reassuring that over 75% of participants did not find it difficult to imagine themselves in a hypothetical scenario when performing the choice tasks and few participants found it 'very difficult' to complete the choice tasks. Few participants based their choices on only a single attribute or demonstrated irrationality by selecting a dominated alternative.

Eye-tracking studies have shown that respondents focus more on attributes at the beginning of a vertical list as they read from top to bottom [339, 340]. However, our results indicated that the final attribute, risk of stomach ulcer bleeding, was highly important. Participants probably incorporated all attributes into their decision-making, regardless of the ordering.

The use of an online market research panel and self-reported diagnoses of osteoarthritis could have raised concerns about the representativeness of the study sample. However, the demographics of the sample were similar to samples of people with osteoarthritis from clinical practice (Section 4.4.2.1). The findings are thus likely to be applicable to clinical populations in the UK.

4.4.5 Limitations

4.4.5.1 Limitations of study sample

Although the demographics of the sample were similar to the clinical population, the respondents to an online survey may not be fully representative, especially given the use of self-reported eligibility questions rather than confirmed clinical diagnosis of osteoarthritis. An online market research panel may be more likely to include younger participants with less severe disease and less likely to include participants with co-existing hand osteoarthritis. The remuneration for the survey may also have attracted respondents who do not work or those with a lower socio-economic status. Some studies have shown that stated preference studies conducted online can produce different results to those using paper-based surveys [341].

The study sample consisted of residents of the United Kingdom. The study results may therefore not be applicable to countries with health-insurance-based systems. Where fee-paying private healthcare is more common, it is likely that medication cost is an important factor in treatment decisions. Studies have also shown that ethnicity and culture can affect treatment choices. Therefore, the study results may not be generalisable outside a UK sample [255, 342, 343, 344]. The interpretation of the different levels may also differ by country because of the terminology used. The wording may need to be altered based on translated versions of the WOMAC measure

if this experiment were reproduced in other countries.

The study results may not be applicable to certain subpopulations of people with osteoarthritis. For the small proportion of people with osteoarthritis who have previously had a heart attack, the increased risk of a second heart attack may exceed the risk level range considered in this study [345]. Very elderly or debilitated people seem to tolerate stomach ulcer bleeding less well than other individuals and, in this group, it is more common for a gastrointestinal issue to lead to a fatality [266]. Therefore, older people (or their clinicians) may consider the risk of stomach ulcer bleeding to be more important.

4.4.5.2 Limitations of study design

A limitation inherent to discrete choice experiments is that only a limited number of attributes can be included without affecting the completion rate [289, 346]. Therefore, this study was not able to fully encompass all potential benefits and risks associated with all medications, such as effects on co-morbid conditions.

The choice of attributes in this study was based on the clinical literature and a sample of 10 patient representatives. While this should have covered the key attributes relevant to the osteoarthritis population, some factors that are important to individual participants might have been omitted. In a real-life situation, there could be differences in other attributes, for example, risk of renal or hepatic events and side effects including nausea or dizziness. It was also assumed that the participant's general practitioner (GP) was indifferent between the two medication choices. In practice, shared decision-making will include a viewpoint from the clinician so a person's care provider may make a recommendation of one medication over another. This has been found to be an important factor in other conditions [347].

The experimental design was not optimised for analysis of interactions between attributes because the primary aim was to explore the trade-off between duration of effect and other factors. To efficiently estimate interaction terms would have added complexity to the experiment, requiring the use of blocking and an increased sample size. Therefore, the results of models with interactions between pain and duration must be interpreted with caution. Future studies would be required to establish whether there is an association between the reduction in pain and the length of the treatment effect.

A key drawback of a discrete choice experiment is that it is based on stated preferences. As such, it is unclear whether the participants would indeed choose to take the selected medication in a real-life scenario. Participants were also required to base their treatment decisions on an imaginary patient scenario. A hypothetical patient scenario was used to increase the homogeneity between respondents in the choice task being considered. However, the results are thus less applicable to people whose symptoms are not severe prior to taking the medication. Participants may also have found it difficult to consider an imaginary patient during the choice tasks, without allowing their own experience to influence their treatment choice. For example, participants who had previously experienced cardiovascular problems would have been more aware of the impact this could have on their daily lives.

Eighteen percent of participants chose different alternatives when presented with the same choice task later in the survey. This could indicate a lack of robustness in the study results. However, sensitivity analysis found that the findings were still similar when those providing inconsistent responses were excluded. In addition, the proportion selecting medication B was 58% when the choice task was first displayed and 56% when it was repeated, so this lack of consistency could be due to uncertainty for this particular choice task. As the identical choice tasks were the 4th and 17th

tasks, the inconsistency could also have been due to learning effects.

4.4.5.3 Limitations of attributes

Another limitation is that the use of attributes is reductive. Pain in general, and osteoarthritis pain specifically, is known to be multi-dimensional. Pain can be described in terms of its duration, intensity, frequency, and impact on mood and physical activity [348]. Other studies have found that pain during movement and pain at rest may relate to different constructs [182, 214, 291, 349, 350]. These different types of pain may be valued differently by participants. Similarly, the risk of heart attack relates to the risk of both fatal and non-fatal events. An NSAID might increase the risk of non-fatal heart attack without increasing the risk of fatal heart attack [351]. It is unclear how varying the different elements of 'composite' attributes could influence the results.

A key limitation is heterogeneity in the interpretation of the individual levels used for the treatment effectiveness attributes. Different people may have different interpretations of what constitutes 'mild pain', for example, due to differing pain thresholds. However, this is an issue for pain measurement in general.

The levels for the duration attribute were chosen to correspond to commonly used assessment time points in randomised trials. However, most of the participants had experienced osteoarthritis symptoms for more than 3 years. This study could thus not quantify preferences for interventions with very long term effects in clinical practice. For similar reasons, preferences for effects measured in hours, such as those arising from a single dose of medication, also could not be estimated.

Previous trade-off studies have highlighted the difficulty in communicating risks [352]. Infographics were included to help participants to interpret the risk levels. It is possible that presenting risk attributes in a different way could have affected partici-

pants' responses [339]. The assumption of linearity for the risk attributes could not be tested because there were insufficient degrees of freedom to include the risk attributes as effects-coded variables and the experiment was not designed to estimate these parameters.

The choice tasks assumed that the treatment effect was constant over time and that once the treatment was no longer effective, the patient will immediately return to having severe symptoms. In reality, symptoms fluctuate and treatment effects are likely to diminish gradually over time. It is also unrealistic to assume that, once the chosen medication is no longer effective, the patient cannot take alternative medications to improve their symptoms. However, this suggestion would be problematic as the effectiveness of the second medication would be uncertain. The attributes related to treatment effectiveness also assumed that all people received the same treatment effect from taking the medication. In practice, there is uncertainty in these estimates and not all people who take the same medication will receive the same level of treatment effect. Previous discrete choice experiments have found that preferences are affected by the imprecision around the effectiveness and harm of different treatments [353, 354].

4.4.6 Implications

The results indicate that the effect on pain and physical function, the duration of treatment effect and risks of serious events are all important factors for people with osteoarthritis when choosing between different medications. The results can be used to estimate which of two medications would be preferred, given the trade-off between the treatment effects on symptoms, the duration of the treatment effect and the associated risks (Section 4.4.3).

4.4.6.1 Implications for clinicians

Clinicians should aim to prescribe osteoarthritis medications that reduce pain to 'mild' levels and eliminate difficulty in carrying out daily activities, whilst minimising risk of serious cardiovascular and gastrointestinal events. Unless medications will eliminate all functional problems or produce sustained effects longer than 6 months, clinicians should prioritise medications that will reduce pain symptoms and have a low risk of serious adverse events. Clinicians should also factor in a person's disease severity when deciding on the most appropriate medication, as risks of serious events may be more important to people with less severe disease.

The presence of preference heterogeneity suggests that clinicians should incorporate patient opinion when choosing an osteoarthritis medication. The trade-off between improvements in pain and the risk of serious events will vary between patients. Clinicians should discuss the risk-benefit profiles of different medications and assess their patients' attitudes to the risk of serious events. The results could inform the development of decision-making support tools to facilitate a shared decision-making process. Involving patients in treatment decision-making and prescribing medications that align with patient preferences could increase adherence and improve patient satisfaction.

4.4.6.2 Implications for trialists

An increase in the duration of treatment effect from 6 month to 12 months was as important as a reduction in pain level from 'moderate' to 'mild'. The importance of the duration of treatment effect suggests that phase III clinical trials of medications for osteoarthritis should include long-term follow-up, ideally 12 months or longer. Extended trial follow-up would allow the measurement of long-term outcomes, to see

whether treatment effects are sustained over time, and facilitate better consideration of rare adverse events. When interpreting clinical trial results, the duration of the treatment effect should be taken into account during the benefit-risk assessment of medications to relieve osteoarthritis symptoms. As the duration of treatment effect was nearly as important as pain and as important as function, duration of treatment benefit should be carefully considered by users of clinical trial results when comparing trials with different assessment time points. Trialists should also consider the optimal time point for patient outcome measurement when designing future trials.

These results could also have implications for the design and interpretation of clinical trials in terms of the appropriateness of MCID (minimum clinical important difference) estimates. The time point of assessment may affect a person's view of what is an 'important difference' in the treatment outcome. Due to the importance of the duration of treatment effect, the clinically important difference may vary depending on the time point of assessment. This could have implications for the sample size required for a randomised trial that uses the MCID estimate as the target difference. The results could also be used to produce MCID estimates that incorporate the risks of treatment, as was done by Hughes *et al.* [355].

When designing clinical trials of pharmacological treatments for osteoarthritis, primary outcomes should focus on pain, potentially in combination with function. Trials using composite outcome measures, such as the WOMAC, should analyse the individual subscales to unpick which domains generate the differences in treatment effect. Although the WOMAC Index incorporates pain, stiffness and function, the results of this study indicate that improvements in the stiffness domain were not valued by participants. Therefore, when reporting the results of clinical trials, authors should report the treatment effect on the individual subscale for each of the three WOMAC domains, as well as the overall WOMAC score. This would help readers to understand

which domain is driving a significant treatment effect and ensure that the treatment will target the mechanisms that are of value to each individual patient. As changes in stiffness were not seen as important, researchers will need to establish whether treatment differences are due to changes in the stiffness domain to fully assess the clinical relevance. The results also suggest that, for some people with osteoarthritis, the changes between different levels are valued differently for different items or domains. This could raise concerns on whether the scaling of the WOMAC outcome measure is appropriate.

4.4.7 Future research

The robustness of these results could be tested by using randomised ordering of the choice tasks, alternatives and attributes. A second study could also be conducted in a more representative sample, for example, recruiting respondents from an outpatient clinic or clinical trial participants. This would allow a comparison to establish whether preferences are similar for an online research panel sample and a sample recruited through clinical pathways.

The results of this experiment could be supplemented by qualitative data. In-person interviews would allow more in-depth assessment of how people with osteoarthritis make their treatment decisions and how well participants understood the background information. This could allow a comparison of using a face-to-face interview to deliver the survey, as opposed to using an online questionnaire. However, an interviewer-based approach would be more costly and time-consuming [356].

Predicted stated preferences could be compared to usage statistics, for example, by analysing the number of prescriptions, to see whether medication prescriptions align with patient preferences. This would allow us to assess the external validity of the results by comparing hypothetical and actual behaviour [356]. Future work could also

explore whether participants' stated preferences align with their revealed preferences. Revealed preferences could be elicited by observing actual medication choices made by people living with osteoarthritis. Future studies could also assess the predicted uptake of osteoarthritis medications to inform cost-effectiveness analysis and economic modelling.

A larger sample could increase the power of subgroup analyses, permitting further investigation of preference heterogeneity. Further subgroup analyses could also be performed if additional data were collected, such as radiographic severity or socio-economic status. Latent class analysis could be used to identify different preference subgroups. Future studies could explore different osteoarthritis subpopulations, for example, whether people with mild and severe symptoms of osteoarthritis have similar preferences. A dynamic questionnaire could be used to avoid using a hypothetical patient scenario. The participant's current disease status could then be incorporated into the choice tasks, making them more relevant to the participant's situation.

Future research could examine whether other key stakeholders, such as commissioners and clinicians, place similar levels of importance on the attributes. It is likely that commissioners would consider financial cost to be a key factor in treatment choice, whereas cost to the NHS was not considered important by the sample of patient representatives during the development of the discrete choice experiment. Clinicians may also make different trade-offs between benefits and risks, for example, placing more importance on side effects that may result in poor adherence.

Future studies could also assess whether the importance of the duration of treatment effect is similar for non-pharmacological interventions, such as exercise or surgical treatments. There could be differences in whether participants would prefer a surgery that offers a moderate improvement in the long-term, as opposed to a large short-term improvement followed by a decline to poor function.

Future studies would be required to evaluate interaction effects between attributes, such as pain and duration. The experimental design should be optimised for the analysis of models including interaction terms. This would require an experimental design with fewer attributes or levels, or the use of blocking, to retain a feasible number of choice tasks. For example, the number of attributes or levels could be decreased by removing stiffness due to the low importance of this attribute and by collapsing levels found to be of similar importance, such as combing the 3 and 6 month levels of the duration attribute.

In this study, respondents placed different levels of importance on a one-level change between adjacent categories of the WOMAC depending on the starting level. For example, respondents placed very little value on a change from 'mild pain' to 'no pain' compared to a reduction from 'moderate pain' to 'mild pain'. Change between these levels would have the same score on the WOMAC Index. This could explain why the minimal important difference of the WOMAC Index is affected by the baseline disease severity of participants [39]. Future research could explore whether the validity and responsiveness of the WOMAC Index can be improved by adjusting the item scoring for changes between levels. This could more accurately reflect participants' views on a change in disease status.

Future research is needed to identify the optimal assessment time points for different patient groups and patterns of recovery. Evaluating the importance of the duration of treatment effects could be facilitated by allowing treatment effects to fluctuate over time, rather than assuming that the treatment effect is constant while it is sustained. A visual representation could be used to illustrate the change in symptom relief over a particular time course. Future research could also explore more explicitly how the duration of treatment effect affects whether participants interpret a change in their condition as clinically important, for example, by presenting multiple scenarios

to participants and asking them whether they feel the treatment effect is clinically important.

Chapter 4 found that the duration of treatment effect was important to participants. Chapter 5 builds on this, using a cohort of people with osteoarthritis to examine whether the MCID estimate varies over the follow-up time period. Following this, Chapter 6 presents a simulation study that examines the statistical properties of different longitudinal methods to analyse the results of randomised trials using data from multiple follow-up time points.

4.4.8 Conclusion

The results of this discrete choice experiment suggest that the duration of the treatment effect is an important factor contributing to medication choice for people living with osteoarthritis, along with the effect on pain and function and the risk of serious cardiovascular and gastrointestinal events. The duration of the treatment effect was viewed with similar importance to the amount of symptom relief provided and risks associated with treatment. Medical research should focus on developing medications that reduce pain symptoms and functional problems in the long term, whilst reducing the risk of serious adverse events. Preference heterogeneity suggests that the importance of the effects on pain, function, duration and risks varies between respondents.

In a shared decision-making process, clinicians and people with osteoarthritis should discuss the benefit-risk profile of osteoarthritis medications, including consideration of the duration of the treatment effect. The duration of treatment effect should also be considered when interpreting the results of clinical trials of osteoarthritis medications. However, future research is needed to test the robustness of these findings to different samples and levels of duration. Future research should also explore whether the dura-

tion of treatment effect is an important factor in treatment decision-making in other disease areas or non-pharmacological interventions. Clinical trials in osteoarthritis populations should be designed to measure long-term clinical outcomes to monitor whether treatment effects are sustained.

Chapter 5

An assessment of the stability of Minimum Clinically Important Difference (MCID) estimates over time: secondary analysis of the Osteoarthritis Initiative (OAI) cohort

Prior publication:

Conferences abstracts for presentations on this chapter have been published (see Appendix F.1 for details).

5.1 Introduction

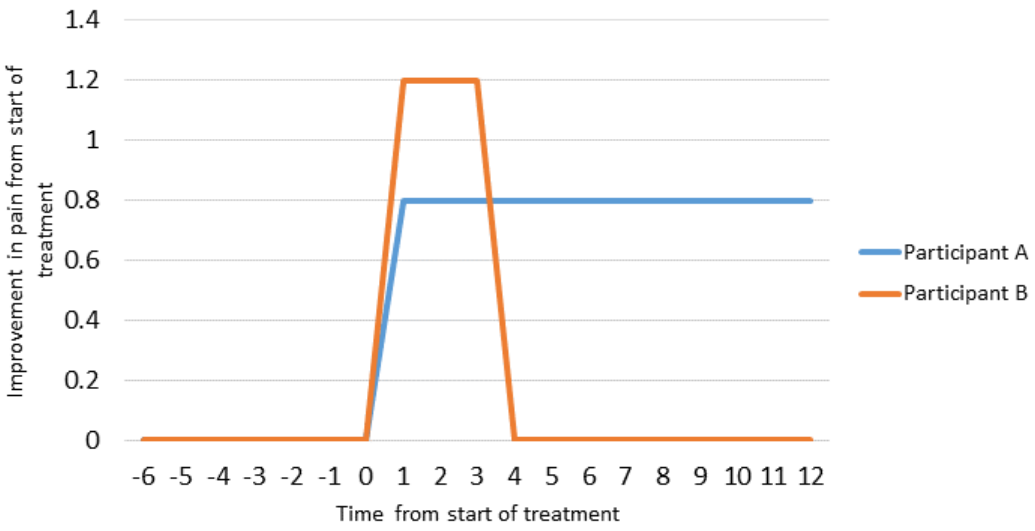
Length of symptom relief and time to recovery are important factors for musculoskeletal conditions. The majority of randomised trials assess outcomes at multiple time points. Other trials in some musculoskeletal conditions, such as acute low back pain and muscle injuries, consider time to recovery (or return to normal function) as a primary outcome [357, 358, 359]. As patient-reported outcomes fluctuate over time, multiple assessment time points can be used to explore when the key changes take place and provide a more accurate representation of the change in the participant's symptoms. Researchers and clinical guidelines have also emphasised the importance of long-term monitoring of patients in clinical studies and clinical practice [71, 226, 230, 360]. As well as detecting potential harms, evaluation of long-term treatment outcomes indicates whether treatment benefits are sustained over time [361]. Trials can sometimes find treatment benefits at short-term follow-up that are not sustained. Trials with only short-term follow-up may also miss treatment benefits that only emerge later [362].

As discussed in the Introduction (Sections 1.2 and 1.3), the minimum clinically important difference (MCID) of the outcome measure is most commonly calculated using an anchor-based approach. However, this standard approach calculates the important difference in a disease-specific patient reported outcome at only one time point; it is unclear whether the time point of assessment affects its value.

A person may view a small difference as unimportant in the short-term but view the same level of difference as important if it is sustained over a long time. For example, participants may prefer a small but sustained improvement (as in participant A in Figure 5.1) over a larger improvement that only lasts for 3 months (as in participant B in Figure 5.1). Participants may consider an improvement of 0.8 that lasts for 12

months to be worthwhile, but not an improvement of 0.8 that lasts for 3 months.

Figure 5.1: Hypothetical examples of a person's improvement in outcome over time



Therefore, an MCID calculated using data from one time point may not be appropriate for use as the target difference in a randomised trial with a different assessment time point. Increased understanding is needed about how the magnitude of difference that participants consider to be important varies over time.

An example in Section 1.4.1 demonstrated that using a different MCID estimate as the target difference for a clinical trial can have a large effect on the required sample size. When comparing a between-group treatment difference to an MCID estimate, the use of a different MCID value could also change whether a treatment provides a clinically meaningful benefit. This could affect whether a treatment is recommended for implementation into clinical practice [363]. Therefore, using an appropriate MCID estimate is critical.

5.1.1 Rationale specific to osteoarthritis

The systematic review in Chapter 2 found that it was common for osteoarthritis trials to specify the target difference in their sample size calculation based on a published estimate of the MCID. The review also found that osteoarthritis trials commonly conducted outcome assessment at more than one follow-up time point. However, the majority of trials did not state the primary time point that they were considering when calculating the sample size and specifying the target difference. In trials that justified the target difference using an MCID estimate, they did not report considering the assessment time point when deciding which published MCID estimate to use. If MCID estimates vary over time, trialists could use inappropriate MCID estimates to calculate the sample size of their trial. For example, trialists could use an MCID estimate calculated at short-term follow-up when their primary time point in the planned trial is long-term participant outcomes.

As well as informing sample size calculations, MCID estimates are also used in interpreting the results of clinical studies. Researchers may compare the summary effect to the MCID to determine whether the between-group treatment difference is clinically meaningful. Comparing the summary effect to the MCID estimate has been used in randomised trials, observational studies and meta-analyses [21, 34, 33, 35, 364, 365].

People can live with musculoskeletal conditions for years or even decades. Therefore, the longevity of treatment benefits is an important consideration when evaluating the effectiveness of treatments for musculoskeletal conditions. The time course of musculoskeletal conditions can also vary substantially between individuals. Some people with musculoskeletal conditions experience persistent severe pain, whereas others have patterns of fluctuating pain in recurrent episodes [366].

The discrete choice experiment in Chapter 4 found that duration of treatment effect

was an important factor for people living with osteoarthritis when choosing between different medications. The results showed that respondents were willing to accept an increased risk of heart attack or smaller improvements in pain symptoms if the treatment had a long-lasting effect.

Two recent reviews by Devji *et al.* and Erdogan *et al.* synthesised existing studies reporting the calculation of MCID estimates for outcome measures in osteoarthritis, including the WOMAC Index [213, 367]. The WOMAC Index is a patient-reported outcome that is commonly assessed in osteoarthritis research. It is described in more detail in Chapter 3, which examined the use of the WOMAC Index in randomised trials of osteoarthritis. From the articles reviewed by Devji *et al.* and Erdogan *et al.*, the majority of the studies calculating the WOMAC MCID estimate used data from only one time point, with the follow-up period ranging from 4 weeks to 5 years. Although the previous reviews did not find an association between the magnitude of MCID estimates and the follow-up assessment time point, they only considered a small number of studies: 10 by Devji *et al.* and 11 by Erdogan *et al.*.

The lack of association could have been due to confounding by other between-study variability. Some of the included studies highlighted that the follow-up time point could be a potential reason for the variability in MCID estimates [368]. Findings among studies that calculated MCID estimates at multiple time points were not consistent. For example, one study found higher MCID values for longer follow-up periods [368] and others found no clear pattern in the values over time [369].

There is therefore a need to explore how MCID estimates vary over time using a longitudinal dataset that has applied the WOMAC Index. The use of a single longitudinal dataset reduces the level of confounding present when multiple datasets and studies are used because it ensures that the population and methodology used are consistent across the different assessment time points.

The aims of the research presented in this chapter are to:

1. Examine whether estimates of the MCID vary depending on the time point of assessment,
2. Assess whether other factors, such as participant characteristics, result in variability in MCID estimates, and to
3. Explore the use of longitudinal methods to calculate MCID estimates, incorporating data from more than one follow-up time point.

5.2 Methods

5.2.1 Description of dataset

The Osteoarthritis Initiative (OAI) is a multi-centre observational study based in North America and includes clinical, imaging and biological outcomes [370, 371]. The OAI database is open access, and anyone is able to access the data by registering for a free account on the OAI website [370]. Since this analysis was completed, the website to access the OAI dataset has changed, and it is now accessible through a site hosted by the National Institutes of Health [372]. The OAI dataset includes a progression cohort of participants with symptomatic knee osteoarthritis at baseline, a development cohort examining participants at high risk of developing symptomatic knee osteoarthritis, and a control cohort of participants with no symptoms or risk factors of osteoarthritis in either knee. Only the progression cohort was used in this analysis.

Participants were eligible for inclusion in the progression cohort if they (i) had at least one knee with symptomatic tibiofemoral knee osteoarthritis defined as definite osteophytes (OARSI grade 1-3) and (ii) had, in the same knee, frequent symptoms (pain, aching or stiffness on most days for at least one month in the past year). Participants were not excluded based on the treatment they received and may have received no treatment for otseoarthritis at all.

There was no primary outcome for the study, however the WOMAC was measured as part of the clinical outcome assessment. The Likert scale version of the WOMAC was used, ranging from 0-96 where 0 represents no symptoms and higher scores indicate more severe disease. This version of the WOMAC was the most commonly used among the randomised trials of treatments for osteoarthritis reviewed in Chapter 3.

When outcome data was measured separately for each knee, the mean value for the left and right knees was used if participants had osteoarthritis in both knees.

Participants were followed up annually. At the time of this analysis, data were available for a 9-year follow-up period (or 108 months). Interim follow-up visits were completed at 18 and 30 months for a subsample of participants (approximately one-third of the progression sub-cohort). However, data from the 18- and 30-month follow-up assessments were not included in this analysis. For all participants in this analysis, the baseline time point refers to the point of study entry and does not relate to the provision of treatment. After the baseline time point, follow-up data were collected at an annual clinical assessment.

5.2.1.1 Description of the anchor measure

The anchor measure was a global assessment of the participant's condition on the day of the clinical assessment. The participant rated their condition on a scale of 0-10 (0 being very good and 10 being very poor). The definition of minimal improvement (or deterioration) was calculated based on the change in the global assessment between that time point and a previous time point.

5.2.2 Assessing the stability of MCID estimates calculated using single time point anchor-based approaches

Existing anchor-based techniques for calculating the MCID of an outcome measure using data from a single time point were applied at different assessment time points (years 1-9 of follow-up at 12-month intervals) [13]. The variability of MCID estimates from different time points was assessed. The MCID for 'deterioration' was calculated using equivalent methods. The MCID for deterioration is not usually symmetric

and is known to differ from estimates for improvement [373, 374]. All analyses were performed on the intention-to-treat population. For the primary analysis, missing data were not imputed. The single time point approaches used are described in Sections 5.2.2.1 to 5.2.2.3.

5.2.2.1 Raw mean difference method

The unadjusted mean difference in the change score was calculated between participants who rated themselves 'minimally improved' and participants who rated themselves 'no change' at that time point.

5.2.2.2 ANCOVA method

Linear regression (equivalent to ANCOVA or analysis of covariance) was used to calculate the between-group difference in the mean change score between participants who rated themselves 'minimally improved' and participants who rated themselves 'no change' at that time point, adjusting for baseline WOMAC total score [375]. A dummy variable for 'minimally important change' was entered into the regression model and the MCID estimate was given by the coefficient for the dummy variable.

5.2.2.3 ROC curve method

Receiver operating characteristic (ROC) curve analysis was used to compare participants who 'minimally improved' with participants who rated themselves 'no change' at the corresponding time point. The MCID based on the ROC curve analysis was estimated to balance sensitivity and specificity at the 'optimal' cut-off value, which minimises: $(1-sensitivity)^2+(1-specificity)^2$ [376]. This method finds the 'optimal' cut-off point at the top left hand corner of a standard ROC curve.

These three methods were also used to calculate the MCID estimate for deterioration, rather than improvement. The equivalent methods were used to examine the difference between the 'minimally worsened' and 'no change' groups at the corresponding time point.

5.2.2.4 Anchor measures and definition of 'improvement' and 'worsening'

For the single time point analyses, participants were defined as:

- *Minimally improved* if the global assessment of their condition reduced by 1 point at the corresponding time point.
- *Minimally worsened* if the global assessment of their condition increased by 1 point at the corresponding time point.
- *No change* if the global assessment of their condition did not change at the corresponding time point.

All of the definitions of improvement and worsening were based on the change in the global anchor measure at the corresponding follow-up time point compared to (i) baseline and (ii) 1 year earlier.

These definitions were designed to capture the 'minimal' concept of the MCID and used the smallest measurable improvement on the global assessment scale.

5.2.3 Subgroup analyses

Subgroup analyses were conducted to highlight other potential sources of variability in MCID estimates based on change from baseline. Subgroup analyses examined differences due to age, sex and baseline WOMAC score. Subgroup analyses were con-

ducted using subsample analyses, for example, performing each analysis separately for men and women. Subgroup differences were tested by comparing the 95% confidence intervals of the estimates in the different subsamples. Age subgroups were categorised into tertiles. The tertile cut-off values for age were 45-56, 57-66 and 67-79 years. The cut-off values for the WOMAC score were 0-12, 13-36 and 37-96 (where a higher score represents more severe disease). The cut-off values were selected based on the Likert scale form of the WOMAC Index: 12 points and 36 points corresponds to an average response of 'none to mild' and 'mild to moderate' symptoms, respectively.

5.2.4 Calculation of MCID estimates using longitudinal methods

This study also explored the use of longitudinal methods to calculate MCID estimates. These methods included (i) an adapted raw mean difference approach, (ii) an AUC approach, (iii) mixed effect regression and (iv) a generalised estimating equation. For the longitudinal measures, only the minimally important improvement was considered. The different methods were summarised in terms of the feasibility, assumptions made, data required, and ease of interpreting the results. The code used for the different methods is given in Appendix D.4.

5.2.4.1 Definition of minimal improvement using longitudinal data

The definition for minimal improvement using longitudinal data was based on the AUC (area under the curve) of the change from baseline in the global anchor measure. The AUC for the change in the anchor measure was calculated and divided by the number of observed time points, giving $\frac{AUC\Delta}{t_{max}}$. t_{max} is the latest assessment time point where the global anchor was recorded for the participant. Participants could thus be

included even if they did not complete the anchor for the full follow-up period. For intermittent missing values before t_{max}, the AUC calculation assumed a linear trend between the observed values.

A participant was defined as 'minimally improved' if $-1.2 \leq \frac{AUC\Delta}{t_{max}} \leq -0.8$.

A participant was defined as 'unchanged' if $-0.2 \leq \frac{AUC\Delta}{t_{max}} \leq 0.2$.

A participant was defined as having an improvement greater than 'minimally improved' if $\frac{AUC\Delta}{t_{max}} < -1.2$.

The cut-off of ± 0.2 was an approximation of 2/9 rounded to 1 decimal place, relating to an average of a 1-point improvement over 2 years or a 2-point improvement over 1 year of the 9-year follow-up period and no change for the remaining years.

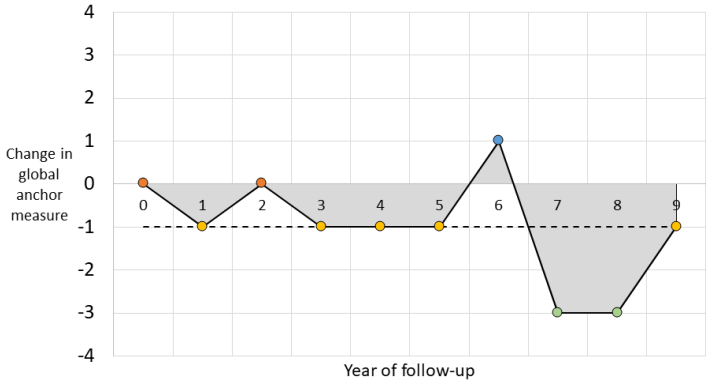
Figure 5.2 shows the change in the anchor measure for a participant who was categorised as 'minimally improved' or 'unchanged', or had an improvement greater than 'minimally improved'.

Figure 5.2 (a) shows a participant who is 'minimally improved' using the longitudinal definition. The AUC of the change in the anchor is -1.056. The yellow markers highlight that, using the single time point definition of 'minimal improvement', the participant had a 1-point improvement from baseline at years 1, 3, 4, 5 and 9. This participant will also have been included in the single time point analysis as 'unchanged' in year 2.

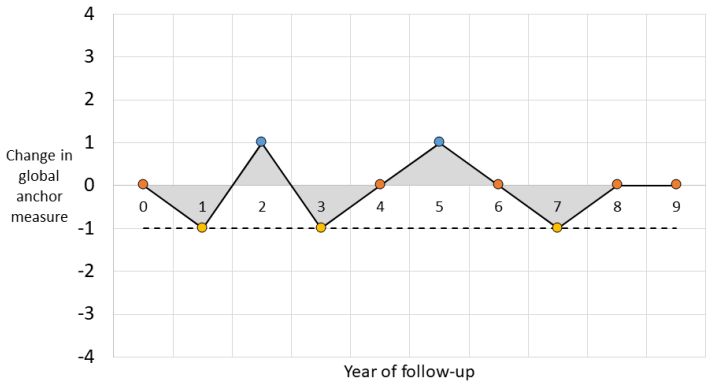
Figure 5.2 (b) shows a participant who is 'unchanged' using the longitudinal definition. The AUC of the change in the anchor is -0.111 and the graph shows that the anchor varies within 1 point either side of their baseline level. The orange markers highlight that, using the single time point definition of 'unchanged', the participant had a 0-point change from baseline at years 4, 6, 8 and 9.

Figure 5.2 (c) shows a participant who has an improvement greater than 'minimal improvement' using the longitudinal definition and therefore was included in the longitudinal analysis since their level of improvement exceeds 'minimal'. The AUC of the change in the anchor is -2.389 and the graph shows that the improvement compared to baseline exceeds 1-point at all time points except year 3. This participant was only included in the single time point analysis in year 3, where they were categorised as 'unchanged'.

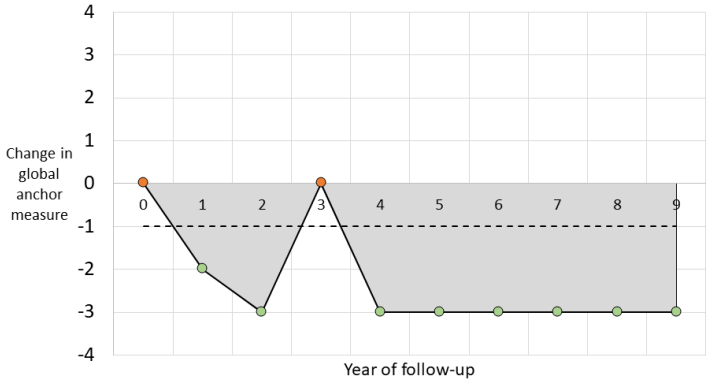
Figure 5.2: Examples of categorisations of participants using longitudinal definitions of improvement



(a) Example of the change in the anchor measure for a 'minimally improved' participant



(b) Example of the change in the anchor measure for an 'unchanged' participant



(c) Example of the change in the anchor measure for a participant with an improvement greater than 'minimally improved'

Note: A negative change in the anchor measure represents an improvement on a scale from 0 (very good) to 10 (very poor). The black dashed line shows a 1-point improvement compared to baseline. The shading of the markers is: Blue - worsening, orange - unchanged, yellow - minimal improvement, and green - large improvement.

5.2.5 Description of longitudinal methods

5.2.5.1 Adapted raw mean difference approach (single-time-point approach using longitudinal definition of improvement)

The adapted raw mean difference approach was applied at multiple follow-up time points to provide a range of possible MCID estimates, as described in Section 5.2.2.

The adapted raw mean difference approach differs from the single time point raw mean difference approach because it uses the longitudinal definition for minimal improvement, as described in Section 5.2.4.1. In the single time point approaches, the same participant could be 'minimally improved' at one time point and 'unchanged' at a different time point. For the longitudinal approaches, the definition of 'minimal improvement' and 'unchanged' were instead based on the average change over the follow-up period (summarised by the AUC). The same participants were thus defined as 'minimally improved' regardless of which time point was analysed.

Although the raw mean difference approach was used here, the ANCOVA and ROC curve methods in Section 5.2.2 could be used with the longitudinal definition of improvement in the same way. Sensitivity analysis of the raw mean difference approach with baseline adjustment (Section 5.2.5.5) is equivalent to using the longitudinal definition of improvement with the ANCOVA method.

Using the longitudinal definition for minimal improvement, the raw mean difference in the change in the WOMAC score between the 'minimally improved' and 'unchanged' groups was calculated at each time point. Separate MCID estimates were produced for the different time points. The median of these values was used to provide an overall summary of the MCID estimate.

5.2.5.2 Area-under-the-curve

AUC approaches quantify the trajectory of a participant's condition into a single value, incorporating data from all time points. A corresponding MCID estimate was calculated as the difference in the mean AUC between the 'minimally improved' group and the 'unchanged' group. The AUC was calculated based on the change score compared to the baseline WOMAC total score, such that zero area corresponds to no change from baseline on average over the follow-up period. The AUC could not be calculated for participants where the WOMAC total score was missing at baseline or was missing for all or all but one time point.

5.2.5.3 Mixed effects linear regression

Mixed effects linear regression allows repeated measures to be incorporated into a model and accounts for intra-participant correlation. A dummy variable was included for 'minimally improved' using the longitudinal definition in Section 5.2.4.1.

The mixed effects model assumed a random intercept model. I had planned to assume an exchange covariance structure for the within-participant residuals. However, as the model did not converge, an identity structure was assumed. I did not adjust for the baseline WOMAC score to correspond with the other methods. The equation for the random intercept model is shown in Equation 5.1.

$$Y_{ij} = \beta_{0,i} + \beta_1 X_i + e_{ij} \tag{5.1}$$

Y_{ij} is the change from baseline in the WOMAC score for participant i at time point j. β_{0i} is the intercept term for participant i. β_1 is the MCID estimate, which is a fixed term coefficient. X_i is an indicator term that is equal to 1 if participant i was categorised as 'minimally improved' and equal to 0 if participant i was categorised as 'unchanged' using the longitudinal definition in Section 5.2.4.1. e_{ij} is the residual error term for participant i at time point j.

5.2.5.4 Generalised estimating equation

Generalised estimating equations (GEE) were used, assuming a normal distribution and exchangeable correlation structure. The GEE approach was used because the mixed effects model did not converge when an exchangeable correlation structure in the residuals was assumed. The GEE method differs in that it is a population-averaged approach to fit a marginal distribution, rather than the conditional approach used in mixed effects methods to estimate the individual specific effect. The GEE was modelled in the same way as the mixed effects method but imposed an exchangeable correlation structure. An exchangeable correlation structure assumes that there is identical correlation between different time points, regardless of the length of time between those time points. For instance, it assumes that the correlation between year 1 and year 2 is the same as the correlation between year 1 and year 9.

5.2.5.5 Sensitivity analyses

Sensitivity analyses were planned to test the robustness of the MCID estimates, including:

i) only the participants who had complete data on the anchor measure,

ii) only the participants who had complete data on the WOMAC measure,

iii) only the participants who had complete data on both the anchor and WOMAC measures, and

iv) using each of the three approaches with adjustment for the baseline WOMAC score.

5.3 Results

5.3.1 Description of study sample

The progression cohort of the OAI dataset included 1390 participants. The baseline demographics of the sample are summarised in Table 5.1. Around three-quarters of the participants were white and one-quarter were black or African American. The age of participants was evenly spread across the range from 45 to 79 years. There were more women than men (57% vs 43%). Around a third of participants had symptomatic knee osteoarthritis in both knees. Non-prescription NSAIDs had been used by 26% of the sample. In terms of prior surgical treatment, one-third of participants had had knee surgery or arthroscopy but very few had had full or partial knee replacement (<1%). Baseline scores for the WOMAC Index indicated that the majority of participants had mild-to-moderate osteoarthritis symptoms.

Table 5.2 shows the follow-up and change scores for the WOMAC total scale and the global anchor measure. WOMAC total scores reduced from 25 to 20 points. There was similar improvement in the global anchor measure, from 3 to 2.5. Histograms for the WOMAC change scores are presented in Appendix D.1. Scores for the WOMAC subscales (pain, function and stiffness) are presented in Appendix D.2. The results in Table 5.2 suggest a gradual improvement in symptoms over time, which suggests that the participants received interventions to treat their osteoarthritis.

The correlation between the change score in the WOMAC total and the change in the global anchor measure was moderate and exceeded 0.4 at all of the time points except the 1-year assessment (Table 5.3 and Figure 5.3). The change in the anchor measure was much more highly correlated with the change in the WOMAC total score than the WOMAC follow-up score, which demonstrates that the change in the anchor measure

was strongly associated with the change in the participants' condition, rather than their current state. The association between the change in the anchor measure and change in the WOMAC was reasonably consistent over time and, at all of the time points, participants with no change in the global anchor measure had a mean change in the WOMAC total score close to zero.

Table 5.1: Demographics (n=1390)

	n	**%**	**N**
Age	Mean 61.36	SD 9.10	1390
Sex:			
Male	597	43.0%	1390
Female	793	57.1%	1390
Ethnicity:			
White or Caucasian	974	70.1%	1389
Black or African American	372	26.8%	1389
Asian	12	0.9%	1389
Other	31	2.2%	1389
WOMAC total (range 0-96)	Mean 24.82	SD 16.97	1378
Global anchor score (range 0-10)	Mean 3.00	SD 2.36	1390
CES-D (range 0-60)	Mean 7.70	SD 7.77	1363
SF-12 physical score (range 0-100)	Mean 44.52	SD 10.15	1368
SF-12 mental score (range 0-100)	Mean 53.44	SD 9.11	1368
Symptomatic knee osteoarthritis:			
Left knee only	427	30.7%	1390
Right knee only	474	34.1%	1390
Both knees	489	35.2%	1390
For included knee(s):			
Prior knee surgery or arthroscopy	451	32.5%	1389
Prior knee replacement (full or partial)	3	0.7%	450
Used on more than half of the days of the last month:			
Non-prescription NSAIDs	363	26.2%	1387
Prescription NSAIDs	124	8.9%	1386
Coxibs	143	10.3%	1385
Steroid injection in last 6 months	60	4.3%	1386

Table 5.1: Demographics (n=1390)

	n	%	N
Osteophytes and JSN (X-ray)*:			
Definite osteophytes and no JSN	388	27.9%	1389
Definite osteophytes and mild-moderate JSN	615	44.3%	1389
Definite osteophytes and severe JSN	386	27.8%	1389
Composite osteoarthritis grade (X-ray)*:			
Mild	387	27.9%	1388
Moderate	615	44.3%	1388
Severe	386	27.8%	1388
Non-knee osteoarthritis sites:			
Hip	113	8.6%	1316
Hand	243	18.3%	1329
Other osteoarthritis	153	11.5%	1327
Rheumatoid arthritis	138	10.3%	1344

* Worst knee grade used if symptomatic osteoarthritis in both knees
JSN: joint space narrowing, CES-D: Center for Epidemiological Studies - Depression, N: denominator

Table 5.2: Follow-up and change scores for WOMAC total and global anchor measure

	Follow up score			Change from baseline		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**
WOMAC total						
Baseline	24.82	16.97	1378			
1 year	21.07	16.76	1259	-2.97	13.60	1249
2 years	20.24	16.87	1188	-3.76	14.44	1178
3 years	19.93	16.89	1154	-3.75	15.73	1148
4 years	19.24	17.02	1163	-4.63	16.02	1156
5 years	20.09	17.02	1081	-3.56	16.97	1075
6 years	19.67	17.29	1035	-4.14	16.59	1030
7 years	20.55	17.74	1027	-3.17	17.98	1021
8 years	18.51	16.77	981	-5.23	17.35	975
9 years	19.71	17.07	881	-3.97	17.74	876
Global anchor measure						
Baseline	3.00	2.36	1390			
1 year	2.66	2.35	1267	-0.26	2.23	1267
2 years	2.47	2.32	1202	-0.39	2.33	1202
3 years	2.50	2.32	1174	-0.37	2.37	1174
4 years	2.52	2.44	1178	-0.36	2.53	1178
5 years	2.46	2.45	1094	-0.45	2.56	1094
6 years	2.44	2.41	1053	-0.44	2.56	1053
7 years	2.44	2.48	1049	-0.45	2.59	1049
8 years	2.29	2.30	1075	-0.63	2.52	1075
9 years	2.57	2.48	901	-0.31	2.59	901

Range of WOMAC total score is 0-96, where 0 is no symptoms and a higher score indicates worse symptoms. Range of global anchor score is 0-10, where 0 is 'very good', 10 is 'very poor' and a higher score indicates worse symptoms.

Figure 5.3: Change in WOMAC total score compared to change in global anchor measure from baseline

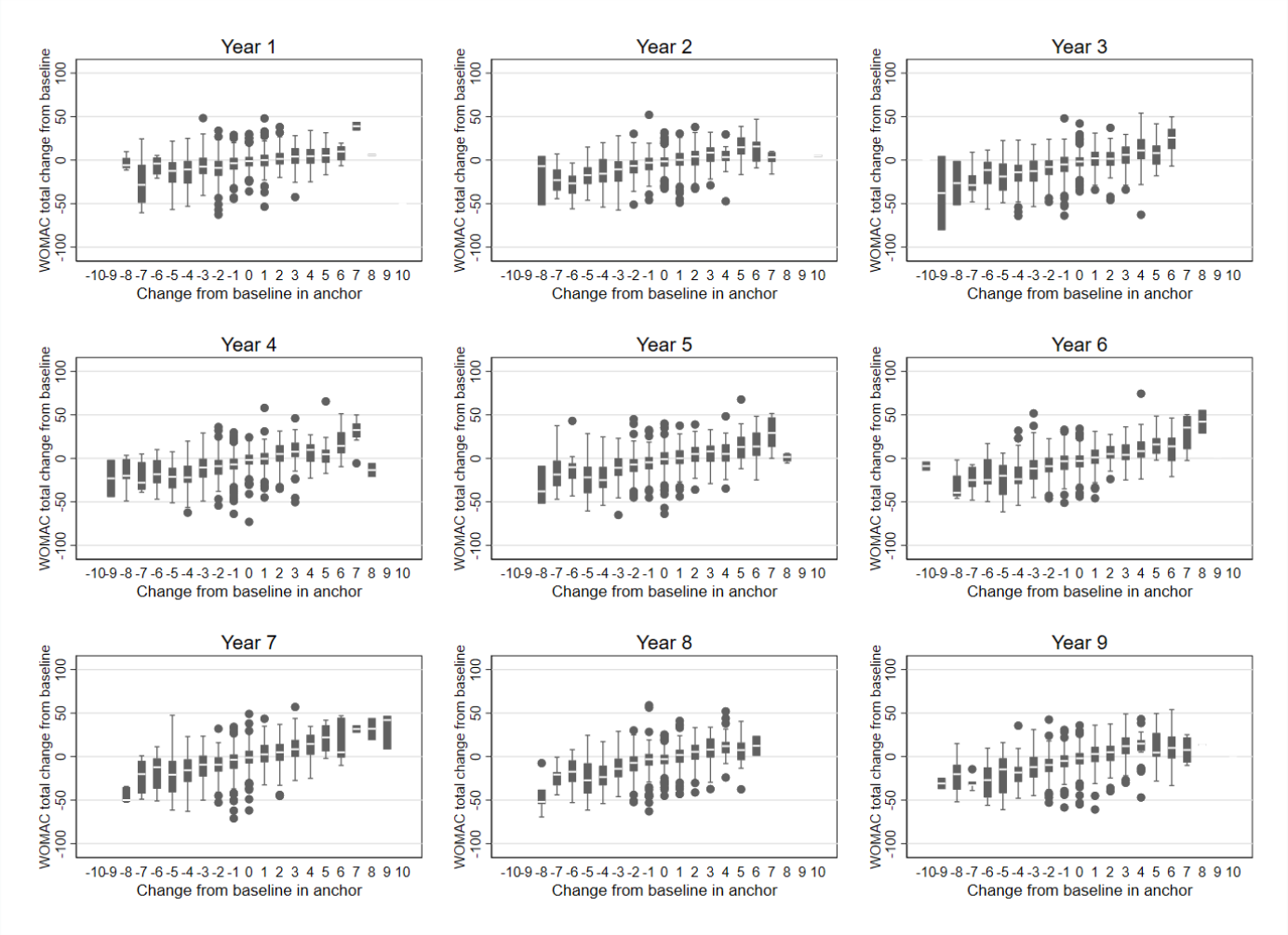


Table 5.3: Correlation between change in global anchor measure and WOMAC total scores from baseline

	Correlation between change in global anchor measure and:	
	WOMAC change score	**WOMAC follow-up score**
1 year	0.34	0.14
2 years	0.42	0.20
3 years	0.45	0.26
4 years	0.48	0.28
5 years	0.46	0.27
6 years	0.51	0.30
7 years	0.51	0.37
8 years	0.50	0.29
9 years	0.47	0.32

5.3.2 Stability of MCID estimates using single time point anchor-based approaches

The results of the single time point approaches described in Sections 5.2.2.1 to 5.2.2.3 are presented in Table 5.4 and shown graphically in Figures 5.4-5.5. The estimates differed depending on the approach used and whether the approach was based on improvement or deterioration in symptoms. In terms of improvement, the ANCOVA methods gave the smallest estimates and the ROC curve method produced the largest estimates.

Within each method, the MCID estimates did vary over time. The MCID estimates were imprecise and there were no statistically significant differences in the MCID estimate by the duration of follow-up. For improvement, there was no pattern between the MCID value and follow-up time point. For deterioration, the raw mean difference and ANCOVA methods suggested that larger MCID estimates were produced for later follow-up time points, although the relationship between the MCID estimate and time was not monotonic.

MCID estimates were also calculated based on the change at each time point from the previous year, artificially splitting the follow-up period into 1-year segments (Table 5.5 and Figures 5.6-5.7). Comparing Figures 5.6-5.7 and Figures 5.4-5.5 suggests that the variability of MCID estimates for a 1-year change over the different segments was similar to the level of variability of MCID estimates for the change from baseline over different follow-up durations.

Below is a summary of the range of the MCID estimates across the different follow up time points for the different single time point methods.

For improvement compared to baseline (Table 5.4 and Figure 5.4):

- Using the raw mean difference: the median MCID value was 3.44 (at year 9). MCID estimates ranged from 1.03 (at year 8) to 5.55 (at year 4).
- Using the ANCOVA method: the median MCID value was 1.97 (at year 1). MCID estimates ranged from -0.17 (at year 8) to 4.10 (at year 4).
- Using the ROC curve method: the median MCID value was 4.10 (at year 3). MCID estimates ranged from 3.55 (at year 2) to 7.80 (at year 6).

For deterioration compared to baseline (Table 5.4 and Figure 5.5):

- Using the raw mean difference: the median MCID value was -3.58 (at year 8). MCID estimates ranged from -4.14 (at year 9) to -1.36 (at year 5).
- Using the ANCOVA method: the median MCID value was -4.20 (at year 3). MCID estimates ranged from -5.95 (at year 9) to -2.05 (at year 1).
- Using the ROC curve method: the median MCID value was -2.00 (at year 1). MCID estimates ranged from -4.05 (at year 9) to -0.40 (at year 5).

For improvement compared to previous year (Table 5.5 and Figure 5.6):

- Using the raw mean difference: the median MCID value was 1.52 (year 3 compared to year 2).

  MCID estimates ranged from 0.42 (year 8 compared to year 7) to 3.31 (year 5 compared to year 4).

- Using the ANCOVA method: the median MCID value was 0.54 (year 2 compared to year 1).

  MCID estimates ranged from -0.88 (year 8 compared to year 7) to 2.02 (year 5 compared to year 4).

- Using the ROC curve method: the median MCID value was 2.10 (at year 5).

  MCID estimates ranged from 1.50 (at year 3) to 3.95 (at year 1).

For deterioration compared to previous year (Table 5.5 and Figure 5.7):

- Using the raw mean difference: the median MCID value was -3.06 (year 4 compared to year 3).

  MCID estimates ranged from -4.66 (year 8 compared to year 7) to -1.25 (year 6 compared to year 5).

- Using the ANCOVA method: the median MCID value was -3.82 (year 7 compared to year 6).

  MCID estimates ranged from -5.24 (year 8 compared to year 7) to -2.05 (year 1 compared to baseline).

- Using the ROC curve method: the median MCID value was -2.40 (at year 3).

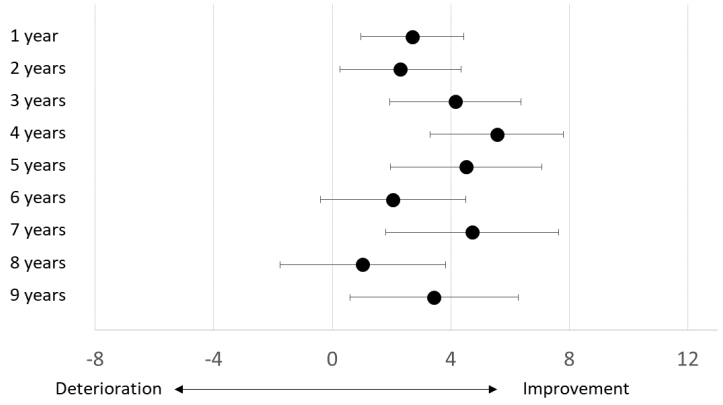
  MCID estimates ranged from -4.15 (at year 5) to -1.05 (at year 2).

Table 5.4: Single time point MCID estimates for WOMAC total score (change from baseline)

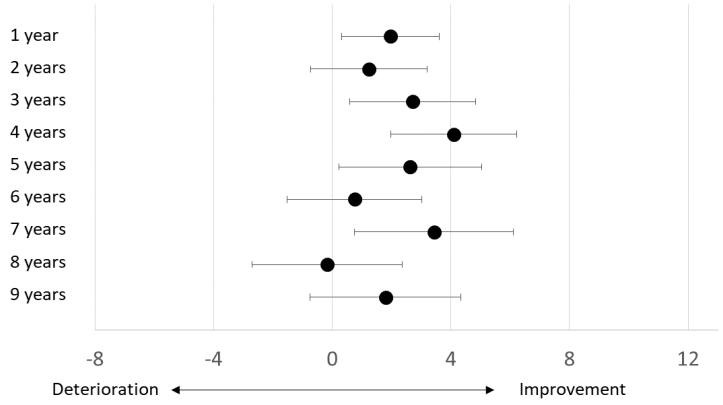
	Raw mean difference			ANCOVA method			ROC curve method		
	Estimate	**95% CI**	**n**	**Estimate**	**95% CI**	**n**	**Estimate**	**95% CI**	**n**
Improvement:									
1 year	2.70	(0.97, 4.43)	562	1.97	(0.32, 3.62)	562	3.95	(1.67, 6.23)	562
2 years	2.30	(0.27, 4.34)	493	1.23	(-0.73, 3.19)	493	3.55	(2.40, 4.70)	493
3 years	4.15	(1.93, 6.38)	484	2.71	(0.59, 4.83)	484	4.10	(1.62, 6.57)	484
4 years	5.55	(3.30, 7.80)	451	4.10	(1.98, 6.21)	451	3.70	(0.55, 6.85)	451
5 years	4.52	(1.97, 7.07)	420	2.63	(0.22, 5.03)	420	3.70	(2.09, 5.31)	420
6 years	2.05	(-0.40, 4.51)	419	0.75	(-1.52, 3.02)	419	7.80	(3.36, 12.24)	419
7 years	4.72	(1.80, 7.63)	403	3.43	(0.75, 6.11)	403	4.60	(1.75, 7.45)	403
8 years	1.03	(-1.77, 3.83)	365	-0.17	(-2.70, 2.36)	365	6.00	(3.18, 8.82)	365
9 years	3.44	(0.61, 6.27)	345	1.79	(-0.76, 4.34)	345	4.60	(2.24, 6.96)	345
Deterioration:									
1 year	-1.55	(-3.45, 0.34)	534	-2.05	(-3.87, -0.23)	534	-2.00	(-5.05, 1.05)	534
2 years	-1.99	(-4.15, 0.17)	483	-3.33	(-5.37, -1.30)	483	-2.70	(-5.46, 0.06)	483
3 years	-3.78	(-6.05, -1.50)	427	-4.20	(-6.38, -2.01)	427	-1.70	(-3.61, 0.21)	427
4 years	-1.90	(-4.19, 0.39)	415	-2.62	(-4.79, -0.44)	415	-2.55	(-5.40, 0.30)	415
5 years	-1.36	(-4.34, 1.62)	372	-2.49	(-5.28, 0.30)	372	-0.40	(-4.11, 3.31)	372
6 years	-4.07	(-6.77, -1.37)	367	-4.92	(-7.42, -2.43)	367	-1.00	(-2.74, 0.74)	367
7 years	-3.81	(-6.79, -0.84)	359	-5.11	(-7.96, -2.26)	359	-2.40	(-1.39, 6.19)	359
8 years	-3.58	(-6.50, -0.65)	317	-5.35	(-8.05, -2.65)	317	-1.00	(-3.70, 1.70)	317
9 years	-4.14	(-7.26, -1.03)	328	-5.95	(-8.79, -3.12)	328	-4.05	(-8.14, 0.04)	328

All 3 methods include the sub-sample of participants where the change in the global anchor was 0 or -1 for improvement and 0 or +1 for deterioration. Estimate gives change score relating to a minimally important improvement or deterioration. A positive estimate score indicates a greater reduction in symptoms.

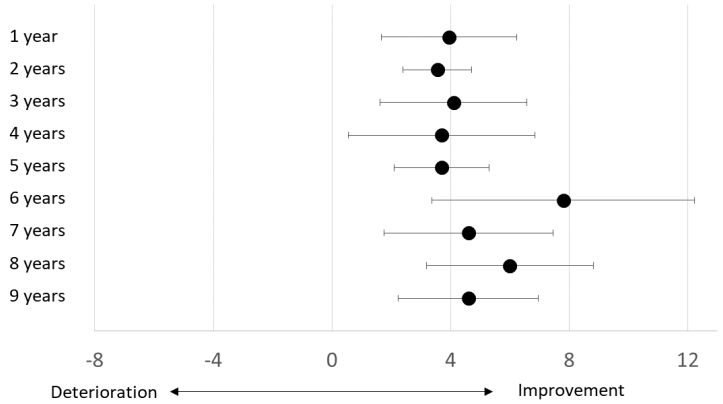
Figure 5.4: MCID estimates (and 95% confidence interval) for improvement compared to baseline



(a) Raw mean difference: Improvement

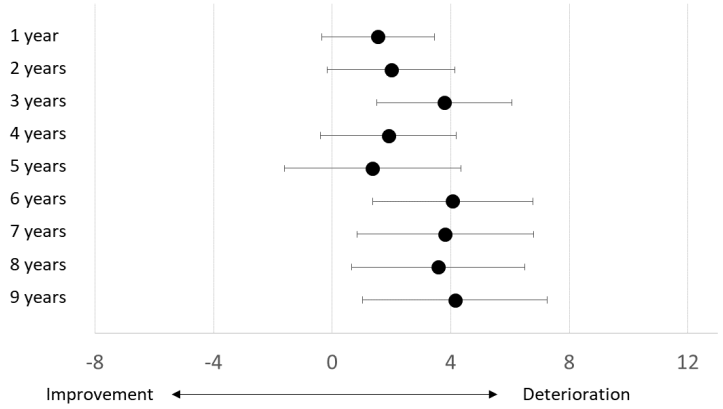


(b) ANCOVA: Improvement

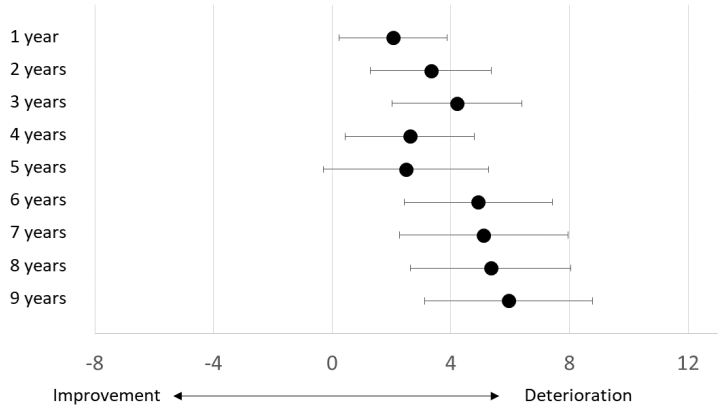


(c) ROC curve: Improvement

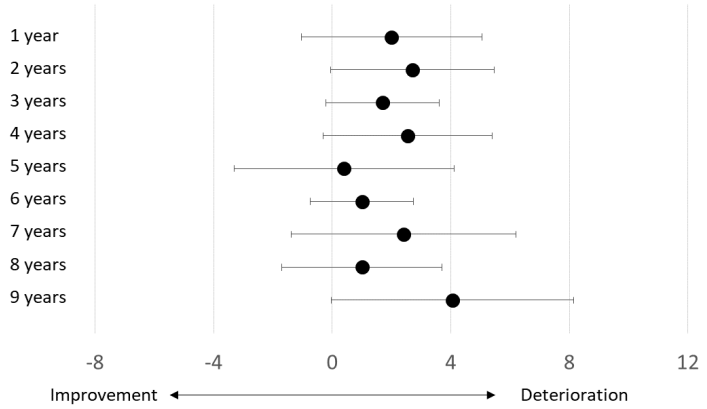
Figure 5.5: MCID estimates (and 95% confidence interval) for deterioration compared to baseline



(a) Raw mean difference: Deterioration



(b) ANCOVA: Deterioration



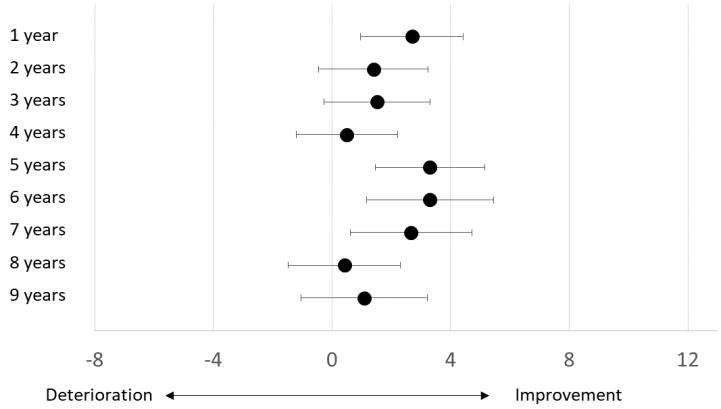
(c) ROC curve: Deterioration

Table 5.5: Single time point MCID estimates for WOMAC total score (change from previous year)

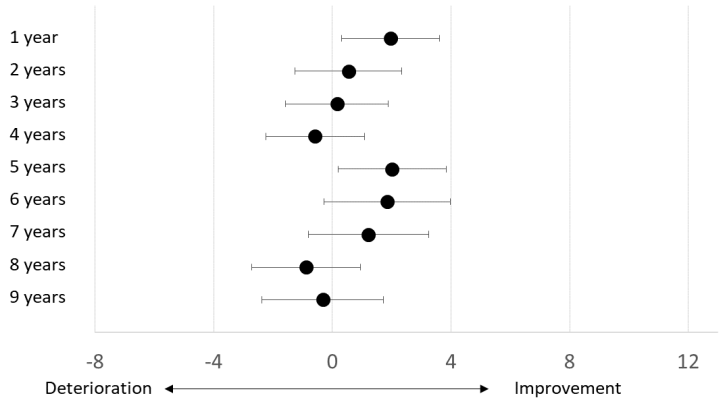
	Raw mean difference			**ANCOVA method**			**ROC curve method**		
	Estimate	**95% CI**	**n**	**Estimate**	**95% CI**	**n**	**Estimate**	**95% CI**	**n**
Improvement:									
1 year	2.70	(0.97, 4.43)	562	1.97	(0.32, 3.62)	562	3.95	(1.66, 6.24)	562
2 years	1.39	(-0.45, 3.24)	515	0.54	(-1.25, 2.33)	515	1.90	(0.59, 3.21)	515
3 years	1.52	(-0.27, 3.30)	489	0.16	(-1.57, 1.89)	489	1.50	(-1.38, 4.38)	489
4 years	0.50	(-1.21, 2.22)	512	-0.58	(-2.25, 1.09)	512	1.55	(-0.13, 3.23)	512
5 years	3.31	(1.45, 5.16)	462	2.02	(0.20, 3.84)	462	2.10	(0.92, 3.28)	462
6 years	3.30	(1.16, 5.45)	445	1.86	(-0.28, 3.99)	445	3.00	(1.78, 4.22)	445
7 years	2.67	(0.62, 4.71)	457	1.22	(-0.80, 3.25)	457	2.90	(0.87, 4.93)	457
8 years	0.42	(-1.47, 2.30)	438	-0.88	(-2.71, 0.95)	438	2.00	(-0.32, 4.32)	438
9 years	1.09	(-1.05, 3.23)	396	-0.32	(-2.37, 1.72)	396	2.50	(-0.26, 5.26)	396
Deterioration:									
1 year	-1.55	(-3.45, 0.34)	534	-2.05	(-3.87, -0.23)	534	-2.00	(-4.90, 0.90)	534
2 years	-2.25	(-4.24, -0.27)	497	-3.44	(-5.39, -1.48)	497	-1.05	(-2.49, 0.39)	497
3 years	-3.44	(-5.27, -1.61)	480	-4.57	(-6.35, -2.79)	480	-2.40	(-3.56, -1.24)	480
4 years	-3.06	(-4.82, -1.31)	483	-3.76	(-5.41, -2.10)	483	-1.80	(-3.33, -0.27)	483
5 years	-3.89	(-5.80, -1.98)	487	-4.74	(-6.62, -2.87)	467	-4.15	(-5.93, -2.37)	467
6 years	-1.25	(-3.29, 0.78)	469	-2.08	(-4.04, -0.12)	469	-3.00	(-5.45, -0.55)	469
7 years	-2.64	(-4.65, -0.63)	455	-3.82	(-5.82, -1.82)	455	-3.10	(-4.73, -1.47)	455
8 years	-4.66	(-6.69, -2.64)	416	-5.24	(-7.16, -3.33)	416	-1.50	(-3.02, 0.02)	416
9 years	-3.85	(-5.82, -1.88)	421	-5.05	(-6.97, -3.12)	421	-3.25	(-4.77, -1.73)	421

All 3 methods include the sub-sample of participants where the change in the global anchor was 0 or -1 for improvement and 0 or +1 for deterioration. Estimate gives change score relating to a minimally important improvement or deterioration. A positive estimate score indicates a greater reduction in symptoms.

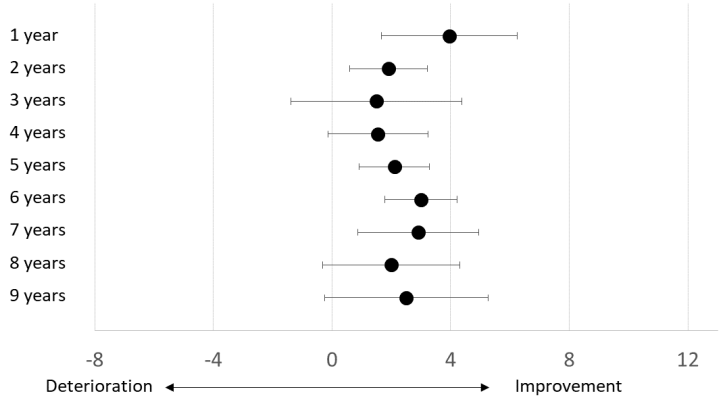
Figure 5.6: MCID estimates (and 95% confidence interval) for improvement compared to previous year



(a) Raw mean difference: Improvement

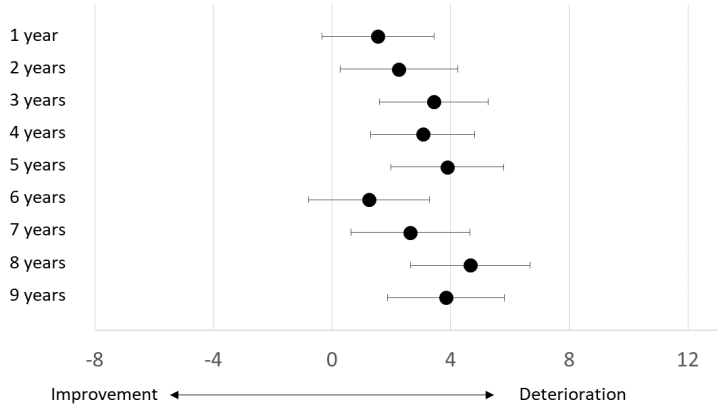


(b) ANCOVA: Improvement

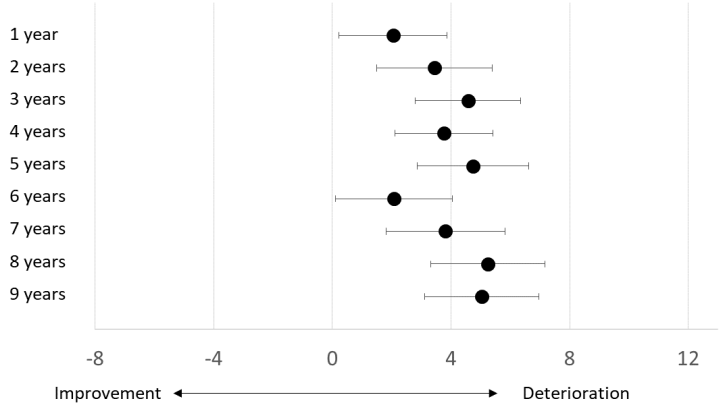


(c) ROC curve: Improvement

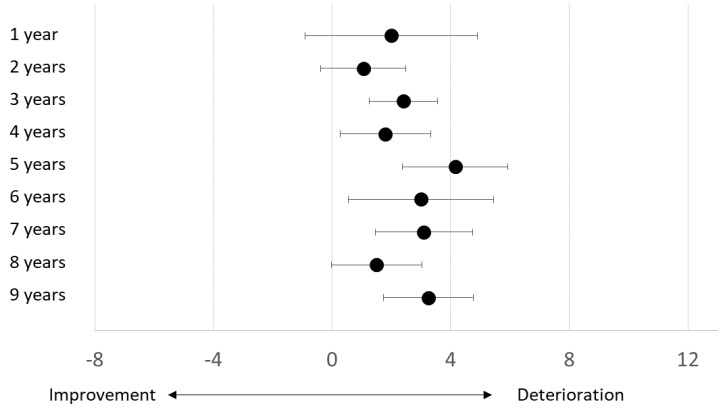
Figure 5.7: MCID estimates (and 95% confidence interval) for deterioration compared to previous year



(a) Raw mean difference: Deterioration



(b) ANCOVA: Deterioration



(c) ROC curve: Deterioration

5.3.3 Subgroup analysis of single time point methods

The results of the subgroup analysis for the raw mean difference method based on improvement compared to baseline are presented in Table 5.6. Within each subgroup category, MCID estimates remained variable over time. However, as for the overall analysis, there was no clear pattern. There was no association between the MCID estimate and age or sex for the single time point methods (Table 5.6 and Appendix D.3). For the ANCOVA and ROC curve methods, MCID estimates were higher for participants with more severe disease at baseline. Estimates for the ROC curve method were much higher in the older age group and participants with more severe disease (Table 5.6 and Appendix D.3).

Table 5.6: Subgroup analysis of MCID estimates for improvement in the WOMAC total score using the raw mean difference method based on change from baseline

	Raw mean difference			**Raw mean difference**			**Raw mean difference**		
	β	**95% CI**	**n**	β	**95% CI**	**n**	β	**95% CI**	**n**
	Age: 45-56			**Age: 57-66**			**Age: 67-79**		
1 year	3.88	(0.99, 6.76)	186	1.52	(-1.51, 4.55)	197	2.94	(-0.21, 6.09)	179
2 years	-0.29	(-3.50, 2.92)	182	3.20	(-0.26, 6.66)	165	4.85	(0.75, 8.94)	146
3 years	5.16	(1.85, 8.46)	170	0.29	(-3.69, 4.27)	152	6.55	(2.27, 10.84)	162
4 years	3.84	(0.55, 7.14)	171	6.35	(2.31, 10.40)	149	6.86	(2.25, 11.47)	131
5 years	7.57	(3.58, 11.56)	144	2.20	(-2.04, 6.44)	154	4.60	(-0.64, 9.84)	122
6 years	-1.37	(-5.22, 2.47)	155	5.95	(1.86, 10.03)	151	1.89	(-3.14, 6.92)	113
7 years	4.65	(0.22, 9.08)	137	10.69	(5.57, 15.82)	151	-2.04	(-7.73, 3.64)	115
8 years	-1.61	(-5.77, 2.55)	140	1.89	(-3.26, 7.05)	121	4.59	(-1.06, 10.24)	104
9 years	2.96	(-1.43, 7.36)	131	2.77	(-1.90, 7.44)	127	5.50	(-0.75, 11.75)	87
	WOMAC: 0-12			**WOMAC: 13-36**			**WOMAC: 37-96**		
1 years	-0.79	(-2.58, 1.00)	207	4.18	(1.44, 6.92)	253	2.68	(-2.06, 7.42)	102
2 years	1.08	(-1.16, 3.33)	186	1.30	(-1.86, 4.46)	234	1.95	(-4.44, 8.34)	73
3 years	1.60	(-1.27, 4.47)	166	0.57	(-2.34, 3.48)	245	12.96	(4.92, 21.00)	73
4 years	1.64	(-0.13, 3.42)	177	5.53	(2.24, 8.82)	197	7.57	(-1.07, 16.20)	77
5 years	0.98	(-2.11, 4.07)	161	6.07	(2.50, 9.64)	193	-0.67	(-9.12, 7.78)	66
6 years	-0.52	(-3.17, 2.12)	163	2.38	(-1.29, 6.07)	195	0.37	(-7.17, 7.92)	61
7 years	0.79	(-2.10, 3.69)	157	4.70	(0.82, 8.60)	184	8.78	(-2.56, 20.13)	62
8 years	-2.42	(-5.54, 0.69)	146	2.21	(-1.53, 5.95)	169	3.95	(-6.25, 14.15)	50
9 years	3.44	(0.36, 6.54)	151	0.07	(-3.94, 4.07)	138	5.32	(-4.39, 15.03)	56

Table 5.6: Subgroup analysis of MCID estimates for improvement in the WOMAC total score using the raw mean difference method based on change from baseline

	Raw mean difference			**Raw mean difference**			**Raw mean difference**		
	β	**95% CI**	**n**	β	**95% CI**	**n**	β	**95% CI**	**n**
	Male			**Female**					
1 year	2.06	(-0.39, 4.51)	245	3.17	(0.75, 5.59)	317			
2 years	3.04	(0.12, 5.96)	224	1.54	(-1.32, 4.41)	269			
3 years	4.95	(1.76, 8.14)	207	3.65	(0.57, 6.73)	277			
4 years	5.62	(2.53, 8.71)	199	5.53	(2.28, 8.77)	252			
5 years	4.06	(0.50, 7.62)	183	4.87	(1.28, 8.47)	237			
6 years	2.37	(-1.04, 5.77)	191	1.89	(-1.63, 5.41)	228			
7 years	5.90	(2.33, 9.46)	193	3.77	(-0.81, 8.35)	210			
8 years	4.38	(0.90, 7.86)	177	-2.12	(-6.45, 2.22)	188			
9 years	1.96	(-1.91, 5.83)	158	5.04	(0.98, 9.11)	187			

5.3.4 Comparison of single time point methods

Table 5.7 compares the three single time point methods in terms of the use of data and ease of analysis and interpretation.

Table 5.7: Comparison of single time point MCID calculation methods

	Single time point method		
	Raw mean difference	**ANCOVA**	**ROC**
Proportion of data used	Only includes people with 'minimal improvement' or 'no change' response to anchor at corresponding time point		
Handling of missing data	Complete case: Excluded if missing data for primary outcome or anchor measure at corresponding time point		
Adjustment for participant characteristics	Cannot adjust for baseline characteristics	Allows adjustment for baseline characteristics	Cannot adjust for baseline characteristics
Ease of analysis	Very easy	Easy	Easy
Ease of interpretation	Provides value on scale of outcome measure with confidence interval		
	May be difficult to interpret if inconsistent at different follow-up time points		
		Need to translate to common baseline score for true comparability over time	May not consider concept of 'minimal' score in choice of optimal cut-off
	Based on group-summary change scores	Based on group-summary change scores	Based on individual-level change scores
Multiplicity	Requires multiple analyses to examine multiple follow-up time points		

5.3.5 Longitudinal methods for MCID calculation

Using the longitudinal definitions, 245 participants were classified as 'minimally improved' or 'unchanged' (n=245, 17.6%). At baseline, the 'minimally improved' and 'unchanged' groups were similar in terms of age and sex. The participants classified as 'minimally improved' had more severe disease symptoms according to the WOMAC total score and worse scores on the global anchor measure (Table 5.8).

Table 5.8: Demographics by longitudinal definition of minimal improvement

	Minimally improved			Unchanged		
	n	**%**	**N**	**n**	**%**	**N**
Age	Mean 60.20	SD 10.05	96	Mean 61.32	SD 8.91	149
Sex:						
Male	44	45.8%	96	66	44.3%	149
Female	52	54.2%	96	83	55.7%	149
Ethnicity:						
White or Caucasian	66	68.8%	96	119	79.9%	149
Black or African American	30	31.2%	96	28	18.8%	149
Asian	0	0%	96	1	0.7%	149
Other	0	0%	96	1	0.7%	149
WOMAC total (range 0-96)	Mean 29.49	SD 17.7	96	Mean 19.13	SD 15.40	149
Global anchor score (range 0-10)	Mean 3.80	SD 1.87	96	Mean 1.80	SD 1.75	149
CES-D (range 0-60)	Mean 7.19	SD 6.54	93	Mean 6.32	SD 7.28	145
SF-12 physical score (range 0-100)	Mean 43.21	SD 10.67	95	Mean 47.46	SD 8.67	148
SF-12 mental score (range 0-100)	Mean 54.28	SD 8.71	95	Mean 54.34	SD 7.86	148
Symptomatic knee osteoarthritis:						
Left knee only	38	39.6%	96	55	36.9%	149
Right knee only	31	32.3%	96	50	33.6%	149
Both knees	27	28.1%	96	44	29.5%	149
For included knee(s):						
Prior knee surgery or arthroscopy	38	39.6%	96	53	35.6%	149
Prior knee replacement (full or partial)	1	2.6%	38	1	1.9%	52

Table 5.8: Demographics by longitudinal definition of minimal improvement

	Minimally improved			Unchanged		
	n	**%**	**N**	**n**	**%**	**N**
Used on more than half of days in last month:						
Non-prescription NSAIDs	22	22.9%	96	47	31.8%	148
Prescription NSAIDs	5	5.2%	96	12	8.1%	148
Coxibs	5	5.1%	96	8	5.4%	148
Steroid injection in last 6 months	3	3.1%	96	6	4.0%	149
Osteophytes and JSN (X-ray)*:						
Definite osteophytes and no JSN	21	21.9%	96	38	25.7%	148
Definite osteophytes and mild-moderate JSN	43	44.8%	61	22	41.2%	148
Definite osteophytes and severe JSN	32	33.3%	96	49	33.1%	148
Non-knee osteoarthritis sites:						
Hip	7	7.6%	92	11	7.6%	144
Hand	19	20.4%	93	26	18.2%	143
Other osteoarthritis	15	16.1%	93	16	11.2%	143
Rheumatoid arthritis	11	11.8%	93	13	8.9%	146

Worst knee grade used if symptomatic osteoarthritis in both knees. JSN: joint space narrowing, CES-D: Center for Epidemiological Studies - Depression, N: denominator.

5.3.5.1 Adapted raw mean difference approach

The MCID estimates for each time point calculated using the raw (unadjusted) mean difference in change in WOMAC score from baseline are shown in Table 5.9.

Table 5.9: MCID estimates calculated using the adapted raw mean difference approach

Year	MCID	(95% CI)
1	-0.82	(-3.71 to 2.06)
2	2.74	(-0.59 to 6.06)
3	4.17	(0.48 to 7.86)
4	3.92	(-0.08 to 7.93)
5	5.78	(1.69 to 9.86)
6	5.12	(1.03 to 8.93)
7	4.07	(-0.06 to 8.20)
8	6.02	(1.80 to 10.25)
9	2.72	(-2.33 to 7.77)

The median value was 4.07 (range -0.82 to 6.02).

5.3.5.2 Area-under-the-curve

Using the mean difference in AUC of the change from baseline in the WOMAC score per year of follow-up produced an MCID estimate of 3.11 (95% CI: 0.60 to 5.62).

5.3.5.3 Mixed effects linear regression

Using mixed effects regression based on a random-intercept model with a time-constant coefficient for minimal improvement produced an MCID estimate of 3.13 (95% CI: 0.56 to 5.70).

5.3.5.4 Generalised estimating equation

Using a GEE with a time-constant coefficient for minimal improvement, assuming an exchangeable correlation structure, produced an MCID estimate of 3.13 (95% CI: 0.70 to 5.56).

5.3.5.5 Sensitivity analyses

Of the participants with complete data on the WOMAC measure, only two participants were missing data on the anchor measure (n=2/726) and neither was classified as 'minimally improved' or 'unchanged' using the longitudinal definitions. Therefore, the sensitivity analysis was conducted on the participants with complete baseline and follow-up data for the WOMAC and global anchor measure. Of those with complete data, 148 participants were defined as 'minimally improved' or 'unchanged' using the longitudinal definitions in Section 5.2.4.1.

The results of the sensitivity analysis based on complete data did not differ significantly from the main analysis, with results differing from the main analysis by approximately 0.5 (Table 5.10).

Adjustment for baseline WOMAC score considerably reduced the MCID estimate from 3.1-4.1 in the main analysis to 0.2-1.0 in the analysis with baseline adjustment (Table 5.11). None of the estimates using baseline adjustment were statistically significant for any of the methods.

Table 5.10: MCID estimates from longitudinal analysis: Sensitivity analysis using complete data

Method	**Main analysis**		**Complete data**	
Adapted raw mean difference	4.07	(range -0.82 to 6.02)	3.57	(range 0.47 to 6.81)
Area under the curve	3.11	(95% CI 0.60 to 5.62)	3.58	(95% CI 0.41 to 6.75)
Mixed effects linear regression	3.13	(95% CI 0.56 to 5.70)	3.35	(95% CI -0.04 to 6.74)
Generalised estimating equation	3.13	(95% CI 0.70 to 5.56)	3.35	(95% CI 0.34 to 6.37)

Table 5.11: MCID estimates from longitudinal analysis: Sensitivity analysis using baseline adjustment

Method	**Main analysis**		**Baseline adjustment**	
Adapted raw mean difference	4.07	(range -0.82 to 6.02)	0.96	(range -3.61 to 2.34)
Area under the curve	3.11	(95% CI 0.60 to 5.62)	0.19	(95% CI -2.13 to 2.51)
Mixed effects linear regression	3.13	(95% CI 0.56 to 5.70)	0.24	(95% CI -2.02 to 2.51)
Generalised estimating equation	3.13	(95% CI 0.70 to 5.56)	0.24	(95% CI -1.97 to 2.45)

5.3.6 Comparison of longitudinal methods

Table 5.12 compares the four longitudinal methods in terms of their use of data and ease of analysis and interpretation.

Table 5.12: Comparison of longitudinal MCID calculation methods

	Longitudinal method:		
	Adapted raw mean difference	**Mixed effects and GEE**	**AUC**
Handling of missing data	Requires follow-up score at corresponding time point	Included if ≥1 follow-up score	Included if ≥1 follow-up score
	Complete case (if no imputation prior to analysis)	Missing observations imputed	Usually assumes linear trend between adjacent discrete time points
	All exclude people if data required for definition of improvement is missing		
Adjustment for participant characteristics	Allows adjustment for baseline characteristics	Allows adjustment for baseline characteristics	Allows adjustment for baseline characteristics
Ease of analysis	Easy	More difficult - requires assumptions on covariance structure Issues with lack of convergence for mixed effects method	More difficult - facilitated by pkexamine command in Stata software

Table 5.12: Comparison of longitudinal MCID calculation methods

	Longitudinal method:		
	Adapted raw mean difference	**Mixed effects and GEE**	**AUC**
Ease of interpretation	Provides value on scale of outcome measure with confidence interval	Provides value on scale of outcome measure with confidence interval	More difficult to translate results on difference in AUC in a way that patients can understand
	May be difficult to interpret if inconsistent at different follow-up time points		Provides a visual representation of change over time
Multiplicity	Requires multiple analyses to examine multiple follow-up time points	Single analysis model	Single analysis model

5.4 Discussion

5.4.1 Summary of findings

This study aimed to assess the variability of MCID estimates over time and explore whether longitudinal methods could be used to calculate MCID estimates, using 1390 participants from the OAI cohort dataset of participants diagnosed with osteoarthritis. The cohort included participants who had received any or no treatment for their osteoarthritis symptoms. However, the improvement in symptoms over time suggests that participants were receiving some form of treatment for their osteoarthritis.

The results of the single time point demonstrated that there was variability in the MCID estimates over the follow-up period. The results for deterioration suggested that later follow-up time points may correspond with larger MCID values. However, because of the imprecision of the estimates, the differences in the MCID estimates by time point were not compared using statistical tests. It was inferred from the overlapping confidence intervals that the MCID estimates were not significantly different. Large confidence intervals also precluded more detailed analysis of time trends. From examining the point estimates, there was no clear pattern in the estimates over time. Moreover, the variation between MCID estimates based on different 1-year periods had similar variability to the MCID estimates for different follow up periods. This suggests that the differences were not based on the length of the follow-up period.

The estimates were more consistent for the ROC curve method. However, it is unclear whether this method measured the intended concept of a *minimal* difference, as was discussed by Angst *et al.* [375]. The ROC curve method may produce an estimate that differentiates well between 'minimally improved' and 'not changed' participants because it classifies only those with very high improvement as 'minimally improved'.

The MCID estimates were not symmetric between improvement and deterioration. As the MCID values were commonly larger for deterioration, the participants likely saw a smaller improvement in symptoms as more important than the equivalent worsening in symptoms. This may be because people living with osteoarthritis expect their symptoms to deteriorate over time. Therefore, a small improvement in their symptoms compared to baseline represents a larger improvement relative to their expected level.

This study has also demonstrated that it is feasible to calculate MCID estimates using longitudinal data using four different methods. However, the longitudinal methods did not improve the precision of the MCID estimates produced. The lack of precision remaining in the longitudinal estimates could be because fewer participants were eligible based on the longitudinal definition of 'minimal improvement' and 'unchanged'. Using the longitudinal definition for improvement, the longitudinal methods produced more precise estimates than corresponding single time point methods.

The MCID estimates produced by the different longitudinal methods were relatively consistent, with an MCID value between 3 and 4 points and a confidence interval around 0 to 6 points. Sensitivity analysis including only participants with complete data produced very similar estimates. The AUC and mixed effects methods produced MCID estimates that were statistically significant (the confidence interval did not include zero). However, after adjustment for the baseline WOMAC score, the point estimates were lower and none were significantly different from zero.

5.4.1.1 Comparison of single time point methods

The different single time point approaches assessed different concepts. The ROC curve method produced the most consistent MCID estimates across the different time points. The ANCOVA and raw mean difference methods consider the difference

in the average group level results, whereas the ROC curve considers differences in the individual-level change scores. The ROC curve method aims to detect the optimal cut-off between the 'minimally improved' and 'not changed' groups. For example, the optimal cut-off to differentiate between the two groups may be very high if this produces very high specificity (few false positives). In this example, a difference higher than the cut-off is likely to be important, yet the minimal clinically important difference estimate is likely to be much lower. Therefore, the ROC curve method may not capture the '*minimum* clinically important difference' concept, as was argued by Angst *et al.* [375].

Angst *et al.* argued that regression modelling could produce less biased estimates because adjustment for participant characteristics, such as differences in age or disease severity, reduces confounding [375]. Adjusting for baseline characteristics is recommended in the analysis of randomised trials, where participants have been allocated at random to the intervention or control arm and baseline differences are due to chance [377]. However, in the calculation of MCID estimates, participants are not assigned randomly to the 'minimally improved' and 'no change' groups. These two groups may be inherently different at baseline, for example, due to ceiling effects. Participants who have the best possible score at baseline (a score of 0 on the global assessment anchor measure, where a lower score is better) cannot improve and, thus, cannot be in the 'minimally improved' group. In the single time point analyses, around one-third of participants in the 'unchanged' group at each time point had the best possible score at baseline. Therefore, adjusting for baseline outcome score would likely attenuate the MCID estimate and produce an MCID estimate that is biased. This view is supported by Miller and Chapman who argue against using ANCOVA when group assignment is correlated with the covariate, because ANCOVA cannot be used to 'unconfound' the relationship [378]. The baseline adjustment removes meaningful variance from the group assignment because it is likely that the difference in baseline

scores between the minimally improved' and no change' groups is inherent and not due to random chance.

As discussed by Fayers and Hays, increasing or decreasing correlation between the baseline score and change scores would change the regression slope [379]. The MCID estimate could differ because the correlation between the baseline score and change scores is varied. Standard baseline adjustment commonly assumes a linear trend between the baseline score and change scores, which may be overly simplistic. Mis-specification of this relationship could lead to a biased MCID estimate. In addition, differences in the estimation of the correlation between baseline and follow-up scores could cause inconsistency over time in the MCID value. For example, if the correlation structure was autoregressive residual (AR(1)), then the correlation between baseline and follow-up scores would be lower for follow-up time points further away from the baseline assessment time point [380, 381].

Attenuation due to ceiling effects and correlation between baseline and change scores could provide biased MCID estimates when using the ANCOVA method. This could explain the very small, and sometimes negative, MCID estimates when the ANCOVA method was used. Moreover, when baseline adjustment is used, the same estimates are produced if change scores or follow-up scores are used as the dependent variable. This could be an argument against the use of baseline adjustment as we are interested in the effect on the change in the participants' condition.

The above arguments support the use of the raw mean difference method to calculate the MCID estimate when data are only available at a single follow-up time point.

5.4.1.2 Comparison of longitudinal methods

All four of the longitudinal methods produced consistent estimates of the MCID. The problems with conducting a statistical test at each time point when analysing

longitudinal data have been widely discussed [382, 383]. It is difficult to interpret results if they are not consistent over time. Conducting separate analyses for each follow-up time point loses information about the correlation between participants' scores over time. Carrying out multiple analyses increases the possibility of a 'chance' finding [383]. Using single time point approaches or the adapted raw mean difference approach may thus be ill-advised due to the resulting multiplicity issues. Consistency over time could also be assessed using longitudinal methods that include interaction terms between the time point and 'minimal improvement' variable.

The AUC approach avoids multiplicity issues and provides a visual representation of the change in the outcome score for an individual participant. However, information is lost by converting the change in a participant's condition over time into a single value. Chapter 4 found that the duration of a treatment effect was important to participants, but that the importance of increasing the duration of effect was not consistent across the scale. For instance, increasing the duration from 3 months to 6 months was not important but increasing duration from 0 to 3 months was highly important. The same problems occurred for the WOMAC Index, in that reducing symptoms from 'mild' to 'none' was not important but reducing symptoms from 'severe' to 'moderate' was highly important. The AUC summarises the change over time into a single value, so that a 10-point reduction lasting 1 year has the same value as a 5-point reduction lasting 2 years. In reality, a participant may value these very differently. Therefore, summarising the change in a participant's condition using the AUC may be too reductive, as using a single value to represent the participant's change over time loses important details about how that change in outcome occurs over the time dimension.

Mixed effects and GEE methods analyse change scores at the individual time point, accounting for the within-participant correlation between the scores and avoiding

multiplicity issues. Similar to the AUC method, the longitudinal definition of 'minimal improvement' used in all of the longitudinal methods was calculated using the AUC and could be viewed as reductive. In addition, the cut-off values used in the definition of 'minimal improvement' based on the longitudinal AUC are subjective. Future research could explore the use of time-varying variables for 'minimal improvement', where a participant could be 'minimally improved' for a period of time and then not 'minimally improved' later in the follow-up time period.

5.4.2 Comparison with existing literature

The findings of this study are consistent with a previous study in knee osteoarthritis by Williams *et al.*, which found some variation in MCID estimates by follow-up time point but did not find an increasing or decreasing trend over time [384]. Williams *et al.* found that MCID estimates for the WOMAC calculated using the ROC curve approach did vary over time, with estimates ranging between 2 and 4 points for one method and 7 and 12 for a second method. However, they did not find a consistent increase or decrease in the MCID estimates over time. The same study examined two other outcome measures of functional ability in knee osteoarthritis and similarly found variation over time but not in a consistent direction.

In musculoskeletal conditions in general, there is no consensus in the existing literature on whether MCID estimates vary by follow-up time point. Tashjian *et al.* found that shorter follow-up was correlated with larger MCID estimates for the American Shoulder and Elbow Surgeons (ASES) score after shoulder arthroplasty using anchor-based methods [41]. However, the same study found no significant correlation for the visual analog scale pain score or Simple Shoulder Test. Similarly, assessing patients after shoulder arthroplasty, Simovitch *et al.* found that "length of follow-up appeared to variably affect the MCID for each of the outcome metrics studied" [49]. In contrast

with Tashjian *et al.*, the results by Simovitch *et al.* suggested larger MCID values for longer follow-up periods. McCreary *et al.* found that the MCID value in an outcome measure for distal radius fractures was larger for the 12-week follow-up than the 6-week follow-up compared to baseline [385]. Similar to the OAI analysis on adjacent 1-year periods, McCreary *et al.* also found very different estimates for the baseline to 6-weeks period compared with the 6-weeks to 12-weeks period. However, other studies in musculoskeletal conditions and other disease areas have found that MCID estimates were consistent across different follow-up periods [51, 386].

The findings of this study also agree with the existing literature that different methods of calculation produce different MCID estimates, even within anchor-based approaches [363, 386, 387, 388]. Many other studies in different conditions have shown that MCID estimates vary by baseline disease severity [38, 389, 390, 391]. This aligns with the differences by baseline WOMAC score in the subgroup analyses for the single time-point methods and the large difference in the MCID estimates in the sensitivity analysis for the longitudinal methods after adjusting for baseline WOMAC score.

In this analysis, the single time point and longitudinal methods suggested that the MCID for improvement in the WOMAC total score was 3-4 points on the 0-96 scale with a confidence interval from approximately 0 to 6. Several reviews have recently examined MCID estimates in lower limb osteoarthritis, with most finding a wide range of estimates [213, 367, 392]. However, many of the included studies only calculated MCID estimates separately for the WOMAC subscales, most commonly examining only the function subscale. The MCID estimates for the WOMAC subscales may not be applicable to the combined total score. Most existing studies found higher MCID estimates than calculated in this analysis. Maratt *et al.* found MCID estimates of 25.0-31.1 on 0-100mm scales for the WOMAC subscales [393]. Hmamouchi *et al.* found MCID estimates of 15.5 on the 0-96 scale for the total score [394]. However,

both of these studies used the ROC curve method, which produced slightly higher estimates in the OAI analysis and does not incorporate the concept of a "minimal" difference.

Angst *et al.* used the ROC curve method to produce estimates of 15 points on a 0-100 scale compared to 7-9 points using regression methods [375]. In earlier studies using a 0-10 scale, Angst *et al.* found MCID estimates of 0.7-0.8, equivalent to 12-18% of baseline scores or 6% of the maximum total score [193, 207]. Using the maximum of 96 points in the Likert scale, this translates to an MCID estimate of 5.8 points, which is slightly higher than the estimates using the OAI dataset. However, this may be due to differences in the disease severity of the population. In the OAI dataset, the mean WOMAC total score at baseline was approximately 25 points. Using 12-18% of baseline scores from Angst *et al.* produces estimates of 3-4.5, which is consistent with the findings of this study.

Other studies have previously used regression methods for MCID calculation. Angst *et al.*, Lee *et al.* and Hwang *et al.* used linear regression with or without adjustment for participant factors in different outcome measures. However, these studies only included data from a single follow-up time point [375, 395, 396, 397]. This study builds on their previous work to include repeated measurements from participants in the same analysis. In a study of people with dementia, O'Connell *et al.* recommended the use of longitudinal methods for calculating MCID estimates, stating that "regression-based approaches are likely best not only for the understanding of whether a change in performance occurred measured with reliable change, but also for determining whether that change is clinically meaningful" [398, 399]. Akaberi *et al.* also used a random-intercept model to calculate MCID estimates using repeated measures on quality of life in patients after a pulmonary embolism [400]. Akaberi *et al.* used a time-varying indicator in the mixed effects analysis to determine important differences

relative to baseline for each time point separately. In contrast, in our study, the definition of minimal improvement was also based on longitudinal data.

5.4.3 Strengths and limitations

This is one of the first studies to compare MCID estimates across multiple time points over a long-term follow-up period. This analysis also used consistent methodology across the different time points in a single sample of people with osteoarthritis, reducing the potential for confounding. The results provide initial insight into whether adjustment for time should be considered in the use and calculation of MCID estimates, using a larger sample than the majority of MCID calculation studies. The patient-reported anchor measure demonstrated favourable properties: it was moderately correlated with the change in the WOMAC and did not exhibit issues with response shift.

However, this analysis is limited as it used only a single dataset. It remains uncertain whether the findings are generalisable to other datasets or disease areas. It is unclear whether the findings hold for different patient groups or different interventions. Although the cohort included 1390 participants, the number included in the MCID estimation was between 300 and 600 participants for single time point methods and only 245 participants for the longitudinal methods. However, this analysis included a larger sample than most MCID calculation studies and the MCID estimates were imprecise with large confidence intervals. The issue of imprecision was exacerbated in the subgroup analysis.

The analysis also used a non-traditional anchor measure, based on a calculated change score derived from participants' assessment of their own current state. This may be less relevant to participants than a direct assessment of the transition in their condition, such as asking participants to compare their current state to a baseline

time point. However, it does avoid issues related to response shift and recall bias [401, 402]. In addition, the change in the anchor measure was correlated with the change in the outcome measure, providing some evidence to support the validity of this non-traditional approach to anchoring (Section 5.3.1, Table 5.3 and Figure 5.3).

The selection of a 1-point change in the anchor measure to represent minimal importance was arbitrary and could not be validated in the dataset. This is especially poignant for osteoarthritis, where there is inherent variability due to the fluctuating nature of the disease and lability of symptoms. As this analysis was retrospective, the results are limited in that we could not ask the original trial participants what level of effect they would feel clinically worthwhile, as was done by Ferreira at al. [403]. In other datasets, the validity could be increased if all levels of the anchor measure were labelled and a level for 'minimally improved' or 'slightly improved' was included.

It could also be questioned whether the longitudinal definition for 'minimal improvement' was appropriate. The choice of cut-off to define the region of minimal improvement was arbitrary. Moreover, the average AUC value could be the same for two participants with very different trajectories of change in the anchor. It was unclear whether a 1-point improvement for 5 years was viewed equivalently to a 5-point improvement for 1 year. This could be combined with evidence of participant preferences to provide a more evidence-based definition of 'minimal improvement'.

The use of an observational cohort with heterogeneity in the treatment received and stage of disease may also limit the interpretation of the results. The MCID estimates could have differed for subgroups of participants receiving specific treatments and the duration, frequency and content of those treatments. Participants' expectations may not have been consistent over time and could have been affected by the differing treatments received and associated risks and burden of treatment. These unmeasured factors could have affects the MCID estimate. For example, treatment-related adverse

events could have affected whether a participant viewed a change in their condition as 'worthwhile'. It has also been suggested that the concept of clinical importance may differ between patients treated in routine clinical practice and participants in experimental trials [404, 405]. Therefore, MCID estimates produced using observational cohort data may not be applicable to clinical trial samples, for example, if they are not representative in terms of disease severity.

5.4.4 Implications

For improvement in the WOMAC score, the duration of follow-up did not appear to be associated with the magnitude of the MCID estimate. The MCID estimates for different lengths of follow-up (comparing baseline to year 2, baseline to year 3, etc.) had a similar level of variability to MCID estimates for different 1-year follow-up periods (comparing year 2 to year 3, etc.). The differences in the MCID estimates were thus likely due to other factors and not due to differences in the length of follow-up period. This suggests that the MCID estimate should not be adjusted based on the assessment time point. Therefore, when considering the sample size calculation of randomised trials, the same MCID estimate can be used as the target difference in the calculation regardless of the assessment time point.

In this dataset, the standard deviation for the follow-up score was very similar across the different follow-up time points and thus could be generally applied. However, for sample size calculations based on the change from baseline, the standard deviation was higher for longer follow-up durations and this should be accounted for when calculating the target effect size.

Even with a moderate-to-large number of participants included in this analysis, the MCID estimates were still imprecise. Therefore, researchers who publish studies on the calculation of MCID estimates should include a measure of the imprecision of the

estimate, such as the 95% confidence interval, to indicate the level of variability. Even if a large sample size was used, trialists and other researchers should be cautious in the use of MCID estimates where no measure of imprecision is reported.

As the MCID estimates did vary by the time point used, it may be worthwhile to calculate MCID estimates based on multiple time points to examine the robustness of the estimate produced and demonstrate the level of variability due to baseline disease severity and assessment time point. This supports the recommendation by McCreary *et al.* to include an anchor measure in a clinical trial such that "the MCID can be determined for the specific patient population included in the study" [385].

The use of longitudinal methods to calculate the MCID could produce more precise estimates. However, this may depend on whether the longitudinal definition of improvement is considered to be relevant to participants and clinicians. In this example, the three longitudinal methods produced very similar results for the MCID estimate. However, it is unclear whether the results would still be consistent when applied to other datasets. Therefore, it seems sensible to choose the method that aligns with the analysis method used in clinical studies where the MCID estimate would be applied. For instance, randomised trials have used AUC approaches to analyse study results by the participant level or to compare the area for different treatment arms calculated based on summary statistics [406, 407]. In this case, the MCID estimate used should be calculated using the AUC estimate.

The results also showed that the estimates varied greatly depending on the baseline WOMAC score in both the single time point and longitudinal methods. Future studies should use an MCID estimate calculated in a sample with similar severity of symptoms to ensure that it is applicable to the study population.

5.4.5 Future research

Future research could expand on the above work by examining the consistency of MCID estimates over time in other datasets and for other conditions. It may be that MCID estimates are more consistent in some disease areas than others. For example, in trauma trials, the trajectory of recovery is important and therefore timing of improvement could be a more important consideration. This would also allow the assessment of consistency of MCID estimates at different time points. The consistency of MCID estimates may be different when considering shorter time periods of less than 1 year, for example, comparing 6 weeks, 3 months and 12 months.

The use of datasets with a larger sample size would provide more precise MCID estimates, which could more easily indicate any trend in MCID estimates over time. Meta-regression could also be used to compare individual trial results and indicate whether duration of follow-up influenced the MCID estimate, whilst accounting for other between-study characteristics, as was done by Terwee *et al.* and Olsen *et al.* [38, 369, 408]. To facilitate analysis using a larger sample size, a future study could combine multiple individual participant datasets and calculate MCID estimates using consistent methodology, for example, by synthesising trial datasets from the OA trial bank, an international consortium of osteoarthritis trials [409, 410].

As this study used a non-traditional anchor measure, future research could assess the validity of this anchor measure. This analysis used a non-traditional method of measuring the participant's current assessment of their global condition and used the change score in this measure as the anchor. A future study could compare this non-traditional anchor to more traditional anchor measures, where the participant provides a direct assessment of how they feel their condition has changed since baseline or the previous time point of assessment, to establish whether these different anchor measures are consistent, similar to the work of Ousmen *et al.* [411]. This would allow

an assessment of whether the MCID estimates produced are robust to different time reference or different wording of the anchor measure.

Further work could examine whether alternative concepts could be used to generate MCID estimates that are more consistent over time, by using relative change, comparing the observed change to the maximum possible change in each participant, or using item-response theory [38, 395, 412]. Future research could also examine whether more specific anchor measures could provide a better assessment of the MCID value by individual WOMAC subscales, rather than an assessment of the participant's overall condition. For example, a pain-specific anchor measure may provide more robust estimates of an MCID in the WOMAC pain subscale [397]. It would also be interesting to explore whether separate domain-specific MCID estimates (e.g., pain, function and stiffness) can be combined to provide an accurate MCID estimate for the total WOMAC scale and other composite outcome measures.

Future research could examine whether MCID estimates vary due to other participant characteristics, such as osteoarthritis phenotype, flare-up patterns or adverse events. Accounting for participant factors is important in MCID calculation as the 'minimally improved' and 'unchanged' groups are not randomly allocated, as in between-treatment comparisons. Therefore, the lack of randomisation means that there could be large differences in the characteristics of the 'minimally improved' and 'unchanged' groups. It has been suggested that MCID estimates calculated using data from randomised trials may not be representative of clinical practice, and vice versa [413]. Future studies could examine whether observational and randomised trial datasets produce similar MCID estimates. More frequent assessment of participant outcomes and anchor measures will become easier as electronic data capture methods become more routinely used in clinical practice settings.

Moreover, simulation studies could be used to examine the use of adjustment for baseline score when calculating MCID estimates. Such studies could assess the level of bias in the MCID estimates caused by different levels of baseline imbalance in the outcome measure between the 'minimally improved' and 'no change' groups [379].

Applying the longitudinal methods in other datasets could indicate whether the consistency of the MCID estimates is generally true or a feature of the dataset used here. Simulation studies could be conducted to assess the statistical properties of the different methods for MCID calculation. The methods could be compared on the level of precision, coverage and type I and type II error. These studies could assess the robustness of these methods to variations in the level of missing data, mechanism of missingness (MAR, MCAR or MNAR), sample size, distribution of the outcome measure, correlation structure between outcome scores over time, and adjustment for participant factors (such as age, sex, disease severity).

Future research could also assess the acceptability of the different methods of MCID calculation in different stakeholder groups, including statisticians, participants, clinicians and commissioners. It is important to examine whether the methods are feasible for use in practice, if stakeholders can interpret the results of these methods and feel that the methods are capturing the correct concept of 'importance' and whether stakeholders would be willing to adopt interventions using the MCID estimates produced by these methods.

Chapter 6 uses a simulation study to examine the statistical implications of using different approaches to analyse the results of longitudinal data from randomised trials, in terms of carrying out the sample size calculation. This will indicate which of the longitudinal methods produce more precise and unbiased estimates of treatment effect, and which is the most appropriate strategy for conducting the sample size calculation. The results will help guide the choice of methods to analyse clinical

trial results under different settings, and how the sample size calculation should be undertaken.

5.4.6 Conclusions

This study has shown that MCID estimates for the WOMAC did not vary by the duration of the follow-up period using a single longitudinal dataset of people living with osteoarthritis. However, there was some variability over time and MCID estimates for the WOMAC were imprecise, even when large numbers of participants were analysed. This study also found that it was feasible to calculate MCID estimates using longitudinal data.

MCID estimates should be calculated for multiple time points or using longitudinal data to ensure that the estimate is not overly extreme due to random fluctuation. Furthermore, the target difference used in the sample size calculation for a randomised trial does not need to be adjusted based on the assessment time point in this condition. However, further research is needed to examine whether the stability of MCID estimates over time is generalisable to other datasets and disease areas.

Chapter 6

Comparing methods to analyse longitudinal data from randomised trials: a simulation study

Prior publication:

Conferences abstracts for presentations on this chapter have been published (see Appendix F.1 for details).

6.1 Introduction

A randomised trial aims to examine whether an intervention produces a better effect than a comparison treatment on the participants' outcomes. A well-designed randomised trial should have high power and low type I error [2, 3]. Power and type I error were described in more detail in Chapter 1. A trial with high power (or low type II error) has a high chance of detecting a treatment effect if a true treatment effect exists. A trial with low type I error has a low chance of detecting a treatment effect if no true treatment effect exists.

Chapter 5 introduced several methods that can be used to calculate minimum clinically important differences (MCIDS) using longitudinal data. Equivalent methods can be used to analyse longitudinal data from randomised trials. In a randomised controlled trial, it is recommended to align the sample size calculation with the method used to analyse the trial results [414]. Chapter 3 reviewed randomised trials in hip or knee osteoarthritis that were published in 2016 and found that several methods were used to analyse the WOMAC as an outcome measure (Section 3.3). Where the same type of outcome data from multiple follow-up time points were available, some trials analysed each time point separately [205, 415], others used a mixed effects model [141, 195], and one trial used a generalised estimating equation (GEE) [416]. However, it is unclear which of these longitudinal methods would perform best under different circumstances, providing the least biased and most precise estimate of the treatment effect.

When comparing different statistical methods, knowledge of the power and type I error of the proposed estimates for the treatment effect can provide information on which methods should be used. It can also inform the sample size required to achieve sufficient power to detect an important treatment effect. If a method produces a

biased or imprecise estimate of the treatment effect, the results of randomised trials may be inaccurate or inconclusive. This could lead to treatments being used in practice that may not be as effective as the results suggest due to bias. Alternatively, an inconclusive result, due to bias or imprecision, could lead to an intervention not being implemented or researched further because of a lack of evidence. The methods used to analyse the results of a randomised trial should ideally provide a precise and unbiased estimate of the treatment effect. However, in practice, the 'true' value of the treatment effect is unknown.

Simulation studies can be used to compare the performance of different statistical methods in a controlled setting [417]. In a simulation study, the 'true' value of the treatment effect is known from the mechanism used to generate the data and, thus, it is possible to quantify the bias in different estimators of the treatment effect. The characteristics of the participant sample and factors related to the data generation process can be changed independently in such a way that any increase or decrease in the performance of the method can be attributed to the characteristic that was changed.

Multiple different factors could affect the performance of the different methods. These factors could include the sample size included in the analysis, the magnitude and consistency of the treatment effect, the number of follow-up time points and the variability in the outcome measure. The highest-performing method is likely to be different for different scenarios where these factors are varied.

This simulation study compared the performance of methods to analyse longitudinal data from a randomised trial, focusing on convergence, power and type I error. The results suggested which method was optimal under different scenarios, including different sample sizes and numbers of follow-up measurements.

6.2 Methods

6.2.1 Data-generating mechanism

This simulation study used parametric simulation, as opposed to re-sampling, to provide more generalisability of the results. The parameters for the distribution of the data-generation mechanism were selected based on the distribution of the WOMAC Index in available datasets, primarily using the TOIB randomised trial, the OAI cohort dataset and the results of the systematic review in Chapter 3 [175, 372].

The varying factors and levels are shown in Table 6.1.

Table 6.1: Factors and levels for the data-generating mechanisms

Factors	**Levels**
Sample size (total for two arms)	100, 200, 400, 600, 800
Number of follow-up time points	2 (3 and 6 months) 3 (3, 6 and 12 months) 4 (3, 6, 12 and 18 months) 5 (3, 6, 12, 18 and 24 months)
Maximum treatment effect ($\beta*$) (on WOMAC scale of 0-96)	0, 4, 8, 12 (for a standard deviation of 16, the standardised effect size is 0, 0.25, 0.5, 0.75)
Pattern of treatment effect	i. Linear improvement ii. Short-term improvement then plateau iii. Temporary improvement

These factors were varied fully factorially, giving 240 (5 x 4 x 4 x 3) scenarios. The maximum treatment effect over the follow-up period was the same for all patterns of improvement ($\beta*$). The scenarios were fixed for all other factors. The allocation between intervention and control was assumed to be 1:1. The treatment effect was assumed to be homogeneous across participants, not varying due to participant characteristics (e.g. disease severity, sex or age). For simplicity, it was assumed that there was no missing data.

The baseline WOMAC scores were assumed to be normally distributed (mean 40, SD 15). The follow-up scores in the WOMAC Index were assumed to be normally distributed with a standard deviation of 12. The mean follow-up score was assumed to be 40 for the control group and 40 minus the treatment difference for the intervention group (as specified for each scenario). The within-participant correlation between follow-up scores at different time points was assumed to be 0.5, following an exchangeable model for the change scores. The correlation between the baseline and follow-up scores was also assumed to be 0.5. The between-participants error term (random intercept, b_i) was assumed to be normally distributed with mean 0 and standard deviation 15. Within each dataset, to model the truncation of the patient-reported outcome measure, participants were re-sampled if any of the WOMAC scores at baseline and follow-up were outside of the 0-96 range. The code to generate the simulated datasets is presented in Appendix E.1.

The maximum treatment effect was reached at different time points for the different patterns. In pattern 1 (linear improvement), the treatment effect was 0 at time 0 and then increased linearly before reaching the maximum treatment effect at the final time point. In pattern 2 (short-term improvement then plateau), the treatment effect was 0 at time 0 and then increased to the maximum treatment effect, and the treatment effect was sustained at the maximum level for the remainder of the follow-

up period. In pattern 3 (temporary improvement), the treatment effect was 0 at time 0, the maximum treatment effect was reached at the first follow-up assessment after time 0 and the treatment effect was then 0 for the remainder of the follow-up period. Figures 6.1 to 6.3 and Table 6.2 show the patterns of the treatment effect over time for different patterns and lengths of follow-up period when the maximum treatment effect was 4, 8 or 12 points.

Figure 6.1: Patterns of improvement for different lengths of follow-up period when the maximum treatment effect was 4 points

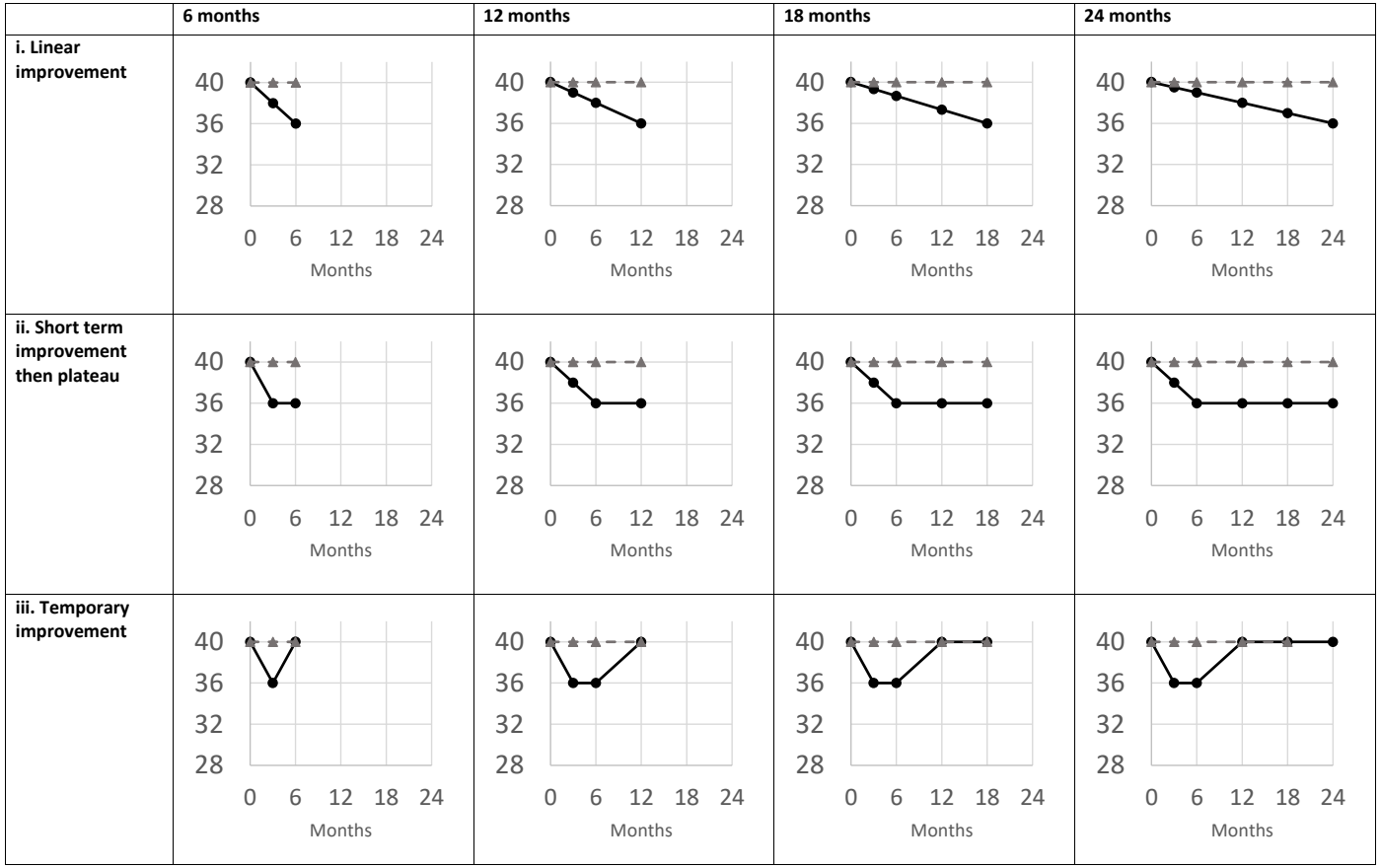


Figure 6.2: Patterns of improvement for different lengths of follow-up period when the maximum treatment effect was 8 points

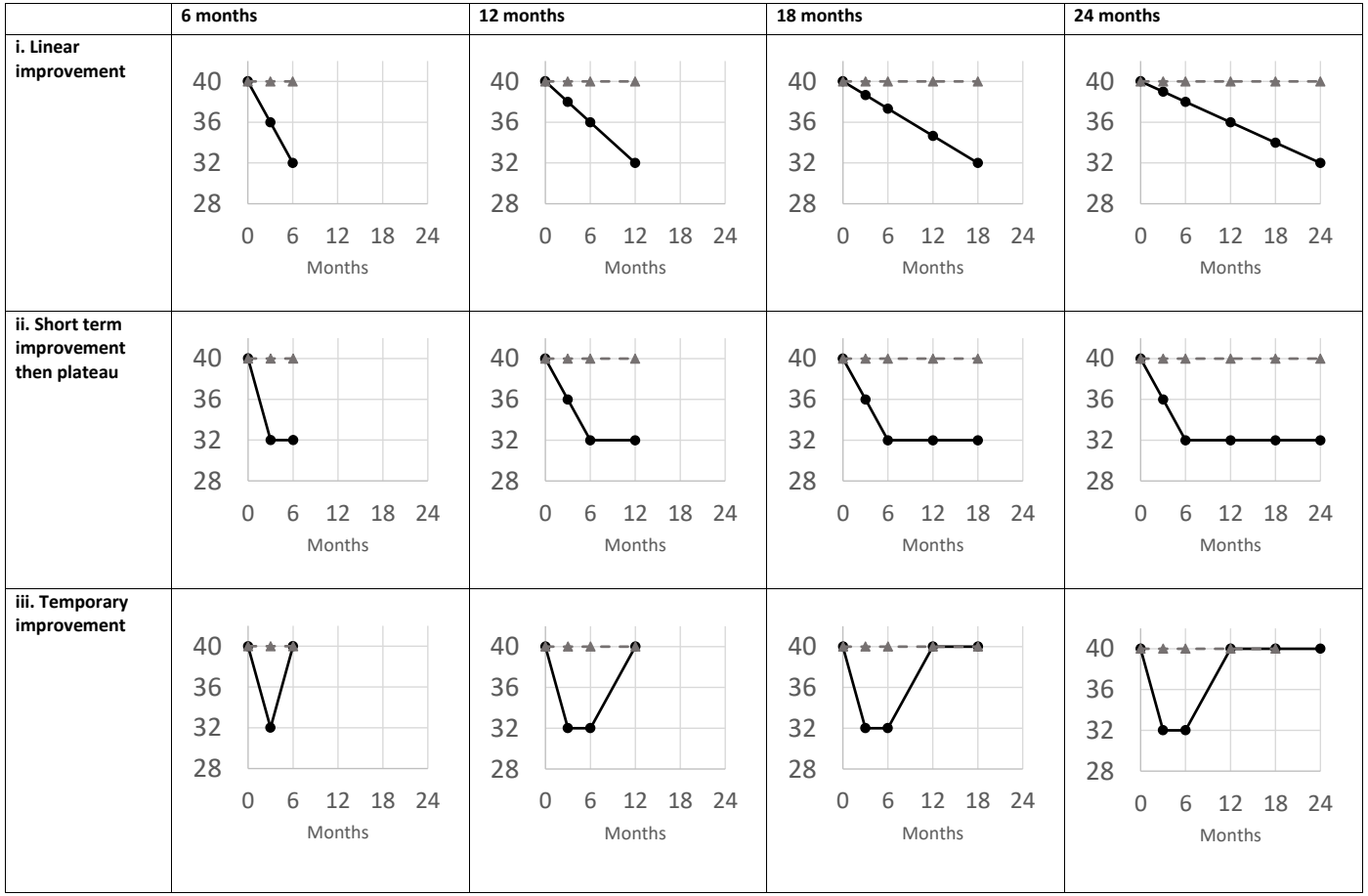


Figure 6.3: Patterns of improvement for different lengths of follow-up period when the maximum treatment effect was 12 points

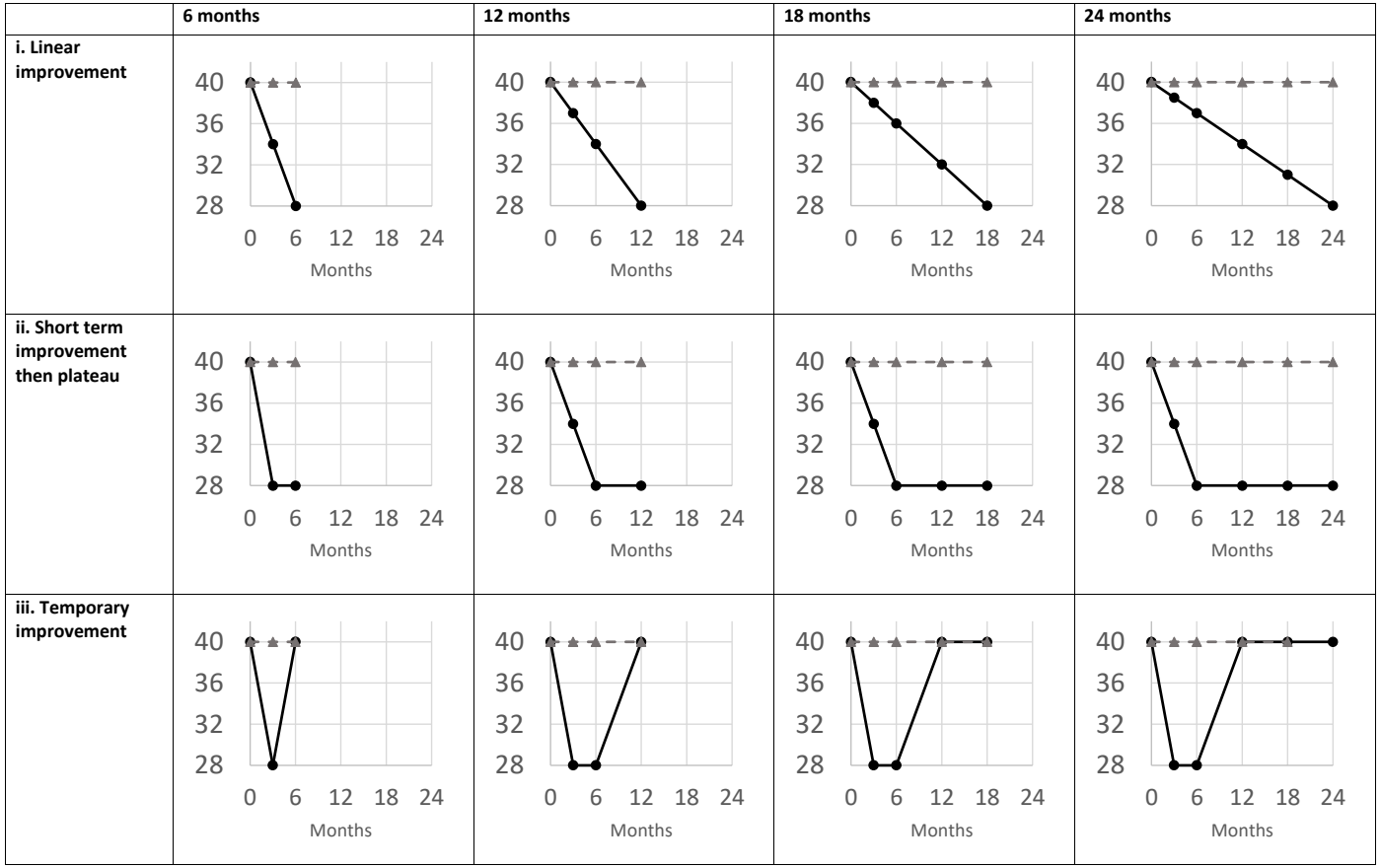


Table 6.2: Treatment effect values for the intervention arm (β_{ij})

		Follow-up length: 6 months		**12 months**			**18 months**				**24 months**				
Pattern	**Maximum treatment effect**	**Time point (j) (months)**		**Time point (j) (months)**			**Time point (j): (months)**				**Time point (j) (months)**				
		3	**6**	**3**	**6**	**12**	**3**	**6**	**12**	**18**	**3**	**6**	**12**	**18**	**24**
i. Linear improvement	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	2	4	1	2	4	0.67	1.33	2.67	4	0.5	1	2	3	4
	8	4	8	2	4	8	1.33	2.67	5.33	8	1	2	4	6	8
	12	6	12	3	6	12	2	4	8	12	1.5	3	6	9	12
ii. Short-term improvement	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	4	4	2	4	4	2	4	4	4	2	4	4	4	4
	8	8	8	4	8	8	4	8	8	8	4	8	8	8	8
	12	12	12	6	12	12	6	12	12	12	6	12	12	12	12
iii. Temporary improvement	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	4	0	4	4	0	4	4	0	0	4	4	0	0	0
	8	8	0	8	8	0	8	8	0	0	8	8	0	0	0
	12	12	0	12	12	0	12	12	0	0	12	12	0	0	0

6.2.2 Estimand

The estimand was the treatment effect, i.e., the difference in the outcome between the intervention and control arms. The value of the estimand for the three methods differed. The study primarily considered the performance of each method in terms of null hypothesis significance testing, by comparing statistical power and type I error, which could be compared across the different methods even when the 'true' values varied.

6.2.3 Statistical methods

Several of the statistical methods compared were described in detail in Section 5.2.5, in the context of calculating MCID estimates. In brief, these methods are:

1. Adapted single time point approach (ANCOVA): Linear regression with baseline adjustment was used to compare the change in the WOMAC score between the intervention and control arms at each individual follow-up time point. The primary time point was set at the final non-zero time point for each pattern, i.e., where the maximum treatment effect was reached.

2. Mixed effects regression: Mixed effects regression models were used on the WOMAC scores assuming a fixed slope and a random intercept by participant and an exchangeable correlation structure for the residuals.

3. GEE regression: The GEE regression model extends the generalised linear model to allow for observations to be correlated within the same participant. GEE models were used on the WOMAC scores, assuming a Gaussian family, identity link and an exchangeable correlation structure.

The code to analyse the simulated datasets using the three methods is presented in Appendix E.2.

The true value of the parameter differed for each of the methods. The true values for the single time point analyses are the same as the treatment effect at the corresponding time point, presented in Table 6.2. The true values for the single time point, mixed effects and GEE methods are presented in Table 6.3. For the mixed effects and GEE methods, it was assumed that the treatment effect did not vary by time point. The estimated value was compared to the average of the true treatment effects at each follow-up time point.

Table 6.3: True values for the estimands for the different methods

		Method 1: ANCOVA (primary) at final non-zero time point:				**Methods 2 and 3: Mixed effects and GEE:**			
		Mean difference at final non-zero time point[a]				**Mean difference averaged over whole period**			
Pattern	**Maximum treatment effect**	**Follow-up period (months)**				**Follow-up period (months)**			
		6	**12**	**18**	**24**	**6**	**12**	**18**	**24**
i. Linear improvement	**0**	0	0	0	0	0	0	0	0
	4	4	4	4	4	3	2.33	2.17	2.1
	8	8	8	8	8	6	4.67	4.33	4.2
	12	12	12	12	12	9	7	6.5	6.3
ii. Short-term improvement	**0**	0	0	0	0	0	0	0	0
	4	4	4	4	4	4	3.33	3.5	3.6
	8	8	8	8	8	8	6.67	7	7.2
	12	12	12	12	12	12	10	10.5	10.8
iii. Temporary improvement	**0**	0	0	0	0	0	0	0	0
	4	4	4	4	4	2	2.67	2	1.6
	8	8	8	8	8	4	5.33	4	3.2
	12	12	12	12	12	6	8	6	4.8

[a] Note: The true values for the interim time points in method 1 (ANCOVA at all time points) are shown in Table 6.2

6.2.4 Performance measures

Several performance measures were used to assess the bias and precision of the estimators. The terms used to define the performance measures were taken from the tutorial by Morris *et al.* [417]. If the true value for the estimand is β, the estimate is given by $\hat{\beta}_i$ with standard error s_i for the i^{th} simulation.

$\bar{\beta} = \frac{1}{n} \sum_i \hat{\beta}_i$

$\bar{s^2} = \frac{1}{n} \sum_i {s_i}^2$

$V_{\bar{\beta}} = \frac{1}{n-1} \sum_i (\hat{\beta}_i - \bar{\beta})^2$

The performance measures assessed were:

Properties of model:

- Convergence

Errors:

- Type I error: Proportion of simulated datasets for which the 90%, 95%, 97.5% and 99% confidence interval did not include zero, i.e., was wrongly found to have a significant treatment effect (note: only applicable for treatment effect of 0).
- Statistical power (or equivalently 1 - type II error): Proportion of simulated datasets for which the 80%, 90% and 95% confidence interval included zero, i.e., was wrongly found an absence of treatment effect (note: not applicable for treatment effect of 0).

Properties of confidence interval:

- Coverage: Proportion of simulated datasets for which the 95% confidence inter-

val included the true parameter ($\hat{\beta}$) used to generate the simulated data.

Properties of estimator:

- Bias: The distance of the estimator from the true value ($E(\hat{\beta}) - \beta$)
- Relative bias: Bias as a proportion of the true value ($\frac{\bar{\beta}-\beta}{\beta} x100\%$)
- Empirical standard error (EmpSE): The standard deviation of $\hat{\beta}$, which provides an assessment of the spread among the estimates ($\sqrt{Var(\hat{\beta})}$).

Properties of standard error:

- Average model-based standard error (ModSE): The average of the standard error from each repetition of the analysis ($\sqrt{E(\hat{Var}(\hat{\beta}))}$).

The code and formulas to calculate the performance measures is presented in Appendix E.3.

For the single time point method, multiple treatment estimates were computed for each of the follow-up time points. The performance of the single time point method was summarised for analysing the final follow-up time point only (ANCOVA primary) for all summary measures. In addition, for the single time point method, the type I error was calculated in the event that all follow-up time points were analysed separately (ANCOVA all), i.e., the probability of a falsely identified statistically significant result at one or more follow-up time points.

Hierarchy of performance measures

The primary performance measures were the convergence of the model, statistical power and type I error. These measures were comparable across the different methods.

The properties of the estimator were not comparable across the methods as the treatment effect was represented in different forms. This allowed us to compare the nec-

essary sample size required for the different methods for a given level of power and type I error. It was possible to compare the properties of the estimator across the different scenarios for the same method.

6.2.5 Analysis

For each scenario, $n_{sim} = 1600$ datasets were produced, achieving a Monte Carlo standard error (MCSE) of 1% for a statistical power of 80% and a MCSE of <0.6% for a coverage of 95%. Stata IC version 14 was used to generate the datasets, analyse the individual datasets and calculate the performance measures [121]. Each simulation scenario included a different random number seed, and the random number generator state was recorded for the repetitions of each scenario. The University of Oxford Advanced Research Computing (ARC) facility was used to carry out this work [418]. The ARC facility was used because of the computational resources required for the large number of simulations and scenarios.

6.2.6 Summary of simulation methods

For 1600 repetitions, repeat steps 1 to 4:

Step 1:

(a) Simulate `n` baseline values using normal distribution $N(40, 15)$.

`n = 100, 200, 400, 600, 800`

(b) Simulate `t` follow-up scores using normal distribution $N(40, 12)$ for each of the `n` participants with a within-participant correlation of 0.5 between baseline and follow-up scores and between two follow-up scores at different time points.

`t = 2, 3, 4, 5`

(c) Add between-participant variation using normal distribution $N(0, 15)$ to the baseline and `t` follow-up scores for each of the `n` participants.

Step 2:

(a) Randomly allocate the `n` to treatment or control arms using 1:1 allocation.

(b) Assign proportion of maximum treatment effect (0-1) at each follow-up time point `t` for the corresponding pattern `p` (see Figures 6.1 to 6.3).

```
p = 1, 2, 3
```

(c) Multiply the treatment effect by maximum treatment effect `b`.

```
b = 0, 4, 8, 12
```

(d) Subtract the corresponding treatment effect from the follow-up scores for participants in the treatment arm.

(e) Round follow-up scores to the nearest whole number to emulate composite outcome measure.

Step 3:

(a) Re-sample by repeating steps 1 and 2 for participants where the baseline score or any of the follow-up scores are less than 0 or greater than 96 to truncate to fit the possible range of the WOMAC (0-96).

Step 4:

(a) Analyse using ANCOVA with baseline adjustment separately for each of the `t` follow-up time points.

(b) Analyse using the GEE method.

(c) Analyse using a mixed effects (random intercept) method.

6.3 Results

6.3.1 Statistical power, type I error, and convergence

The statistical power for each of the scenarios for a 95% confidence interval is presented in Figures 6.4 to 6.6 and Tables 6.4 to 6.6. For all of the patterns, in scenarios with larger sample sizes and treatment effects, all of the methods provided high power ($\beta \geq 8$ and $n \geq 400$).

When the treatment effect was not consistent over time (in patterns 1 and 3), ANCOVA using the primary time point had greater power than the GEE or mixed effects methods using the time-averaged treatment effect, especially when the treatment effect was small (Figures 6.4 and 6.6, Tables 6.4 and 6.6).

However, when the treatment effect was consistent over time (in pattern 2), the GEE and mixed effects methods demonstrated higher power than the ANCOVA method at the primary time point (Figure 6.5 and Table 6.5). The GEE method provided slightly greater power than the mixed effects method.

The findings on power for 95% confidence intervals were consistent with the results for 80% and 90% confidence intervals (Appendix E.4, Tables E.1 to E.6).

The method with the lowest type I error varied across the scenarios and the optimal method did not seem to be associated with the sample size or number of time points. In the majority of scenarios, the type I error for a 95% confidence interval was higher for the GEE and mixed effects methods than ANCOVA using the final non-zero time point as the primary follow-up assessment (Tables 6.7 to 6.9 and Figures 6.7 to 6.9). However, the type I error was much larger for the ANCOVA method if the primary time point was not pre-specified and analysis was conducted for all of the follow-up time points. The type I error for the ANCOVA method conducted at each follow-up

time point separately increased as the number of follow-up assessments increased. In some cases for the ANCOVA method with five follow-up assessments, the probability of a falsely-identified significant treatment effect at any follow-up time point was as high as 20% for the 95% confidence interval.

Findings were similar for the 95%, 97.5% and 99% confidence intervals (Appendix E.5). However, for the 90% confidence interval, type I error was not consistently higher for the GEE method than the ANCOVA approach at the primary time point.

The performance measures for the ANCOVA method at each time point is presented in Appendix E.6.

For the ANCOVA and GEE methods, the models converged for all of the possible scenarios and repetitions. However, convergence was poor for the mixed effects method (Table 6.10). Convergence was achieved within 100 iterations for between 14% and 89% of the repetitions for each scenario. Convergence was less likely in scenarios with a larger sample size and more follow-up assessments.

Figure 6.4: Statistical power: Pattern 1 (linear improvement)

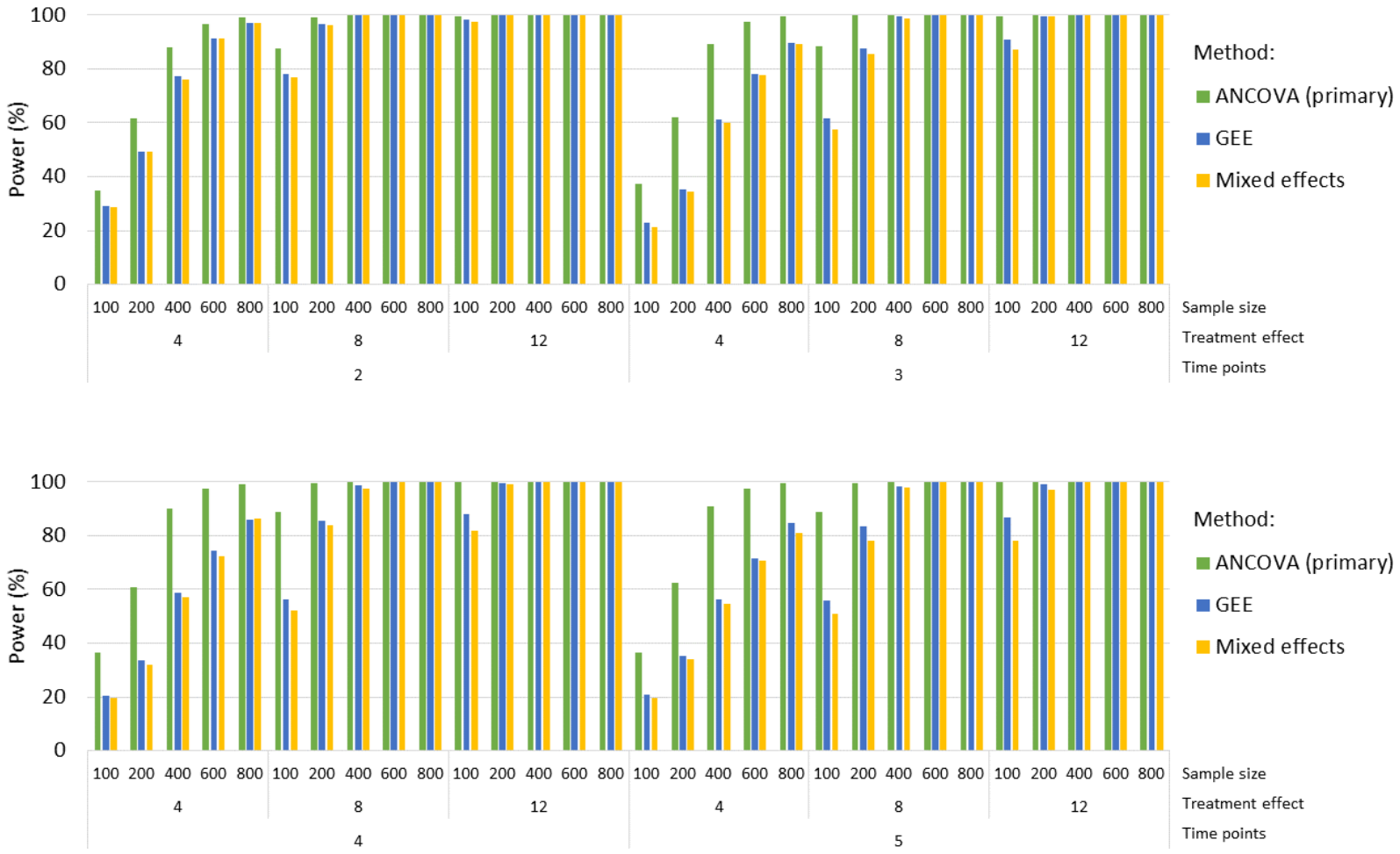


Table 6.4: Statistical power (for 95% confidence interval): Pattern 1 (linear improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Power	34.8	87.7	99.7	37.3	88.3	99.6	36.4	88.9	99.8	36.3	88.8	99.9
	(primary)	MCSE	(1.19)	(0.82)	(0.14)	(1.21)	(0.81)	(0.15)	(1.20)	(0.79)	(0.11)	(1.20)	(0.79)	(0.09)
	GEE	Power	29.2	77.9	98.1	23.0	61.6	90.7	20.5	56.4	87.8	20.9	56.0	86.9
		MCSE	(1.14)	(1.04)	(0.34)	(1.05)	(1.22)	(0.73)	(1.01)	(1.24)	(0.82)	(1.02)	(1.24)	(0.84)
	Mixed	Power	28.5	76.8	97.5	21.4	57.5	87.0	19.8	52.3	81.7	19.5	51.1	77.9
	effects	MCSE	(1.20)	(1.10)	(0.41)	(1.16)	(1.41)	(0.96)	(1.23)	(1.53)	(1.21)	(1.27)	(1.61)	(1.37)
200	ANCOVA	Power	61.5	99.1	100.0	62.1	99.8	100.0	60.9	99.5	100.0	62.3	99.5	100.0
	(primary)	MCSE	(1.22)	(0.23)	(0.00)	(1.21)	(0.13)	(0.00)	(1.22)	(0.18)	(0.00)	(1.21)	(0.18)	(0.00)
	GEE	Power	49.4	96.8	99.9	35.1	87.7	99.6	33.6	85.4	99.5	35.4	83.4	99.3
		MCSE	(1.25)	(0.44)	(0.06)	(1.19)	(0.82)	(0.15)	(1.18)	(0.88)	(0.18)	(1.20)	(0.93)	(0.22)
	Mixed	Power	49.3	96.2	99.9	34.6	85.3	99.4	32.1	84.0	99.0	33.9	78.0	97.1
	effects	MCSE	(1.38)	(0.53)	(0.08)	(1.43)	(1.05)	(0.23)	(1.50)	(1.19)	(0.32)	(1.59)	(1.42)	(0.59)
400	ANCOVA	Power	88.1	100.0	100.0	89.1	100.0	100.0	90.1	100.0	100.0	90.8	100.0	100.0
	(primary)	MCSE	(0.81)	(0.00)	(0.00)	(0.78)	(0.00)	(0.00)	(0.75)	(0.00)	(0.00)	(0.72)	(0.00)	(0.00)
	GEE	Power	77.1	99.9	100.0	61.2	99.4	100.0	58.8	98.8	100.0	56.2	98.4	100.0
		MCSE	(1.05)	(0.06)	(0.00)	(1.22)	(0.19)	(0.00)	(1.23)	(0.27)	(0.00)	(1.24)	(0.31)	(0.00)
	Mixed	Power	76.0	99.9	100.0	59.8	98.5	100.0	57.0	97.5	100.0	54.5	98.0	100.0
	effects	MCSE	(1.25)	(0.09)	(0.00)	(1.55)	(0.38)	(0.00)	(1.66)	(0.52)	(0.00)	(1.78)	(0.52)	(0.00)

Table 6.4: Statistical power (for 95% confidence interval): Pattern 1 (linear improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Power	96.6	100.0	100.0	97.3	100.0	100.0	97.6	100.0	100.0	97.4	100.0	100.0
	(primary)	MCSE	(0.45)	(0.00)	(0.00)	(0.40)	(0.00)	(0.00)	(0.38)	(0.00)	(0.00)	(0.40)	(0.00)	(0.00)
	GEE	Power	91.2	100.0	100.0	78.2	100.0	100.0	74.4	99.9	100.0	71.4	99.9	100.0
		MCSE	(0.71)	(0.00)	(0.00)	(1.03)	(0.00)	(0.00)	(1.09)	(0.06)	(0.00)	(1.13)	(0.06)	(0.00)
	Mixed	Power	91.4	100.0	100.0	77.8	100.0	100.0	72.3	99.7	100.0	70.7	99.7	100.0
	effects	MCSE	(0.87)	(0.00)	(0.00)	(1.43)	(0.00)	(0.00)	(1.58)	(0.18)	(0.00)	(1.74)	(0.23)	(0.00)
800	ANCOVA	Power	99.3	100.0	100.0	99.7	100.0	100.0	99.2	100.0	100.0	99.3	100.0	100.0
	(primary)	MCSE	(0.22)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)	(0.22)	(0.00)	(0.00)	(0.21)	(0.00)	(0.00)
	GEE	Power	97.2	100.0	100.0	89.8	100.0	100.0	85.8	100.0	100.0	84.7	100.0	100.0
		MCSE	(0.41)	(0.00)	(0.00)	(0.76)	(0.00)	(0.00)	(0.87)	(0.00)	(0.00)	(0.90)	(0.00)	(0.00)
	Mixed	97.1	100.0	100.0	89.3	100.0	100.0	86.5	100.0	100.0	81.1	100.0	100.0	
	effects	MCSE	(0.55)	(0.00)	(0.00)	(1.15)	(0.00)	(0.00)	(1.30)	(0.00)	(0.00)	(1.61)	(0.00)	(0.00)

$\beta*$: Maximum treatment effect.

Figure 6.5: Statistical power: Pattern 2 (short-term improvement then plateau)

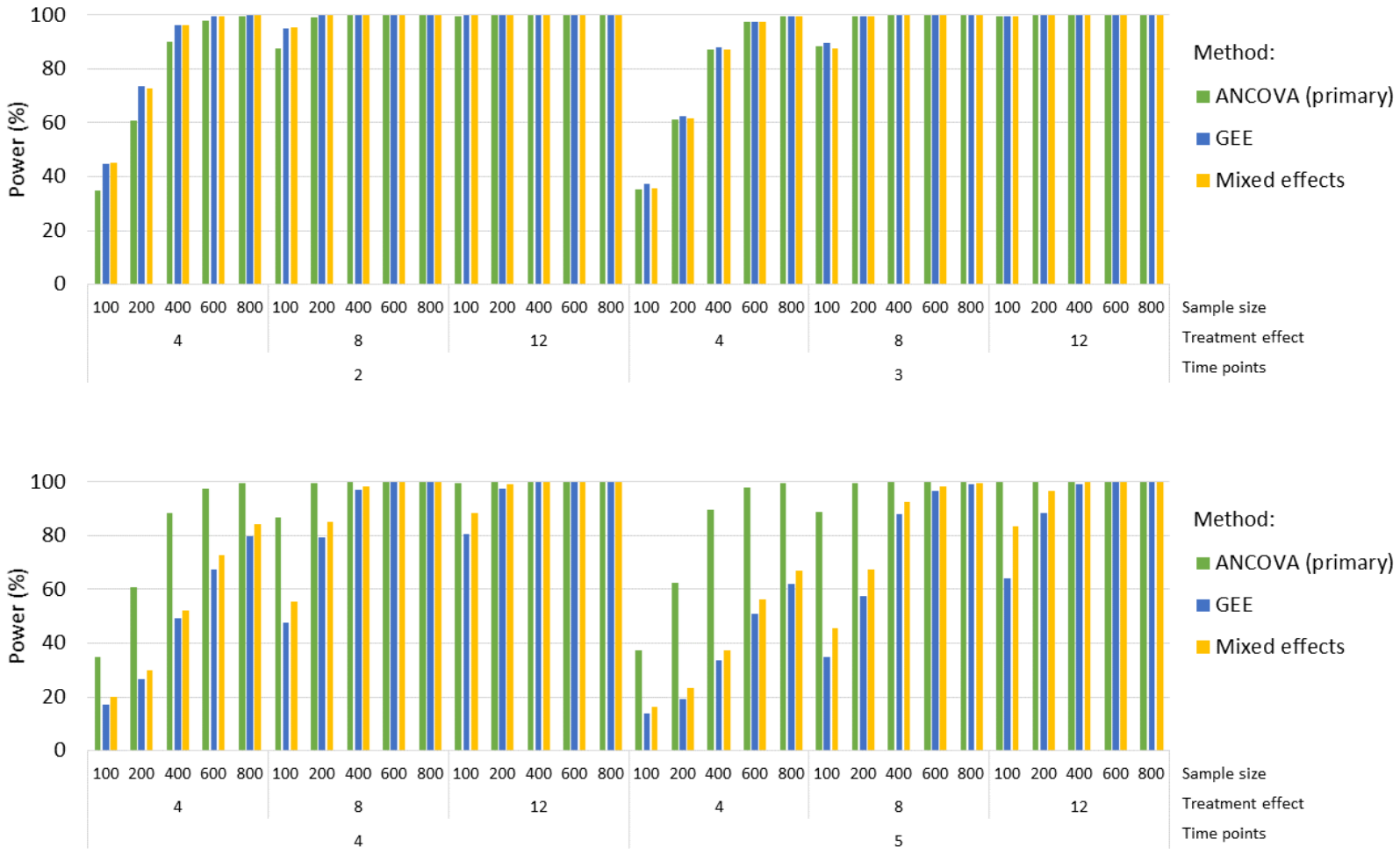


Table 6.5: Statistical power (for 95% confidence interval): Pattern 2 (short-term improvement then plateau)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Power	34.8	87.6	99.6	35.1	88.3	99.5	36.7	88.3	99.9	33.9	86.3	99.8
	(primary)	MCSE	(1.19)	(0.82)	(0.15)	(1.19)	(0.81)	(0.18)	(1.21)	(0.80)	(0.06)	(1.18)	(0.86)	(0.13)
	GEE	Power	44.7	95.1	100.0	37.1	89.7	99.5	43.6	93.3	100.0	44.3	94.5	100.0
		MCSE	(1.24)	(0.54)	(0.00)	(1.21)	(0.76)	(0.18)	(1.24)	(0.63)	(0.00)	(1.24)	(0.57)	(0.00)
	Mixed	Power	45.2	95.2	100.0	35.7	87.6	99.3	43.1	92.6	100.0	43.2	93.9	100.0
	effects	MCSE	(1.31)	(0.56)	(0.00)	(1.36)	(0.95)	(0.23)	(1.54)	(0.82)	(0.00)	(1.60)	(0.76)	(0.00)
200	ANCOVA	Power	60.8	99.1	100.0	61.1	99.5	100.0	62.5	99.2	100.0	59.5	99.1	100.0
	(primary)	MCSE	(1.22)	(0.24)	(0.00)	(1.22)	(0.18)	(0.00)	(1.21)	(0.22)	(0.00)	(1.23)	(0.23)	(0.00)
	GEE	Power	73.5	99.9	100.0	62.6	99.4	100.0	69.3	99.8	100.0	73.4	99.9	100.0
		MCSE	(1.10)	(0.06)	(0.00)	(1.21)	(0.19)	(0.00)	(1.15)	(0.11)	(0.00)	(1.10)	(0.06)	(0.00)
	Mixed	Power	72.9	99.9	100.0	61.7	99.3	100.0	68.1	99.7	100.0	72.6	99.9	100.0
	effects	MCSE	(1.22)	(0.08)	(0.00)	(1.44)	(0.25)	(0.00)	(1.48)	(0.17)	(0.00)	(1.49)	(0.11)	(0.00)
400	ANCOVA	Power	90.0	100.0	100.0	87.3	100.0	100.0	88.3	100.0	100.0	87.0	100.0	100.0
	(primary)	MCSE	(0.75)	(0.00)	(0.00)	(0.83)	(0.00)	(0.00)	(0.81)	(0.00)	(0.00)	(0.84)	(0.00)	(0.00)
	GEE	Power	96.1	100.0	100.0	88.1	100.0	100.0	93.8	100.0	100.0	94.2	100.0	100.0
		MCSE	(0.49)	(0.00)	(0.00)	(0.81)	(0.00)	(0.00)	(0.60)	(0.00)	(0.00)	(0.58)	(0.00)	(0.00)
	Mixed	Power	96.4	100.0	100.0	87.3	100.0	100.0	94.4	100.0	100.0	92.9	100.0	100.0
	effects	MCSE	(0.55)	(0.00)	(0.00)	(1.07)	(0.00)	(0.00)	(0.78)	(0.00)	(0.00)	(0.89)	(0.00)	(0.00)

Table 6.5: Statistical power (for 95% confidence interval): Pattern 2 (short-term improvement then plateau)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Power	97.8	100.0	100.0	97.5	100.0	100.0	96.9	100.0	100.0	97.0	100.0	100.0
	(primary)	MCSE	(0.37)	(0.00)	(0.00)	(0.39)	(0.00)	(0.00)	(0.43)	(0.00)	(0.00)	(0.43)	(0.00)	(0.00)
	GEE	Power	99.6	100.0	100.0	97.6	100.0	100.0	98.9	100.0	100.0	99.3	100.0	100.0
		MCSE	(0.17)	(0.00)	(0.00)	(0.38)	(0.00)	(0.00)	(0.26)	(0.00)	(0.00)	(0.21)	(0.00)	(0.00)
	Mixed	Power	99.4	100.0	100.0	97.3	100.0	100.0	98.7	100.0	100.0	99.3	100.0	100.0
	effects	MCSE	(0.24)	(0.00)	(0.00)	(0.54)	(0.00)	(0.00)	(0.42)	(0.00)	(0.00)	(0.32)	(0.00)	(0.00)
800	ANCOVA	Power	99.4	100.0	100.0	99.6	100.0	100.0	99.4	100.0	100.0	99.3	100.0	100.0
	(primary)	MCSE	(0.20)	(0.00)	(0.00)	(0.17)	(0.00)	(0.00)	(0.20)	(0.00)	(0.00)	(0.21)	(0.00)	(0.00)
	GEE	Power	100.0	100.0	100.0	99.6	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
		MCSE	(0.00)	(0.00)	(0.00)	(0.15)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Mixed	Power	100.0	100.0	100.0	99.6	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0
	effects	MCSE	(0.00)	(0.00)	(0.00)	(0.22)	(0.00)	(0.00)	(0.16)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

$\beta*$: Maximum treatment effect.

Figure 6.6: Statistical power: Pattern 3 (temporary improvement)

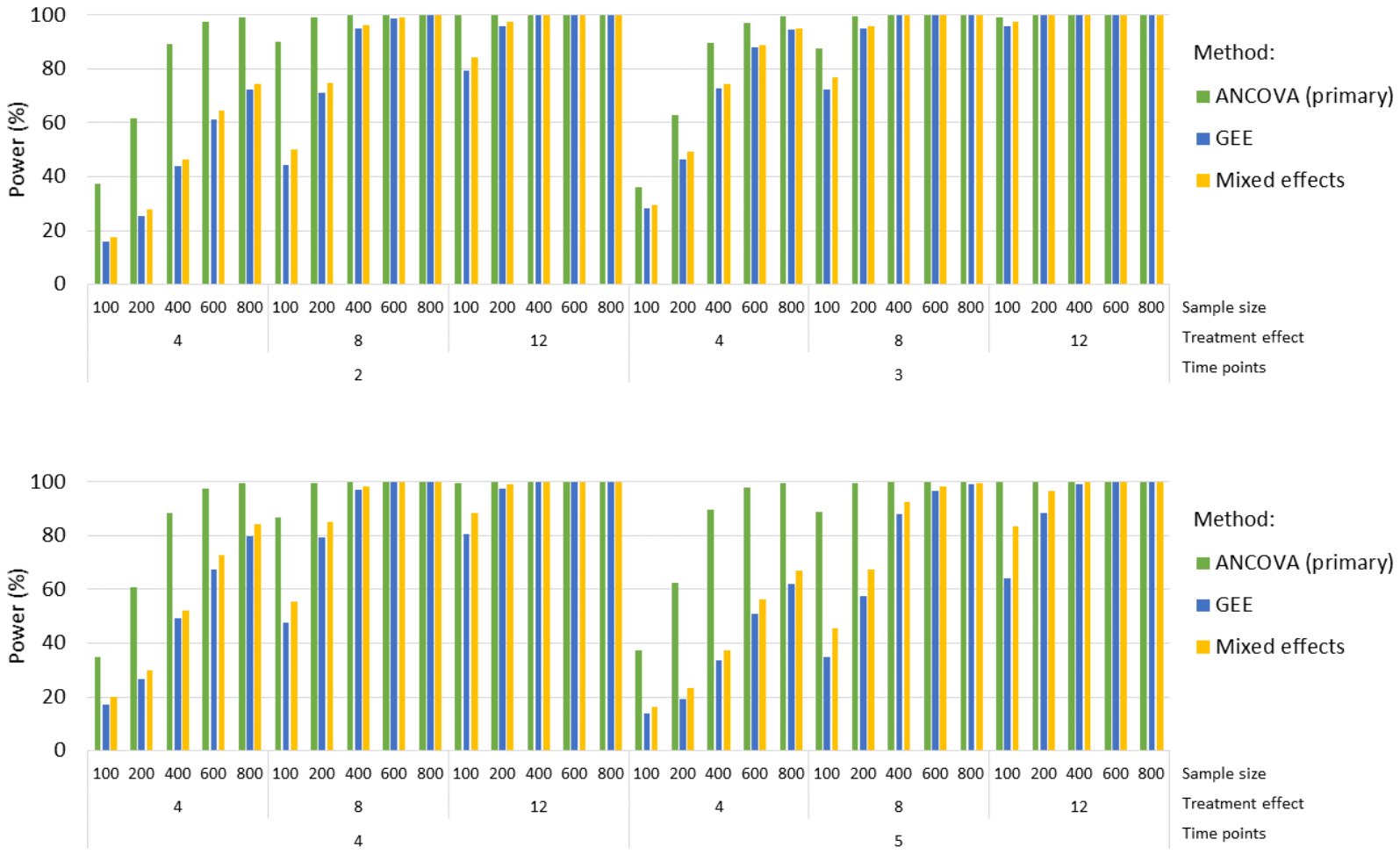


Table 6.6: Statistical power (for 95% confidence interval): Pattern 3 (temporary improvement)

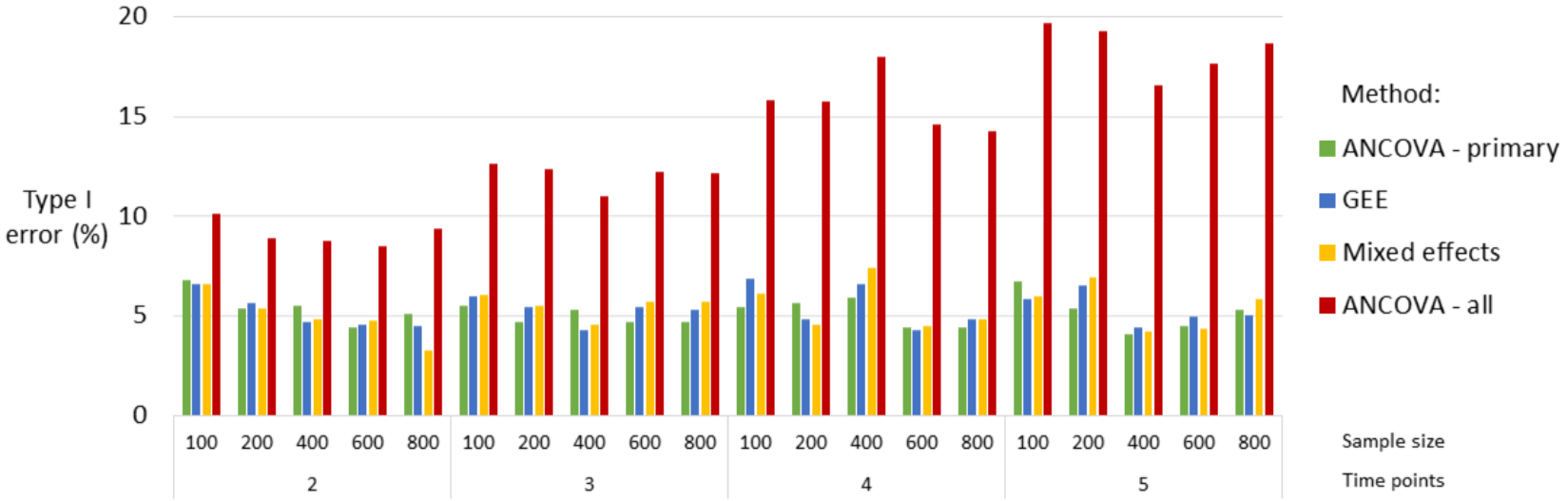
			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Power	37.5	90.0	99.9	36.2	87.7	99.1	34.8	86.7	99.6	37.3	88.8	99.8
	(primary)	MCSE	(1.21)	(0.75)	(0.06)	(1.20)	(0.82)	(0.23)	(1.19)	(0.85)	(0.17)	(1.21)	(0.79)	(0.13)
	GEE	Power	16.0	44.3	79.3	28.1	72.2	95.9	17.2	47.8	80.7	14.0	34.9	64.1
		MCSE	(0.92)	(1.24)	(1.01)	(1.12)	(1.12)	(0.49)	(0.94)	(1.25)	(0.99)	(0.87)	(1.19)	(1.20)
	Mixed	Power	17.6	49.9	84.4	29.5	76.7	97.4	20.1	55.3	88.5	16.2	45.7	83.6
	effects	MCSE	(1.01)	(1.31)	(0.95)	(1.30)	(1.22)	(0.46)	(1.25)	(1.57)	(1.02)	(1.21)	(1.63)	(1.24)
200	ANCOVA	Power	61.6	99.3	100.0	62.9	99.5	100.0	60.9	99.5	100.0	62.6	99.3	100.0
	(primary)	MCSE	(1.22)	(0.22)	(0.00)	(1.21)	(0.18)	(0.00)	(1.22)	(0.18)	(0.00)	(1.21)	(0.21)	(0.00)
	GEE	Power	25.3	71.2	95.9	46.6	94.9	100.0	26.8	79.5	97.6	19.4	57.6	88.4
		MCSE	(1.09)	(1.13)	(0.50)	(1.25)	(0.55)	(0.00)	(1.11)	(1.01)	(0.39)	(0.99)	(1.24)	(0.80)
	Mixed	Power	27.8	74.9	97.3	49.2	95.7	100.0	29.8	85.1	98.9	23.1	67.2	96.8
	effects	MCSE	(1.22)	(1.19)	(0.44)	(1.50)	(0.62)	(0.00)	(1.47)	(1.15)	(0.35)	(1.42)	(1.66)	(0.65)
400	ANCOVA	Power	89.0	100.0	100.0	89.4	100.0	100.0	88.6	100.0	100.0	89.5	100.0	100.0
	(primary)	MCSE	(0.78)	(0.00)	(0.00)	(0.77)	(0.00)	(0.00)	(0.80)	(0.00)	(0.00)	(0.77)	(0.00)	(0.00)
	GEE	Power	43.8	94.8	100.0	72.8	99.9	100.0	49.3	97.0	100.0	33.6	88.1	99.3
		MCSE	(1.24)	(0.55)	(0.00)	(1.11)	(0.06)	(0.00)	(1.25)	(0.43)	(0.00)	(1.18)	(0.81)	(0.22)
	Mixed	Power	46.5	96.1	100.0	74.5	99.9	100.0	52.1	98.4	100.0	37.3	92.4	99.8
	effects	MCSE	(1.47)	(0.57)	(0.00)	(1.38)	(0.10)	(0.00)	(1.69)	(0.43)	(0.00)	(1.75)	(1.03)	(0.23)

Table 6.6: Statistical power (for 95% confidence interval): Pattern 3 (temporary improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Power	97.6	100.0	100.0	97.2	100.0	100.0	97.5	100.0	100.0	97.8	100.0	100.0
	(primary)	MCSE	(0.38)	(0.00)	(0.00)	(0.41)	(0.00)	(0.00)	(0.39)	(0.00)	(0.00)	(0.37)	(0.00)	(0.00)
	GEE	Power	61.3	98.8	100.0	87.8	100.0	100.0	67.3	99.9	100.0	50.8	96.6	100.0
		MCSE	(1.22)	(0.27)	(0.00)	(0.82)	(0.00)	(0.00)	(1.17)	(0.09)	(0.00)	(1.25)	(0.45)	(0.00)
	Mixed	Power	64.6	99.0	100.0	88.9	100.0	100.0	72.7	99.9	100.0	56.2	98.3	100.0
	effects	MCSE	(1.48)	(0.30)	(0.00)	(1.08)	(0.00)	(0.00)	(1.57)	(0.14)	(0.00)	(1.89)	(0.55)	(0.00)
800	ANCOVA	Power	99.1	100.0	100.0	99.4	100.0	100.0	99.6	100.0	100.0	99.4	100.0	100.0
	(primary)	MCSE	(0.24)	(0.00)	(0.00)	(0.20)	(0.00)	(0.00)	(0.15)	(0.00)	(0.00)	(0.20)	(0.00)	(0.00)
	GEE	Power	72.4	99.9	100.0	94.6	100.0	100.0	79.8	100.0	100.0	61.9	99.1	100.0
		MCSE	(1.12)	(0.06)	(0.00)	(0.57)	(0.00)	(0.00)	(1.00)	(0.00)	(0.00)	(1.21)	(0.24)	(0.00)
	Mixed	Power	74.4	99.9	100.0	94.9	100.0	100.0	84.1	100.0	100.0	67.1	99.5	100.0
	effects	MCSE	(1.43)	(0.11)	(0.00)	(0.80)	(0.00)	(0.00)	(1.44)	(0.00)	(0.00)	(2.04)	(0.32)	(0.00)

$\beta*$: Maximum treatment effect.

Figure 6.7: Type I error: Pattern 1 (linear improvement)



Note: Treatment effect (β) is 0 for type I error calculation.

Table 6.7: Type I error (for 95% confidence interval): Pattern 1 (linear improvement)

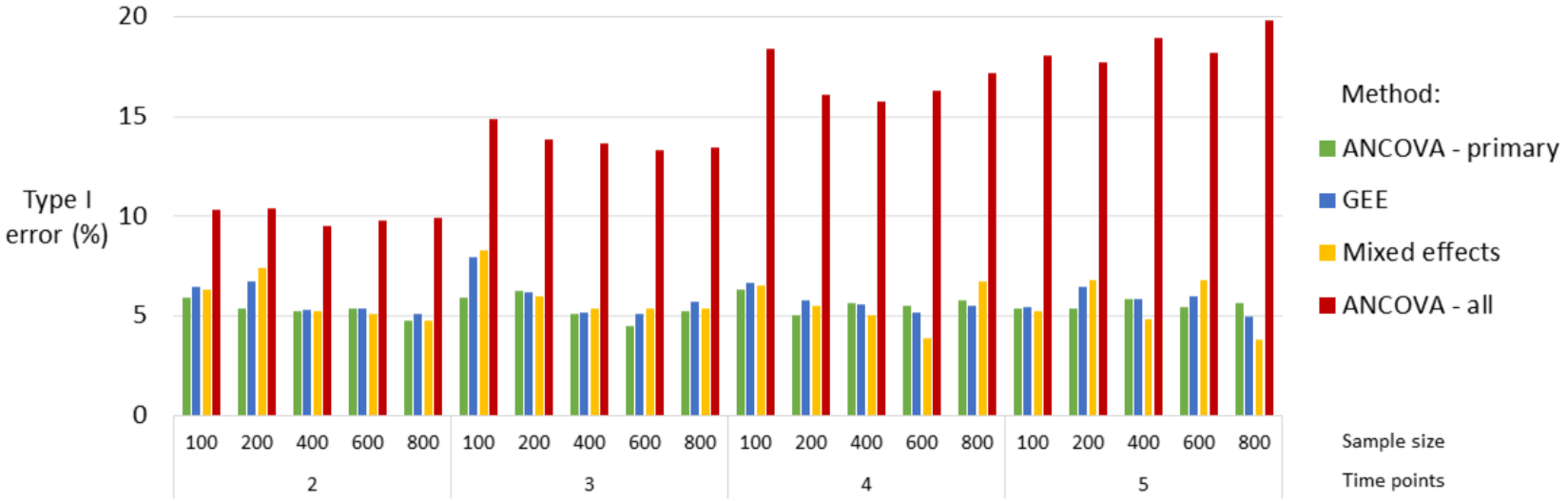
			Number of time points:			
n	**Method**		**2**	**3**	**4**	**5**
100	ANCOVA (primary)	Type I error	6.8	5.5	5.4	6.7
		MCSE	(0.63)	(0.57)	(0.57)	(0.63)
	GEE	Type I error	6.6	6.0	6.9	5.8
		MCSE	(0.62)	(0.59)	(0.63)	(0.58)
	Mixed effects	Type I error	6.6	6.0	6.1	6.0
		MCSE	(0.65)	(0.69)	(0.72)	(0.76)
	ANCOVA (all time points)	Type I error	10.1	12.6	15.8	19.7
		MCSE	(0.75)	(0.83)	(0.91)	(0.99)
200	ANCOVA (primary)	Type I error	5.4	4.7	5.6	5.4
		MCSE	(0.56)	(0.53)	(0.58)	(0.56)
	GEE	Type I error	5.6	5.4	4.8	6.5
		MCSE	(0.58)	(0.57)	(0.54)	(0.62)
	Mixed effects	Type I error	5.4	5.5	4.5	6.9
		MCSE	(0.62)	(0.68)	(0.65)	(0.86)
	ANCOVA (all time points)	Type I error	8.8	12.4	15.8	19.3
		MCSE	(0.71)	(0.82)	(0.91)	(0.99)
400	ANCOVA (primary)	Type I error	5.5	5.3	5.9	4.1
		MCSE	(0.57)	(0.56)	(0.59)	(0.49)
	GEE	Type I error	4.7	4.3	6.6	4.4
		MCSE	(0.53)	(0.51)	(0.62)	(0.52)
	Mixed effects	Type I error	4.8	4.5	7.4	4.2
		MCSE	(0.63)	(0.66)	(0.87)	(0.70)
	ANCOVA (all time points)	Type I error	8.7	11.0	18.0	16.6
		MCSE	(0.71)	(0.78)	(0.96)	(0.93)
600	ANCOVA (primary)	Type I error	4.4	4.7	4.4	4.5
		MCSE	(0.51)	(0.53)	(0.51)	(0.52)
	GEE	Type I error	4.6	5.4	4.3	4.9
		MCSE	(0.52)	(0.57)	(0.50)	(0.54)
	Mixed effects	Type I error	4.8	5.7	4.5	4.4
		MCSE	(0.66)	(0.78)	(0.76)	(0.77)
	ANCOVA (all time points)	Type I error	8.5	12.2	14.6	17.6
		MCSE	(0.70)	(0.82)	(0.88)	(0.95)

Table 6.7: Type I error (for 95% confidence interval): Pattern 1 (linear improvement)

			Number of time points:			
n	**Method**		**2**	**3**	**4**	**5**
800	ANCOVA (primary)	Type I error	5.1	4.7	4.4	5.3
		MCSE	(0.55)	(0.53)	(0.51)	(0.56)
	GEE	Type I error	4.5	5.3	4.8	5.0
		MCSE	(0.52)	(0.56)	(0.54)	(0.55)
	Mixed effects	Type I error	3.3	5.7	4.9	5.9
		MCSE	(0.58)	(0.81)	(0.84)	(0.99)
	ANCOVA (all time points)	Type I error	9.4	12.1	14.3	18.7
		MCSE	(0.73)	(0.82)	(0.87)	(0.97)

Note: Treatment effect (β) is 0 for type I error calculation.

Figure 6.8: Type I error: Pattern 2 (short-term improvement then plateau)



Note: Treatment effect (β) is 0 for type I error calculation.

Table 6.8: Type I error (for 95% confidence interval): Pattern 2 (short-term improvement then plateau)

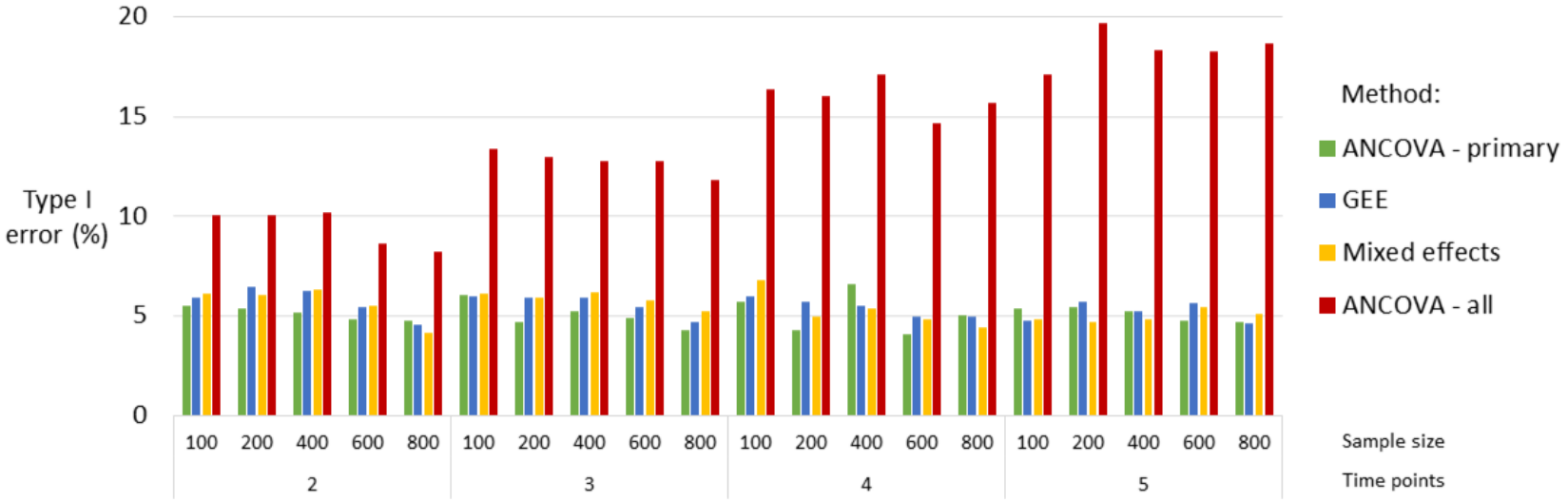
			Number of time points:			
n	**Method**		**2**	**3**	**4**	**5**
100	ANCOVA (primary)	Type I error	5.9	5.9	6.3	5.4
		MCSE	(0.59)	(0.59)	(0.61)	(0.56)
	GEE	Type I error	6.4	7.9	6.6	5.4
		MCSE	(0.61)	(0.68)	(0.62)	(0.57)
	Mixed effects	Type I error	6.3	8.3	6.5	5.2
		MCSE	(0.64)	(0.79)	(0.76)	(0.72)
	ANCOVA (all time points)	Type I error	10.3	14.9	18.4	18.1
		MCSE	(0.76)	(0.89)	(0.97)	(0.96)
200	ANCOVA (primary)	Type I error	5.4	6.3	5.0	5.4
		MCSE	(0.56)	(0.60)	(0.55)	(0.56)
	GEE	Type I error	6.8	6.2	5.8	6.4
		MCSE	(0.63)	(0.60)	(0.58)	(0.61)
	Mixed effects	Type I error	7.4	6.0	5.5	6.8
		MCSE	(0.72)	(0.71)	(0.72)	(0.83)
	ANCOVA (all time points)	Type I error	10.4	13.9	16.1	17.8
		MCSE	(0.76)	(0.86)	(0.92)	(0.96)
400	ANCOVA (primary)	Type I error	5.3	5.1	5.6	5.8
		MCSE	(0.56)	(0.55)	(0.58)	(0.58)
	GEE	Type I error	5.3	5.2	5.6	5.8
		MCSE	(0.56)	(0.55)	(0.57)	(0.58)
	Mixed effects	Type I error	5.3	5.4	5.0	4.8
		MCSE	(0.66)	(0.72)	(0.74)	(0.74)
	ANCOVA (all time points)	Type I error	9.5	13.6	15.8	18.9
		MCSE	(0.73)	(0.86)	(0.91)	(0.98)
600	ANCOVA (primary)	Type I error	5.4	4.5	5.5	5.4
		MCSE	(0.56)	(0.52)	(0.57)	(0.57)
	GEE	Type I error	5.4	5.1	5.2	6.0
		MCSE	(0.56)	(0.55)	(0.55)	(0.59)
	Mixed effects	Type I error	5.1	5.4	3.9	6.8
		MCSE	(0.68)	(0.76)	(0.68)	(0.97)
	ANCOVA (all time points)	Type I error	9.8	13.3	16.3	18.2
		MCSE	(0.74)	(0.85)	(0.92)	(0.96)

Table 6.8: Type I error (for 95% confidence interval): Pattern 2 (short-term improvement then plateau)

n	**Method**		**Number of time points: 2**	**3**	**4**	**5**
800	ANCOVA (primary)	Type I error	4.8	5.3	5.8	5.6
		MCSE	(0.53)	(0.56)	(0.58)	(0.58)
	GEE	Type I error	5.1	5.7	5.5	4.9
		MCSE	(0.55)	(0.58)	(0.57)	(0.54)
	Mixed effects	Type I error	4.8	5.3	6.8	3.8
		MCSE	(0.69)	(0.80)	(0.97)	(0.81)
	ANCOVA (all time points)	Type I error	9.9	13.4	17.2	19.8
		MCSE	(0.75)	(0.85)	(0.94)	(1.00)

Note: Treatment effect (β) is 0 for type I error calculation.

Figure 6.9: Type I error: Pattern 3 (temporary improvement)



Note: Treatment effect (β) is 0 for type I error calculation.

Table 6.9: Type I error (for 95% confidence interval): Pattern 3 (temporary improvement)

			Number of time points:			
n	**Method**		**2**	**3**	**4**	**5**
100	ANCOVA (primary)	Type I error	5.5	6.1	5.7	5.4
		MCSE	(0.57)	(0.60)	(0.58)	(0.56)
	GEE	Type I error	5.9	6.0	6.0	4.8
		MCSE	(0.59)	(0.59)	(0.59)	(0.53)
	Mixed effects	Type I error	6.1	6.1	6.8	4.8
		MCSE	(0.63)	(0.69)	(0.78)	(0.67)
	ANCOVA (all time points)	Type I error	10.1	13.4	16.4	17.1
		MCSE	(0.75)	(0.85)	(0.93)	(0.94)
200	ANCOVA (primary)	Type I error	5.4	4.7	4.3	5.4
		MCSE	(0.56)	(0.53)	(0.50)	(0.57)
	GEE	Type I error	6.4	5.9	5.7	5.7
		MCSE	(0.61)	(0.59)	(0.58)	(0.58)
	Mixed effects	Type I error	6.1	5.9	4.9	4.7
		MCSE	(0.65)	(0.70)	(0.69)	(0.69)
	ANCOVA (all time points)	Type I error	10.1	12.9	16.0	19.7
		MCSE	(0.75)	(0.84)	(0.92)	(0.99)
400	ANCOVA (primary)	Type I error	5.1	5.3	6.6	5.3
		MCSE	(0.55)	(0.56)	(0.62)	(0.56)
	GEE	Type I error	6.3	5.9	5.5	5.3
		MCSE	(0.60)	(0.59)	(0.57)	(0.56)
	Mixed effects	Type I error	6.3	6.2	5.4	4.8
		MCSE	(0.73)	(0.76)	(0.75)	(0.76)
	ANCOVA (all time points)	Type I error	10.2	12.8	17.1	18.3
		MCSE	(0.76)	(0.83)	(0.94)	(0.97)
600	ANCOVA (primary)	Type I error	4.8	4.9	4.1	4.8
		MCSE	(0.54)	(0.54)	(0.49)	(0.53)
	GEE	Type I error	5.4	5.4	4.9	5.6
		MCSE	(0.57)	(0.57)	(0.54)	(0.58)
	Mixed effects	Type I error	5.5	5.7	4.8	5.4
		MCSE	(0.72)	(0.79)	(0.76)	(0.87)
	ANCOVA (all time points)	Type I error	8.6	12.8	14.7	18.3
		MCSE	(0.70)	(0.83)	(0.89)	(0.97)

Table 6.9: Type I error (for 95% confidence interval): Pattern 3 (temporary improvement)

n	**Method**		**Number of time points: 2**	**3**	**4**	**5**
800	ANCOVA (primary)	Type I error	4.8	4.3	5.0	4.7
		MCSE	(0.53)	(0.51)	(0.55)	(0.53)
	GEE	Type I error	4.6	4.7	4.9	4.6
		MCSE	(0.52)	(0.53)	(0.54)	(0.53)
	Mixed effects	Type I error	4.1	5.3	4.4	5.1
		MCSE	(0.65)	(0.82)	(0.80)	(0.90)
	ANCOVA (all time points)	Type I error	8.2	11.8	15.7	18.7
		MCSE	(0.69)	(0.81)	(0.91)	(0.97)

Note: Treatment effect (β) is 0 for type I error calculation.

Table 6.10: Convergence for mixed effects method

Pattern	**n**		**Time points: 2**			**Time points: 3**			**Time points: 4**			**Time points: 5**		
		$\beta*$:	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100		89%	91%	89%	79%	77%	77%	66%	67%	64%	60%	60%	57%
1	200		82%	83%	81%	69%	71%	70%	61%	60%	58%	55%	54%	50%
1	400		73%	73%	72%	62%	63%	60%	56%	55%	53%	49%	44%	37%
1	600		65%	67%	64%	53%	52%	53%	50%	49%	46%	43%	39%	31%
1	800		59%	61%	58%	46%	47%	45%	43%	41%	40%	37%	34%	26%
2	100		90%	90%	89%	77%	75%	76%	65%	64%	64%	60%	62%	60%
2	200		82%	81%	82%	72%	71%	68%	62%	62%	62%	56%	57%	55%
2	400		71%	71%	71%	61%	62%	62%	54%	55%	57%	51%	48%	48%
2	600		64%	64%	66%	56%	56%	52%	47%	49%	49%	43%	43%	41%
2	800		59%	57%	58%	50%	48%	46%	40%	40%	40%	34%	34%	35%
3	100		89%	91%	91%	76%	75%	76%	65%	63%	61%	58%	59%	55%
3	200		84%	84%	86%	69%	67%	70%	61%	60%	56%	55%	50%	46%
3	400		72%	71%	79%	62%	61%	58%	55%	52%	43%	47%	41%	28%
3	600		65%	68%	72%	53%	54%	52%	50%	45%	36%	43%	34%	18%
3	800		59%	59%	70%	47%	47%	44%	40%	41%	29%	33%	27%	14%

$\beta*$: Maximum treatment effect.

6.3.2 Properties of the confidence interval: coverage

The magnitude of the treatment effect was the main factor associated with differences in the level of coverage (Tables 6.11 to 6.13). For all methods, coverage was <90% primarily when the treatment effect was 12 points. However, such a large treatment effect is unlikely to be present in practice. Coverage was closer to 95% when the treatment effect was 4 points. However, it is important to bear in mind that the true value in each scenario differed between the different methods, which could have influenced the level of coverage (Table 6.3).

Poor coverage for larger treatment effects was especially apparent in pattern 2 when the treatment effect was sustained over time (Table 6.12). In pattern 2, coverage was <90% when the treatment effect was 8 or 12 points for larger sample sizes. For pattern 2, coverage was better for the ANCOVA method at the primary time point for almost all scenarios. The improved coverage using the ANCOVA (primary) method was more apparent for larger treatment effects. The mixed effects method never produced the best coverage of the three methods across all scenarios in pattern 2.

For pattern 1 (linear improvement in treatment effect), the ANCOVA (primary) method commonly showed better coverage than the GEE or mixed effects methods for smaller sample sizes ($n \leq 400$). However, the GEE method had better coverage for large sample size ($n \geq 600$), especially when the treatment effect was large. The mixed effects method very rarely produced optimal coverage for pattern 1.

For pattern 3 (temporary improvement), no method was consistently preferred across the different scenarios. Coverage was reasonable and did not differ greatly between the different methods when the sample size was low. The ANCOVA (primary) method produced the best coverage for the smallest sample size ($n = 100$). For larger sample sizes ($n \geq 600$), the mixed effects method often provided the best coverage, especially

if treatment effects were large.

Table 6.11: Coverage (for 95% confidence interval): Pattern 1 (linear improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Cover.	94.8	94.9	93.3	94.9	93.8	93.8	95.4	95.1	93.1	94.5	95.4	94.7
	(primary)	MCSE	(0.55)	(0.55)	(0.63)	(0.55)	(0.60)	(0.60)	(0.52)	(0.54)	(0.63)	(0.57)	(0.53)	(0.56)
	GEE	Cover.	94.4	94.2	92.9	94.8	93.9	93.5	94.7	94.0	92.3	93.8	94.5	93.0
		MCSE	(0.57)	(0.58)	(0.64)	(0.56)	(0.60)	(0.62)	(0.56)	(0.59)	(0.67)	(0.60)	(0.57)	(0.64)
	Mixed	Cover.	94.6	93.2	91.2	94.5	93.4	91.3	94.7	92.7	88.6	93.6	94.1	85.7
	effects	MCSE	(0.60)	(0.66)	(0.75)	(0.64)	(0.71)	(0.80)	(0.69)	(0.80)	(0.99)	(0.79)	(0.76)	(1.16)
200	ANCOVA	Cover.	95.0	93.8	91.9	95.1	93.6	92.6	93.6	93.1	92.3	94.6	93.9	91.0
	(primary)	MCSE	(0.55)	(0.60)	(0.68)	(0.54)	(0.61)	(0.66)	(0.61)	(0.63)	(0.67)	(0.56)	(0.60)	(0.71)
	GEE	Cover.	94.0	93.4	92.5	94.5	93.1	93.4	94.1	93.3	91.9	94.1	93.1	92.1
		MCSE	(0.59)	(0.62)	(0.66)	(0.57)	(0.63)	(0.62)	(0.59)	(0.63)	(0.68)	(0.59)	(0.63)	(0.67)
	Mixed	Cover.	94.0	92.3	91.5	94.1	92.2	91.0	94.5	92.8	87.2	94.0	91.2	84.3
	effects	MCSE	(0.66)	(0.73)	(0.78)	(0.71)	(0.80)	(0.85)	(0.73)	(0.84)	(1.10)	(0.80)	(0.97)	(1.29)
400	ANCOVA	Cover.	94.4	94.4	90.0	94.4	93.4	88.3	95.3	93.5	88.4	95.2	92.3	89.5
	(primary)	MCSE	(0.58)	(0.57)	(0.75)	(0.58)	(0.62)	(0.80)	(0.53)	(0.62)	(0.80)	(0.54)	(0.67)	(0.77)
	GEE	Cover.	93.8	94.0	89.5	94.9	93.3	90.1	93.4	93.3	88.3	95.0	93.3	89.3
		MCSE	(0.60)	(0.59)	(0.77)	(0.55)	(0.63)	(0.75)	(0.62)	(0.63)	(0.81)	(0.55)	(0.63)	(0.77)
	Mixed	Cover.	93.3	92.8	88.0	94.8	91.7	86.4	93.0	90.7	81.0	95.5	90.7	78.3
	effects	MCSE	(0.73)	(0.76)	(0.96)	(0.70)	(0.87)	(1.11)	(0.85)	(0.97)	(1.35)	(0.74)	(1.09)	(1.69)

Table 6.11: Coverage (for 95% confidence interval): Pattern 1 (linear improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Cover.	93.9	91.6	83.9	94.3	92.1	86.3	94.3	91.4	84.4	94.3	91.5	86.1
	(primary)	MCSE	(0.60)	(0.70)	(0.92)	(0.58)	(0.67)	(0.86)	(0.58)	(0.70)	(0.91)	(0.58)	(0.70)	(0.87)
	GEE	Cover.	94.2	90.5	85.8	95.6	92.3	88.3	94.6	92.4	86.1	94.9	92.3	87.3
		MCSE	(0.58)	(0.73)	(0.87)	(0.52)	(0.67)	(0.80)	(0.57)	(0.66)	(0.87)	(0.55)	(0.67)	(0.83)
	Mixed	Cover.	94.4	90.8	82.8	94.8	92.1	84.2	94.1	89.1	76.4	93.4	86.5	71.1
	effects	MCSE	(0.71)	(0.88)	(1.18)	(0.77)	(0.94)	(1.26)	(0.83)	(1.11)	(1.57)	(0.95)	(1.38)	(2.03)
800	ANCOVA	Cover.	93.9	90.6	82.6	94.5	91.3	81.7	94.3	91.4	82.3	94.5	90.8	81.1
	(primary)	MCSE	(0.60)	(0.73)	(0.95)	(0.57)	(0.70)	(0.97)	(0.58)	(0.70)	(0.96)	(0.57)	(0.72)	(0.98)
	GEE	Cover.	94.7	91.6	83.3	94.8	91.9	84.9	93.9	91.4	84.5	94.4	90.9	81.9
		MCSE	(0.56)	(0.70)	(0.93)	(0.55)	(0.68)	(0.90)	(0.60)	(0.70)	(0.91)	(0.57)	(0.72)	(0.96)
	Mixed	Cover.	95.0	90.0	79.9	95.6	89.3	81.0	94.0	86.3	71.3	93.8	88.9	56.6
	effects	MCSE	(0.71)	(0.96)	(1.31)	(0.76)	(1.13)	(1.47)	(0.91)	(1.34)	(1.79)	(0.99)	(1.35)	(2.45)

$\beta*$: Maximum treatment effect.

Table 6.12: Coverage (for 95% confidence interval): Pattern 2 (short-term improvement then plateau)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Cover.	94.8	94.1	92.6	94.8	94.6	93.4	93.4	93.5	91.8	94.9	93.8	93.5
	(primary)	MCSE	(0.55)	(0.59)	(0.66)	(0.55)	(0.57)	(0.62)	(0.62)	(0.62)	(0.69)	(0.55)	(0.60)	(0.62)
	GEE	Cover.	93.6	93.7	91.5	94.3	93.4	91.8	93.4	93.8	90.2	93.7	93.0	91.4
		MCSE	(0.61)	(0.61)	(0.70)	(0.58)	(0.62)	(0.69)	(0.62)	(0.60)	(0.74)	(0.61)	(0.64)	(0.70)
	Mixed	Cover.	94.2	93.6	91.0	93.3	92.4	90.6	92.8	92.5	89.2	94.3	92.7	89.3
	effects	MCSE	(0.62)	(0.65)	(0.76)	(0.71)	(0.77)	(0.84)	(0.80)	(0.82)	(0.97)	(0.75)	(0.83)	(0.99)
200	ANCOVA	Cover.	94.4	92.9	90.1	94.8	93.8	90.2	94.3	93.1	90.0	94.8	92.8	89.7
	(primary)	MCSE	(0.58)	(0.64)	(0.75)	(0.56)	(0.60)	(0.74)	(0.58)	(0.63)	(0.75)	(0.56)	(0.65)	(0.76)
	GEE	Cover.	94.3	93.1	88.5	94.9	92.6	89.2	94.3	92.4	86.0	94.3	91.4	84.8
		MCSE	(0.58)	(0.63)	(0.80)	(0.55)	(0.66)	(0.78)	(0.58)	(0.66)	(0.87)	(0.58)	(0.70)	(0.90)
	Mixed	Cover.	93.7	92.8	89.5	95.2	91.6	87.5	94.2	91.9	85.6	94.8	91.5	83.3
	effects	MCSE	(0.67)	(0.72)	(0.85)	(0.63)	(0.82)	(1.00)	(0.74)	(0.86)	(1.11)	(0.74)	(0.92)	(1.26)
400	ANCOVA	Cover.	94.6	92.3	86.4	93.6	92.8	85.5	94.4	90.3	86.9	93.4	91.4	84.1
	(primary)	MCSE	(0.57)	(0.67)	(0.86)	(0.61)	(0.65)	(0.88)	(0.58)	(0.74)	(0.84)	(0.62)	(0.70)	(0.91)
	GEE	Cover.	94.4	91.1	83.5	93.6	91.6	82.2	94.3	89.6	81.4	92.4	89.4	78.6
		MCSE	(0.57)	(0.71)	(0.93)	(0.61)	(0.69)	(0.96)	(0.58)	(0.76)	(0.97)	(0.66)	(0.77)	(1.02)
	Mixed	Cover.	94.5	91.2	82.4	92.9	91.8	79.5	94.4	89.0	77.5	91.0	88.3	76.4
	effects	MCSE	(0.67)	(0.84)	(1.13)	(0.82)	(0.87)	(1.28)	(0.78)	(1.05)	(1.39)	(1.00)	(1.16)	(1.53)

Table 6.12: Coverage (for 95% confidence interval): Pattern 2 (short-term improvement then plateau)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Cover.	94.9	90.8	83.4	94.9	91.1	81.4	93.8	90.6	80.8	94.3	89.4	78.7
	(primary)	MCSE	(0.55)	(0.72)	(0.93)	(0.55)	(0.71)	(0.97)	(0.60)	(0.73)	(0.98)	(0.58)	(0.77)	(1.02)
	GEE	Cover.	94.4	89.3	77.8	93.2	88.0	79.3	94.0	88.8	72.4	93.1	86.8	66.3
		MCSE	(0.57)	(0.77)	(1.04)	(0.63)	(0.81)	(1.01)	(0.59)	(0.79)	(1.12)	(0.63)	(0.85)	(1.18)
	Mixed	Cover.	94.2	89.4	78.3	93.5	86.3	77.1	92.8	87.3	69.7	91.8	86.2	61.4
	effects	MCSE	(0.73)	(0.96)	(1.27)	(0.83)	(1.15)	(1.45)	(0.95)	(1.19)	(1.65)	(1.05)	(1.31)	(1.90)
800	ANCOVA	Cover.	93.7	89.5	78.2	94.3	89.4	78.5	94.2	88.3	73.6	92.9	86.9	71.8
	(primary)	MCSE	(0.61)	(0.77)	(1.03)	(0.58)	(0.77)	(1.03)	(0.58)	(0.80)	(1.10)	(0.64)	(0.84)	(1.13)
	GEE	Cover.	94.1	87.9	72.4	94.3	86.8	73.7	93.2	85.4	65.3	91.3	83.0	59.9
		MCSE	(0.59)	(0.82)	(1.12)	(0.58)	(0.85)	(1.10)	(0.63)	(0.88)	(1.19)	(0.70)	(0.94)	(1.23)
	Mixed	Cover.	93.2	87.9	72.0	93.7	85.2	70.8	93.8	85.1	63.6	91.2	80.9	56.4
	effects	MCSE	(0.83)	(1.08)	(1.48)	(0.86)	(1.28)	(1.67)	(0.96)	(1.40)	(1.91)	(1.21)	(1.69)	(2.10)

$\beta*$: Maximum treatment effect.

Table 6.13: Coverage (for 95% confidence interval): Pattern 3 (temporary improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Cover.	94.6	93.9	93.3	93.6	93.9	93.4	95.3	93.5	92.5	95.8	94.8	93.5
	(primary)	MCSE	(0.57)	(0.60)	(0.63)	(0.61)	(0.60)	(0.62)	(0.53)	(0.62)	(0.66)	(0.50)	(0.56)	(0.62)
	GEE	Cover.	94.6	93.9	93.7	93.0	93.3	92.5	95.1	92.3	92.3	94.7	94.8	93.4
		MCSE	(0.57)	(0.60)	(0.61)	(0.64)	(0.63)	(0.66)	(0.54)	(0.67)	(0.67)	(0.56)	(0.56)	(0.62)
	Mixed	Cover.	93.8	93.2	93.0	92.8	93.4	92.5	95.0	93.1	90.8	93.9	95.3	88.5
	effects	MCSE	(0.64)	(0.66)	(0.67)	(0.74)	(0.72)	(0.76)	(0.68)	(0.80)	(0.92)	(0.78)	(0.69)	(1.07)
200	ANCOVA	Cover.	95.1	93.6	91.4	95.1	94.1	90.6	94.6	93.6	89.6	95.0	92.1	91.4
	(primary)	MCSE	(0.54)	(0.61)	(0.70)	(0.54)	(0.59)	(0.73)	(0.56)	(0.61)	(0.76)	(0.55)	(0.67)	(0.70)
	GEE	Cover.	94.7	94.0	91.8	94.1	94.1	90.0	94.6	94.3	90.8	94.4	92.6	90.2
		MCSE	(0.56)	(0.59)	(0.69)	(0.59)	(0.59)	(0.75)	(0.57)	(0.58)	(0.72)	(0.57)	(0.65)	(0.74)
	Mixed	Cover.	95.2	93.9	92.1	94.6	94.9	92.4	95.8	93.6	91.5	95.2	91.5	86.8
	effects	MCSE	(0.59)	(0.66)	(0.73)	(0.68)	(0.67)	(0.79)	(0.65)	(0.79)	(0.93)	(0.72)	(0.99)	(1.24)
400	ANCOVA	Cover.	94.1	94.0	89.5	95.3	93.1	88.4	95.2	92.6	90.8	95.9	93.5	87.9
	(primary)	MCSE	(0.59)	(0.59)	(0.77)	(0.53)	(0.64)	(0.80)	(0.54)	(0.66)	(0.72)	(0.49)	(0.62)	(0.82)
	GEE	Cover.	95.2	94.2	90.9	94.9	92.1	85.9	95.3	92.1	90.0	95.0	92.0	87.9
		MCSE	(0.54)	(0.58)	(0.72)	(0.55)	(0.67)	(0.87)	(0.53)	(0.68)	(0.75)	(0.55)	(0.68)	(0.82)
	Mixed	Cover.	95.4	94.9	91.8	95.3	93.5	91.4	95.7	94.0	91.8	95.3	92.3	81.9
	effects	MCSE	(0.62)	(0.65)	(0.77)	(0.67)	(0.79)	(0.92)	(0.69)	(0.82)	(1.05)	(0.77)	(1.04)	(1.83)

Table 6.13: Coverage (for 95% confidence interval): Pattern 3 (temporary improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Cover.	94.8	91.6	84.9	95.0	91.5	84.4	93.9	92.8	85.1	94.7	91.0	85.1
	(primary)	MCSE	(0.55)	(0.70)	(0.89)	(0.55)	(0.70)	(0.91)	(0.60)	(0.65)	(0.89)	(0.56)	(0.71)	(0.89)
	GEE	Cover.	94.6	93.6	88.2	95.3	90.8	81.7	94.2	91.6	83.7	95.0	92.2	85.4
		MCSE	(0.56)	(0.61)	(0.81)	(0.53)	(0.72)	(0.97)	(0.58)	(0.70)	(0.92)	(0.55)	(0.67)	(0.88)
	Mixed	Cover.	94.6	93.7	90.2	95.1	93.0	89.5	93.9	93.1	91.5	94.7	91.6	80.4
	effects	MCSE	(0.70)	(0.74)	(0.88)	(0.74)	(0.87)	(1.06)	(0.85)	(0.94)	(1.17)	(0.86)	(1.20)	(2.33)
800	ANCOVA	Cover.	94.8	90.8	81.6	94.1	90.8	80.3	95.1	91.4	80.9	94.8	90.6	81.0
	(primary)	MCSE	(0.56)	(0.72)	(0.97)	(0.59)	(0.72)	(0.99)	(0.54)	(0.70)	(0.98)	(0.55)	(0.73)	(0.98)
	GEE	Cover.	94.8	93.3	85.6	93.6	90.2	78.7	94.8	90.3	79.7	95.0	90.8	80.5
		MCSE	(0.55)	(0.63)	(0.88)	(0.61)	(0.74)	(1.02)	(0.56)	(0.74)	(1.01)	(0.55)	(0.72)	(0.99)
	Mixed	Cover.	93.9	94.4	90.1	93.3	91.3	88.0	95.4	94.0	91.4	94.1	93.8	74.6
	effects	MCSE	(0.78)	(0.75)	(0.90)	(0.91)	(1.03)	(1.23)	(0.83)	(0.93)	(1.30)	(1.02)	(1.15)	(2.88)

$\beta*$: Maximum treatment effect.

6.3.3 Properties of the estimator: bias and empirical standard error

Although the true values varied, the level of absolute bias was similar across the different methods (Tables 6.14 to 6.16). The levels of absolute bias and relative bias were greater for larger treatment effects. However, the level of bias did not differ greatly when the sample sizes or number of assessment time points varied.

For pattern 1 (linear improvement), the GEE method produced the least biased estimates (with the smallest MCSEs) compared with the mixed effects and ANCOVA (primary) methods in terms of absolute bias (Table 6.14). However, it should be taken into account that the true value for the ANCOVA (primary) method was higher than the true value for the GEE and mixed effects methods in pattern 1, around twice the size when the treatment effect was large (Table 6.3). The GEE and mixed effects methods had the same true values and demonstrated that the GEE method would produce less biased estimates than the mixed effects method for pattern 1. In terms of relative bias, ANCOVA at the primary time point produced the least biased estimates in all scenarios, followed by the GEE method and, lastly, the mixed effects method produced the most biased estimates (Table 6.17).

For pattern 2 (short-term improvement then plateau), the true values were the most similar for the different methods (Table 6.3). The level of bias was also very similar across the different methods for pattern 2, especially for larger sample sizes (Table 6.15). In terms of absolute bias, the GEE method produced slightly less biased estimates when there were three or more follow-up time points. In terms of relative bias, the ANCOVA method at the primary time point produced the least biased estimates in almost all scenarios (Table 6.18).

For pattern 3 (temporary improvement), the mixed effects method produces much less biased results than the GEE method in terms of absolute and relative bias (Tables 6.16 and 6.19). The level of absolute bias was even greater for the ANCOVA (primary) method, although the true value was also much larger. In the majority of scenarios, the level of relative bias was lowest for the mixed effects method and highest for the GEE approach. The mixed effects analysis may have been more likely to reach convergence when the treatment effect estimate was closer to the true value.

The empirical standard error was highest for the ANCOVA (primary) method for all three patterns (Tables 6.20 to 6.22). The increased precision for the GEE and mixed effects methods was more apparent when the number of follow-up time points was greater. The empirical standard error did not differ greatly for different levels of the treatment effect. As would be expected, the empirical standard error was much smaller for larger sample sizes. For the GEE and mixed effects methods, the empirical standard error reduced as the number of follow-up time points increased. For all patterns, the empirical standard error was very similar for the GEE and mixed effects methods when the sample size was large ($n \geq 400$). For smaller sample sizes, the GEE method often produced the most precise estimates, especially for pattern 3 (Table 6.22).

Table 6.14: Bias: Pattern 1 (linear improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Bias	-0.20	-0.49	-0.84	-0.12	-0.43	-0.78	-0.21	-0.51	-0.80	-0.16	-0.40	-0.82
	(primary)	MCSE	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	GEE	Bias	-0.17	-0.43	-0.69	-0.09	-0.33	-0.60	-0.15	-0.37	-0.60	-0.14	-0.33	-0.65
		MCSE	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)
	Mixed	Bias	-0.22	-0.55	-0.85	-0.18	-0.51	-0.91	-0.20	-0.55	-1.02	-0.27	-0.57	-1.18
	effects	MCSE	(0.06)	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)
200	ANCOVA	Bias	-0.20	-0.57	-0.86	-0.18	-0.45	-0.87	-0.21	-0.47	-0.85	-0.15	-0.49	-0.91
	(primary)	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	GEE	Bias	-0.18	-0.43	-0.70	-0.16	-0.34	-0.63	-0.14	-0.36	-0.64	-0.10	-0.41	-0.67
		MCSE	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	Mixed	Bias	-0.20	-0.50	-0.83	-0.18	-0.48	-0.82	-0.20	-0.46	-0.94	-0.16	-0.60	-1.12
	effects	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)
400	ANCOVA	Bias	-0.23	-0.48	-0.85	-0.19	-0.47	-0.84	-0.19	-0.47	-0.82	-0.15	-0.48	-0.88
	(primary)	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	GEE	Bias	-0.21	-0.41	-0.72	-0.15	-0.38	-0.62	-0.14	-0.36	-0.62	-0.14	-0.38	-0.65
		MCSE	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	Bias	-0.24	-0.48	-0.81	-0.19	-0.50	-0.75	-0.18	-0.50	-0.89	-0.19	-0.55	-1.06
	effects	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)

Table 6.14: Bias: Pattern 1 (linear improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Bias	-0.20	-0.51	-0.87	-0.24	-0.45	-0.83	-0.21	-0.49	-0.82	-0.23	-0.52	-0.83
	(primary)	MCSE	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	Bias	-0.17	-0.44	-0.73	-0.17	-0.35	-0.59	-0.17	-0.37	-0.63	-0.18	-0.41	-0.64
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	Bias	-0.17	-0.49	-0.80	-0.18	-0.42	-0.73	-0.21	-0.48	-0.93	-0.17	-0.54	-1.07
	effects	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
800	ANCOVA	Bias	-0.21	-0.47	-0.89	-0.22	-0.49	-0.84	-0.21	-0.49	-0.86	-0.18	-0.45	-0.90
	(primary)	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	Bias	-0.19	-0.39	-0.72	-0.16	-0.37	-0.62	-0.17	-0.38	-0.62	-0.15	-0.34	-0.68
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	Bias	-0.20	-0.43	-0.79	-0.19	-0.47	-0.73	-0.18	-0.53	-0.91	-0.18	-0.49	-1.16
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)

$\beta*$: Maximum treatment effect.

Table 6.15: Bias: Pattern 2 (short-term improvement then plateau)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Bias	-0.24	-0.57	-1.06	-0.28	-0.49	-0.93	-0.29	-0.62	-1.14	-0.30	-0.75	-1.06
	(primary)	MCSE	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	GEE	Bias	-0.27	-0.62	-1.07	-0.22	-0.49	-0.83	-0.23	-0.60	-1.05	-0.30	-0.71	-1.04
		MCSE	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)
	Mixed	Bias	-0.27	-0.61	-1.04	-0.25	-0.59	-0.98	-0.27	-0.70	-1.19	-0.27	-0.83	-1.15
	effects	MCSE	(0.05)	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
200	ANCOVA	Bias	-0.25	-0.54	-1.07	-0.21	-0.51	-0.96	-0.27	-0.61	-1.04	-0.27	-0.66	-1.22
	(primary)	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	GEE	Bias	-0.27	-0.56	-1.04	-0.20	-0.46	-0.89	-0.26	-0.59	-1.01	-0.25	-0.64	-1.14
		MCSE	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	Mixed	Bias	-0.29	-0.57	-1.01	-0.23	-0.52	-0.98	-0.30	-0.64	-1.08	-0.27	-0.67	-1.19
	effects	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
400	ANCOVA	Bias	-0.22	-0.58	-1.01	-0.29	-0.56	-1.01	-0.24	-0.66	-1.08	-0.30	-0.66	-1.16
	(primary)	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	GEE	Bias	-0.20	-0.56	-1.01	-0.24	-0.52	-0.94	-0.25	-0.62	-1.03	-0.29	-0.66	-1.09
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	Bias	-0.20	-0.55	-1.02	-0.27	-0.55	-1.02	-0.24	-0.65	-1.10	-0.34	-0.69	-1.16
	effects	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)

Table 6.15: Bias: Pattern 2 (short-term improvement then plateau)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Bias	-0.25	-0.56	-0.98	-0.27	-0.59	-1.01	-0.25	-0.61	-1.07	-0.26	-0.67	-1.14
	(primary)	MCSE	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	Bias	-0.23	-0.56	-0.98	-0.25	-0.56	-0.91	-0.25	-0.59	-1.04	-0.24	-0.64	-1.11
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	Bias	-0.22	-0.59	-0.97	-0.27	-0.62	-0.99	-0.27	-0.63	-1.09	-0.29	-0.67	-1.18
	effects	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
800	ANCOVA	Bias	-0.24	-0.57	-0.99	-0.23	-0.56	-0.99	-0.29	-0.59	-1.10	-0.30	-0.66	-1.15
	(primary)	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	Bias	-0.22	-0.58	-0.99	-0.20	-0.52	-0.90	-0.27	-0.56	-1.04	-0.29	-0.65	-1.10
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	Bias	-0.25	-0.59	-0.98	-0.20	-0.59	-0.98	-0.27	-0.59	-1.08	-0.34	-0.68	-1.15
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)

$\beta*$: Maximum treatment effect.

Table 6.16: Bias: Pattern 3 (temporary improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Bias	-0.09	-0.41	-0.73	-0.17	-0.51	-0.98	-0.18	-0.59	-0.89	-0.15	-0.43	-0.68
	(primary)	MCSE	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	GEE	Bias	-0.04	-0.29	-0.50	-0.13	-0.41	-0.89	-0.13	-0.49	-0.76	-0.14	-0.36	-0.57
		MCSE	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)
	Mixed	Bias	0.08	0.01	0.04	-0.06	-0.07	-0.32	0.04	-0.12	0.29	0.00	0.11	0.75
	effects	MCSE	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)	(0.07)
200	ANCOVA	Bias	-0.20	-0.53	-0.80	-0.14	-0.57	-0.99	-0.23	-0.45	-1.00	-0.24	-0.50	-0.83
	(primary)	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	GEE	Bias	-0.14	-0.40	-0.56	-0.14	-0.48	-0.80	-0.22	-0.36	-0.81	-0.17	-0.39	-0.66
		MCSE	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	Mixed	Bias	0.00	-0.17	-0.14	-0.05	-0.30	-0.37	-0.11	0.02	0.19	-0.03	0.01	0.62
	effects	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.06)
400	ANCOVA	Bias	-0.21	-0.47	-0.79	-0.23	-0.52	-0.93	-0.22	-0.49	-0.80	-0.23	-0.49	-0.87
	(primary)	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	GEE	Bias	-0.15	-0.34	-0.57	-0.18	-0.46	-0.79	-0.17	-0.36	-0.66	-0.19	-0.34	-0.68
		MCSE	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	Bias	-0.08	-0.17	-0.24	-0.13	-0.29	-0.44	-0.10	-0.04	0.33	-0.09	-0.01	0.55
	effects	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.06)

Table 6.16: Bias: Pattern 3 (temporary improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Bias	-0.20	-0.44	-0.86	-0.23	-0.51	-0.94	-0.21	-0.49	-0.87	-0.17	-0.49	-0.84
	(primary)	MCSE	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	Bias	-0.13	-0.32	-0.62	-0.21	-0.45	-0.80	-0.15	-0.38	-0.71	-0.11	-0.35	-0.64
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	Bias	-0.07	-0.14	-0.33	-0.12	-0.26	-0.43	-0.06	-0.06	0.16	-0.02	0.03	0.60
	effects	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.06)
800	ANCOVA	Bias	-0.16	-0.46	-0.83	-0.19	-0.51	-0.93	-0.17	-0.49	-0.88	-0.18	-0.47	-0.88
	(primary)	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	Bias	-0.13	-0.31	-0.60	-0.18	-0.43	-0.80	-0.14	-0.40	-0.72	-0.14	-0.36	-0.67
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	Bias	-0.07	-0.14	-0.32	-0.17	-0.26	-0.42	-0.06	-0.08	0.22	-0.06	0.01	0.63
	effects	MCSE	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.06)

$\beta*$: Maximum treatment effect.

Table 6.17: Relative bias (%): Pattern 1 (linear improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA (primary)		-5.0	-6.1	-7.0	-2.9	-5.3	-6.5	-5.3	-6.4	-6.7	-4.1	-5.1	-6.8
	GEE		-5.8	-7.2	-7.7	-4.0	-7.0	-8.6	-6.8	-8.4	-9.2	-6.6	-7.8	-10.3
	Mixed effects		-7.2	-9.1	-9.5	-7.6	-10.9	-13.0	-9.0	-12.8	-15.7	-12.7	-13.6	-18.7
200	ANCOVA (primary)		-5.0	-7.1	-7.1	-4.5	-5.6	-7.2	-5.3	-5.9	-7.1	-3.7	-6.2	-7.6
	GEE		-6.1	-7.2	-7.8	-6.9	-7.3	-9.1	-6.7	-8.2	-9.9	-4.8	-9.7	-10.7
	Mixed effects		-6.8	-8.4	-9.2	-7.7	-10.3	-11.7	-9.4	-10.6	-14.4	-7.6	-14.2	-17.7
400	ANCOVA (primary)		-5.8	-6.0	-7.1	-4.7	-5.8	-7.0	-4.8	-5.9	-6.8	-3.8	-6.0	-7.4
	GEE		-7.0	-6.8	-8.1	-6.6	-8.2	-8.9	-6.5	-8.4	-9.5	-6.7	-9.0	-10.4
	Mixed effects		-8.1	-7.9	-9.0	-8.2	-10.8	-10.8	-8.5	-11.6	-13.6	-8.9	-13.1	-16.9
600	ANCOVA (primary)		-5.0	-6.4	-7.3	-6.0	-5.6	-6.9	-5.3	-6.2	-6.9	-5.7	-6.5	-6.9
	GEE		-5.7	-7.4	-8.1	-7.1	-7.4	-8.4	-7.7	-8.6	-9.7	-8.4	-9.7	-10.1
	Mixed effects		-5.8	-8.2	-8.9	-7.6	-9.0	-10.4	-9.9	-11.2	-14.3	-8.3	-12.8	-17.0
800	ANCOVA (primary)		-5.3	-5.9	-7.4	-5.6	-6.1	-7.0	-5.4	-6.1	-7.2	-4.5	-5.6	-7.5
	GEE		-6.4	-6.6	-8.0	-7.0	-7.9	-8.8	-8.0	-8.8	-9.6	-7.2	-8.2	-10.8
	Mixed effects		-6.8	-7.2	-8.7	-8.2	-10.2	-10.4	-8.4	-12.3	-14.0	-8.4	-11.6	-18.4

$\beta*$: Maximum treatment effect.

Table 6.18: Relative bias (%): Pattern 2 (short term improvement then plateau)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA (primary)		-6.1	-7.2	-8.8	-7.1	-6.2	-7.7	-7.3	-7.7	-9.5	-7.6	-9.4	-8.8
	GEE		-6.9	-7.8	-8.9	-6.7	-7.4	-8.3	-6.6	-8.5	-10.0	-8.2	-9.9	-9.6
	Mixed effects		-6.8	-7.6	-8.7	-7.5	-8.8	-9.8	-7.6	-10.0	-11.3	-7.6	-11.5	-10.7
200	ANCOVA (primary)		-6.2	-6.8	-8.9	-5.4	-6.4	-8.0	-6.6	-7.6	-8.6	-6.8	-8.3	-10.2
	GEE		-6.7	-7.0	-8.7	-5.9	-6.9	-8.9	-7.5	-8.5	-9.7	-7.0	-8.9	-10.6
	Mixed effects		-7.3	-7.1	-8.4	-6.8	-7.8	-9.8	-8.5	-9.1	-10.3	-7.5	-9.3	-11.0
400	ANCOVA (primary)		-5.6	-7.2	-8.4	-7.1	-7.0	-8.4	-6.0	-8.2	-9.0	-7.6	-8.3	-9.7
	GEE		-5.0	-7.1	-8.4	-7.3	-7.8	-9.4	-7.1	-8.8	-9.8	-8.0	-9.1	-10.1
	Mixed effects		-5.1	-6.9	-8.5	-8.0	-8.2	-10.2	-6.9	-9.3	-10.4	-9.4	-9.5	-10.7
600	ANCOVA (primary)		-6.2	-7.0	-8.2	-6.8	-7.4	-8.4	-6.2	-7.6	-8.9	-6.5	-8.4	-9.5
	GEE		-5.7	-7.0	-8.1	-7.4	-8.3	-9.1	-7.1	-8.4	-9.9	-6.8	-8.8	-10.3
	Mixed effects		-5.5	-7.4	-8.1	-8.2	-9.4	-9.9	-7.6	-9.0	-10.4	-8.2	-9.3	-10.9
800	ANCOVA (primary)		-5.9	-7.1	-8.2	-5.8	-7.0	-8.3	-7.2	-7.3	-9.1	-7.5	-8.3	-9.6
	GEE		-5.5	-7.2	-8.3	-6.0	-7.7	-9.0	-7.7	-8.0	-9.9	-8.0	-9.0	-10.2
	Mixed effects		-6.2	-7.3	-8.2	-5.9	-8.8	-9.8	-7.7	-8.4	-10.3	-9.4	-9.5	-10.6

$\beta*$: Maximum treatment effect.

Table 6.19: Relative bias (%): Pattern 3 (temporary improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$:	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA (primary)		-2.3	-5.2	-6.0	-4.2	-6.3	-8.2	-4.5	-7.4	-7.4	-3.7	-5.4	-5.7
	GEE		-1.9	-7.2	-8.4	-4.8	-7.8	-11.1	-6.5	-12.3	-12.6	-8.6	-11.1	-11.8
	Mixed effects		4.2	0.2	0.6	-2.4	-1.4	-4.0	1.8	-3.1	4.8	-0.2	3.4	15.6
200	ANCOVA (primary)		-5.0	-6.6	-6.6	-3.5	-7.1	-8.3	-5.8	-5.6	-8.3	-5.9	-6.3	-6.9
	GEE		-7.1	-9.9	-9.4	-5.4	-9.1	-10.0	-11.2	-9.0	-13.4	-10.5	-12.2	-13.7
	Mixed effects		-0.2	-4.2	-2.3	-1.7	-5.6	-4.6	-5.4	0.4	3.2	-2.0	0.4	12.8
400	ANCOVA (primary)		-5.2	-5.9	-6.6	-5.6	-6.5	-7.7	-5.4	-6.1	-6.7	-5.9	-6.1	-7.2
	GEE		-7.6	-8.4	-9.4	-6.7	-8.7	-9.9	-8.4	-9.1	-11.0	-11.7	-10.7	-14.2
	Mixed effects		-3.9	-4.4	-4.0	-5.0	-5.5	-5.5	-4.8	-0.9	5.5	-5.6	-0.4	11.4
600	ANCOVA (primary)		-5.0	-5.6	-7.2	-5.8	-6.4	-7.8	-5.2	-6.1	-7.3	-4.3	-6.2	-7.0
	GEE		-6.7	-8.0	-10.3	-7.7	-8.5	-10.1	-7.4	-9.6	-11.9	-6.9	-11.1	-13.3
	Mixed effects		-3.7	-3.6	-5.5	-4.7	-5.0	-5.3	-2.9	-1.5	2.7	-1.5	0.9	12.6
800	ANCOVA (primary)		-4.1	-5.7	-6.9	-4.8	-6.4	-7.7	-4.2	-6.1	-7.4	-4.5	-5.8	-7.3
	GEE		-6.5	-7.7	-9.9	-6.7	-8.1	-10.0	-7.2	-10.0	-11.9	-8.5	-11.1	-13.9
	Mixed effects		-3.6	-3.5	-5.3	-6.3	-4.8	-5.3	-2.9	-1.9	3.8	-3.5	0.4	13.2

$\beta*$: Maximum treatment effect.

Table 6.20: Empirical standard error (EmpSE): Pattern 1 (linear improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	EmpSE	2.35	2.39	2.40	2.40	2.41	2.33	2.43	2.33	2.33	2.36	2.30	2.29
	(primary)	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	GEE	EmpSE	2.07	2.05	2.09	1.94	1.99	1.92	1.90	1.88	1.92	1.86	1.83	1.82
		MCSE	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	Mixed	EmpSE	2.07	2.08	2.13	1.95	2.02	1.98	1.88	1.92	2.02	1.87	1.82	2.01
	effects	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)
200	ANCOVA	EmpSE	1.66	1.68	1.68	1.67	1.72	1.63	1.71	1.69	1.67	1.72	1.66	1.63
	(primary)	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	GEE	EmpSE	1.47	1.46	1.46	1.37	1.41	1.34	1.33	1.33	1.32	1.33	1.33	1.28
		MCSE	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	EmpSE	1.47	1.48	1.47	1.37	1.41	1.38	1.32	1.35	1.38	1.35	1.36	1.37
	effects	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
400	ANCOVA	EmpSE	1.18	1.12	1.15	1.20	1.18	1.20	1.18	1.18	1.22	1.16	1.20	1.15
	(primary)	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	EmpSE	1.05	1.00	1.02	0.97	0.95	0.97	0.95	0.94	0.95	0.91	0.92	0.91
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	EmpSE	1.05	1.01	1.01	0.95	0.98	0.98	0.97	0.96	1.00	0.89	0.91	0.97
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)

Table 6.20: Empirical standard error (EmpSE): Pattern 1 (linear improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	EmpSE	0.99	0.99	0.95	0.95	0.96	0.94	0.97	0.98	0.96	0.96	0.95	0.98
	(primary)	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	EmpSE	0.85	0.87	0.84	0.78	0.78	0.77	0.75	0.77	0.78	0.76	0.74	0.72
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Mixed	EmpSE	0.86	0.86	0.85	0.78	0.77	0.79	0.77	0.81	0.82	0.80	0.77	0.74
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
800	ANCOVA	EmpSE	0.84	0.84	0.81	0.81	0.83	0.84	0.85	0.83	0.81	0.85	0.85	0.80
	(primary)	MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	GEE	EmpSE	0.71	0.73	0.71	0.66	0.68	0.68	0.67	0.65	0.65	0.66	0.66	0.64
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Mixed	EmpSE	0.71	0.74	0.72	0.65	0.68	0.71	0.67	0.67	0.68	0.67	0.66	0.65
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

$\beta*$: Maximum treatment effect.

Table 6.21: Empirical standard error (EmpSE): Pattern 2 (short-term improvement then plateau)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	EmpSE	2.37	2.41	2.39	2.36	2.39	2.30	2.50	2.37	2.36	2.37	2.39	2.27
	(primary)	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	GEE	EmpSE	2.06	2.05	2.07	1.95	1.94	1.92	2.01	1.84	1.87	1.82	1.83	1.79
		MCSE	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	Mixed	EmpSE	2.06	2.05	2.10	1.97	1.98	1.94	2.03	1.86	1.86	1.79	1.81	1.84
	effects	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
200	ANCOVA	EmpSE	1.70	1.69	1.65	1.68	1.68	1.67	1.67	1.70	1.69	1.70	1.68	1.65
	(primary)	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	GEE	EmpSE	1.46	1.45	1.43	1.35	1.40	1.38	1.32	1.34	1.33	1.29	1.32	1.29
		MCSE	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	EmpSE	1.48	1.46	1.43	1.34	1.42	1.37	1.32	1.37	1.32	1.28	1.31	1.32
	effects	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
400	ANCOVA	EmpSE	1.17	1.18	1.19	1.24	1.18	1.21	1.20	1.23	1.13	1.21	1.19	1.15
	(primary)	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	EmpSE	1.03	1.03	1.01	1.01	0.95	0.99	0.91	0.94	0.92	0.94	0.92	0.88
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	EmpSE	1.01	1.03	1.02	1.03	0.94	1.01	0.91	0.95	0.95	0.96	0.94	0.89
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

Table 6.21: Empirical standard error (EmpSE): Pattern 2 (short-term improvement then plateau)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	EmpSE	0.95	0.98	0.93	0.98	0.96	0.96	0.99	0.95	0.94	0.96	0.96	0.94
	(primary)	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	EmpSE	0.82	0.85	0.81	0.80	0.80	0.78	0.77	0.76	0.74	0.76	0.74	0.72
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Mixed	EmpSE	0.84	0.82	0.81	0.80	0.81	0.79	0.79	0.76	0.75	0.76	0.72	0.75
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
800	ANCOVA	EmpSE	0.84	0.83	0.82	0.85	0.84	0.80	0.82	0.86	0.81	0.84	0.83	0.85
	(primary)	MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	GEE	EmpSE	0.72	0.73	0.72	0.69	0.70	0.67	0.64	0.68	0.63	0.66	0.64	0.64
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Mixed	EmpSE	0.73	0.72	0.73	0.70	0.71	0.65	0.63	0.67	0.62	0.66	0.65	0.66
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

$\beta*$: Maximum treatment effect.

Table 6.22: Empirical standard error (EmpSE): Pattern 3 (temporary improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	EmpSE	2.41	2.37	2.38	2.51	2.37	2.34	2.36	2.45	2.37	2.35	2.35	2.40
	(primary)	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	GEE	EmpSE	2.05	2.09	2.03	2.06	1.98	1.90	1.85	1.92	1.91	1.85	1.80	1.82
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	Mixed	EmpSE	2.08	2.13	2.16	2.08	2.00	2.04	1.88	1.99	2.17	1.91	1.81	2.08
	effects	MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)
200	ANCOVA	EmpSE	1.67	1.68	1.71	1.68	1.62	1.67	1.68	1.66	1.70	1.61	1.74	1.70
	(primary)	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	GEE	EmpSE	1.44	1.44	1.48	1.38	1.31	1.38	1.32	1.30	1.31	1.30	1.34	1.34
		MCSE	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	EmpSE	1.44	1.49	1.56	1.37	1.33	1.46	1.29	1.36	1.50	1.31	1.43	1.53
	effects	MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.04)	(0.04)
400	ANCOVA	EmpSE	1.20	1.14	1.18	1.16	1.17	1.17	1.17	1.18	1.14	1.13	1.19	1.19
	(primary)	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	EmpSE	1.03	1.00	1.05	0.96	0.96	0.97	0.92	0.95	0.89	0.89	0.93	0.92
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	Mixed	EmpSE	1.03	1.03	1.12	0.95	0.97	1.01	0.93	0.97	1.02	0.90	0.96	1.23
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.04)

Table 6.22: Empirical standard error (EmpSE): Pattern 3 (temporary improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	EmpSE	0.95	0.98	0.99	0.97	0.98	0.96	0.98	0.97	0.95	0.98	0.98	0.98
	(primary)	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	GEE	EmpSE	0.83	0.84	0.85	0.77	0.81	0.80	0.76	0.76	0.76	0.74	0.76	0.77
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Mixed	EmpSE	0.85	0.88	0.91	0.78	0.81	0.85	0.75	0.82	0.86	0.75	0.82	0.95
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.04)
800	ANCOVA	EmpSE	0.85	0.83	0.87	0.86	0.83	0.82	0.82	0.84	0.81	0.82	0.85	0.83
	(primary)	MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	GEE	EmpSE	0.73	0.72	0.75	0.70	0.69	0.67	0.66	0.68	0.66	0.64	0.66	0.65
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Mixed	EmpSE	0.76	0.74	0.83	0.70	0.74	0.72	0.67	0.70	0.72	0.67	0.70	0.89
	effects	MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)

$\beta*$: Maximum treatment effect.

6.3.4 Properties of the standard error: model-based standard error

The empirical standard error was very similar to the model-based standard error for all methods (Tables 6.23 to 6.25). This was true across the different patterns, numbers of time points and treatment effect sizes. As expected, the model-based standard error was smaller for the GEE and mixed effects methods than the ANCOVA (primary) method and was smaller for larger sample sizes.

Table 6.23: Average model-based standard error (ModSE): Pattern 1 (linear improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA (primary)	ModSE	2.40	2.39	2.38	2.40	2.39	2.38	2.39	2.38	2.37	2.40	2.39	2.37
	GEE	ModSE	2.05	2.04	2.04	1.94	1.92	1.92	1.87	1.86	1.85	1.83	1.82	1.82
	Mixed effects	ModSE	2.05	2.04	2.03	1.93	1.92	1.92	1.87	1.86	1.86	1.82	1.81	1.81
200	ANCOVA (primary)	ModSE	1.69	1.68	1.67	1.68	1.68	1.67	1.68	1.67	1.67	1.68	1.68	1.66
	GEE	ModSE	1.45	1.45	1.44	1.37	1.36	1.36	1.32	1.32	1.32	1.29	1.29	1.29
	Mixed effects	ModSE	1.45	1.45	1.44	1.37	1.36	1.36	1.32	1.32	1.32	1.29	1.29	1.28
400	ANCOVA (primary)	ModSE	1.19	1.19	1.18	1.19	1.18	1.18	1.19	1.18	1.18	1.19	1.18	1.18
	GEE	ModSE	1.03	1.03	1.02	0.97	0.97	0.96	0.94	0.93	0.93	0.92	0.91	0.91
	Mixed effects	ModSE	1.03	1.03	1.02	0.97	0.97	0.96	0.94	0.93	0.93	0.92	0.91	0.91
600	ANCOVA (primary)	ModSE	0.97	0.96	0.96	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.96	0.96
	GEE	ModSE	0.84	0.84	0.83	0.79	0.79	0.79	0.76	0.76	0.76	0.75	0.75	0.74
	Mixed effects	ModSE	0.84	0.83	0.83	0.79	0.79	0.79	0.77	0.76	0.76	0.75	0.75	0.75
800	ANCOVA (primary)	ModSE	0.84	0.83	0.83	0.84	0.84	0.83	0.84	0.83	0.83	0.84	0.83	0.83
	GEE	ModSE	0.73	0.72	0.72	0.69	0.68	0.68	0.66	0.66	0.66	0.65	0.65	0.64
	Mixed effects	ModSE	0.73	0.72	0.72	0.69	0.68	0.68	0.66	0.66	0.66	0.65	0.65	0.64

$\beta*$: Maximum treatment effect.

Table 6.24: Average model-based standard error (ModSE): Pattern 2 (short-term improvement then plateau)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA (primary)	ModSE	2.40	2.39	2.38	2.39	2.39	2.37	2.40	2.38	2.38	2.39	2.39	2.38
	GEE	ModSE	2.05	2.03	2.03	1.92	1.92	1.91	1.86	1.86	1.85	1.82	1.81	1.81
	Mixed effects	ModSE	2.04	2.03	2.02	1.92	1.92	1.90	1.87	1.85	1.85	1.82	1.81	1.81
200	ANCOVA (primary)	ModSE	1.69	1.68	1.68	1.68	1.68	1.67	1.68	1.68	1.67	1.68	1.68	1.67
	GEE	ModSE	1.45	1.44	1.44	1.37	1.36	1.36	1.32	1.32	1.31	1.29	1.29	1.28
	Mixed effects	ModSE	1.45	1.44	1.43	1.37	1.36	1.36	1.32	1.31	1.31	1.29	1.28	1.28
400	ANCOVA (primary)	ModSE	1.19	1.19	1.18	1.19	1.18	1.18	1.19	1.18	1.18	1.19	1.18	1.18
	GEE	ModSE	1.03	1.02	1.02	0.97	0.96	0.96	0.94	0.93	0.93	0.91	0.91	0.91
	Mixed effects	ModSE	1.03	1.02	1.01	0.97	0.96	0.96	0.94	0.93	0.93	0.92	0.91	0.91
600	ANCOVA (primary)	ModSE	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.96	0.96
	GEE	ModSE	0.84	0.84	0.83	0.79	0.79	0.78	0.76	0.76	0.76	0.75	0.74	0.74
	Mixed effects	ModSE	0.84	0.83	0.83	0.79	0.79	0.78	0.77	0.76	0.76	0.75	0.74	0.74
800	ANCOVA (primary)	ModSE	0.84	0.84	0.83	0.84	0.84	0.83	0.84	0.83	0.83	0.84	0.84	0.83
	GEE	ModSE	0.73	0.72	0.72	0.68	0.68	0.68	0.66	0.66	0.66	0.65	0.65	0.64
	Mixed effects	ModSE	0.73	0.72	0.72	0.68	0.68	0.68	0.66	0.66	0.66	0.65	0.65	0.64

$\beta*$: Maximum treatment effect.

Table 6.25: Average model-based standard error (ModSE): Pattern 3 (temporary improvement)

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA (primary)	ModSE	2.40	2.39	2.37	2.40	2.38	2.37	2.39	2.38	2.37	2.39	2.37	2.37
	GEE	ModSE	2.05	2.04	2.04	1.93	1.92	1.91	1.87	1.86	1.85	1.82	1.82	1.81
	Mixed effects	ModSE	2.04	2.03	2.03	1.92	1.91	1.91	1.87	1.86	1.85	1.82	1.82	1.81
200	ANCOVA (primary)	ModSE	1.69	1.68	1.66	1.69	1.68	1.67	1.68	1.67	1.67	1.68	1.68	1.66
	GEE	ModSE	1.46	1.45	1.44	1.37	1.36	1.35	1.32	1.32	1.31	1.30	1.29	1.28
	Mixed effects	ModSE	1.45	1.45	1.44	1.37	1.36	1.35	1.32	1.32	1.31	1.29	1.29	1.28
400	ANCOVA (primary)	ModSE	1.19	1.18	1.18	1.19	1.18	1.18	1.19	1.19	1.17	1.19	1.18	1.18
	GEE	ModSE	1.03	1.02	1.02	0.97	0.96	0.96	0.94	0.93	0.93	0.92	0.91	0.91
	Mixed effects	ModSE	1.03	1.02	1.02	0.97	0.96	0.96	0.94	0.94	0.93	0.92	0.91	0.91
600	ANCOVA (primary)	ModSE	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.96	0.96	0.97	0.96	0.96
	GEE	ModSE	0.84	0.84	0.83	0.79	0.79	0.78	0.77	0.76	0.76	0.75	0.75	0.74
	Mixed effects	ModSE	0.84	0.84	0.83	0.79	0.79	0.78	0.77	0.76	0.76	0.75	0.75	0.75
800	ANCOVA (primary)	ModSE	0.84	0.84	0.83	0.84	0.84	0.83	0.84	0.84	0.83	0.84	0.83	0.83
	GEE	ModSE	0.73	0.73	0.72	0.68	0.68	0.68	0.66	0.66	0.66	0.65	0.65	0.64
	Mixed effects	ModSE	0.73	0.72	0.72	0.69	0.68	0.68	0.66	0.66	0.66	0.65	0.65	0.64

$\beta*$: Maximum treatment effect.

6.4 Discussion

6.4.1 Summary of findings

This simulation study aimed to determine the optimal method for analysing a randomised trial of treatments for osteoarthritis under different scenarios, based on the sample size, number of follow-up assessment time points, and pattern and size of the treatment effect.

All of the methods produced high power when the sample included at least 400 participants and when the treatment effect was at least 8 points on the WOMAC. However, with a large sample size and large treatment effect, the analysis often did not converge when using the mixed effects method. The method with the optimal performance varied based on the sample size and treatment effect pattern over time.

When the treatment effect was consistent over time (pattern 2), all of the methods produced unbiased estimates. The GEE and mixed effects methods produced higher statistical power and more precise estimates. However, the level of coverage was better for the ANCOVA method when using the final non-zero time point as the primary follow-up assessment.

When the treatment effect was not consistent over time (patterns 1 and 3), the ANCOVA (primary) method produced the highest power. When the sample size was small, the ANCOVA (primary) method had the best coverage. However, the ANCOVA (primary) method was based on selecting the time point with the maximum treatment effect. In reality, the primary time point will have been pre-specified and may not correspond to the highest treatment effect over the follow-up period. In addition, the type I error would be very high if a primary time point was not specified because testing the treatment difference at each time point separately would increase

the risk of falsely identifying a beneficial treatment effect.

When the treatment effect increased over time (pattern 1), the GEE method produced the least biased estimates. When the sample size was large, the GEE method also had the best coverage. When the treatment effect was short-term and not sustained over time (pattern 3),the mixed effects method produced the least biased estimates and had the best coverage for larger sample sizes. However, the performance measures were based on the mixed effects method where the analysis did converge.

6.4.2 Comparison with existing literature

Previous research on the choice of analysis method for continuous outcomes in a randomised trial has largely focused on improved efficiency using longitudinal methods. Previous findings have indicated that mixed models or GEE methods are preferable to conducting ANCOVA at multiple time points, due to the increased type I error [382, 419, 420]. This simulation study aligns with this finding when considering the risk of falsely identifying a statistically significant treatment effect at any of the follow-up time points.

Most existing studies have focused on the optimal performance of different methods with regards to handling missing data. In the presence of missing data, Mallinckrod *et al.* and Prakash *et al.* found mixed models were preferable to single time point methods with simplistic imputation processes for missing outcome data (for example, carrying forward the last observation or baseline observation) [421, 422]. Further work would be needed to establish whether the performance of the methods in this simulation study would be different if outcome data were missing. Chu *et al.* also found that mixed effects was the preferred method for multi-centre trials as it attained high power and more accurate estimation of standard error than GEE [423]. As Chu *et al.* included an additional random effect for centre in both models, clustering could

have affected method performance.

Existing research has also highlighted the issue of lack of convergence in mixed effects models, in line with the findings in this simulation study [424]. Zhang *et al.* used an adaptive fitting procedure that improved convergent properties. However, the improvement in convergence was greatest for small sample sizes, and our findings indicated that convergence of the mixed effects model was poor for larger sample sizes. Evans *et al.* found convergence improved when using generalised linear mixed model using restricted maximum likelihood [425]. Further exploration of techniques to improve convergence of linear mixed models is needed.

Along with the results of this simulation study, Mayer-Hamblett and Kronmal also found that the pattern of treatment effect affected the choice of analysis method. Overall, their results indicated that it was beneficial to examine the full follow-up period and not only focus on the final time point, especially if the treatment only provided a short-term temporary effect [426]. This aligns with our findings that the optimal method depended on the pattern of treatment effect and that selecting the incorrect primary time point when using ANCOVA at each separate follow-up time point could result in low power.

The potential variability of the true treatment effects over time could indicate that time-averaged effects are not appropriate given that the treatment effect could be inconsistent over the follow-up period. This was supported by Hallgren *et al.*, who argued that the aggregation of outcomes over time may not be appropriate [427]. As an alternative, Hallgren *et al.* suggested that linear mixed models with repeated measures "would allow treatment effects to be pinpointed to a specific time when the effect may be largest, which may improve power." Others have argued that mixed models are preferred when individual trajectories are important because subject-specific inference is feasible, but that the GEE approach is favoured when the population-averaged

effect is important [426, 428, 429, 430]. As randomised trials are often designed to assimilate treatment effects in the clinical population, the GEE approach may be the most appropriate. However, future research is needed to examine the performance of the GEE method when treatment-by-time interactions are included in the model to allow the measurement of time-specific treatment effects.

6.4.3 Strengths and limitations

The main strength of a simulation study is the controlled way in which the data are generated and analysed. As the properties of the datasets are known from the data-generating mechanism, we can compare the estimates to a 'true' value. Each of the analysis methods is used to analyse the same group of datasets and thus the results can be compared between the different settings and methods. The parameters used in the distributions to generate the WOMAC scores were based on the properties of samples of people living with osteoarthritis, and therefore likely to reflect the population of people with osteoarthritis participating in research studies.

However, in this study, only a limited number of data-generating mechanisms were considered. The results of this study may not be generalisable to other patterns of treatment effect, for example, if the treatment effect was high in the short term and gradually reduced to zero. As the data-generating mechanism was based primarily on a single randomised trial, the results may not be generalisable to other populations, outcome measures or disease areas. For instance, a sample of people with osteoarthritis with less severe symptoms may have less variability in their outcomes, resulting in datasets with different distributional properties.

There are also limitations in the data-generating mechanisms used. This study assumed that the treatment effect was identical for all participants. In reality, it is likely that the treatment effect would depend on the participant's disease severity, age or

other characteristics. Future simulation studies could incorporate heterogeneity into the treatment effect to make the scenarios more realistic. This study also assumed that there would be no missing data. In practice, randomised controlled trials almost always have missing data, even for the primary outcome and especially for longer follow-up periods. Before recommending using one of these methods over another to analyse clinical trial data, future work is needed to examine the performance of the methods with different amounts of missing data.

To replicate the bounded nature of the WOMAC Index, re-sampling was used when the generated data point was outside the possible range of the WOMAC Index. This could have biased the results because it is more likely that observations in the intervention arm would be below zero (reaching the best possible score) due to the treatment effect. However, the impact of this is likely to have been small as the distributional properties of the datasets still corresponded with the data-generating mechanism.

A key limitation of this simulation study is that the different statistical methods produce different estimators and different true values for the estimand. It is not possible to compare the performance of these different methods in terms of the bias or precision of the estimator when the estimators use different scales. However, the results of this study can still be used to assess the performance of each method across the different scenarios, and the different methods can be compared on their power, type I error and coverage.

The performance of the mixed effects and GEE methods was examined based on the treatment effect averaged across all follow-up time points. However, the performance of these methods based on the average treatment effect may not reflect how well each method estimates the treatment effect at each individual time point. Within the mixed effects and GEE methods, there are many different variations that could

have been tested to explore their properties further. This study focused solely on the exchangeable correlation structure for the residuals and fitting the mixed effects model using maximum likelihood estimation.

6.4.4 Implications

The choice of method to analyse a randomised trial where the primary outcome is the WOMAC should depend primarily on the expected pattern of treatment effect. Although the number of time points had little effect on the performance of the different methods, repeated measurements should be used to deal with the potential patterns of treatment effects.

The GEE method is preferred if the treatment effect is consistent over time. When the treatment effect is not consistent over time and there is evidence to support when the maximum treatment effect is likely to occur, the ANCOVA (primary) method is preferred in terms of power and coverage unless the sample size is large. However, if the time-averaged treatment effect is of interest, GEE or mixed effects methods will produce the least biased and most precise treatment effect estimates. The GEE method is preferred if the treatment effect is expected to increase over time and the mixed effects method is preferred when the treatment effect is expected to be maintained in the short-term only and then disappear.

For randomised trials with large sample size ($n \geq 400$) aiming to detect large treatment effects (8 points or greater, compared to an MCID of approximately 4 points), the choice of method seems largely irrelevant in terms of power as all methods had high power. However, it could be argued that the mixed effects method should be avoided due to poor convergence. If the mixed model does not converge, the treatment effect estimate is not reliable and an alternative analysis method should be used. When the sample size is large, GEE or mixed effects methods would be optimal

as they provide good coverage and unbiased treatment effect estimates. However, randomised trials of treatments with sample sizes of 600 or more are rare (only 2 of 116 trials in Chapter 2 recruited more than 600 participants) and mixed models were less likely to converge with a large sample size, especially when several follow-up assessments were analysed. If the mixed model is chosen as the primary method of analysis for a randomised trial, the statistical analysis plan and protocol should provide alternative strategies in the event that the mixed model analysis does not converge. If the mixed effects analysis does not reach convergence, the GEE method is likely to be the best alternative approach as it can incorporate the same modelling assumptions for the correlation structure.

An advantage of the GEE and mixed effects methods is that they can easily incorporate a hierarchical structure. However, the correlation structure must also be specified in the model. This simulation study assumed an exchangeable correlation structure. However, misspecification of this correlation structure could result in much poorer model performance. If there is a considerable amount of missing data, mixed models may be preferred as participants with intermittent outcome data can still be included in the analysis, whereas participants with missing outcome data at the relevant follow-up time point would be excluded from an ANCOVA analysis.

The other main disadvantages of the ANCOVA method are the reduced precision of the treatment effect estimates and the increased type I error when analysing separate time points. The outcomes of the same participant sample at different time points will likely be correlated, and there is no consensus on how to adjust the p-value to account for this dependence. In practice, it is difficult to interpret the results when a significant treatment effect is detected at one follow-up time point, but not at the primary time point, especially when it may not be a 'true' treatment effect. More research is needed to compare the performance of the ANCOVA method at separate

time points to conditional time-specific treatment effects generated using GEE and mixed effects methods that include treatment-by-time interactions.

The data-generating mechanism for this simulation study showed when the treatment effect would be maximised. In practice, it could be difficult to predict at which follow-up time point the maximum treatment effect would be reached. If the treatment effect is not maximised at the primary time point, the power of the ANCOVA method may thus be lower than anticipated.

The results show that the best-performing method depends on the pattern of the treatment effect. This causes issues when planning a randomised trial as the treatment effect pattern is unknown before the start of the trial in most situations. The method of analysis should be pre-specified and the sample size calculation should align with the statistical analysis used in the trial. Sample size calculations for longitudinal methods are more complex and rarely used in practice (Chapter 2). The standard Neyman-Pearson method for the sample size calculation is not appropriate for any of the methods used in this simulation study; even when the ANCOVA method is used, the baseline adjustment should be accounted for [414]. However, the sample size calculation for ANCOVA does require additional assumptions on the correlation between the baseline and follow-up scores, which must be specified *a priori* [431]. Zhang and Ahn provide sample size calculations for differences in time-averaged treatment effects using GEEs. Similarly, methods for sample size calculations for linear mixed models have been published [432, 433].

More consideration should be given to the expected pattern of treatment effect over the follow-up period, informed by evidence from previous trials and observational studies. The choice of analysis strategy should be based on the anticipated pattern of treatment effect, and the sample size calculation should be appropriately aligned with the planned analysis strategy [414]. The assumptions made should be checked during

the analysis of the trial. Sensitivity analysis should also be conducted on the sample size calculation in case the pattern of treatment effect is different to that which was assumed when the trial was designed.

6.4.5 Future research

Future research should examine different scenarios incorporating missingness in data generation. The incorporation of missing data and the use of multiple imputation would be more reflective of the analysis methods used in practice in randomised trials. Future studies should examine the performance of the different methods when varying the proportion of missing data, the pattern of missing data across the different time points (e.g., monotone or intermittent missingness) and the method used to handle missing data. The effect of using multiple imputation to handle missing data in the simulated datasets should be examined. Simulation studies could also use different data-generating mechanisms to test the robustness of the analyses to model misspecifications, such as missing-not-at-random.

In this study, data were generated based on the distribution of the total WOMAC scale only. In reality, the WOMAC score is calculated from multiple item scores. For the WOMAC Index and other composite outcome measures, future research should explore whether missing values should be imputed for the WOMAC total scale, individual subscale scores or item scores [434, 435]. Generating data (and imputing missing data) at the item level could increase the representativeness of the simulated datasets. However, it would require additional assumptions on the items more greatly affected by osteoarthritis, which could make the results less generalisable, depending on the symptoms of the sample of participants with osteoarthritis or the mediating factors that the intervention was designed to target.

This study could be extended in future by examining different data-generating mechanisms. Additional research could examine the robustness of the different methods to model misspecification, including different correlation structures in the residuals and non-normal distributions for the outcome scores. Additional data-generating mechanisms could include a comparison of smaller treatment effect sizes that are closer to the estimated MCID in the outcome measure. Future studies could incorporate heterogeneity in the treatment effect, which would be more realistic as most interventions have different effects in different subgroups of participants. Data-generating mechanisms could be extended to include increased variation when the treatment effect is large. Data-generating mechanisms could also incorporate clustering effects, for example if interventions were delivered by GP practices or hospital settings. It would be informative to compare the different methods of analysis in future simulation studies of different outcome measures and health conditions. This would test the robustness of the results of this study and indicate whether the results are generally applicable for datasets with different distributional properties.

The mixed effects model often did not converge within 100 iterations using Stata software [121]. Further research could examine whether the convergence could be improved by using more iterations or different statistical software, such as R or SAS, or using different types of mixed effects methods, for example, using the restricted maximum likelihood approach or estimating the variance-covariance matrix using cluster-robust standard errors. Due to the lack of convergence, simulation studies could be used to examine the performance of 'staged' analysis plans, for example, using a mixed effects method if the analysis converges within 100 iterations and otherwise using a GEE method.

Future simulation studies could focus on the effect on efficiency when using mixed effects models with treatment-by-time interaction terms compared to ANCOVA at

each separate time point or time-averaged treatment effects [436]. This would allow the bias and coverage to be compared more robustly between methods as the true effects at each time point would be the same for each of the different methods. In future, researchers could also explore different methods to correct for the type I error associated with multiple testing across different time points. Alternative statistical analysis methods could also be compared, such as the area-under-the-curve (AUC) approach [437].

Future research could also explore methods to predict the pattern of treatment effect and the correlation structure of longitudinal data to inform the choice of analysis method and corresponding sample size calculation. Additional simulation studies could examine the variation in the target sample size between the different analysis methods and sample size calculation formulas, especially when the pattern of treatment effect differs to what was expected. Efforts should be made to improve the accessibility of sample size calculations for longitudinal data, including commands in statistical software packages and accessible journal articles.

6.4.6 Conclusions

This simulation study has shown that the preferred method of analysis for a randomised trial depends on the pattern of the treatment effect. GEE is recommended when the treatment effect is consistent over the follow-up period. When the treatment effect is not consistent over time, ANCOVA at the primary time point has the greatest power when the pattern of the treatment effect and the time point of the maximum treatment effect can be predicted accurately. However, before the trial is conducted, it is often unclear how the treatment effect will change over time. When there are several follow-up assessments, the results of the ANCOVA method should be interpreted with caution when no primary time point has been specified due to the

high type I error, even when a large sample size is used. If the focus is on estimating the time-averaged marginal treatment effect with minimum bias and maximum precision (rather than on coverage and statistical significance), the GEE method is preferred when the treatment effect increases over time and the mixed effects method is preferred when the treatment effect is a short-term temporary effect. These were the methods that produced less biased estimates and better coverage with the same sample size. However, future research is needed to examine whether longitudinal methods including treatment-by-time interactions or cluster-robust variance, such as GEE or mixed effects, can out-perform ANCOVA methods in detecting time-specific conditional treatment effects or when data are missing.

Chapter 7

Discussion

7.1 Summary of thesis findings

This thesis aimed to improve methods for specifying target differences in the sample size calculation of randomised trials examining treatments for osteoarthritis. I aimed to explore whether and how time should be incorporated into sample size calculations for randomised trials in long-term conditions. I examined whether target differences, in particular, Minimum Clinically Important Difference (MCID) estimates, should vary based on the time point of assessment, how they might be addressed in the planned statistical analysis and the implications for sample size calculations.

Chapter 2 reviewed 116 randomised trials of osteoarthritis published in 2016 and found that sample size calculations were only included in the published article for two-thirds of trials and that these calculations were often poorly reported. The sample size calculations were rarely reproducible, raising concerns about the accuracy of the reported information and validity of the methodology in the sample size calculations. The target differences used were justified based on the results of previously published

trials or MCID estimates calculated using distribution or anchor-based methods. Although most trials used more than one follow-up time point and often analysed the results using longitudinal methods, the sample size calculation did not take this into account. Almost all of the sample size calculations used the Neyman-Pearson approach and considered only one time point, without specifying the primary time point if outcomes were measured at multiple follow-up time points. This highlighted that there is scope for improvement in the reporting of sample size calculations in osteoarthritis trials.

Chapter 3 extended this review by examining in more depth the 62 studies reporting the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), a commonly used outcome measure in osteoarthritis trials. The focus on the WOMAC was to support the work in the following chapters, where I intended to calculate the MCID estimates for the WOMAC and examine statistical methods to analyse results for the WOMAC. The trials used several different versions of the WOMAC due to variations in how the items were collected and combined to create the summary score. Trials often did not report which version they used, which could cause difficulties in interpreting trial results and issues when comparing or combining trial results. The version of the WOMAC used was unclear in several trials that used it as their primary outcome measure. This makes it difficult for readers to assess the validity of the specification of the target difference. Indeed, many trials specified their target difference based on an MCID estimate that was calculated for a different version of the WOMAC measure. The most common version of the WOMAC used was the Likert scale version combined using summation to give a total range of 0 to 96; this version was used in the subsequent chapters.

Chapter 4 examined the importance of the duration of treatment effect from the participant perspective. A discrete choice experiment showed that duration of treatment effect was an important factor in the medication choices of people living with osteoarthritis. Participants were willing to accept less symptom relief on the three domains of the WOMAC (pain, stiffness and function) and higher risks of serious adverse events if a treatment had a long-lasting effect. For example, on average, participants were equally likely to choose a medication that would reduce their pain from 'severe' to 'moderate' for 12 months as to choose a medication that would reduce their pain from 'severe' to 'mild' for 6 months. These results support that the assessment time point when the treatment effect is measured should be an important part of interpreting the results of a randomised trial and comparing different treatments.

Chapter 5 used secondary analysis of a cohort dataset to assess the consistency of MCID estimates across different follow-up time points. Applying single time point methods at different assessment time points showed that the MCID estimates did vary across the follow-up time points. However, there was no clear visual trend in how the MCID estimates changed over time. The results confirmed previous findings that showed that baseline disease severity affected the MCID estimate, with higher MCID values in participants with more severe disease symptoms. However, there were high levels of imprecision for all of the MCID estimates, as shown by wide confidence intervals. Adjustment for baseline outcome score attenuated the MCID estimates and may introduce issues due to correlation with the anchor measure when interaction terms are not included. I demonstrated that longitudinal methods could be used to calculate MCID estimates across multiple follow-up time points and the different methods produced consistent estimates. However, longitudinal methods did not improve the precision of the MCID estimates.

Chapter 6 presented the results of a simulation study to compare the statistical properties of longitudinal methods used to analyse randomised trials, based on the distributional properties of the WOMAC from the cohort in Chapter 5 and previous randomised trials, including those reviewed in Chapter 3. The simulation study found that the optimal method of analysis depended on the pattern of the treatment effect over time, including the duration and stability of the treatment effect.

When the treatment effect was consistent and sustained over the follow-up period, the GEE (Generalised Estimating Equation) method produced the highest power and most precise estimates with low bias. When the treatment effect increased over time or a short-term benefit was not sustained over time and the sample size was small, the ANCOVA (Analysis of Covariance) method repeated at multiple time points had the highest power. However, if no primary time point is specified, the results of the ANCOVA method should be interpreted with caution due to the high type I error from multiple testing. For larger sample sizes ($n \geq 400$), the GEE method was preferred if the treatment effect increased over time and the mixed effects method was preferred if the treatment effect was temporary and lasted only 3-6 months. However, the mixed effects analysis often did not reach convergence, especially for a larger sample size or a greater number of follow-up assessments. Further research is needed to explore the optimal method for conditional time-specific effects using models with interaction terms, different data-generating mechanisms and different assumptions on missing data.

7.2 Implications of this research

Chapters 2 and 3 indicated that there is a clear need for improvement in the reporting of sample size calculations in osteoarthritis trials. This would provide a more accurate representation of the objectives when trials are designed and allow readers to interpret trial results with this information in mind. Sample size calculations using the Neyman-Pearson approach rely on assumptions about the variability of the outcome measure, subjective decisions on what level of type I and type II errors are permissible and the target difference that the trial is designed to detect. Researchers, clinicians and patients reading articles reporting the results of clinical trials should be able to see what trialists specified for each of these elements of the sample size calculation, allowing them to decide for themselves whether they believe these specifications and decisions were appropriate. Improved reporting and increased transparency in the design of randomised trials, including the sample size calculation, would aid the interpretation of clinical trial results.

Trialists should ensure that the target difference is specified based on evidence for the same version of the outcome measure used in the randomised trial, where this is available. For example, trials using the 0-100 VAS (Visual Analogue Scale) version of the WOMAC should try to find previous trials and MCID estimates using the 0-100 VAS version of the WOMAC to inform the specification of the target difference because this will provide the most applicable evidence. Trials that use longitudinal methods should also consider accounting for this in the sample size calculation, if there is evidence to inform the predicted correlation between the outcome measurements at different time points [414, 438].

The discrete choice experiment in Chapter 4 showed that the duration of treatment effect was important to people living with osteoarthritis when choosing between different medications. This implies that the assessment time point should be considered when interpreting between-group differences in randomised trials for osteoarthritis treatments. In randomised trials and observational studies, data on outcome measures should be collected at time points relevant to the participants, based on when duration of benefit is important to them and what fits into their treatment pathway. Duration of treatment benefit should be incorporated into the benefit-risk assessment when comparing different treatments, as well as the level of treatment benefit and risks, side effects and costs associated with the treatments.

In Chapter 5, secondary analysis of the cohort study suggested that MCID estimates were variable over different follow-up time points. The high variability in the MCID estimates suggests that a single study with one follow-up time point should not be the sole basis for an MCID estimate. Studies that aim to calculate MCID estimates should include multiple follow-up time points if possible, to show how variable the MCID estimates are. A single MCID estimate from one time point could result in the use of an estimate that is too small or too large, resulting in overly large sample sizes or a higher possibility of missing clinically important treatment effects. Sensitivity analysis should be conducted to assess the effect on the sample size of changing the assumptions and inputs of the calculation, for example, if different MCID estimates were used to specify target difference [439].

However, there was no strong visual pattern between the MCID estimate and follow-up time point. This implies that MCID estimates should not be adjusted based on the assessment time point used for studies of people with osteoarthritis. This also suggests that it is not necessary to select an MCID estimate that was calculated at a similar assessment time point to the primary time point when designing a randomised trial.

The association between the baseline outcome score and MCID estimate suggests that, if multiple MCID estimates are available, trialists should focus on the MCID estimates calculated in samples of participants with the most similar disease severity to the target population for the proposed randomised trial.

As the time point did not influence the MCID, clinical trialists could incorporate an anchor measure into their data collection. This would allow calculation of an MCID specific to the trial sample in the short-term and could verify whether the target difference specified in the sample size calculation was appropriate. For trials with long-term follow-up assessment, this could allow time for sample size re-estimation once the short-term results have been collected, if the target difference that was originally used in the sample size calculation was found to be inappropriate in the trial sample.

Chapter 5 also showed that it is feasible to calculate MCID estimates using longitudinal methods on data from multiple follow-up time points. However, this did not improve the precision of the MCID estimates in this example. Further research is needed to explore whether this is true generally in other participant samples and different outcome measures. In this case, the findings suggest that the increased complexity of longitudinal methods do not provide additional information beyond using single time point methods separately at each of the different follow-up time points.

The simulation study in Chapter 6 found that the optimal method to analyse a randomised trial using the WOMAC as the primary outcome depends on the pattern of the treatment effect. In most circumstances when the time-averaged treatment effect is of interest, the GEE method is preferred due to the type I error associated with using ANCOVA repeatedly at multiple time points and the lack of convergence of the mixed effects method. However, ANCOVA methods may be preferred in situations with inconsistent treatment effects, smaller sample sizes, fewer assessment time points

(to reduce type I error), and when there is sufficient evidence to suggest at which time point the maximum treatment effect will occur. The mixed effects method may be preferred if the treatment effect is expected to only last for 3-6 months before disappearing and the sample size is large. However, due to problems with convergence, if the mixed effects method is chosen as the primary analysis method, the protocol and statistical analysis plan for the trial should specify secondary analysis methods in case the mixed effects analysis does not converge.

More consideration should be given to the expected pattern of the treatment effect over the follow-up period during the design of a randomised trial. This should inform the choice of analysis method and the corresponding sample size calculation. Sensitivity analysis on the sample size calculation should be carried out to ensure that the target sample size will provide sufficient sufficient statistical power if the pattern of treatment effect or analysis method used differs to what was anticipated.

Trialists should consider whether the pattern of treatment effect would affect the clinical importance of the trial results. Although the number of assessment time points did not considerably improve the performance of different statistical analysis methods, trialists should aim to measure outcomes with sufficient frequency and duration to show how the treatment effect changes over time and, where it is important, whether the treatment effects are maintained in the long-term.

Although this thesis focused on randomised trials of osteoarthritis, the findings could also have implications for trials in other long-term conditions that use continuous measures as the primary outcome. As well as the implications for sample size calculations discussed above, the results have implications for other situations where MCID estimates are used. This includes any study assesses the clinical importance of the results, including examining the between-group difference in a study outcome for randomised trials, meta-analyses and observational studies. It also has relevance for

post-trial follow-up, where duration might be more important [362]. As registries and routinely collected data become more widespread, data on extended follow-up will be more common, especially as new technologies such as mobile phone applications make observational and extended follow-up less arduous [440, 441]. This thesis has shown that extended follow-up measurement could be used to calculate MCID estimates to inform future trial design.

It is important that appropriate MCID values are used because MCID estimation has implications for whether different treatments are seen as clinically worthwhile. Olsen et al. stated that the MCID estimate "is central for clinical guideline development, interpretation of results of randomised clinical trials or meta-analyses, and for choosing an appropriate sample size for a clinical study, but the measure is potentially misguiding if estimated, applied or interpreted inappropriately."

7.3 Limitations of this research

This thesis focused on two-arm superiority randomised trials that used a continuous patient-reported primary outcome measure. The findings therefore have limited relevance for binary or time-to-event outcomes, laboratory outcomes or Bayesian designs [442]. For example, the findings are not applicable to oncology trials, where the primary outcome is usually mortality or disease progression [443, 444]. Sample size calculations were assumed to use a Neyman-Pearson approach focusing on a single primary time point. The results may therefore be less relevant for sample size calculations that account for additional design features, such as adaptive designs, stepped-wedge or cluster randomised trials, non-inferiority or equivalence hypotheses, or sample size calculations accounting for repeated measures [414, 438, 445, 446, 447]. The reviews in Chapters 2 and 3 only considered the information reported on the sample size calculation. We cannot know whether the reported information is an accurate representation of the reasoning behind the sample size calculation, or even whether the calculation was performed before the start of the trial.

There is no consensus on the optimal method to specify a target difference or MCID [9]. This thesis focused on the use of anchor-based methods, which are some of the most widely-known methods and are commonly used in osteoarthritis trials. Reviews in other conditions have found little use of anchor-based MCID estimates in sample size calculations for randomised trials [82]. This limits the applicability of the findings of this thesis. The findings do not provide insight into the use of alternative methods for target difference specification, for example, approaches based on cost-effectiveness, value of information or opinion-seeking methods [448].

This thesis examined the generalisability of MCID estimates, with a particular focus on the follow-up time point. The applicability of MCID estimates to a particular sample will be of little relevance if there is limited MCID literature for a specific condition or outcome measure. If there are only one or two published MCID estimates for a condition, the intervention and population characteristics are likely to be more important considerations, compared to the assessment time point. Further knowledge of the applicability of MCID estimates will become more useful as more published MCID estimates become available.

Even if consensus has been achieved on the MCID in an outcome measure, there may be other key factors to consider. When interpreting the results of a randomised trial, researchers, clinicians and other stakeholders will consider other features beyond whether the treatment effect exceeds the MCID. It is also important to consider safety risks, side effects and financial cost, which could affect whether a treatment is prescribed at an individual-level or implemented at the system-level [449, 450, 451]. As well as negative aspects, when choosing between treatments, clinicians, patients and commissioners will consider the full profile of the treatment options, including effects on secondary outcomes, subgroup effects, levels of compliance and the potential effect on co-morbid conditions.

A key limitation of this thesis is the focus on hip and knee osteoarthritis, with particular attention on the use of the WOMAC Index. Focusing on a single condition ensured the relevance of treatment effect domains and reduced the level of heterogeneity in the samples of participants and articles considered. However, the results of this thesis may be related to specific aspects of osteoarthritis, such as the flare-up nature of the condition. It is unclear whether the findings of this thesis are generalisable to other long-term conditions, such as respiratory disorders. In Chapters 4 and 5, each experiment was based on a single sample of participants. Therefore, further research

in other samples of people with osteoarthritis is needed to ensure that the results are generalisable to populations with differing levels of severity. Future research is also required to explore whether the findings are consistent for other outcome measures and health conditions.

7.4 Future research

This thesis focused on target differences and MCID estimates in osteoarthritis, particularly for the WOMAC outcome measure. It would be useful to explore whether the findings are consistent across other conditions. The evidence on the association between MCID estimates and assessment time point would be strengthened by additional studies examining the consistency of MCID estimates over time in other outcome measures, interventions and disease areas.

Future research could consider interventions to improve the reporting of sample size calculations in randomised trials. Additional guidance could be incorporated into the CONSORT statement with recommendations for the specific components that should be reported in the description of the sample size calculation, including the target difference, primary outcome measure and primary time point [157]. The recently published DELTA^2 guidance, which includes recommendations on reporting of sample size calculations in manuscripts, may improve reporting and reproducibility [439, 448].

In future, researchers should also endeavour to improve the consistency in the terminology used for MCID estimates. This should focus on differentiating between important changes within a single participant over time and important differences between treatment arms or different participant groups. Further work establishing greater consensus on the 'optimal' methods for MCID calculation would be valuable.

Further insight into the appropriateness of anchor-based MCID estimates could be facilitated by including anchor measures in the data collected in randomised trials. This would allow us to calculate an MCID estimate using measurements from the trial sample and compare this to the target difference specified in the sample size calculation. This would provide information on the generalisability of MCID estimates and whether researchers should apply MCID estimates generally across different samples

and populations.

The secondary analysis in Chapter 5 used a non-traditional anchor measure, based on a change score in a patient-reported global assessment. While this does not allow the participant to self-assess the change in their condition, it avoids issues with traditional anchors, such as recall bias. Future work should compare the MCID estimates produced using non-traditional anchors, such as that used in Chapter 5, with those produced using traditional anchor measures.Future research could also examine whether domain-specific anchor measures could be more appropriate for certain participants or disease populations.

Future studies on variability across MCID estimates will be facilitated with the introduction of automated data collection in some disease areas [452]. In areas with a lack of studies on MCID estimation, the collection of anchor measures could be incorporated into disease registries and large-scale observation studies. For osteoarthritis specifically, this could include cohort studies (such as the Swedish population-based National Quality Register for Better Management of Patients with Osteoarthritis) and arthroplasty registries (including the National Joint Registry in the United Kingdom) [453, 454, 455]. The incorporation of anchor measures into routine data collection could increase the number of participants used to calculate MCID estimates and thus lead to increased precision of MCID estimates.

As more studies on the calculation of MCID estimates are published, additional reviews of MCID estimates would be useful to synthesise evidence for particular outcome measures, such as the review by Bohannon *et al.* [456]. Reviews could also separate MCID estimates by different characteristics, such as different populations, time points and interventions. Meta-analyses and meta-regression would facilitate the comparison of MCID estimates and allow exploration into factors that may cause differences in MCID estimates. For example, in osteoarthritis populations, the OA trial bank

collects data from multiple randomised trials in osteoarthritis [409, 410]. These data could be combined to explore differences in MCID estimates across different trials. The use of individual participant data from multiple trials would permit the use of consistent methodology in the MCID calculation, reducing some of the confounding present when comparing MCID estimates that have been calculated in different studies. The use of pooled datasets would also increase the precision of MCID estimates for particular subgroups. This type of research could provide information on which factors are more important when using an MCID estimate to specify the target difference in a future randomised trial, for example, whether it is more important for the disease severity of the participants or the interventions used in the sample used to calculate the MCID estimate to be similar to those in the proposed randomised trials.

Although no association between the assessment time point and the MCID estimate was found, the discrete choice experiment in Chapter 4 found that duration of treatment effect was important to participants. Future research should examine how duration of treatment effect can be incorporated into benefit-risk assessments when comparing different treatments. This could also involve further research to compare longitudinal methods for the analysis of randomised trials, building on the simulation study in Chapter 6.

Although several trials in the review in Chapter 2 used longitudinal methods to analyse the results, the longitudinal nature of the data and the analysis was not considered in the sample size calculation. Therefore, this thesis focused on the commonly-used Neyman-Pearson approach in sample size calculation. However, further efforts should be made to increase education and awareness and promote methods for sample size calculation that incorporate the multilevel structure of longitudinal data collected in randomised trials. This could build on the research conducted by Hooper *et al.*,

Harden *et al.*, Kasza *et al.* and Hemming *et al.* on sample size calculation for longitudinal cluster-randomised trials [438, 446, 447, 457]. It is hoped that these methods will become more commonly used in practice in future trials. Web-based applications could facilitate the use of more complex methodology for sample size calculations in longitudinal cluster trials [458]. This should better align the methods used to calculate the sample size and analyse the results of randomised trials.

Further research should aim to improve methods for sample size calculation in longitudinal trials. For example, future research could explore how to estimate the correlation structure of longitudinal data in randomised trials during trial design. Improved prediction of the correlation structure would increase the accuracy of assumptions in sample size calculations that account for the longitudinal nature of the outcome measurements. More complex trial designs and methodologies are increasingly used in medical research, such as adaptive designs and Bayesian methods [459, 460]. Future research should also examine how to account for duration of treatment effect and the longitudinal nature of the data collected in the analysis and sample size calculation for these more complex trials.

7.5 Conclusion

This thesis has shown that sample size calculations in randomised trials of osteoarthritis are poorly reported and often not reproducible. Additional efforts should be made by researchers, peer reviewers, journal editors and grant funding bodies to ensure that the sample size for a randomised trial is justified *a priori* in an accurate manner. Moreover, trialists should clearly report the version of the outcome measure that has been used in their trial, especially when the primary outcome measure is a composite outcome made up of a combination of multiple items, such as the WOMAC.

The results of the preference study showed that duration of treatment effect is important to people with osteoarthritis when comparing different pharmacological treatments. Therefore, duration of treatment effect should be accounted for in the cost-benefit assessments of interventions. Trials should measure outcomes at assessment time points that are relevant to the target population and when outcomes could have a greater effect on participants' treatment decision-making.

In a single cohort study, there was no clear trend in the MCID estimate for the WOMAC across different follow-up time points. This indicates that the target difference in the sample size calculation does not need to be adjusted due to the assessment time points. Further research is needed using different samples, outcome measures, anchor measures and disease areas to indicate whether these findings are generalisable.

Although there was no apparent relationship between the MCID estimate and follow-up time point, the MCID estimates were imprecise and varied across different follow-up time points. Therefore, future studies calculating MCID estimates using anchor-based measures should, where possible, use large sample sizes and multiple assessment time points and report the level of imprecision of the MCID estimate. Single time point methods for MCID calculation were adequate and precision was not improved

using longitudinal methods. However, future work should be conducted to develop consensus on the optimal methods for MCID calculation and terminology used for MCID estimates.

Randomised trials justifying the target difference using an anchor-based MCID estimate should consider all available MCID estimates in similar populations, rather than focusing solely on a single MCID estimate. Incorporating an anchor measure into the data collection of a randomised trial could allow for within-sample validation of the target difference used or sample size re-estimation for trials with a sufficiently long follow-up period.

A simulation study showed that the optimal analysis method for a randomised trial depends on the pattern of the treatment effect over the follow-up period, assuming no missing data. For example, when the treatment effect is consistent over time, the GEE method is preferred when estimating the time-averaged treatment effect, whereas the mixed effects method is preferred for a trial with a large sample size estimating a short-term treatment effect that is not maintained. Using ANCOVA at multiple time points separately produced higher power for small sample sizes but also high type I error due to the multiple testing involved. Mixed effects methods often did not converge and thus an alternative method should be pre-specified in the event that mixed effects methods do not converge. Because it can be difficult to predict the pattern of the treatment effect at the start of the trial, sensitivity analyses should be conducted for the sample size calculation under different scenarios. Software packages and more accessible journal articles should be produced to facilitate sample size calculations for longitudinal methods. Future simulation studies should examine the performance of the GEE and mixed effects methods when time-by-treatment interactions are included in the model and when there are no missing data, and should explore methods to improve the convergence of mixed effects analyses.

Overall, this thesis has shown that sample size calculations of randomised trials of treatments for osteoarthritis are poorly reported and do not account for the longitudinal nature of the outcome measurement. Duration of the treatment effect is important to people living with osteoarthritis when considering which medications to take. One case study found that the MCID of the WOMAC measure was not associated with the timing of the follow-up assessment. However, a simulation study showed that the pattern of the treatment effect over time in a randomised trial affected the preferred method of analysis, thus affecting the corresponding sample size calculation.

Appendix A

Appendices for Chapter 2

A.1 Search strategy example (MEDLINE Ovid)

1 randomized controlled trial.pt.

2 controlled clinical trial.pt.

3 randomized.ab.

4 placebo.ab.

5 clinical trial/

6 randomly.ab.

7 trial.ti.

8 1 or 2 or 3 or 4 or 5 or 6 or 7

9 humans/

10 8 and 9

11 exp Osteoarthritis/

12 osteoarthr$.tw.

13 (degenerative adj3 (arthr$ or joint$ or disease$)).tw.

14 arthros?s.tw.

15 11 or 12 or 13 or 14

16 10 and 15

17 limit 16 to yr="2016"

A.2 Included studies in systematic review

First author:	Year:	Title:	Use of WOMAC
Abdalbary, S. A.	2016	Ultrasound with mineral water or aqua gel to reduce pain and improve the WOMAC of knee osteoarthritis. Future Science OA 2 (1) (no pagination)(FSO110).	Yes
Ali, A.	2016	Similar patient-reported outcomes and performance after total knee arthroplasty with or without patellar resurfacing: A randomized study of 74 patients with 6 years of follow-up. Acta Orthopaedica 87(3): 274-279.	No
Allen, K. (Allen 2016 a)	2016	A combined patient and provider intervention for management of osteoarthritis in veterans. Annals of Internal Medicine 164(2): 73-83.	Yes
Allen, K. D. (Allen 2016 b)	2016	Group Versus Individual Physical Therapy for Veterans With Knee Osteoarthritis: Randomized Clinical Trial. Physical therapy 96(5): 597-608.	Yes
Arden, N.	2016	The effect of vitamin D supplementation on knee osteoarthritis, the VIDEO study: a randomised controlled trial. Osteoarthritis and Cartilage 24(11): 1858-1866.	Yes
Arendt-Nielsen, L.	2016	Intra-articular onabotulinumtoxinA in osteoarthritis knee pain: effect on human mechanistic pain biomarkers and clinical pain. Scandinavian Journal of Rheumatology: 1-14.	Yes
Aunan, E.	2016	Patellar resurfacing in total knee arthroplasty: Functional outcome differs with different outcome scores. Acta Orthopaedica 87(2): 158-164.	No
Aydogan, N.	2016	The clinical effect of platelet-rich plasma prepared through different activation methods on patients with knee osteoarthritis. Journal of Clinical and Analytical Medicine 7(6): 767-771.	Yes
Bagnato, G.	2016	Pulsed electromagnetic fields in knee osteoarthritis: a double blind, placebo-controlled, randomized clinical trial. Rheumatology (Oxford, England) 55, 755-762.	Yes
Baktir, A.	2016	Mobile- versus fixed-bearing total knee arthroplasty: a prospective randomized controlled trial featuring 6-10-year follow-up. Acta Orthop Traumatol Turc. 2016;50(1):1-9.	No
Banerjee, M.	2016	Comparative study of efficacy and safety of tapentadol versus etoricoxib in mild to moderate grades of chronic osteorthritis of knee. Indian Journal of Rheumatology 11, 21-25.	Yes
Bellamy, J.	2016	Economic Impact of Ketorolac vs Corticosteroid Intra-Articular Knee Injections for Osteoarthritis: A Randomized, Double-Blind, Prospective Study. Journal of Arthroplasty 31(9 Supplement): 293-297.	Yes
Beselga, C.	2016	Immediate effects of hip mobilization with movement in patients with hip osteoarthritis: A randomised controlled trial. Manual Therapy 22: 80-85.	No
Bily, W.	2016	Effects of leg-press training with moderate vibration on muscle strength, pain, and function after total knee arthroplasty: a randomized controlled trial [with consumer summary]. Archives of Physical Medicine and Rehabilitation 2016 Jun;97(6):857-865.	Yes
Bisicchia, S.	2016	HYADD 4 versus methylprednisolone acetate in symptomatic knee osteoarthritis: A single-centre single blind prospective randomised controlled clinical study with a 1-year follow-up. Clinical and Experimental Rheumatology 34(5): 857-863.	Yes
Bokaeian, H.	2016	The effect of adding whole body vibration training to strengthening training in the treatment of knee osteoarthritis: A randomized clinical trial. Journal of Bodywork and Movement Therapies 20, 334-340	Yes

Bryk, F.	2016	Exercises with partial vascular occlusion in patients with knee osteoarthritis: a randomized clinical trial. Knee Surgery, Sports Traumatology, Arthroscopy 2016 May;24(5);1580-1586.	No
Cakir, T.	2016	Isokinetic exercise improves concentric knee flexion torque better than isometric exercise in patients with advanced osteoarthritis. Isokinetics and Exercise Science 24(1): 7-15.	Yes
Calatayud, J.	2016	High-intensity preoperative training improves physical and functional recovery in the early post-operative periods after total knee arthroplasty: a randomized controlled trial. Knee Surgery, Sports Traumatology, Arthroscopy 2016 Jan 14:Epub ahead of print.	Yes
Chen, J. Y.	2016	Intravenous versus intra-articular tranexamic acid in total knee arthroplasty: A double-blinded randomised controlled noninferiority trial. Knee 23(1): 152-156.	No
Cherian, J. (Cherian 2016 a)	2016	Knee Osteoarthritis: Does Transcutaneous Electrical Nerve Stimulation Work? Orthopedics 39, e180-186	No
Cherian, J. J. (Cherian 2016 b)	2016	Do the Effects of Transcutaneous Electrical Nerve Stimulation on Knee Osteoarthritis Pain and Function Last? Journal of Knee Surgery 29(6): 497-501.	No
Cho, J. J.	2016	An MRI Evaluation of Patients Who Underwent Treatment with a Cell-Mediated Gene Therapy for Degenerative Knee Arthritis: A Phase IIa Clinical Trial. J Knee Surg. 2017 Sep;30(7):694-703.	No
Chughtai, M.	2016	Clinical Outcomes of a Pneumatic Unloader Brace for Kellgren-Lawrence Grades 3 to 4 Osteoarthritis: A Minimum 1-Year Follow-Up Study. Journal of Knee Surgery 29(8): 634-638.	No
Conrozier, T.	2016	Safety and efficacy of intra-articular injections of a combination of hyaluronic acid and mannitol (HAnOX-M) in patients with symptomatic knee osteoarthritis: Results of a double-blind, controlled, multicenter, randomized trial. Knee 23(5): 842-848.	Yes
Cosman, F.	2016	Effect of teriparatide on bone formation in the human femoral neck. J Clin Endocrinol Metab. 2016 Apr;101(4):1498-505.	No
Da Graca-Tarrago, M.	2016	Electrical intramuscular stimulation in osteoarthritis enhances the inhibitory systems in pain processing at cortical and cortical spinal system. Pain Medicine (United States) 17(5): 877-891.	No
de Rooij, M.	2016	Efficacy of tailored exercise therapy on physical functioning in patients with knee osteoarthritis and comorbidity: a randomized controlled trial [with consumer summary]. Arthritis Care & Research 2016 Aug 26:Epub ahead of print.	Yes
Di Sante, L.	2016	Intra-articular hyaluronic acid vs platelet-rich plasma in the treatment of hip osteoarthritis. Med Ultrason. 2016 Dec 5;18(4):463-468.	Yes
Dincer, U.	2016	The effects of closed kinetic chain exercise on articular cartilage morphology: Myth or reality? A randomized controlled clinical trial. Turkiye Fiziksel Tip ve Rehabilitasyon Dergisi 62(1): 28-36.	Yes
Diracoglu, D.	2016	Single versus multiple dose hyaluronic acid: Comparison of the results. Journal of Back and Musculoskeletal Rehabilitation 29(4): 881-886.	Yes
Dundar, U.	2016	Assessment of pulsed electromagnetic field therapy with Serum YKL-40 and ultrasonography in patients with knee osteoarthritis. International Journal of Rheumatic Diseases 19(3): 287-293.	Yes
Eker, H. E.	2016	The efficacy of intra-articular lidocaine administration in chronic knee pain due to osteoarthritis: A randomized, double-blind, controlled study	Yes
Elbadawy, M. A.	2016	Effectiveness of Periosteal Stimulation Therapy and Home Exercise Program in the Rehabilitation of Patients with Advanced Knee Osteoarthritis. Clinical Journal of Pain. 10.	No
Feczko, P.	2016	Computer-assisted total knee arthroplasty using mini midvastus or medial parapatellar approach technique: A prospective, randomized, international multicentre trial. BMC Musculoskeletal Disorders 17: 19.	Yes

Freitag, T.	2016	Bone remodelling after femoral short stem implantation in total hip arthroplasty: 1-year results from a randomized DEXA study. Archives of Orthopaedic & Trauma Surgery 136(1): 125-130.	Yes
Gigis, I.	2016	Comparison of two different molecular weight intra-articular injections of hyaluronic acid for the treatment of knee osteoarthritis. Hippokratia 20(1): 26-31.	Yes
Goksen, N.	2016	Magnetic resonance therapy for knee osteoarthritis: a randomized, double blind placebo controlled trial. Eur J Phys Rehabil Med. 2016 Aug;52(4):431-9.	Yes
Gook-Joo, K.	2016	The effects of high intensity laser therapy on pain and function in patients with knee osteoarthritis. J Phys Ther Sci. 2016 Nov;28(11):3197-3199.	Yes
Gor, A.	2016	A comparative study of efficacy and safety of oral diclofenac and decreased dose of diclofenac plus topical diclofenac in treatment of knee osteoarthritis. International Journal of Pharmaceutical Sciences and Research 7(5): 2083-2089.	No
Guner, S.	2016	Effectiveness of etofenamate for treatment of knee osteoarthritis: A randomized controlled trial. Therapeutics and Clinical Risk Management 12: 1693-1699.	Yes
Gungen, G. O.	2016	Effect of mud compress therapy on cartilage destruction detected by CTX-II in patients with knee osteoarthritis. Journal of Back and Musculoskeletal Rehabilitation 29(3): 429-438.	Yes
Ha, C.	2016	Prospective, randomized, double-blinded, double-dummy and multicenter phase IV clinical study comparing the efficacy and safety of PG201 (Layla) and SKI306X in patients with osteoarthritis. Journal of Ethnopharmacology 181: 1-7.	Yes
Helianthi, D. R.	2016	Pain reduction after laser acupuncture treatment in geriatric patients with knee osteoarthritis: a randomized controlled trial. Acta Medica Indonesiana 2016 Apr;48(2):114-121.	No
Hermann, A.	2016	Preoperative progressive explosive-type resistance training is feasible and effective in patients with hip osteoarthritis scheduled for total hip arthroplasty - a randomized controlled trial. Osteoarthritis and Cartilage 24(1): 91-98.	No
Hill, C.	2016	Fish oil in knee osteoarthritis: A randomised clinical trial of low dose versus high dose. Annals of the Rheumatic Diseases 75(1): 23-29.	Yes
Hinman, R. S.	2016	Unloading shoes for self-management of knee osteoarthritis a randomized trial. Annals of Internal Medicine 165(6): 381-389.	Yes
Hochberg, M.	2016	Combined chondroitin sulfate and glucosamine for painful knee osteoarthritis: A multicentre, randomised, double-blind, non-inferiority trial versus celecoxib. Annals of the Rheumatic Diseases 75(1): 37-44.	Yes
Holsgaard-Larsen, A.	2016	The effect of instruction in analgesic use compared with neuromuscular exercise on knee-joint load in patients with knee osteoarthritis: a randomized, single-blind, controlled trial. Osteoarthritis and Cartilage 2016 Nov 9:Epub ahead of print.	No
Hommel, H.	2016	Kinematic femoral alignment with gap balancing and patient-specific instrumentation in total knee arthroplasty: a randomized clinical trial. European Journal of Orthopaedic Surgery and Traumatology: 1-6.	Yes
Hsieh, R. (Hsieh 2016 a)	2016	Clinical effects of lateral wedge arch support insoles in knee osteoarthritis: A prospective double-blind randomized study. Medicine (United States) 95 (27) (no pagination)(e3952).	No
Hsieh, L. F. (Hsieh 2016 b)	2016	Effects of Botulinum Toxin Landmark-Guided Intra-articular Injection in Subjects With Knee Osteoarthritis. PM R. 2016 Dec;8(12):1127-1135.	Yes
Isik, M.	2016	Comparison of the effectiveness of medicinal leech and TENS therapy in the treatment of primary osteoarthritis of the knee: A randomized controlled trial. Zeitschrift fur Rheumatologie: 1-8.	Yes

Jame Bozorgi, A.	2016	The effectiveness of occupational therapy supervised usage of adaptive devices on functional outcomes and independence after total hip replacement in Iranian elderly: A randomized controlled trial. Occupational Therapy International 23(2): 143-153.	Yes
Jin, X.	2016	Effect of Vitamin D supplementation on Tibial cartilage volume and knee pain among patients with symptomatic knee osteoarthritis: A randomized clinical trial. JAMA - Journal of the American Medical Association 315(10): 1005-1013.	Yes
Jorgensen, P. B.	2016	The efficacy of early initiated, supervised, progressive resistance training compared to unsupervised, home-based exercise after unicompartmental knee arthroplasty: a single-blinded randomized controlled trial [with consumer summary]. Clinical Rehabilitation 2016 Mar 30:Epub ahead of print.	No
Kadic, L.	2016	The effect of pregabalin and s-ketamine in total knee arthroplasty patients: A randomized trial. J Anaesthesiol Clin Pharmacol. 2016 Oct-Dec;32(4):476-482.	No
Kapadia, B.	2016	Gait Using Pneumatic Brace for End-Stage Knee Osteoarthritis. The journal of knee surgery 29, 218-223.	No
Kaya Mutlu, E.	2016	Does Kinesio taping of the knee improve pain and functionality in patients with knee osteoarthritis? A randomized controlled clinical trial. American Journal of Physical Medicine & Rehabilitation 2016 May 4:Epub ahead of print.	Yes
Koh, I. J.	2016	The Patient's Perception Does Not Differ Following Subvastus and Medial Parapatellar Approaches in Total Knee Arthroplasty: A Simultaneous Bilateral Randomized Study. Journal of Arthroplasty 31(1): 112-117.	Yes
Kongtharvonskul, J.	2016	Efficacy of glucosamine plus diacerein versus monotherapy of glucosamine: A double-blind, parallel randomized clinical trial. Arthritis Research and Therapy 18 (1) (no pagination)(233).	Yes
Kulshrestha, V.	2016	Outcome of Unicondylar Knee Arthroplasty vs Total Knee Arthroplasty for Early Medial Compartment Arthritis: A Randomized Study. J Arthroplasty. 2017 May;32(5):1460-1469.	No
Li, Z.	2016	Effects of cold irrigation on early results after total knee arthroplasty a randomized, double-blind, controlled study. Medicine (United States) 95 (24) (no pagination)(e3563).	No
Lim, S. H.	2016	Effects of joint effusion on quadriceps muscles in patients with knee osteoarthritis. Physical Therapy in Sport 17: 14-18.	No
Losina, E.	2016	Postoperative Care Navigation for Total Knee Arthroplasty Patients: A Randomized Controlled Trial. Arthritis Care and Research 68(9): 1252-1259.	Yes
Maghsoumi-Norouzabad, L.	2016	Effects of Arctium lappa L. (Burdock) root tea on inflammatory status and oxidative stress in patients with knee osteoarthritis. International Journal of Rheumatic Diseases 19(3): 255-261.	No
Majeed, M.	2016	Evaluating adjunctive role of PPI as anti inflammatory agent in knee osteoartharitis: A prospective study in tertiary care hospital. International Journal of Pharmaceutical Sciences and Research 7, 2492-2498.	No
Martin Martin, L. S.	2016	A double blind randomized active-controlled clinical trial on the intra-articular use of Md-Knee versus sodium hyaluronate in patients with knee osteoarthritis (Joint). BMC Musculoskeletal Disorders 17 (1) (no pagination)(948).	No
Meinardi, J.	2016	Palacos compared to Palamed bone cement in total hip replacement: a randomized controlled trial: RSA migration similar at 10-year follow-up. Acta Orthopaedica 87, 473-478	No
Mihaljevic, Z.	2016	Influence of fondaparinux versus nadroparin calcium thromboprophylaxis on clinical parameters following total knee arthroplasty. Acta Clin Croat. 2016 Sep;55(3):414-421.	No
Monaghan, B.	2016	Randomised controlled trial to evaluate a physiotherapy-led functional exercise programme after total hip replacement. Physiotherapy. 2017 Sep;103(3):283-288.	Yes

Montanez-Heredia, E.	2016	Intra-articular injections of platelet-rich plasma versus hyaluronic acid in the treatment of osteoarthritic knee pain: A randomized clinical trial in the context of the Spanish national health care system. International Journal of Molecular Sciences 17.	No
Moorthy, S.	2016	Comparison of the efficacy and safety of tramadol versus tapentadol in acute osteoarthritic knee pain: A randomized, controlled trial. Asian Journal of Pharmaceutical and Clinical Research 9(3).	Yes
Mu, R.	2016	Efficacy and safety of loxoprofen hydrogel patch versus loxoprofen tablet in patients with knee osteoarthritis: a randomized controlled non-inferiority trial. Clinical Rheumatology 35(1): 165-173.	No
Munukka, M.	2016	Efficacy of progressive aquatic resistance training for tibiofemoral cartilage in postmenopausal women with mild knee osteoarthritis: a randomised controlled trial. Osteoarthritis and Cartilage 2016 Oct;24(10):1708-1717.	No
Myers, S.	2016	Effects of fucoidan from Fucus vesiculosus in reducing symptoms of osteoarthritis: A randomized placebo-controlled trial. Biologics: Targets and Therapy 10: 81-88.	No
Notarnicola, A.	2016	Methylsulfonylmethane and boswellic acids versus glucosamine sulfate in the treatment of knee arthritis: Randomized trial. International Journal of Immunopathology and Pharmacology 29, 140-146.	No
Ojoawo, A.	2016	Comparative effects of proprioceptive and isometric exercises on pain intensity and difficulty in patients with knee osteoarthritis: A randomised control study. Technology and Health Care 24(6): 853-863.	Yes
Ollivier, M.	2016	The John Insall Award: No Functional Benefit After Unicompartmental Knee Arthroplasty Performed With Patient-specific Instrumentation: A Randomized Trial. Clinical Orthopaedics & Related Research 474(1): 60-68.	No
Ozturk, A.	2016	Posterior cruciate-substituting total knee replacement recovers the flexion arc faster in the early postoperative period in knees with high varus deformity: a prospective randomized study. Archives of Orthopaedic and Trauma Surgery 136, 999-1006.	No
Parvizi, J.	2016	Total Hip Arthroplasty Performed Through Direct Anterior Approach Provides Superior Early Outcome: Results of a Randomized, Prospective Study. Orthop Clin North Am. 2016 Jul;47(3):497-504.	No
Pfitzner, T.	2016	Influence of the tourniquet on tibial cement mantle thickness in primary total knee arthroplasty. Knee Surgery, Sports Traumatology, Arthroscopy 24(1): 96-101.	No
Rahimzadeh, P.	2016	Adding intra-articular growth hormone to platelet rich plasma under ultrasound guidance in knee osteoarthritis: A comparative double-blind clinical trial. Anesth Pain Med. 2016 Oct 19;6(6):e41719.	Yes
Gopal, S. N.	2016	Radiological and biochemical effects (CTX-II, MMP-3, 8, and 13) of low-level laser therapy (LLLT) in chronic osteoarthritis in Al-Kharj, Saudi Arabia. Lasers in Medical Science: 1-7.	No
Saghafi, M.	2016	Oral glucosamine effect on blood glucose and insulin levels in patients with non-diabetic osteoarthritis: A double-blind, placebo-controlled clinical trial. Arch Rheumatol. 2016 Oct 1;31(4):340-345.	No
Sari, S. (Sari 2016 a)	2016	Which one is more effective for the clinical treatment of chronic pain in knee osteoarthritis: Radiofrequency neurotomy of the genicular nerves or intra-articular injection? Int J Rheum Dis. 2018 Oct;21(10):1772-1778.	Yes
Sari, S. (Sari 2016 b)	2016	Which imaging method should be used for genicular nerve radio frequency thermocoagulation in chronic knee osteoarthritis? Journal of Clinical Monitoring and Computing: 1-7.	Yes
Saw, M.	2016	Significant improvements in pain after a six-week physiotherapist-led exercise and education intervention, in patients with osteoarthritis awaiting arthroplasty, in South Africa: A randomised controlled trial. BMC Musculoskeletal Disorders 17.	No

Schinsky, M. F.	2016	Multifaceted Comparison of Two Cryotherapy Devices Used After Total Knee Arthroplasty. Orthopaedic Nursing 35(5): 309-316.	No
Schotanus, M. G.	2016	A radiological analysis of the difference between MRI- and CT-based patient-specific matched guides for total knee arthroplasty from the same manufacturer: a randomised controlled trial. Bone Joint J. 2016 Jun;98-B(6):786-92.	No
Simental-Mendia, M.	2016	Leukocyte-poor platelet-rich plasma is more effective than the conventional therapy with acetaminophen for the treatment of early knee osteoarthritis. Archives of Orthopaedic and Trauma Surgery 136(12): 1723-1732.	Yes
Singh, S..	2016	Effectiveness of hip abductor strengthening on health status, strength, endurance and six minute walk test in participants with medial compartment symptomatic knee osteoarthritis. Journal of Back and Musculoskeletal Rehabilitation 29, 65-75.	Yes
Skoffer, B.	2016	Efficacy of Preoperative Progressive Resistance Training on Postoperative Outcomes in Patients Undergoing Total Knee Arthroplasty. Arthritis Care and Research 68(9): 1239-1251.	No
Srivastava, S.	2016	Curcuma longa extract reduces inflammatory and oxidative stress biomarkers in osteoarthritis of knee: a four-month, double-blind, randomized, placebo-controlled trial. Inflammopharmacology 24(6): 377-388.	Yes
Stucinskas, J.	2016	Measuring long radiographs affects the positioning of femoral components in total knee arthroplasty: a randomized controlled trial. Arch Orthop Trauma Surg. 2016 May;136(5):693-700.	No
Tammachote, N.	2016	Intra-articular, single-shot hylan G-F 20 hyaluronic acid injection compared with corticosteroid in knee osteoarthritis:a double-blind, randomized controlled trial. Journal of Bone and Joint Surgery - American Volume 98(11): 885-892.	Yes
Teirlinck, C.	2016	Effectiveness of exercise therapy added to general practitioner care in patients with hip osteoarthritis: a pragmatic randomized controlled trial. Osteoarthritis and Cartilage 24, 82-90.	No
Vaishya, R.	2016	Intra-articular hyaluronic acid is superior to steroids in knee osteoarthritis: A comparative, randomized study. J Clin Orthop Trauma. 2017 Jan-Mar;8(1):85-88.	No
van der Voort, P.	2016	Comparison of femoral component migration between Refobacin bone cement R and Palacos R + G in cemented total hip arthroplasty: A randomised controlled roentgen stereophotogrammetric analysis and clinical study. Bone Joint J. 2016 Oct;98-B(10):1333-1341.	No
Verburg, H.	2016	Comparison of mini-midvastus and conventional total knee arthroplasty with clinical and radiographic evaluation a prospective randomized clinical trial with 5-year follow-up. Journal of Bone and Joint Surgery - American Volume 98, 1014-1022.	No
Wadsworth, L. T.	2016	Efficacy and safety of diclofenac sodium 2% topical solution for osteoarthritis of the knee: a randomized, double-blind, vehicle-controlled, 4 week study. Current Medical Research & Opinion 32(2): 241-250.	Yes
Wageck, B.	2016	Kinesio Taping does not improve the symptoms or function of older people with knee osteoarthritis: a randomised trial [with consumer summary]. Journal of Physiotherapy 2016 Jul;62(3):153-158.	Yes
Wallis, J. A.	2016	A walking program for people with severe knee osteoarthritis did not reduce pain but may have benefits for cardiovascular health: a phase II randomised controlled trial. Osteoarthritis Cartilage. 2017 Dec;25(12):1969-1979.	Yes
Wang, P. (Wang 2016 a)	2016	Effects of Whole Body Vibration Exercise associated with Quadriceps Resistance Exercise on functioning and quality of life in patients with knee osteoarthritis: A randomized controlled trial. Clin Rehabil. 2016 Nov;30(11):1074-1087.	Yes
Wang, C. (Wang 2016 b)	2016	Comparative effectiveness of Tai Chi versus physical therapy for knee osteoarthritis: A randomized trial. Annals of Internal Medicine 165, 77-86.	Yes

Wang, P. (Wang 2016 c)	2016	Effects of whole-body vibration training with quadriceps strengthening exercise on functioning and gait parameters in patients with medial compartment knee osteoarthritis: A randomised controlled preliminary study. Physiotherapy (United Kingdom) 102, 86-92.	Yes
Waterson, H. B.	2016	The early outcome of kinematic versus mechanical alignment in total knee arthroplasty: a prospective randomised control trial. Bone Joint J. 2016 Oct;98-B(10):1360-1368.	No
Winther, N. S.	2016	Comparison of a novel porous titanium construct (Regenerex) to a well proven porous coated tibial surface in cementless total knee arthroplasty - A prospective randomized RSA study with two-year follow-up. Knee. 2016 Dec;23(6):1002-1011.	No
Xin, Y.	2016	The efficacy and safety of sodium hyaluronate injection (Adant) in treating degenerative osteoarthritis: A multi-center, randomized, double-blind, positive-drug parallel-controlled and non-inferiority clinical study. International Journal of Rheumatic Diseases 19(3): 271-278.	No
Yataba, I.	2016	Efficacy of S-flurbiprofen plaster in knee osteoarthritis treatment: Results from a phase III, randomized, active-controlled, adequate, and well-controlled trial. Modern Rheumatology: 1-7.	No
Yegin, T.	2016	The Effect of Therapeutic Ultrasound on Pain and Physical Function in Patients with Knee Osteoarthritis. Ultrasound Med Biol. 2017 Jan;43(1):187-194. Epub 2016 Oct 7.	Yes
Yuan, Y.	2016	Clinical observation of pulsed radiofrequency in treatment of knee osteoarthritis. International Journal of Clinical and Experimental Medicine 9(10): 20050-20055.	Yes
Zhang, Y. (Zhang 2016 a)	2016	Influence of acupuncture in treatment of knee osteoarthritis and cartilage repairing. American Journal of Translational Research 8(9): 3995-4002.	Yes
Zhang, B. (Zhang 2016 b)	2016	Partial versus Intact Posterior Cruciate Ligament-retaining Total Knee Arthroplasty: a Comparative Study of Early Clinical Outcomes. Orthop Surg. 2016 Aug;8(3):331-7.	No
Zhang, Q. (Zhang 2016 c)	2016	Effect on Pain and Symptoms of Aspiration Before Hyaluronan Injection for Knee Osteoarthritis: A Prospective, Randomized, Single-blind Study. American journal of physical medicine & rehabilitation / Association of Academic Physiatrists 95(5): 366-371.	Yes
Zhu, Q.	2016	Effects of Tai Ji Quan training on gait kinematics in older Chinese women with knee osteoarthritis: A randomized controlled trial. Journal of Sport and Health Science 5(3): 297-303.	Yes

Note: Some cited as 2017 or 2018 due to delays between online publication and print publication.

Appendix B

Appendices for Chapter 3

B.1 WOMAC questionnaire (Likert version)

WOMAC Osteoarthritis Index LK3.1 (IK)

INSTRUCTIONS TO PATIENTS

In Sections A, B, and C questions are asked in the following format. Please mark your answers by putting an " ✗ " in one of the boxes.

EXAMPLES:

1. If you put your " ✗ " in the box on the far left as shown below,

none	mild	moderate	severe	extreme
☒	☐	☐	☐	☐

then you are indicating that you feel **no** pain.

2. If you put your " ✗ " in the box on the far right as shown below,

none	mild	moderate	severe	extreme
☐	☐	☐	☐	☒

then you are indicating that you feel **extreme** pain.

3. Please note:
   a) that the further to the right you place your " ✗ ", the **more** pain you feel.
   b) that the further to the left you place your " ✗ ", the **less** pain you feel.
   c) **please do not** place your " ✗ " **outside any of the boxes**.

You will be asked to indicate on this type of scale the amount of pain, stiffness or disability you have felt during the last 48 hours.

Think about your knee to be injected when answering the questions. Indicate the severity of your pain and stiffness and the difficulty you have in doing daily activities that you feel are caused by the arthritis in your knee to be injected.

Your knee to be injected has been identified for you by your health care professional. If you are unsure which knee is to be injected, please ask before completing the questionnaire.

V3 - English for USA
(at baseline)

WOMAC Osteoarthritis Index LK3.1 (IK)

Section A

PAIN

Think about the pain you felt during the last 48 hours caused by the arthritis in your knee to be injected.

(Please mark your answers with an " ✗ ".)

QUESTION: **How much pain have you had . . .**

	none	mild	moderate	severe	extreme	Study Coordinator Use Only
1. when walking on a flat surface?	❑	❑	❑	❑	❑	PAIN1 _____
2. when going up or down stairs?	❑	❑	❑	❑	❑	PAIN2 _____
3. at night while in bed? (that is - pain that disturbs your sleep)	❑	❑	❑	❑	❑	PAIN3 _____
4. while sitting or lying down?	❑	❑	❑	❑	❑	PAIN4 _____
5. while standing?	❑	❑	❑	❑	❑	PAIN5 _____

V3 - English for USA
(at baseline)

WOMAC Osteoarthritis Index LK3.1 (IK)

Section B

STIFFNESS

Think about the stiffness (not pain) you felt during the last 48 hours caused by the arthritis in your knee to be injected.

Stiffness is a sensation of **decreased** ease in moving your joint.

(Please mark your answers with an " ✗ ".)

6. How **severe** has your stiffness been **after you first woke up** in the morning?

none	mild	moderate	severe	extreme
❑	❑	❑	❑	❑

7. How **severe** has your stiffness been after sitting or lying down or while resting **later in the day**?

none	mild	moderate	severe	extreme
❑	❑	❑	❑	❑

Study Coordinator Use Only

STIFF6 _____

STIFF7 _____

WOMAC Osteoarthritis Index LK3.1 (IK)

Section C

DIFFICULTY PERFORMING DAILY ACTIVITIES

Think about the difficulty you had in doing the following daily physical activities during the last 48 hours caused by the arthritis in your knee to be injected. By this we mean **your ability to move around and take care of yourself**.

(Please mark your answers with an " ✗ ".)

QUESTION: **How much difficulty have you had . . .**	Study Coordinator Use Only
8. when going down the stairs? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN8 _____
9. when going up the stairs? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN9 _____
10. when getting up from a sitting position? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN10 _____
11. while standing? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN11 _____
12. when bending to the floor? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN12 _____
13. when walking on a flat surface? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN13 _____

V3 - English for USA
(at baseline)

WOMAC Osteoarthritis Index LK3.1 (IK)

DIFFICULTY PERFORMING DAILY ACTIVITIES

Think about the difficulty you had in doing the following daily physical activities during the last 48 hours caused by the arthritis in your knee to be injected. By this we mean **your ability to move around and take care of yourself**.

(Please mark your answers with an " ✗ ".)

QUESTION: **How much difficulty have you had . . .**	Study Coordinator Use Only
14. getting in or out of a car, or getting on or off a bus?	
none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN14 _____
15. while going shopping?	
none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN15 _____
16. when putting on your socks or panty hose or stockings?	
none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN16 _____
17. when getting out of bed?	
none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN17 _____
18. when taking off your socks or panty hose or stockings?	
none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN18 _____
19. while lying in bed?	
none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN19 _____

WOMAC Osteoarthritis Index LK3.1 (IK)

DIFFICULTY PERFORMING DAILY ACTIVITIES

Think about the difficulty you had in doing the following daily physical activities during the last 48 hours caused by the arthritis in your knee to be injected. By this we mean **your ability to move around and take care of yourself**.

(Please mark your answers with an " ✗ ".)

QUESTION: **How much difficulty have you had . . .**	Study Coordinator Use Only
20. when getting in or out of the bathtub? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN20 _____
21. while sitting? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN21 _____
22. when getting on or off the toilet? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN22 _____
23. while doing heavy household chores? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN23 _____
24. while doing light household chores? none ❑ mild ❑ moderate ❑ severe ❑ extreme ❑	PFTN24 _____

V3 - English for USA
(at baseline)

Appendix C

Appendices for Chapter 4

C.1 Search strategies for patient preference studies in osteoarthritis

MEDLINE

1. “discrete choice*”.tw.
2. DCE.tw.
3. “conjoint analysis”.tw.
4. “best-worst*”.tw.
5. BWS.tw.
6. attributes.tw.
7. “*preferences”.tw.
8. 1 or 2 or 3 or 4 or 5 or 6 or 7
9. exp osteoarthritis/
10. osteoarthr$.tw.
11. (degenerative adj3 (arthr$ or joint$ or disease$)).tw.
12. arthros?s.tw.
13. 9 or 10 or 11 or 12
14. 8 and 13

EMBASE

1. “discrete choice*”.tw.
2. DCE.tw.
3. “conjoint analysis”.tw.
4. “best-worst*”.tw.
5. BWS.tw.
6. attributes.tw.
7. “*preferences”.tw.
8. 1 or 2 or 3 or 4 or 5 or 6 or 7
9. exp osteoarthritis/
10. osteoarthr$.tw.
11. (degenerative adj3 (arthr$ or joint$ or disease$)).tw.
12. arthros?s.tw.
13. 9 or 10 or 11 or 12
14. 8 and 13

C.2 Details on ranking and rating exercises

The results of a literature search of stated preference studies for osteoarthritis treatments were used to create a long list of potential attributes that could be included, excluding those relating specifically to non-pharmacological treatments. Discrete choice experiments can include only a limited number of attributes. Therefore, patient input was used to identify which of the potential attributes were most important. A group of 10 osteoarthritis patients were recruited using an online advert placed on the PAIRS (Patients Active in Research) Thames Valley website. These 10 patient representatives conducted rating and ranking exercises on the importance of the potential attributes. The attributes were selected based on the results of the patient input and existing literature on patient preferences, whilst avoiding the selection of attributes with overlapping constructs.

Table C.1: Results of ranking exercises

Potential attribute:	**Ranking: 1 to 15 (1 is most important)**			
	Mean	**Median**	**IQR**	**Range**
The effect of the treatment on your pain	3.4	2	1 - 4.5	1 - 9
The effect of the treatment on your stiffness	6.4	7	1 - 9.5	1 - 12
The effect of the treatment on your ability to do daily activities, such as putting on socks or doing chores	6.9	7.5	1 - 9.75	1 - 13
The effect of the treatment on your level of fatigue	9.2	10	4 - 11.75	4 - 14
The effect of the treatment on your quality of sleep	6.2	6.5	1 - 7.75	1 - 11
Speed of symptom relief (how quickly it acts)	7	8	3 - 8	3 - 12
Length of symptom relief (how long the treatment effect lasts)	7.8	7.5	2 - 10	2 - 15
Whether you need a prescription or can purchase it over-the-counter	11.2	11.5	6 - 12.75	6 - 15
How often you need to take the medication	9.5	10.5	3 - 13.75	3 - 15
Cost to the NHS	12.9	13	11 - 14	11 - 15
The effect of the treatment on your anxiety	11.1	12	3 - 14	3 - 15
Common side effects, such as stomach ache, headache or diarrhoea	5.4	5	1 - 7	1 - 13
Risk of rare serious events, e.g., stomach ulcer, heart attack, stroke	5.2	5	1 - 5.75	1 - 14
The impact of the treatment on your work	10.9	11.5	3 - 14.75	3 - 15
The impact of the treatment on your social activities, e.g., going out with friends	6.8	7	2 - 10	2 - 12

Table C.2: Results of rating exercises

Potential attribute:	**Rating: 0 to 5 (5 is very important)**			
	Mean	**Median**	**IQR**	**Range**
The effect of the treatment on your pain	4.6	5	3 - 5	3 - 5
The effect of the treatment on your stiffness	3.9	4	2 - 5	2 - 5
The effect of the treatment on your ability to do daily activities, such as putting on socks or doing chores	4.2	4.5	3 - 5	3 - 5
The effect of the treatment on your level of fatigue	2.5	3	1 - 3	1 - 3
The effect of the treatment on your sleep quality	3.9	4	2 - 4.75	2 - 5
The effect of the treatment on your level of anxiety	2	1.5	0 - 3.5	0 - 5
The effect of the treatment, on your social activities e.g., going out with friends	3.8	4	2 - 5	2 - 5
The effect of the treatment on your work	2.4	2.5	0 - 3	0 - 5
Speed of symptom relief (how quickly the medication acts)	3.4	3.5	1 - 4	1 - 5
Length of symptom relief (how long the treatment effect lasts)	4	4	2 - 5	2 - 5
How often you need to take the medication	3.8	4.5	1 - 5	1 - 5
Whether you need a prescription or can buy it over-the-counter	2	2	0 - 3	0 - 4
Cost to you	1.6	1	0 - 2.75	0 - 4
Cost to the NHS	2.3	2.5	0 - 3.75	0 - 4
Stomach ache	3.9	4	2 - 5	2 - 5
Constipation	4	4.5	2 - 5	2 - 5
Diarrhoea	3.8	4	2 - 5	2 - 5
Headaches	3.5	4	1 - 4.75	1 - 5
Dizziness	3.9	4.5	1 - 5	1 - 5
Drowsiness	3.8	4	2 - 5	2 - 5
Risk of stomach ulcer	4.1	4.5	2 - 5	2 - 5
Risk of high blood pressure	4.2	5	2 - 5	2 - 5
Risk of heart attack	4.3	5	2 - 5	2 - 5
Risk of stroke	4.5	5	2 - 5	2 - 5
Risk of kidney and liver problems problems	4.4	5	2 - 5	2 - 5

C.3 Detailed example of choice task

Choice question 1

You should answer the choice questions while imagining you are the osteoarthritis patient with the symptoms described below. You should not focus on your own current or past condition.

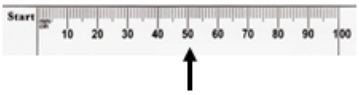
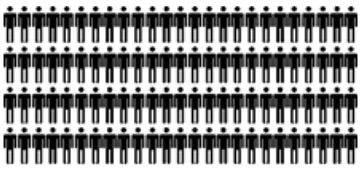
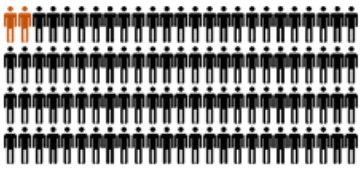
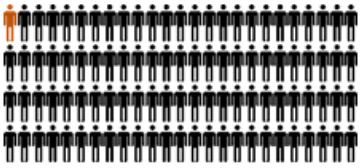
Imaginary patient (this is the same for each choice question):

Imagine that…

The patient is living with osteoarthritis in their knee or hip which causes severe pain. The osteoarthritis also causes severe stiffness and severe difficulty doing daily activities. Severe is marked as 75 of out 100 on a scale from 0 (none) to 100 (extreme). Please imagine that the patient has not had any heart problems or serious stomach problems in the past.

The only medication (gel, cream or tablet) that the patient is taking for their osteoarthritis is paracetamol. Taking paracetamol has not improved their symptoms. Their doctor has recommended 2 possible medications that could improve their symptoms. However, these medications might also increase their risk of a heart attack or stomach ulcer bleed. Their doctor has said that both medications are potential options for the patient and it is up to the patient to decide which they would prefer. Imagining that you are this patient, you must choose one of the 2 medications to take.

Benefits and risks for question 1 (this is different for each choice question):

	Medication A	Medication B
The level of pain	Moderate (50 / 100) 	Mild (25 / 100) 
The level of stiffness	Severe (75 / 100) 	None (0 / 100) 
The level of difficulty doing daily activities	Moderate (50 / 100) 	None (0 / 100) 
The length of the symptom relief	1 month	6 months
The risk of heart attack in the next year	0% (0 out of every 100 people) 	2% (2 out of every 100 people) 
The risk of stomach ulcer bleed in the next year	1% (1 out of every 100 people) 	2% (2 out of every 100 people) 

If you were this patient, which of these medications would you prefer to take?

Choose one of the following answers

Please choose only one of the following:

☐ Medication A

☐ Medication B

C.4 Ngene code

Ngene code for generating the pilot design:

```
Design
;alts(M1) = medA, medB
;rows = 16
;eff = M1(mnl, d)
;model(M1):
U(medA) = b0[0] + b1[0|0|0].effects * A[0,1,2,3]
+ b2[0|0|0].effects * B[0,1,2,3] + b3[0|0|0].effects * C[0,1,2,3]
+ b4[0|0|0].effects * D[1,3,6,12] + b5[0] * E[0,0.5,1,2]
+ b6[0] * F[0,1,2,4] /
U(medB) = b1 * A  + b2 * B + b3 * C + b4 * D + b5 * E + b6 * F
$
```

Ngene code for generating the main design:

```
Design
;alts(M1) = medA, medB
;alts(M2) = medA, medB
;rows = 16
;eff = 0.5*M1(mnl, d) + 0.5*M2(mnl,d)
? Note: Base level is 3 i.e. severe
? Note: Set intercept as 0 since unlabelled alternatives
;model(M1):
U(medA) = b1[0.258|0.222|-0.060].effects * A[0,1,2,3]
+ b2[0.260|0.108|-0.196].effects * B[0,1,2,3]
+ b3[-0.023|0.133|0.035].effects * C[0,1,2,3]
+ b4[-0.039|0.230|0.070].effects * D[12,6,3,1]
+ b5[-0.292] * E[0,0.5,1,2] + b6[-0.200] * F[0,1,2,4] /
U(medB) = b0[0] + b1 * A  + b2 * B + b3 * C + b4 * D
+ b5 * E + b6 * F
? Model 2 - Remove coefficients in illogical direction
? for function and duration so all priors for function are zero
? A - pain
? B - stiffness
? C - function
? D - duration
? E - heart attack risk
? F - stomach bleed risk
;model(M2):
U(medA) = b1[0.258|0.222|-0.060].effects * A[0,1,2,3]
+ b2[0.260|0.108|-0.196].effects * B[0,1,2,3]
+ b3[0|0|0].effects * C[0,1,2,3] + b4[0|0|0].effects * D[12,6,3,1]
+ b5[-0.292] * E[0,0.5,1,2] + b6[-0.200] * F[0,1,2,4] /
U(medB) = b0[0] + b1 * A  + b2 * B + b3 * C + b4 * D +
b5 * E + b6 * F
$
```

C.5 Experimental designs

Experimental design for the pilot study:

Choice task	**Pain (out of 100)**		**Stiffness (out of 100)**		**Difficulty doing daily activities (out of 100)**		**Duration (months)**		**Risk of heart attack (%)**		**Risk of stomach ulcer (%)**	
	A	B	A	B	A	B	A	B	A	B	A	B
1	75	50	0	50	25	75	12	1	1	0.5	4	0
2	0	75	0	25	75	0	6	1	0.5	1	1	2
3	25	75	25	75	50	25	3	12	2	0	4	0
4	0	50	25	50	0	50	12	3	0.5	1	2	1
5	25	0	0	25	0	50	1	12	0	2	2	1
6	0	25	75	25	0	25	3	6	1	0.5	2	1
7	50	25	25	50	25	75	3	12	0	2	0	4
8	50	25	25	0	75	50	6	3	0.5	1	4	0
9	25	0	50	0	25	50	3	1	2	0	1	2
10	0	75	50	75	0	75	1	3	1	0.5	0	4
11	75	50	75	50	75	0	1	12	2	0	0	4
12	75	25	50	75	25	0	1	6	0.5	1	4	0
13	25	75	75	0	50	0	12	6	0	2	1	2
14	75	50	50	0	50	25	6	1	0	2	1	2
15	50	0	75	25	50	75	6	3	2	0	2	1
16	50	0	0	75	75	25	12	6	1	0.5	0	4

Experimental design for the main study:

Choice task	**Pain (out of 100)**		**Stiffness (out of 100)**		**Difficulty doing daily activities (out of 100)**		**Duration (months)**		**Risk of heart attack (%)**		**Risk of stomach ulcer (%)**	
	A	B	A	B	A	B	A	B	A	B	A	B
1	0	75	0	75	25	75	1	3	0.5	1	2	0
2	0	75	25	50	75	25	6	12	1	0.5	1	2
3	75	50	0	50	50	25	1	12	2	0	2	1
4	0	75	50	25	25	0	3	1	2	0	4	0
5	50	75	75	25	0	25	12	1	1	0.5	0	2
6	75	25	50	25	0	75	3	12	0.5	1	0	4
7	50	25	25	0	0	50	6	3	0.5	1	4	1
8	50	25	75	0	50	0	1	6	0	2	1	2
9	25	0	50	75	50	0	6	3	2	0	0	4
10	25	0	0	75	0	25	12	6	0.5	1	4	0
11	75	50	75	50	25	75	12	3	1	0.5	1	2
12	0	50	50	0	75	50	1	6	1	0.5	0	4
13	25	0	75	25	75	50	6	12	0	2	2	1
14	75	25	0	75	75	0	12	1	0	2	2	1
15	50	0	25	50	25	50	3	6	2	0	1	4
16	25	50	25	0	50	75	3	1	0	2	4	0

C.6 Interpreting the model coefficients

The formula for the probability of choosing medication A is:

Probability of choosing medication A =

$$\frac{exp(\sum \beta_{iA})}{exp(\sum_i \beta_{iA}) + exp(\sum_i \beta_{iB})} \tag{C.1}$$

β_{iA} is the coefficient in Table 4.4 for the level of medication A for the i^{th} attribute. The levels and corresponding coefficients for choice task 2 are presented in Table C.3.

Table C.3: Levels and coefficients of choice task 2

	Medication A		**Medication B**	
Attribute	**Level**	**Coefficient**	**Level**	**Coefficient**
Pain	None	0.28865	Severe	-0.52971
Stiffness	Mild	0.03014	Moderate	0.03809
Function	Severe	-0.30675	Mild	0.04310
Duration	6 months	0.03287	12 months	0.35157
Heart attack risk	1%	-0.23884	0.5%	-0.11942
Stomach bleed risk	1%	-0.25894	2%	-0.51788
Total		-0.45287		-0.72226
Exp (total)		0.63580		0.48565

In choice task 2, the exponent of the sum of coefficients is 0.63580 for medication A and 0.48565 for medication B. According to the model, the probability of choosing medication A is:

$$\frac{0.63580}{0.63580 + 48565} = 0.567 \tag{C.2}$$

This corresponds to the value in Table 4.5. This is also similar to the proportion of respondents in our sample who selected medication A in choice task 2 (n=186/300, 62%).

Focusing on the difference in the sums of the coefficients, Equation C.1 can be re-arranged to:

Probability of choosing medication A =

$$\frac{exp(\sum \beta_{iA})}{exp(\sum_i \beta_{iA}) + exp(\sum_i \beta_{iB})} = \frac{1}{exp(\sum_i \beta_{iA} - \beta_{iB}) + 1} \tag{C.3}$$

C.7 Marginal rates of substitution (fixed effects model)

Table C.4: Marginal rates of substituton for the fixed effects model

		Marginal rates of substitution for:			
		Risk of heart attack		**Risk of stomach ulcer bleeding**	
		WTP	**95% CI***	**WTP**	**95% CI***
Pain	None	1.21	0.86 to 1.56	1.11	0.86 to 1.37
	Mild	1.31	0.97 to 1.64	1.21	0.95 to 1.47
	Moderate	-0.30	-0.59 to -0.01	-0.27	-0.54 to -0.01
	Severe	-2.22	-2.68 to -1.76	-2.05	-2.33 to -1.76
Stiffness	None	0.27	-0.03 to 0.56	0.25	-0.03 to 0.52
	Mild	0.13	-0.15 to 0.40	0.12	-0.14 to 0.37
	Moderate	0.16	-0.14 to 0.46	0.15	-0.12 to 0.42
	Severe	-0.55	-0.84 to -0.26	-0.51	-0.77 to -0.25
Function	None	0.87	0.54 to 1.21	0.81	0.54 to 1.07
	Mild	0.23	-0.05 to 0.52	0.21	-0.06 to 0.48
	Moderate	0.18	-0.11 to 0.47	0.17	-0.09 to 0.43
	Severe	-1.28	-1.64 to -0.93	-1.18	-1.44 to -0.93
Duration	12 months	1.47	1.04 to 1.90	1.36	1.07 to 1.64
	6 months	0.14	-0.13 to 0.41	0.13	-0.12 to 0.37
	3 months	-0.08	-0.35 to 0.19	-0.07	-0.32 to 0.18
	1 month	-1.52	-1.95 to -1.11	-1.41	-1.68 to -1.15
Heart attack risk		-	-	-0.92	-1.09 to -0.75
Stomach bleed risk		-1.08	-1.29 to -0.88	-	-

* CI: Confidence interval.

C.8 Effects on the probability of medication choice due to a one-level improvement in a single attribute

The effect of changing one attribute to the adjacent level is shown in Table C.5. If the medications were identical for all attributes, the probability of choosing either medication is 0.5. Table C.5 assumes that two medications, A and B, are identical in all attributes, except for one. As an example, in the first row, it assumes medication A reduces pain to 'mild' and medication B reduces pain to 'none' and indicates that the fixed effects model suggests that the probability of choosing medication B is 0.494. The probability of choosing medication A is therefore 0.506 in this example, highlighting that reducing pain from 'mild' to 'none' does not have a large effect on the choice of medication. Conversely, in the last row of Table C.5, comparing medication A with a 4% risk of stomach ulcer bleeding to medication B with a 2% risk, the probability of choosing medication B becomes 0.627.

Table C.5: Changes in the probability of medication selection due to one-level improvements in a single attribute

Attribute	**Change in level (level A to level B)**	**Coefficient for level A**	**Coefficient for level B**	**Difference in coefficients**	**Probability of selecting medication B**	**Increase in selection probability***
Pain	Mild to none	0.31266	0.28865	-0.02400	0.494	-0.6%
	Moderate to mild	-0.07161	0.31266	0.38426	0.595	9.5%
	Severe to moderate	-0.52971	-0.07161	0.45810	0.613	11.3%
Stiffness	Mild to none	0.03014	0.06354	0.03340	0.508	0.8%
	Moderate to mild	0.03809	0.03014	-0.00795	0.498	-0.2%
	Severe to moderate	-0.13177	0.03809	0.16986	0.542	4.2%
Function	Mild to none	0.05509	0.20856	0.15347	0.538	3.8%
	Moderate to mild	0.04310	0.05509	0.01199	0.503	0.3%
	Severe to moderate	-0.30675	0.04310	0.34984	0.587	8.7%
Duration	6 to 12 months	0.03287	0.35157	0.31870	0.579	7.9%
	3 to 6 months	-0.01916	0.03287	0.05203	0.513	1.3%
	1 to 3 months	-0.36529	-0.01916	0.34614	0.586	8.6%
Risk of heart attack	0.5% to 0%	-0.11942	0.00000	0.11942	0.530	3.0%
	1% to 0.5%	-0.23884	-0.11942	0.11942	0.530	3.0%
	2% to 1%	-0.47769	-0.23884	0.23884	0.559	5.9%
Risk of stomach ulcer bleeding	1% to 0%	-0.25894	0.00000	0.25894	0.564	6.4%
	2% to 1%	-0.51788	-0.25894	0.25894	0.564	6.4%
	4% to 2%	-1.03576	-0.51788	0.51788	0.627	12.7%

* Absolute increase in probability of selecting medication B compared to 50%. All other attributes are equal in the two medications.

C.9 Fixed effects model with interactions

Table C.6: Coefficients for fixed effects model with interactions

Main effects:		**Coefficient**	**LCI**	**UCI***
Pain	None	0.30418	0.23652	0.37185
	Mild	0.33115	0.26094	0.40135
	Moderate	-0.05233	-0.12139	0.01674
Stiffness	None	0.06028	-0.01300	0.13357
	Mild	0.02224	-0.04527	0.08976
	Moderate	0.04380	-0.02578	0.11338
Function	None	0.22758	0.16086	0.29432
	Mild	0.05881	-0.01283	0.13046
	Moderate	0.03445	-0.03382	0.10273
Duration	12 months	0.46891	0.37684	0.56097
	6 months	0.03454	-0.02996	0.09904
	3 months	-0.02360	-0.08835	0.04116
Heart attack risk		-0.25954	-0.31338	-0.20571
Stomach bleed risk		-0.28896	-0.32070	-0.25721
Interactions:				
Age * Pain	None	0.00539	0.00089	0.00990
	Mild	0.00554	0.00062	0.01045
Age * Duration	12 months	0.01051	0.00594	0.01508
Age * Heart attack risk		-0.00723	-0.01081	-0.00364
Age * Stomach bleed risk		-0.00384	-0.00595	-0.00173
WOMAC * Pain	None	-0.00568	-0.00840	-0.00296
	Mild	-0.00326	-0.00610	-0.00043
WOMAC * Function	None	-0.00438	-0.00683	-0.00194
WOMAC * Duration	12 months	-0.00295	-0.00575	-0.00015
WOMAC * Stomach bleed risk		0.00198	0.00088	0.00309
Sex * Duration	12 months	-0.16220	-0.27294	-0.05147
Risk loving * Stomach bleed risk		0.09691	0.03219	0.16163

* LCI: Lower Confidence interval, UCI: Upper Confidence interval.

C.10 Coefficients for models with interactions between pain and duration (excluding stiffness)

Table C.7: Mixed effects model removing stiffness attribute (to show similarity of coefficients)

		Coefficient	95% LCI	95% UCI*
Main effects:				
Pain	None	0.47709	0.35979	0.59439
	Mild	0.57590	0.45289	0.69892
	Moderate	-0.00031	-0.09178	0.09116
Function	None	0.50066	0.36838	0.63295
	Mild	0.05271	-0.0369	0.14233
	Moderate	0.02147	-0.06158	0.10453
Duration	12 months	0.68368	0.53353	0.83384
	6 months	0.06332	-0.01456	0.14119
	3 months	-0.08426	-0.16167	-0.00685
Risk of heart attack		-0.44929	-0.56309	-0.33548
Risk of stomach ulcer bleeding		-0.45467	-0.53416	-0.37518

* LCI: Lower confidence interval, UCI: Upper confidence interval.

Table C.8: Interactions between pain (none) and duration

		Coefficient	95% LCI	95% UCI*
Main effects:				
Pain	None	1.26448	0.99488	1.53408
(1 month duration base)	Mild	0.38322	0.21449	0.55195
	Moderate	-0.25965	-0.45696	-0.06235
	Severe	-1.38805	-1.74034	-1.03575
Function	None	0.35695	0.19243	0.52148
	Mild	-0.15580	-0.33877	0.02716
	Moderate	0.26782	0.09752	0.43811
	Severe	-0.46897	-0.61338	-0.32456
Duration	12 months	0.50876	0.29553	0.72199
	6 months	0.57397	0.20741	0.94053
	3 months	-0.18296	-0.31820	-0.04771
	1 month	-0.89977	-1.13716	-0.66239
Heart attack risk	0%	0		
	0.5%	-0.26635	-0.34172	-0.19097
	1%	-0.53269	-0.68344	-0.38195
	2%	-1.06538	-1.36687	-0.76389
Stomach bleed risk	0%	0		
	1%	-0.52420	-0.66039	-0.38801
	2%	-1.04841	-1.32079	-0.77603
	4%	-2.09681	-2.64157	-1.55206
Interactions:				
Pain (none) x Duration	3 months	-0.26556	-1.03811	0.50699
	6 months	-2.02240	-2.97695	-1.06785
	12 months	-0.19951	-0.72231	0.32328

* LCI: Lower confidence interval, UCI: Upper confidence interval.

Table C.9: Interactions between pain (mild) and duration

		Coefficient	95% LCI	95% UCI*
Main effects:				
Pain	None	0.91240	0.71875	1.10606
(1 month duration base)	Mild	-0.45048	-0.86914	-0.03182
	Moderate	0.10047	-0.02312	0.22407
	Severe	-0.56240	-0.91904	-0.20575
Function	None	0.48331	0.31110	0.65552
	Mild	0.09653	-0.08876	0.28182
	Moderate	0.25622	0.13398	0.37845
	Severe	-0.83606	-1.03823	-0.63388
Duration	12 months	0.85297	0.66585	1.04010
	6 months	-0.37326	-0.54611	-0.20040
	3 months	0.11074	0.01092	0.21055
	1 month	-0.59045	-0.77887	-0.40203
Heart attack risk	0%	0		
	0.5%	-0.23805	-0.30654	-0.16957
	1%	-0.47611	-0.61308	-0.33913
	2%	-0.95221	-1.22617	-0.67826
Stomach bleed risk	0%	0		
	1%	-0.43335	-0.52273	-0.34397
	2%	-0.86670	-1.04547	-0.68793
	4%	-1.73340	-2.09093	-1.37586
Interactions:				
Pain (mild) x Duration	3 months	0.38225	-0.02153	0.78603
	6 months	2.29726	1.53366	3.06086
	12 months	0.94848	0.29614	1.60081

* LCI: Lower confidence interval, UCI: Upper confidence interval.

Table C.10: Interactions between pain (moderate) and duration

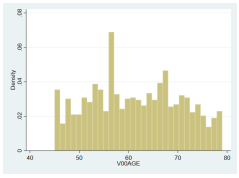
		Coefficient	95% LCI	95% UCI*
Main effects:				
Pain	None	0.90700	0.58584	1.22816
(1 month duration base)	Mild	1.10381	0.67443	1.53320
	Moderate	0.29542	-0.01662	0.60745
	Severe	-2.30623	-3.06727	-1.54518
Function	None	0.82894	0.60975	1.04813
	Mild	-0.39003	-0.62036	-0.15969
	Moderate	0.07789	-0.01948	0.17525
	Severe	-0.51680	-0.71161	-0.32198
Duration	12 months	0.59935	0.42874	0.76996
	6 months	-0.03282	-0.32619	0.26055
	3 months	0.25889	0.02043	0.49735
	1 month	-0.82542	-1.02217	-0.62867
Heart attack risk	0%	0		
	0.5%	-0.29634	-0.37753	-0.21515
	1%	-0.59267	-0.75505	-0.43029
	2%	-1.18535	-1.51011	-0.86059
Stomach bleed risk	0%	0		
	1%	-0.69469	-0.88501	-0.50436
	2%	-1.38938	-1.77003	-1.00873
	4%	-2.77875	-3.54005	-2.01745
Interactions:				
Pain (moderate) x Duration	3 months	-0.13518	-0.44297	0.17260
	6 months	0.72525	-0.76897	2.21948
	12 months	-2.29003	-3.44286	-1.13721

* LCI: Lower confidence interval, UCI: Upper confidence interval.

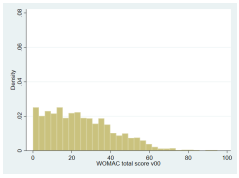
Appendix D

Appendices for Chapter 5

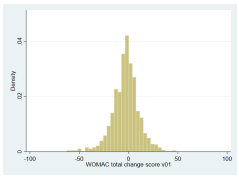
D.1 Histograms of age and WOMAC scores



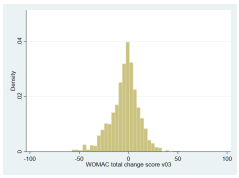
(a) Age



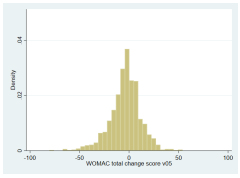
(b) WOMAC total at baseline



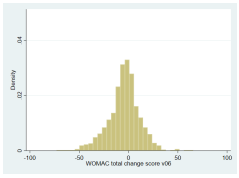
(c) WOMAC total change at 1 year



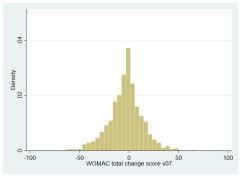
(d) WOMAC total change at 2 years



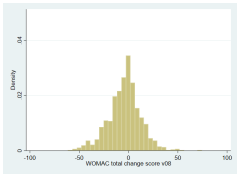
(e) WOMAC total change at 3 years



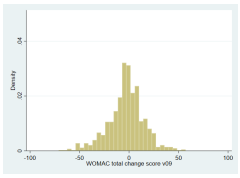
(f) WOMAC total change at 4 years



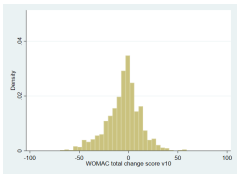
(g) WOMAC total change at 5 years



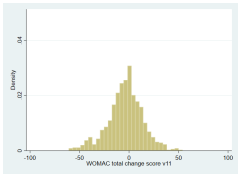
(h) WOMAC total change at 6 years



(i) WOMAC total change at 7 years



(j) WOMAC total change at 8 years



(k) WOMAC total change at 9 years

Note: Change scores are the change from baseline to the relevant follow-up time point

D.2 Follow-up and change scores for WOMAC subscales

	Follow up score			Change from baseline		
	Mean	**SD**	**n**	**Mean**	**SD**	**n**
WOMAC pain						
Baseline	5.17	3.73	1390			
1 year	4.31	3.68	1267	-0.70	3.19	1267
2 years	4.19	3.70	1204	-0.81	3.48	1204
3 years	4.12	3.73	1174	-0.86	3.65	1174
4 years	3.99	3.74	1175	-1.02	3.77	1175
5 years	4.12	3.78	1094	-0.85	3.92	1094
6 years	4.05	3.79	1053	-0.94	3.87	1053
7 years	4.12	3.92	1046	-0.86	4.18	1046
8 years	3.76	3.68	1074	-1.25	4.06	1074
9 years	3.90	3.67	900	-1.05	4.02	900
WOMAC function						
Baseline	16.90	12.55	1379			
1 year	14.39	12.25	1259	-1.94	10.12	1250
2 years	13.79	12.28	1188	-2.54	10.59	1179
3 years	13.65	12.33	1156	-2.40	11.56	1151
4 years	13.12	12.38	1163	-3.05	11.72	1157
5 years	13.87	12.37	1081	-2.19	12.50	1076
6 years	13.54	12.59	1036	-2.60	12.14	1032
7 years	14.33	12.93	1027	-1.74	13.22	1022
8 years	12.73	12.10	982	-3.39	12.54	977
9 years	13.74	12.42	882	-2.34	12.91	878
WOMAC stiffness						
Baseline	2.78	1.67	1389			
1 year	2.38	1.68	1267	-0.33	1.63	1266
2 years	2.29	1.73	1204	-0.43	1.75	1203
3 years	2.25	1.69	1172	-0.45	1.76	1171
4 years	2.17	1.70	1176	-0.52	1.81	1175
5 years	2.14	1.66	1094	-0.53	1.86	1093
6 years	2.10	1.67	1053	-0.58	1.84	1052
7 years	2.11	1.66	1048	-0.58	1.89	1047
8 years	2.14	1.74	998	-0.55	1.95	997
9 years	2.11	1.69	900	-0.58	1.94	899

WOMAC pain subscale ranges from 0-20, function 0-68 and stiffness 0-8.
Higher scores indicate worse symptoms for all subscales.

D.3 Subgroup analysis: ANCOVA and ROC curve method

Table D.1: Subgroup analysis of MCID estimates for improvement in the WOMAC total score using the ANCOVA method based on change from baseline

	ANCOVA method			**ANCOVA method**			**ANCOVA method**		
	β	**95% CI**	**n**	β	**95% CI**	**n**	β	**95% CI**	**n**
	Age: 45-56			**Age: 57-66**			**Age: 67-79**		
1 year	3.78	(1.04, 6.53)	186	0.65	(-2.22, 3.53)	197	1.49	(-1.57, 4.56)	179
2 years	-0.52	(-3.64, 2.59)	182	1.24	(-2.11, 4.59)	165	3.43	(-0.52, 7.37)	146
3 years	4.43	(1.30, 7.56)	170	-0.72	(-4.61, 3.17)	152	3.52	(-0.54, 7.58)	162
4 years	2.85	(-0.15, 5.86)	171	5.12	(1.20, 9.05)	149	3.99	(-0.37, 8.34)	131
5 years	6.28	(2.44, 10.12)	144	-0.76	(-4.74, 3.22)	154	3.17	(-1.66, 7.99)	122
6 years	-1.95	(-5.58, 1.68)	155	3.44	(-0.31, 7.19)	151	0.86	(-3.67, 5.40)	113
7 years	4.05	(-0.23, 8.34)	137	8.45	(3.64, 13.26)	151	-2.90	(-7.65, 1.84)	115
8 years	-2.26	(-6.16, 1.65)	140	-0.23	(-4.55, 4.09)	121	3.42	(-1.86, 8.69)	104
9 years	1.42	(-2.68, 5.51)	131	0.96	(-3.26, 5.17)	127	4.20	(-1.09, 9.48)	87
	WOMAC: 0-12			**WOMAC: 13-36**			**WOMAC: 37-96**		
1 year	-1.25	(-3.06, 0.56)	207	4.16	(1.43, 6.90)	253	2.27	(-2.44, 6.97)	102
2 years	0.71	(-1.55, 2.98)	186	1.29	(-1.85, 4.42)	234	2.05	(-4.34, 8.45)	73
3 years	0.85	(-2.02, 3.73)	166	0.97	(-1.89, 3.83)	245	12.63	(4.80, 20.46)	73
4 years	1.27	(-0.49, 3.04)	177	4.85	(1.68, 8.03)	197	8.48	(0.22, 16.74)	77
5 years	-0.05	(-3.07, 2.97)	161	5.66	(2.10, 9.23)	193	-0.05	(-8.54, 8.44)	66
6 years	-1.12	(-3.80, 1.56)	163	2.06	(-1.61, 5.72)	195	0.48	(-7.13, 8.09)	61
7 years	0.22	(-2.69, 3.14)	157	4.59	(0.72, 8.46)	184	8.61	(-1.95, 19.18)	62
8 years	-3.58	(-6.67, -0.50)	146	2.13	(-1.63, 5.89)	169	3.92	(-6.22, 14.05)	50

Table D.1: Subgroup analysis of MCID estimates for improvement in the WOMAC total score using the ANCOVA method based on change from baseline

	ANCOVA method			**ANCOVA method**			**ANCOVA method**		
	β	**95% CI**	**n**	β	**95% CI**	**n**	β	**95% CI**	**n**
9 years	2.78	(-0.31, 5.87)	151	0.01	(-4.00, 4.02)	138	2.24	(-7.33, 11.80)	56
	Male			**Female**					
1 year	1.89	(-0.43, 4.20)	245	2.05	(-0.28, 4.39)	317			
2 years	1.77	(-1.04, 4.57)	224	0.47	(-2.30, 3.23)	269			
3 years	3.38	(0.35, 6.41)	207	2.16	(-0.81, 5.13)	277			
4 years	3.93	(1.13, 6.74)	199	4.02	(0.90, 7.14)	252			
5 years	1.00	(-2.39, 4.38)	183	3.69	(0.29, 7.10)	237			
6 years	0.81	(-2.31, 3.93)	191	0.64	(-2.65, 3.93)	228			
7 years	5.28	(2.04, 8.52)	193	1.49	(-2.79, 5.76)	210			
8 years	3.13	(-0.09, 6.35)	177	-3.27	(-7.12, 0.58)	188			
9 years	0.79	(-2.80, 4.38)	158	2.72	(-0.95, 6.39)	187			

Table D.2: Subgroup analysis of MCID estimates for improvement in the WOMAC total score using the ROC curve method based on change from baseline

	ROC curve			**ROC curve**			**ROC curve**		
	β	**95% CI**	**n**	β	**95% CI**	**n**	β	**95% CI**	**n**
	Age: 45-56			**Age: 57-66**			**Age: 67-79**		
1 year	3.00	(-0.88, 6.88)	186	7.00	(2.13, 11.87)	197	3.70	(1.42, 5.98)	179
2 years	1.80	(-1.58, 5.18)	182	3.00	(1.03, 4.97)	165	3.15	(-1.91, 8.21)	146
3 years	2.40	(-0.19, 4.99)	170	6.00	(0.16, 11.84)	152	7.10	(2.45, 11.75)	162
4 years	2.50	(-1.81, 6.81)	171	5.90	(3.78, 8.02)	149	6.15	(3.12, 9.18)	131
5 years	2.10	(-2.31, 6.51)	144	5.00	(1.45, 8.55)	154	3.20	(-1.71, 8.11)	122
6 years	3.50	(-0.67, 7.67)	155	10.00	(4.77, 15.23)	151	7.20	(1.15, 13.25)	113
7 years	1.90	(-1.31, 5.11)	137	5.20	(2.65, 7.75)	151	4.20	(-0.52, 8.92)	115
8 years	3.00	(-0.57, 6.57)	140	4.10	(-0.38, 8.58)	121	6.00	(2.03, 9.97)	104
9 years	1.50	(-2.41, 5.41)	131	4.50	(-0.07, 9.07)	127	4.60	(-1.05, 10.25)	87
	WOMAC: 0-12			**WOMAC: 13-36**			**WOMAC: 37-96**		
1 year	0.00	(-3.12, 3.12)	207	2.10	(-2.87, 7.07)	253	14.70	(7.50, 21.90)	102
2 years	2.50	(0.94, 4.06)	186	3.55	(-0.94, 8.04)	234	3.40	(-7.08, 13.88)	73
3 years	2.50	(0.18, 4.82)	166	7.10	(1.04, 13.16)	245	16.70	(8.26, 25.14)	73
4 years	3.50	(2.07, 4.93)	177	8.60	(4.56, 12.64)	197	7.70	(-1.94, 17.34)	77
5 years	3.20	(0.32, 6.08)	161	-1.00	(-6.76, 4.76)	193	14.95	(5.10, 24.81)	66
6 years	1.00	(-1.11, 3.11)	163	11.50	(2.94, 20.06)	195	12.40	(2.13, 22.67)	61
7 years	1.00	(-1.44, 3.44)	157	3.20	(-0.46, 6.86)	184	24.00	(11.18, 36.82)	62
8 years	2.50	(0.16, 4.84)	146	5.55	(-0.82, 11.92)	169	33.00	(12.70, 53.30)	50
9 years	2.50	(0.03, 4.97)	151	6.30	(-0.46, 13.06)	138	10.50	(-0.08, 21.08)	56

Table D.2: Subgroup analysis of MCID estimates for improvement in the WOMAC total score using the ROC curve method based on change from baseline

	ROC curve			**ROC curve**			**ROC curve**		
	β	**95% CI**	**n**	β	**95% CI**	**n**	β	**95% CI**	**n**
	Male			**Female**					
1 year	3.70	(0.11, 7.29)	245	3.10	(-0.10, 6.30)	317			
2 years	3.25	(1.82, 4.68)	224	4.00	(1.11, 6.89)	269			
3 years	5.50	(2.99, 8.01)	207	8.00	(3.68, 12.32)	277			
4 years	4.20	(1.42, 6.98)	199	8.30	(4.31, 12.29)	252			
5 years	3.70	(1.21, 6.19)	183	5.30	(2.70, 7.90)	237			
6 years	3.50	(-3.95, 10.95)	191	7.30	(2.99, 11.61)	228			
7 years	1.10	(-2.39, 4.59)	193	3.10	(-1.82, 8.03)	210			
8 years	5.55	(2.68, 8.42)	177	2.50	(-1.15, 6.15)	188			
9 years	4.60	(0.39, 8.81)	158	4.50	(1.52, 7.48)	187			

D.4 Stata code for longitudinal methods

D.4.1 Code for longitudinal definition of minimal improvement

```
gen auc_kglrs=.
gen tmax_kglrs=.
gen tomc_kglrs=.

levelsof id, local(levels)
foreach l of local levels {
display "‘l’"
sum obsn if id==‘l’
global fit = npreskglrs[‘=r(min)’]
sum obsn if id==‘l’
global zeron = nzermiskglrs[‘=r(min)’]
display "$fit"
if $fit > 2 & $zeron < 9 {
display "pkexamine year kglrsy if id==‘l’, fit($fit) trapezoid"
pkexamine year kglrsy if id==‘l’, fit($fit) trapezoid
replace auc_kglrs = r(auc) if id==‘l’
replace tmax_kglrs = r(tmax) if id==‘l’
replace tomc_kglrs = r(tomc) if id==‘l’
}
}

gen auc_kglrs_change = auc_kglrs - (tmax_kglrs*v00kglrs)
gen auc_kglrs_chperyr = auc_kglrs_change/tmax_kglrs

gen long_minimproved = -1 if auc_kglrs_chperyr<=-0.8 & ///
auc_kglrs_chperyr>=-1.2
replace long_minimproved = 0 if auc_kglrs_chperyr<=0.2 & ///
auc_kglrs_chperyr>=-0.2
```

Note: Where possible, AUC values were calculated manually for participants with 0 or 1 observed non-zero values for the change in the anchor measure.

D.4.2 Code for adapted raw mean difference method

```
foreach x in v01 v03 v05 v06 v07 v08 v09 v10 v11 {
regress ‘x’womtsc_ch_bsl long_minimproved
}
```

Note: The median values for the MCID estimates at different follow-up time points was calculated manually.

D.4.3 Code for area-under-the-curve method

```
regress auc_wmc_chperyr long_minimproved
```

Note: The AUC for the WOMAC score was calculated as for the anchor measure in the longitudinal definition of minimal improvement.

D.4.4 Code for mixed effects regression method

```
xtmixed womtsc_ch_y long_minimproved || id:, covariance(identity) ///
mle vce(robust)
```

D.4.5 Code for generalised estimating equation method

```
xtset id year
xtgee womtsc_ch_y long_minimproved v00womtsc, family(gaussian) ///
link(identity) corr(exchangeable)
```

Appendix E

Appendices for Chapter 6

E.1 Example of Stata code to generate datasets

```
clear
set more off

local nreps = 1600
local runfile = 1
local i = 1
local s = 1


* Linear improvement
matrix pattern1 = ( 0.5, 1, ., ., . \ ///
0.25, 0.5, 1, ., . \ ///
1/6, 1/3, 2/3, 1, . \ ///
0.125, 0.25, 0.5, 0.75, 1)

* Short term improvement then plateau
matrix pattern2 = (1, 1, ., ., . \ ///
0.5, 1, 1, ., . \ ///
0.5, 1, 1, 1, . \ ///
0.5, 1, 1, 1, 1)

* Short term temporary improvement
matrix pattern3 = ( 1, 0, ., ., . \ ///
1, 1, 0, ., . \ ///
1, 1, 0, 0, . \ ///
1, 1, 0, 0, 0)


tempname sim_‘runfile’

postfile ‘sim_‘runfile’’ ///
int(i) float(seed) int(t p b n r) ///
float(trteffect_time1 trteffect_time2 trteffect_time3 ///
trteffect_time4 trteffect_time5) ///
str100(dataset) ///
str2000(rngstate_start1 rngstate_start2 rngstate_start3) ///
str2000(rngstate_end1 rngstate_end2 rngstate_end3) ///
scenario ///
using "Sim_datasets_15Mar2019_run‘runfile’.dta", replace

forvalues time = 1/4 {
```

```
local followups = ‘time’+1
local assessments = ‘time’+2

local wom_terms_t‘time’ wom_0

forvalues t = 1/‘followups’ {
local wom_terms_t‘time’ ‘wom_terms_t‘time’’ wom_‘t’
}

* Note: Time is row in pattern matrix.
* Follow ups (time+1) is number of follow up time points.
* Assessments (time+2) is number of time points
* including baseline.


forvalues p = 1/3 {
foreach b in 0 4 8 12 {
foreach n in 100 200 400 600 800 {

clear
set seed ‘seed’

matrix M = (40, J(1,‘followups’,40))
matrix S = (15, J(1,‘followups’,12))
matrix i1 = 0.5*I(‘assessments’)
matrix i2 = J(‘assessments’,‘assessments’,0.5)
matrix C = i1 + i2

forvalues r = 1/1600 {

clear
set obs ‘n’

local rngstate_start1 = substr(c(rngstate),1,2000)
local rngstate_start2 = substr(c(rngstate),2001,4000)
local rngstate_start3 = substr(c(rngstate),4001,.)
gen rndstate = c(rngstate)

drawnorm ‘wom_terms_t‘time’’, means(M) sds(S) corr(C)

* Generate variable for treatment arm
gen treat = runiform()<0.50

* Generate random intercept by participant
```

```
gen withinperson_var = rnormal(0, 15)
gen wom_bsl = wom_0 + withinperson_var

* Generate treatment effects

forvalues t = 1/`followups' {
gen trteffect`t' = treat*`b'*pattern`p'[`time',`t']
local trteff`t' `b'*pattern`p'[`time',`t']
gen wom_fu_`t' = wom_`t' - trteffect`t' + withinperson_var
gen wom_ch_`t' = wom_fu_`t' - wom_bsl
}

* Differentiate whether from original dataset or resampled

gen append = "original"

* Count and drop if one or more WOMAC scores are
* outside 0-96 range

gen resample = 0
replace resample = 1 if wom_bsl<0 | wom_bsl>96
forvalues t = 1/`followups' {
replace resample = 1 if wom_fu_`t'<0 | wom_fu_`t'>96
}

count if resample==1
local numoutsiderange = `r(N)'
display `numoutsiderange'

drop if resample==1

save "sim_time`time'_pat`p'_trt`b'_n`n'_rep`r'_id`i'_
07Mar2019_BC_shortened.dta", replace


* Re-sample if one or more WOMAC scores are outside
* 0-96 range until all scores are within the
* correct range

local repeat = 1
while `numoutsiderange' > 0 {

clear
set obs `numoutsiderange'
```

```
gen rndstate = c(rngstate)

drawnorm ‘wom_terms_t‘time’’, means(M) sds(S) corr(C)

* Generate variable for treatment arm
gen treat = runiform()<0.50

* Generate random intercept by participant
gen withinperson_var = rnormal(0, 15)
gen wom_bsl = wom_0 + withinperson_var

forvalues t = 1/‘followups’ {
gen trteffect‘t’ = treat*‘b’*pattern‘p’[‘time’,‘t’]
local trteff‘t’ ‘b’*pattern‘p’[‘time’,‘t’]
gen wom_fu_‘t’ = wom_‘t’ - trteffect‘t’ + withinperson_var
gen wom_ch_‘t’ = wom_fu_‘t’ - wom_bsl
}

gen resample = 0
replace resample = 1 if wom_bsl<0 | wom_bsl>96
forvalues t = 1/‘followups’ {
replace resample = 1 if wom_fu_‘t’<0 | wom_fu_‘t’>96
}

count if resample==1
local numoutsiderange = ‘r(N)’
display "Number outside range is: " ‘numoutsiderange’
drop if resample==1

gen append = "append_‘repeat’"

save "sim_time‘time’_pat‘p’_trt‘b’_n‘n’_rep‘r’_id‘i’_
07Mar2019_BC_toappend‘repeat’.dta", replace

use "sim_time‘time’_pat‘p’_trt‘b’_n‘n’_rep‘r’_id‘i’_
07Mar2019_BC_shortened.dta", clear

append using "sim_time‘time’_pat‘p’_trt‘b’_n‘n’_rep‘r’_
id‘i’_07Mar2019_BC_toappend‘repeat’.dta"

save "sim_time‘time’_pat‘p’_trt‘b’_n‘n’_rep‘r’_id‘i’_
07Mar2019_BC_shortened.dta", replace

erase "sim_time‘time’_pat‘p’_trt‘b’_n‘n’_rep‘r’_id‘i’_
07Mar2019_BC_toappend‘repeat’.dta"
```

```
local repeat = `repeat'+1

}

local rngstate_end = c(rngstate)

local rngstate_end1 = substr(c(rngstate),1,2000)
local rngstate_end2 = substr(c(rngstate),2001,4000)
local rngstate_end3 = substr(c(rngstate),4001,.)

* Save final dataset and check whether all WOMAC scores
* are within range 0-96

gen pattern = `p'
gen time = `time'
gen assessments = `assessments'
gen maxtrteffect = `b'
gen samplesize = `n'
gen scenario_id = `s'
gen dataset_id = `i'


* Round WOMAC values to get integer

forvalues t = 1/`followups' {
gen wom_fu_notround_`t' = wom_fu_`t'
gen wom_ch_notround_`t' = wom_ch_`t'
replace wom_fu_`t' = round(wom_fu_`t')
replace wom_ch_`t' = round(wom_ch_`t')
}

gen wom_bsl_notround = wom_bsl
replace wom_bsl = round(wom_bsl)

save "sim_final_time`time'_pat`p'_trt`b'_n`n'_rep`r'_
07Mar2019_BC.dta", replace


forvalues k = 1/5 {
local trteff`k' = pattern`p'[`time',`k']*`b'
}

post `sim_`runfile'' (`i') (`seed') (`followups') ///
(`p') (`b') (`n') (`r') ///
```

```
(‘trteff1’) (‘trteff2’) (‘trteff3’) ///
(‘trteff4’) (‘trteff5’) ///
("sim_time‘time’_pat‘p’_trt‘b’_n‘n’_rep‘r’_id‘i’_
07Mar2019_BC.dta") ///
("‘rngstate_start1’") ("‘rngstate_start2’") ///
("‘rngstate_start3’") ("‘rngstate_end1’") ///
("‘rngstate_end2’") ("‘rngstate_end3’") (‘s’)


local i = ‘i’+1


}

local s = ‘s’+1

}

}

}

}

postclose ‘sim_‘runfile’’
```

E.2 Example of Stata code to analyse datasets

Single time point ANCOVA code:

```
forvalues t = 1/4 {

forvalues p = 1/3 {

foreach b in 0 4 8 12 {

foreach n in 100 200 400 600 800 {

forvalues r = 1/1600 {

use "sim_final_time‘t’_pat‘p’_trt‘b’_n‘n’_rep‘r’_
07Mar2019_BC.dta", clear

gen pat_id = _n

local folups = ‘t’+1

forvalues tpoint = 1/‘folups’ {

regress wom_fu_‘tpoint’ i.treat wom_bsl

}
}
}
}
}
}
```

GEE code:

```
use "sim_final_time‘t’_pat‘p’_trt‘b’_n‘n’_rep‘r’_
07Mar2019_BC.dta", clear

gen pat_id = _n

reshape long wom_fu_ wom_ch_ wom_ trteffect, ///
i(pat_id) j(timepoint)

xtset pat_id timepoint

xtgee wom_fu_ i.treat wom_bsl, family(gaussian) ///
link(identity) corr(exchangeable)
```

Mixed effects code:

```
use "sim_final_time‘t’_pat‘p’_trt‘b’_n‘n’_rep‘r’_
07Mar2019_BC.dta", clear

gen pat_id = _n

reshape long wom_fu_ wom_ch_ wom_ trteffect, ///
i(pat_id) j(timepoint)

xtmixed wom_fu_ i.treat i.timepoint wom_bsl || ///
pat_id: timepoint, residuals(exchangeable) mle ///
stddeviations iter(100)
```

E.3 Example of Stata code and underlying formulas to calculate performance measures

```
forvalues t = 1/4 {
forvalues p = 1/3 {
foreach b in 0 4 8 12 {
foreach n in 100 200 400 600 800 {
foreach level in 80 90 95 97.5 99 {

summarize true if t==‘t’ & p==‘p’ & b==‘b’ & n==‘n’
local truevalue = r(mean)
simsum coeff_neg if t==‘t’ & p==‘p’ & b==‘b’ & n==‘n’, ///
se(se_coeff) true(‘truevalue’) level(‘level’) ///
mcse transpose ///
saving("Simsum_output_GEE_t‘t’_p‘p’_b‘b’_n‘n’_
level‘level’_29Apr2019_BC.dta")

}
}
}
}
}
```

Note: Variable ‘true’ was imported from an Excel file containing the true values for each scenario

E.4 Power for 80% and 90% confidence intervals

Table E.1: Power (for 80% confidence interval): Pattern 1

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Power	62.7	97.1	99.9	63.9	96.8	100.0	61.9	97.1	100.0	62.7	97.2	99.9
	(primary)	MCSE	(1.21)	(0.42)	(0.09)	(1.20)	(0.44)	(0.00)	(1.21)	(0.42)	(0.00)	(1.21)	(0.41)	(0.06)
	GEE	Power	53.8	92.8	99.7	46.1	81.7	97.7	43.2	78.9	96.8	41.8	79.3	96.6
		MCSE	(1.25)	(0.65)	(0.14)	(1.25)	(0.97)	(0.38)	(1.24)	(1.02)	(0.44)	(1.23)	(1.01)	(0.45)
	Mixed	Power	51.9	91.7	99.7	45.5	79.2	96.8	41.4	75.4	93.5	39.9	74.8	90.4
	effects	MCSE	(1.32)	(0.72)	(0.16)	(1.41)	(1.16)	(0.50)	(1.52)	(1.32)	(0.77)	(1.58)	(1.40)	(0.97)
200	ANCOVA	Power	82.9	99.9	100.0	84.8	100.0	100.0	82.4	99.9	100.0	83.2	100.0	100.0
	(primary)	MCSE	(0.94)	(0.06)	(0.00)	(0.90)	(0.00)	(0.00)	(0.95)	(0.06)	(0.00)	(0.94)	(0.00)	(0.00)
	GEE	Power	74.6	99.5	100.0	61.1	96.9	99.9	60.3	95.9	100.0	59.8	94.8	99.8
		MCSE	(1.09)	(0.18)	(0.00)	(1.22)	(0.43)	(0.09)	(1.22)	(0.49)	(0.00)	(1.23)	(0.55)	(0.11)
	Mixed	Power	73.7	99.3	100.0	60.8	95.9	99.8	58.6	95.4	100.0	58.9	92.4	99.6
	effects	MCSE	(1.22)	(0.23)	(0.00)	(1.47)	(0.59)	(0.13)	(1.58)	(0.68)	(0.00)	(1.66)	(0.91)	(0.22)
400	ANCOVA	Power	97.3	100.0	100.0	97.6	100.0	100.0	97.6	100.0	100.0	97.4	100.0	100.0
	(primary)	MCSE	(0.41)	(0.00)	(0.00)	(0.38)	(0.00)	(0.00)	(0.39)	(0.00)	(0.00)	(0.40)	(0.00)	(0.00)
	GEE	Power	91.9	100.0	100.0	82.5	99.9	100.0	81.1	99.8	100.0	80.5	99.7	100.0
		MCSE	(0.68)	(0.00)	(0.00)	(0.95)	(0.06)	(0.00)	(0.98)	(0.13)	(0.00)	(0.99)	(0.14)	(0.00)
	Mixed	Power	91.3	100.0	100.0	82.6	99.8	100.0	79.8	99.7	100.0	78.5	99.4	100.0
	effects	MCSE	(0.83)	(0.00)	(0.00)	(1.20)	(0.14)	(0.00)	(1.34)	(0.20)	(0.00)	(1.47)	(0.28)	(0.00)

Table E.1: Power (for 80% confidence interval): Pattern 1

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Power	99.1	100.0	100.0	99.9	100.0	100.0	99.6	100.0	100.0	99.7	100.0	100.0
	(primary)	MCSE	(0.23)	(0.00)	(0.00)	(0.09)	(0.00)	(0.00)	(0.15)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)
	GEE	Power	97.9	100.0	100.0	93.6	100.0	100.0	91.5	100.0	100.0	90.1	100.0	100.0
		MCSE	(0.35)	(0.00)	(0.00)	(0.61)	(0.00)	(0.00)	(0.70)	(0.00)	(0.00)	(0.75)	(0.00)	(0.00)
	Mixed	Power	98.3	100.0	100.0	93.2	100.0	100.0	90.4	100.0	100.0	89.0	100.0	100.0
	effects	MCSE	(0.40)	(0.00)	(0.00)	(0.87)	(0.00)	(0.00)	(1.04)	(0.00)	(0.00)	(1.19)	(0.00)	(0.00)
800	ANCOVA	Power	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	(primary)	MCSE	(0.00)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	GEE	Power	99.8	100.0	100.0	97.3	100.0	100.0	95.2	100.0	100.0	95.8	100.0	100.0
		MCSE	(0.13)	(0.00)	(0.00)	(0.41)	(0.00)	(0.00)	(0.54)	(0.00)	(0.00)	(0.50)	(0.00)	(0.00)
	Mixed	Power	99.7	100.0	100.0	97.3	100.0	100.0	94.9	100.0	100.0	95.3	100.0	100.0
	effects	MCSE	(0.18)	(0.00)	(0.00)	(0.60)	(0.00)	(0.00)	(0.84)	(0.00)	(0.00)	(0.87)	(0.00)	(0.00)

Table E.2: Power (for 90% confidence interval): Pattern 1

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Power	47.8	93.8	99.8	49.2	92.9	99.9	48.0	93.9	99.9	49.9	94.4	99.9
	(primary)	MCSE	(1.25)	(0.60)	(0.13)	(1.25)	(0.64)	(0.06)	(1.25)	(0.60)	(0.09)	(1.25)	(0.57)	(0.06)
	GEE	Power	39.1	86.2	98.8	31.9	71.7	95.1	29.2	68.1	93.1	30.5	68.3	92.7
		MCSE	(1.22)	(0.86)	(0.27)	(1.17)	(1.13)	(0.54)	(1.14)	(1.17)	(0.64)	(1.15)	(1.16)	(0.65)
	Mixed	Power	38.0	84.8	98.4	30.3	68.9	92.2	28.2	64.0	87.8	28.5	63.2	84.7
	effects	MCSE	(1.29)	(0.94)	(0.33)	(1.30)	(1.32)	(0.77)	(1.39)	(1.47)	(1.02)	(1.45)	(1.55)	(1.19)
200	ANCOVA	Power	73.7	99.6	100.0	73.5	99.9	100.0	72.3	99.9	100.0	73.3	100.0	100.0
	(primary)	MCSE	(1.10)	(0.15)	(0.00)	(1.10)	(0.06)	(0.00)	(1.12)	(0.09)	(0.00)	(1.11)	(0.00)	(0.00)
	GEE	Power	61.1	98.8	99.9	47.9	92.9	99.8	44.8	91.0	99.9	47.3	89.6	99.6
		MCSE	(1.22)	(0.28)	(0.06)	(1.25)	(0.64)	(0.13)	(1.24)	(0.71)	(0.06)	(1.25)	(0.76)	(0.17)
	Mixed	Power	61.1	98.5	99.9	47.9	91.2	99.6	42.6	89.6	99.6	45.5	86.1	98.4
	effects	MCSE	(1.35)	(0.34)	(0.08)	(1.50)	(0.84)	(0.20)	(1.58)	(0.99)	(0.22)	(1.68)	(1.18)	(0.45)
400	ANCOVA	Power	93.4	100.0	100.0	94.3	100.0	100.0	94.6	100.0	100.0	94.6	100.0	100.0
	(primary)	MCSE	(0.62)	(0.00)	(0.00)	(0.58)	(0.00)	(0.00)	(0.56)	(0.00)	(0.00)	(0.57)	(0.00)	(0.00)
	GEE	Power	85.6	99.9	100.0	72.9	99.8	100.0	70.4	99.4	100.0	67.6	99.1	100.0
		MCSE	(0.88)	(0.06)	(0.00)	(1.11)	(0.13)	(0.00)	(1.14)	(0.20)	(0.00)	(1.17)	(0.24)	(0.00)
	Mixed	Power	84.7	99.9	100.0	72.4	99.4	100.0	68.8	99.0	100.0	66.2	99.0	100.0
	effects	MCSE	(1.06)	(0.09)	(0.00)	(1.42)	(0.24)	(0.00)	(1.55)	(0.34)	(0.00)	(1.70)	(0.37)	(0.00)

Table E.2: Power (for 90% confidence interval): Pattern 1

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Power	98.2	100.0	100.0	99.1	100.0	100.0	99.0	100.0	100.0	99.1	100.0	100.0
	(primary)	MCSE	(0.33)	(0.00)	(0.00)	(0.24)	(0.00)	(0.00)	(0.25)	(0.00)	(0.00)	(0.24)	(0.00)	(0.00)
	GEE	Power	95.2	100.0	100.0	86.1	100.0	100.0	84.2	100.0	100.0	81.9	100.0	100.0
		MCSE	(0.54)	(0.00)	(0.00)	(0.87)	(0.00)	(0.00)	(0.91)	(0.00)	(0.00)	(0.96)	(0.00)	(0.00)
	Mixed	Power	95.3	100.0	100.0	85.0	100.0	100.0	82.7	100.0	100.0	80.4	99.8	100.0
	effects	MCSE	(0.65)	(0.00)	(0.00)	(1.23)	(0.00)	(0.00)	(1.34)	(0.00)	(0.00)	(1.52)	(0.16)	(0.00)
800	ANCOVA	Power	99.7	100.0	100.0	99.9	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0
	(primary)	MCSE	(0.14)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)
	GEE	Power	98.7	100.0	100.0	94.3	100.0	100.0	91.3	100.0	100.0	90.9	100.0	100.0
		MCSE	(0.28)	(0.00)	(0.00)	(0.58)	(0.00)	(0.00)	(0.71)	(0.00)	(0.00)	(0.72)	(0.00)	(0.00)
	Mixed	Power	98.8	100.0	100.0	94.2	100.0	100.0	91.2	100.0	100.0	90.7	100.0	100.0
	effects	MCSE	(0.35)	(0.00)	(0.00)	(0.86)	(0.00)	(0.00)	(1.08)	(0.00)	(0.00)	(1.19)	(0.00)	(0.00)

Table E.3: Power (for 80% confidence interval): Pattern 2

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Power	62.9	95.9	99.9	60.9	96.9	99.9	61.6	96.9	100.0	61.4	96.1	100.0
	(primary)	MCSE	(1.21)	(0.49)	(0.09)	(1.22)	(0.43)	(0.06)	(1.22)	(0.43)	(0.00)	(1.22)	(0.49)	(0.00)
	GEE	Power	71.8	99.0	100.0	61.7	97.4	99.9	66.6	98.8	100.0	71.1	98.9	100.0
		MCSE	(1.13)	(0.25)	(0.00)	(1.22)	(0.40)	(0.06)	(1.18)	(0.28)	(0.00)	(1.13)	(0.26)	(0.00)
	Mixed	Power	71.9	99.0	100.0	61.0	96.8	99.9	64.7	98.5	100.0	71.9	99.0	100.0
	effects	MCSE	(1.19)	(0.26)	(0.00)	(1.39)	(0.51)	(0.08)	(1.49)	(0.37)	(0.00)	(1.45)	(0.32)	(0.00)
200	ANCOVA	Power	82.0	99.9	100.0	84.3	99.9	100.0	84.0	100.0	100.0	82.2	99.8	100.0
	(primary)	MCSE	(0.96)	(0.09)	(0.00)	(0.91)	(0.06)	(0.00)	(0.92)	(0.00)	(0.00)	(0.96)	(0.13)	(0.00)
	GEE	Power	89.0	100.0	100.0	84.7	100.0	100.0	88.9	100.0	100.0	90.7	99.9	100.0
		MCSE	(0.78)	(0.00)	(0.00)	(0.90)	(0.00)	(0.00)	(0.79)	(0.00)	(0.00)	(0.73)	(0.06)	(0.00)
	Mixed	Power	88.4	100.0	100.0	84.0	100.0	100.0	88.0	100.0	100.0	90.4	100.0	100.0
	effects	MCSE	(0.88)	(0.00)	(0.00)	(1.08)	(0.00)	(0.00)	(1.03)	(0.00)	(0.00)	(0.98)	(0.00)	(0.00)
400	ANCOVA	Power	97.5	100.0	100.0	96.1	100.0	100.0	96.8	100.0	100.0	95.6	100.0	100.0
	(primary)	MCSE	(0.39)	(0.00)	(0.00)	(0.49)	(0.00)	(0.00)	(0.44)	(0.00)	(0.00)	(0.52)	(0.00)	(0.00)
	GEE	Power	99.6	100.0	100.0	96.6	100.0	100.0	98.5	100.0	100.0	98.9	100.0	100.0
		MCSE	(0.17)	(0.00)	(0.00)	(0.45)	(0.00)	(0.00)	(0.30)	(0.00)	(0.00)	(0.26)	(0.00)	(0.00)
	Mixed	Power	99.8	100.0	100.0	96.6	100.0	100.0	98.6	100.0	100.0	98.9	100.0	100.0
	effects	MCSE	(0.12)	(0.00)	(0.00)	(0.58)	(0.00)	(0.00)	(0.40)	(0.00)	(0.00)	(0.36)	(0.00)	(0.00)

Table E.3: Power (for 80% confidence interval): Pattern 2

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Power	99.8	100.0	100.0	99.7	100.0	100.0	99.6	100.0	100.0	99.6	100.0	100.0
	(primary)	MCSE	(0.11)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)	(0.17)	(0.00)	(0.00)	(0.15)	(0.00)	(0.00)
	GEE	Power	99.9	100.0	100.0	99.9	100.0	100.0	99.7	100.0	100.0	99.9	100.0	100.0
		MCSE	(0.06)	(0.00)	(0.00)	(0.09)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)
	Mixed	Power	99.9	100.0	100.0	100.0	100.0	100.0	99.5	100.0	100.0	100.0	100.0	100.0
	effects	MCSE	(0.10)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.27)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
800	ANCOVA	Power	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.8	100.0	100.0
	(primary)	MCSE	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.11)	(0.00)	(0.00)
	GEE	Power	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		MCSE	(0.00)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Mixed	Power	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	effects	MCSE	(0.00)	(0.00)	(0.00)	(0.13)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Table E.4: Power (for 90% confidence interval): Pattern 2

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Power	47.5	93.1	99.8	46.9	93.6	99.8	47.1	93.1	100.0	46.5	91.7	100.0
	(primary)	MCSE	(1.25)	(0.63)	(0.13)	(1.25)	(0.61)	(0.13)	(1.25)	(0.63)	(0.00)	(1.25)	(0.69)	(0.00)
	GEE	Power	57.3	97.4	100.0	48.8	94.3	99.8	53.9	96.6	100.0	57.9	97.1	100.0
		MCSE	(1.24)	(0.40)	(0.00)	(1.25)	(0.58)	(0.13)	(1.25)	(0.45)	(0.00)	(1.23)	(0.42)	(0.00)
	Mixed	Power	57.9	97.5	100.0	47.7	93.2	99.7	52.5	96.5	100.0	58.9	96.9	100.0
	effects	MCSE	(1.30)	(0.41)	(0.00)	(1.42)	(0.72)	(0.17)	(1.55)	(0.57)	(0.00)	(1.59)	(0.55)	(0.00)
200	ANCOVA	Power	72.8	99.6	100.0	73.3	99.8	100.0	73.4	99.9	100.0	71.3	99.5	100.0
	(primary)	MCSE	(1.11)	(0.15)	(0.00)	(1.11)	(0.13)	(0.00)	(1.10)	(0.06)	(0.00)	(1.13)	(0.18)	(0.00)
	GEE	Power	81.8	99.9	100.0	74.3	99.8	100.0	80.1	99.9	100.0	83.9	99.9	100.0
		MCSE	(0.96)	(0.06)	(0.00)	(1.09)	(0.13)	(0.00)	(1.00)	(0.06)	(0.00)	(0.92)	(0.06)	(0.00)
	Mixed	Power	80.9	99.9	100.0	73.5	99.7	100.0	78.7	99.9	100.0	84.7	100.0	100.0
	effects	MCSE	(1.08)	(0.08)	(0.00)	(1.30)	(0.18)	(0.00)	(1.30)	(0.10)	(0.00)	(1.20)	(0.00)	(0.00)
400	ANCOVA	Power	94.4	100.0	100.0	91.9	100.0	100.0	93.0	100.0	100.0	92.1	100.0	100.0
	(primary)	MCSE	(0.58)	(0.00)	(0.00)	(0.68)	(0.00)	(0.00)	(0.64)	(0.00)	(0.00)	(0.67)	(0.00)	(0.00)
	GEE	Power	98.3	100.0	100.0	92.9	100.0	100.0	96.6	100.0	100.0	97.1	100.0	100.0
		MCSE	(0.33)	(0.00)	(0.00)	(0.64)	(0.00)	(0.00)	(0.46)	(0.00)	(0.00)	(0.42)	(0.00)	(0.00)
	Mixed	Power	98.2	100.0	100.0	92.3	100.0	100.0	96.4	100.0	100.0	96.5	100.0	100.0
	effects	MCSE	(0.39)	(0.00)	(0.00)	(0.85)	(0.00)	(0.00)	(0.63)	(0.00)	(0.00)	(0.64)	(0.00)	(0.00)

Table E.4: Power (for 90% confidence interval): Pattern 2

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Power	98.8	100.0	100.0	98.8	100.0	100.0	98.6	100.0	100.0	98.9	100.0	100.0
	(primary)	MCSE	(0.27)	(0.00)	(0.00)	(0.28)	(0.00)	(0.00)	(0.30)	(0.00)	(0.00)	(0.26)	(0.00)	(0.00)
	GEE	Power	99.8	100.0	100.0	99.3	100.0	100.0	99.4	100.0	100.0	99.5	100.0	100.0
		MCSE	(0.11)	(0.00)	(0.00)	(0.21)	(0.00)	(0.00)	(0.19)	(0.00)	(0.00)	(0.18)	(0.00)	(0.00)
	Mixed	Power	99.6	100.0	100.0	99.2	100.0	100.0	99.3	100.0	100.0	99.6	100.0	100.0
	effects	MCSE	(0.20)	(0.00)	(0.00)	(0.29)	(0.00)	(0.00)	(0.30)	(0.00)	(0.00)	(0.25)	(0.00)	(0.00)
800	ANCOVA	Power	99.8	100.0	100.0	99.9	100.0	100.0	99.8	100.0	100.0	99.6	100.0	100.0
	(primary)	MCSE	(0.11)	(0.00)	(0.00)	(0.09)	(0.00)	(0.00)	(0.11)	(0.00)	(0.00)	(0.15)	(0.00)	(0.00)
	GEE	Power	100.0	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		MCSE	(0.00)	(0.00)	(0.00)	(0.11)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	Mixed	Power	100.0	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	effects	MCSE	(0.00)	(0.00)	(0.00)	(0.18)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Table E.5: Power (for 80% confidence interval): Pattern 3

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Power	64.4	97.1	100.0	60.1	96.6	99.8	62.9	95.9	100.0	63.9	97.3	100.0
	(primary)	MCSE	(1.20)	(0.42)	(0.00)	(1.22)	(0.46)	(0.13)	(1.21)	(0.50)	(0.00)	(1.20)	(0.40)	(0.00)
	GEE	Power	38.8	69.9	91.4	51.8	88.9	99.1	38.3	72.3	93.9	35.4	59.9	86.1
		MCSE	(1.22)	(1.15)	(0.70)	(1.25)	(0.79)	(0.23)	(1.22)	(1.12)	(0.60)	(1.20)	(1.23)	(0.86)
	Mixed	Power	42.0	75.5	94.0	52.7	92.8	99.3	41.9	77.4	96.2	39.5	68.7	94.7
	effects	MCSE	(1.31)	(1.13)	(0.62)	(1.43)	(0.75)	(0.23)	(1.53)	(1.32)	(0.61)	(1.60)	(1.51)	(0.75)
200	ANCOVA	Power	83.9	99.9	100.0	84.4	99.9	100.0	83.0	99.8	100.0	83.3	99.9	100.0
	(primary)	MCSE	(0.92)	(0.09)	(0.00)	(0.91)	(0.06)	(0.00)	(0.94)	(0.11)	(0.00)	(0.93)	(0.06)	(0.00)
	GEE	Power	50.1	88.4	98.9	71.4	99.4	100.0	52.8	93.6	99.6	45.7	80.1	97.1
		MCSE	(1.25)	(0.80)	(0.26)	(1.13)	(0.19)	(0.00)	(1.25)	(0.61)	(0.17)	(1.25)	(1.00)	(0.42)
	Mixed	Power	54.3	90.8	99.3	73.5	99.5	100.0	55.0	95.8	100.0	48.5	85.4	99.6
	effects	MCSE	(1.36)	(0.79)	(0.22)	(1.33)	(0.21)	(0.00)	(1.60)	(0.65)	(0.00)	(1.68)	(1.25)	(0.23)
400	ANCOVA	Power	96.9	100.0	100.0	97.1	100.0	100.0	97.9	100.0	100.0	97.8	100.0	100.0
	(primary)	MCSE	(0.43)	(0.00)	(0.00)	(0.42)	(0.00)	(0.00)	(0.35)	(0.00)	(0.00)	(0.37)	(0.00)	(0.00)
	GEE	Power	69.6	99.1	100.0	90.7	100.0	100.0	74.8	99.4	100.0	60.3	95.7	99.8
		MCSE	(1.15)	(0.23)	(0.00)	(0.73)	(0.00)	(0.00)	(1.08)	(0.19)	(0.00)	(1.22)	(0.51)	(0.11)
	Mixed	Power	72.4	99.4	100.0	91.7	100.0	100.0	78.0	99.6	100.0	64.7	97.9	100.0
	effects	MCSE	(1.32)	(0.23)	(0.00)	(0.88)	(0.00)	(0.00)	(1.40)	(0.21)	(0.00)	(1.74)	(0.56)	(0.00)

Table E.5: Power (for 80% confidence interval): Pattern 3

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Power	99.5	100.0	100.0	99.6	100.0	100.0	99.6	100.0	100.0	99.7	100.0	100.0
	(primary)	MCSE	(0.18)	(0.00)	(0.00)	(0.17)	(0.00)	(0.00)	(0.17)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)
	GEE	Power	82.5	99.6	100.0	97.2	100.0	100.0	87.6	100.0	100.0	76.2	99.4	100.0
		MCSE	(0.95)	(0.17)	(0.00)	(0.41)	(0.00)	(0.00)	(0.82)	(0.00)	(0.00)	(1.06)	(0.20)	(0.00)
	Mixed	Power	84.0	99.5	100.0	97.7	100.0	100.0	90.5	100.0	100.0	79.6	99.6	100.0
	effects	MCSE	(1.14)	(0.23)	(0.00)	(0.52)	(0.00)	(0.00)	(1.03)	(0.00)	(0.00)	(1.53)	(0.26)	(0.00)
800	ANCOVA	Power	100.0	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0
	(primary)	MCSE	(0.00)	(0.00)	(0.00)	(0.09)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.09)	(0.00)	(0.00)
	GEE	Power	89.9	100.0	100.0	98.7	100.0	100.0	93.8	100.0	100.0	84.0	99.9	100.0
		MCSE	(0.75)	(0.00)	(0.00)	(0.28)	(0.00)	(0.00)	(0.60)	(0.00)	(0.00)	(0.92)	(0.06)	(0.00)
	Mixed	Power	90.3	100.0	100.0	98.5	100.0	100.0	95.2	100.0	100.0	86.8	100.0	100.0
	effects	MCSE	(0.97)	(0.00)	(0.00)	(0.43)	(0.00)	(0.00)	(0.84)	(0.00)	(0.00)	(1.47)	(0.00)	(0.00)

Table E.6: Power (for 90% confidence interval): Pattern 3

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
100	ANCOVA	Power	50.6	94.3	100.0	47.2	92.9	99.6	47.7	92.4	99.9	49.1	94.3	99.8
	(primary)	MCSE	(1.25)	(0.58)	(0.00)	(1.25)	(0.64)	(0.17)	(1.25)	(0.66)	(0.06)	(1.25)	(0.58)	(0.11)
	GEE	Power	25.3	55.6	86.3	37.5	80.9	98.1	25.1	59.4	88.4	21.8	47.4	75.8
		MCSE	(1.09)	(1.24)	(0.86)	(1.21)	(0.98)	(0.34)	(1.08)	(1.23)	(0.80)	(1.03)	(1.25)	(1.07)
	Mixed	Power	28.1	62.0	89.7	39.2	85.7	98.9	28.0	65.3	93.2	24.5	56.5	89.0
	effects	MCSE	(1.19)	(1.27)	(0.80)	(1.40)	(1.01)	(0.30)	(1.40)	(1.50)	(0.81)	(1.41)	(1.62)	(1.05)
200	ANCOVA	Power	72.7	99.8	100.0	73.0	99.9	100.0	72.4	99.8	100.0	73.7	99.8	100.0
	(primary)	MCSE	(1.11)	(0.11)	(0.00)	(1.11)	(0.09)	(0.00)	(1.12)	(0.11)	(0.00)	(1.10)	(0.13)	(0.00)
	GEE	Power	36.3	79.8	97.6	58.3	97.9	100.0	37.6	87.8	99.3	30.9	68.2	93.2
		MCSE	(1.20)	(1.00)	(0.39)	(1.23)	(0.35)	(0.00)	(1.21)	(0.82)	(0.21)	(1.16)	(1.16)	(0.63)
	Mixed	Power	38.8	83.9	98.5	61.5	98.9	100.0	42.4	91.3	99.8	35.3	76.1	98.7
	effects	MCSE	(1.33)	(1.00)	(0.32)	(1.46)	(0.32)	(0.00)	(1.59)	(0.91)	(0.16)	(1.60)	(1.51)	(0.42)
400	ANCOVA	Power	93.8	100.0	100.0	93.8	100.0	100.0	94.3	100.0	100.0	94.8	100.0	100.0
	(primary)	MCSE	(0.60)	(0.00)	(0.00)	(0.60)	(0.00)	(0.00)	(0.58)	(0.00)	(0.00)	(0.56)	(0.00)	(0.00)
	GEE	Power	55.6	97.6	100.0	82.8	99.9	100.0	61.8	98.4	100.0	45.6	92.4	99.6
		MCSE	(1.24)	(0.38)	(0.00)	(0.94)	(0.06)	(0.00)	(1.22)	(0.31)	(0.00)	(1.25)	(0.66)	(0.15)
	Mixed	Power	58.6	98.6	100.0	83.4	99.9	100.0	65.3	99.3	100.0	50.2	94.5	100.0
	effects	MCSE	(1.45)	(0.35)	(0.00)	(1.18)	(0.10)	(0.00)	(1.61)	(0.29)	(0.00)	(1.81)	(0.89)	(0.00)

Table E.6: Power (for 90% confidence interval): Pattern 3

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
n	**Method**	$\beta*$**:**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
600	ANCOVA	Power	98.8	100.0	100.0	98.4	100.0	100.0	99.0	100.0	100.0	98.9	100.0	100.0
	(primary)	MCSE	(0.27)	(0.00)	(0.00)	(0.31)	(0.00)	(0.00)	(0.25)	(0.00)	(0.00)	(0.26)	(0.00)	(0.00)
	GEE	Power	71.8	99.3	100.0	93.9	100.0	100.0	79.2	99.9	100.0	63.3	98.3	100.0
		MCSE	(1.13)	(0.21)	(0.00)	(0.60)	(0.00)	(0.00)	(1.01)	(0.06)	(0.00)	(1.21)	(0.33)	(0.00)
	Mixed	Power	73.5	99.3	100.0	94.7	100.0	100.0	83.3	100.0	100.0	67.7	99.1	100.0
	effects	MCSE	(1.37)	(0.26)	(0.00)	(0.77)	(0.00)	(0.00)	(1.32)	(0.00)	(0.00)	(1.78)	(0.41)	(0.00)
800	ANCOVA	Power	99.7	100.0	100.0	99.8	100.0	100.0	99.8	100.0	100.0	99.8	100.0	100.0
	(primary)	MCSE	(0.14)	(0.00)	(0.00)	(0.13)	(0.00)	(0.00)	(0.11)	(0.00)	(0.00)	(0.13)	(0.00)	(0.00)
	GEE	Power	81.9	100.0	100.0	97.1	100.0	100.0	87.4	100.0	100.0	73.4	99.8	100.0
		MCSE	(0.96)	(0.00)	(0.00)	(0.42)	(0.00)	(0.00)	(0.83)	(0.00)	(0.00)	(1.10)	(0.13)	(0.00)
	Mixed	Power	83.3	100.0	100.0	97.5	100.0	100.0	90.3	100.0	100.0	76.9	99.8	100.0
	effects	MCSE	(1.22)	(0.00)	(0.00)	(0.57)	(0.00)	(0.00)	(1.17)	(0.00)	(0.00)	(1.83)	(0.23)	(0.00)

E.5 Type I error for 90%, 97.5% and 99% confidence intervals

Table E.7: Type I error for 90%, 97.5% and 99% confidence intervals: Pattern 1

			90% confidence interval Number of time points:				**97.5% confidence interval Number of time points:**				**99% confidence interval Number of time points:**			
n	**Method**		**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
100	ANCOVA	Type I	12.8	10.1	11.0	10.9	2.9	3.0	2.9	3.4	1.5	0.9	1.4	1.6
	(primary)	MCSE	(0.83)	(0.75)	(0.78)	(0.78)	(0.42)	(0.43)	(0.42)	(0.45)	(0.30)	(0.23)	(0.30)	(0.31)
	GEE	Type I	11.6	11.2	12.4	10.8	3.1	3.6	3.4	3.6	1.8	1.4	1.4	1.9
		MCSE	(0.80)	(0.79)	(0.83)	(0.77)	(0.43)	(0.47)	(0.45)	(0.46)	(0.33)	(0.29)	(0.29)	(0.34)
	Mixed	Type I	11.6	11.4	11.6	11.4	3.5	3.4	3.1	3.9	1.8	1.4	1.4	2.3
	effects	MCSE	(0.84)	(0.92)	(0.97)	(1.02)	(0.48)	(0.52)	(0.53)	(0.62)	(0.35)	(0.34)	(0.35)	(0.48)
	ANCOVA	Type I	18.9	23.7	30.8	35.1	4.8	7.3	9.1	10.4	2.3	2.7	4.6	4.8
	(all)	MCSE	(0.98)	(1.06)	(1.15)	(1.19)	(0.53)	(0.65)	(0.72)	(0.76)	(0.38)	(0.40)	(0.53)	(0.53)
200	ANCOVA	Type I	10.9	10.3	10.4	10.1	2.6	2.1	3.8	2.4	0.8	0.9	1.9	1.0
	(primary)	MCSE	(0.78)	(0.76)	(0.76)	(0.75)	(0.40)	(0.36)	(0.48)	(0.38)	(0.22)	(0.23)	(0.34)	(0.25)
	GEE	Type I	10.8	10.9	9.4	11.3	2.9	2.9	2.8	3.4	1.4	1.0	1.3	1.4
		MCSE	(0.77)	(0.78)	(0.73)	(0.79)	(0.42)	(0.42)	(0.41)	(0.45)	(0.30)	(0.25)	(0.28)	(0.30)
	Mixed	Type I	10.2	10.9	8.9	11.8	2.9	2.7	2.5	3.6	1.5	0.9	1.3	1.5
	effects	MCSE	(0.83)	(0.93)	(0.89)	(1.09)	(0.46)	(0.48)	(0.48)	(0.63)	(0.34)	(0.28)	(0.35)	(0.41)
	ANCOVA	Type I	17.4	24.0	28.9	34.8	4.5	6.2	8.9	11.3	2.0	2.6	4.1	4.9
	(all)	MCSE	(0.95)	(1.07)	(1.13)	(1.19)	(0.52)	(0.60)	(0.71)	(0.79)	(0.35)	(0.40)	(0.50)	(0.54)

Table E.7: Type I error for 90%, 97.5% and 99% confidence intervals: Pattern 1

			90% confidence interval Number of time points:				**97.5% confidence interval Number of time points:**				**99% confidence interval Number of time points:**			
n	**Method**		**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
400	ANCOVA	Type I	11.0	10.6	9.8	9.2	2.6	2.4	3.3	2.1	1.0	1.1	1.6	1.0
	(primary)	MCSE	(0.78)	(0.77)	(0.74)	(0.72)	(0.40)	(0.39)	(0.44)	(0.36)	(0.25)	(0.26)	(0.32)	(0.25)
	GEE	Type I	9.5	8.9	12.1	9.1	2.6	2.5	3.4	2.4	1.2	1.1	1.7	0.9
		MCSE	(0.73)	(0.71)	(0.81)	(0.72)	(0.40)	(0.39)	(0.45)	(0.38)	(0.27)	(0.26)	(0.32)	(0.24)
	Mixed	Type I	9.2	9.2	12.1	9.6	2.8	2.6	3.9	2.5	1.1	0.9	1.9	1.2
	effects	MCSE	(0.85)	(0.92)	(1.08)	(1.02)	(0.49)	(0.51)	(0.64)	(0.55)	(0.31)	(0.30)	(0.45)	(0.38)
	ANCOVA	Type I	17.5	21.7	30.4	31.1	4.1	5.3	9.3	9.3	2.1	2.6	4.7	4.3
	(all)	MCSE	(0.95)	(1.03)	(1.15)	(1.16)	(0.49)	(0.56)	(0.73)	(0.72)	(0.35)	(0.40)	(0.53)	(0.51)
600	ANCOVA	Type I	8.7	9.5	9.3	10.4	2.7	2.6	2.2	2.1	1.1	1.2	0.9	0.6
	(primary)	MCSE	(0.70)	(0.73)	(0.73)	(0.76)	(0.40)	(0.40)	(0.37)	(0.36)	(0.26)	(0.27)	(0.23)	(0.20)
	GEE	Type I	10.5	10.6	9.3	9.4	2.4	3.4	2.1	2.9	1.3	1.4	0.7	1.3
		MCSE	(0.77)	(0.77)	(0.72)	(0.73)	(0.39)	(0.45)	(0.35)	(0.42)	(0.28)	(0.30)	(0.21)	(0.28)
	Mixed	Type I	10.3	10.8	9.8	9.6	2.3	3.3	2.2	2.0	1.2	1.4	0.4	0.7
	effects	MCSE	(0.94)	(1.05)	(1.09)	(1.10)	(0.46)	(0.60)	(0.54)	(0.52)	(0.34)	(0.39)	(0.23)	(0.31)
	ANCOVA	Type I	16.8	22.9	27.6	33.1	4.6	6.8	7.8	9.3	1.9	3.1	3.4	4.0
	(all)	MCSE	(0.94)	(1.05)	(1.12)	(1.18)	(0.53)	(0.63)	(0.67)	(0.73)	(0.34)	(0.43)	(0.45)	(0.49)

Table E.7: Type I error for 90%, 97.5% and 99% confidence intervals: Pattern 1

n	Method		90% confidence interval Number of time points: 2	3	4	5	97.5% confidence interval Number of time points: 2	3	4	5	99% confidence interval Number of time points: 2	3	4	5
800	ANCOVA	Type I	10.4	10.3	9.7	10.1	2.7	2.1	1.6	2.9	1.4	1.0	0.6	0.6
	(primary)	MCSE	(0.76)	(0.76)	(0.74)	(0.75)	(0.40)	(0.36)	(0.32)	(0.42)	(0.30)	(0.25)	(0.20)	(0.20)
	GEE	Type I	9.4	9.8	8.9	10.3	2.3	2.6	2.2	2.9	1.0	0.6	1.0	1.0
		MCSE	(0.73)	(0.74)	(0.71)	(0.76)	(0.38)	(0.40)	(0.37)	(0.42)	(0.25)	(0.20)	(0.25)	(0.25)
	Mixed	Type I	8.5	10.0	9.6	10.8	1.5	2.8	2.1	3.5	0.6	0.7	1.4	1.6
	effects	MCSE	(0.90)	(1.05)	(1.14)	(1.31)	(0.39)	(0.58)	(0.56)	(0.78)	(0.26)	(0.30)	(0.45)	(0.53)
	ANCOVA	Type I	18.1	22.8	28.8	33.6	4.4	6.1	6.8	10.3	2.0	2.9	2.9	4.0
	(all)	MCSE	(0.96)	(1.05)	(1.13)	(1.18)	(0.52)	(0.60)	(0.63)	(0.76)	(0.35)	(0.42)	(0.42)	(0.49)

Table E.8: Type I error for 90%, 97.5% and 99% confidence intervals: Pattern 2

n	Method		90% confidence interval Number of time points: 2	3	4	5	97.5% confidence interval Number of time points: 2	3	4	5	99% confidence interval Number of time points: 2	3	4	5
100	ANCOVA	Type I	11.5	11.5	12.1	10.4	2.7	3.6	3.2	2.8	1.1	1.9	1.1	1.4
	(primary)	MCSE	(0.80)	(0.80)	(0.81)	(0.76)	(0.40)	(0.46)	(0.44)	(0.41)	(0.26)	(0.34)	(0.26)	(0.29)
	GEE	Type I	12.4	12.3	12.4	10.7	3.4	4.1	3.8	2.9	1.6	1.8	1.4	1.3
		MCSE	(0.82)	(0.82)	(0.83)	(0.77)	(0.46)	(0.50)	(0.48)	(0.42)	(0.31)	(0.33)	(0.29)	(0.28)
	Mixed	Type I	12.3	12.3	12.1	10.0	3.7	4.0	4.0	2.4	1.7	1.8	1.3	1.0
	effects	MCSE	(0.86)	(0.94)	(1.00)	(0.97)	(0.50)	(0.56)	(0.60)	(0.39)	(0.34)	(0.38)	(0.35)	(0.31)
	ANCOVA	Type I	19.1	24.6	30.8	32.0	5.3	8.8	9.6	9.4	2.2	3.9	4.3	4.1
	(all)	MCSE	(0.98)	(1.08)	(1.15)	(1.17)	(0.56)	(0.71)	(0.74)	(0.73)	(0.37)	(0.48)	(0.50)	(0.50)
200	ANCOVA	Type I	10.1	10.6	10.9	10.7	2.8	3.4	2.6	2.4	1.2	1.4	1.2	1.1
	(primary)	MCSE	(0.75)	(0.77)	(0.78)	(0.77)	(0.41)	(0.46)	(0.40)	(0.39)	(0.27)	(0.30)	(0.27)	(0.26)
	GEE	Type I	12.0	12.0	10.9	10.0	3.0	3.5	3.5	3.4	1.2	1.7	1.1	1.4
		MCSE	(0.81)	(0.81)	(0.78)	(0.75)	(0.43)	(0.46)	(0.46)	(0.46)	(0.27)	(0.32)	(0.26)	(0.29)
	Mixed	Type I	12.7	11.7	10.4	10.1	3.0	3.1	3.2	3.5	1.1	1.5	1.2	1.5
	effects	MCSE	(0.92)	(0.96)	(0.97)	(1.00)	(0.47)	(0.52)	(0.56)	(0.61)	(0.29)	(0.37)	(0.34)	(0.41)
	ANCOVA	Type I	18.1	24.2	30.0	32.2	5.5	7.3	8.7	9.6	2.4	2.5	3.9	4.4
	(all)	MCSE	(0.96)	(1.07)	(1.15)	(1.17)	(0.57)	(0.65)	(0.70)	(0.73)	(0.39)	(0.39)	(0.48)	(0.52)

Table E.8: Type I error for 90%, 97.5% and 99% confidence intervals: Pattern 2

			90% confidence interval Number of time points:				**97.5% confidence interval Number of time points:**				**99% confidence interval Number of time points:**			
n	**Method**		**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
400	ANCOVA	Type I	10.6	9.8	10.8	10.5	2.8	2.6	2.8	2.9	1.0	1.1	1.1	1.1
	(primary)	MCSE	(0.77)	(0.74)	(0.78)	(0.77)	(0.41)	(0.40)	(0.41)	(0.42)	(0.25)	(0.26)	(0.26)	(0.26)
	GEE	Type I	11.1	10.8	9.9	11.0	2.8	2.9	2.9	2.8	1.1	1.0	1.2	0.9
		MCSE	(0.78)	(0.77)	(0.75)	(0.78)	(0.41)	(0.42)	(0.42)	(0.41)	(0.26)	(0.25)	(0.27)	(0.24)
	Mixed	Type I	10.9	10.9	9.6	8.9	2.5	2.8	2.4	2.0	1.0	1.1	0.9	0.7
	effects	MCSE	(0.92)	(0.99)	(0.99)	(0.99)	(0.46)	(0.52)	(0.52)	(0.49)	(0.28)	(0.34)	(0.32)	(0.29)
	ANCOVA	Type I	19.1	23.9	28.3	33.3	4.8	6.9	8.1	11.1	2.0	2.6	3.3	4.4
	(all)	MCSE	(0.98)	(1.07)	(1.13)	(1.18)	(0.53)	(0.63)	(0.68)	(0.79)	(0.35)	(0.40)	(0.44)	(0.51)
600	ANCOVA	Type I	10.3	10.4	10.7	10.9	2.4	2.4	3.3	3.1	1.1	0.6	1.1	1.4
	(primary)	MCSE	(0.76)	(0.76)	(0.77)	(0.78)	(0.39)	(0.38)	(0.45)	(0.43)	(0.26)	(0.19)	(0.26)	(0.29)
	GEE	Type I	10.9	10.8	11.3	10.9	2.3	2.9	2.6	3.0	0.9	1.0	1.0	1.4
		MCSE	(0.78)	(0.77)	(0.79)	(0.78)	(0.37)	(0.42)	(0.40)	(0.43)	(0.23)	(0.25)	(0.25)	(0.30)
	Mixed	Type I	10.0	10.6	9.9	12.3	2.1	3.4	1.9	3.0	0.8	1.1	0.8	1.0
	effects	MCSE	(0.93)	(1.04)	(1.06)	(1.26)	(0.45)	(0.61)	(0.48)	(0.65)	(0.27)	(0.36)	(0.31)	(0.39)
	ANCOVA	Type I	18.2	23.9	28.9	34.2	4.9	7.1	8.5	10.3	2.2	2.3	3.5	4.2
	(all)	MCSE	(0.96)	(1.07)	(1.13)	(1.19)	(0.54)	(0.64)	(0.70)	(0.76)	(0.37)	(0.38)	(0.46)	(0.50)

Table E.8: Type I error for 90%, 97.5% and 99% confidence intervals: Pattern 2

n	Method		90% confidence interval Number of time points: 2	3	4	5	97.5% confidence interval Number of time points: 2	3	4	5	99% confidence interval Number of time points: 2	3	4	5
800	ANCOVA	Type I	10.1	9.9	11.4	10.6	2.6	2.9	3.1	2.8	1.2	1.4	1.4	1.2
	(primary)	MCSE	(0.75)	(0.75)	(0.80)	(0.77)	(0.40)	(0.42)	(0.43)	(0.41)	(0.27)	(0.29)	(0.30)	(0.27)
	GEE	Type I	10.6	10.4	10.9	9.8	2.6	2.3	2.8	2.7	1.1	0.9	1.4	1.4
		MCSE	(0.77)	(0.76)	(0.78)	(0.74)	(0.40)	(0.37)	(0.41)	(0.40)	(0.26)	(0.24)	(0.29)	(0.29)
	Mixed	Type I	10.7	9.9	12.4	8.6	2.5	2.3	3.5	2.5	0.8	1.0	1.8	1.1
	effects	MCSE	(1.00)	(1.07)	(1.28)	(1.19)	(0.51)	(0.53)	(0.71)	(0.66)	(0.30)	(0.36)	(0.52)	(0.44)
	ANCOVA	Type I	18.6	24.6	30.4	34.4	5.3	6.6	9.5	11.2	2.1	2.7	4.5	4.8
	(all)	MCSE	(0.97)	(1.08)	(1.15)	(1.19)	(0.56)	(0.62)	(0.73)	(0.79)	(0.35)	(0.40)	(0.52)	(0.53)

Table E.9: Type I error for 90%, 97.5% and 99% confidence intervals: Pattern 3

			90% confidence interval Number of time points:				**97.5% confidence interval Number of time points:**				**99% confidence interval Number of time points:**			
n	**Method**		**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
100	ANCOVA	Type I	10.4	11.3	11.3	10.6	2.8	3.1	2.9	3.0	1.4	1.1	1.3	0.9
	(primary)	MCSE	(0.76)	(0.79)	(0.79)	(0.77)	(0.41)	(0.43)	(0.42)	(0.43)	(0.30)	(0.26)	(0.28)	(0.24)
	GEE	Type I	11.2	11.1	11.0	10.7	3.6	3.4	3.3	2.2	1.7	1.6	1.6	0.9
		MCSE	(0.79)	(0.79)	(0.78)	(0.77)	(0.46)	(0.46)	(0.45)	(0.37)	(0.32)	(0.31)	(0.32)	(0.24)
	Mixed	Type I	11.8	10.5	11.1	10.2	3.4	3.7	3.5	2.2	1.8	1.3	1.7	0.6
	effects	MCSE	(0.85)	(0.88)	(0.98)	(0.95)	(0.47)	(0.54)	(0.57)	(0.46)	(0.35)	(0.33)	(0.41)	(0.24)
	ANCOVA	Type I	17.1	23.7	29.3	31.8	5.3	7.6	9.1	9.5	2.2	3.2	3.9	3.9
	(all)	MCSE	(0.94)	(1.06)	(1.14)	(1.16)	(0.56)	(0.66)	(0.72)	(0.73)	(0.37)	(0.44)	(0.49)	(0.48)
200	ANCOVA	Type I	10.9	10.2	8.7	11.4	3.0	2.2	1.9	2.7	1.2	1.0	1.2	1.3
	(primary)	MCSE	(0.78)	(0.76)	(0.70)	(0.79)	(0.43)	(0.37)	(0.34)	(0.40)	(0.27)	(0.25)	(0.27)	(0.28)
	GEE	Type I	11.9	10.8	10.7	11.1	3.4	2.8	2.7	2.9	1.1	0.9	1.3	1.4
		MCSE	(0.81)	(0.78)	(0.77)	(0.79)	(0.46)	(0.41)	(0.40)	(0.42)	(0.26)	(0.24)	(0.28)	(0.29)
	Mixed	Type I	11.5	11.3	9.9	10.6	3.2	2.6	2.3	2.7	0.9	0.9	1.3	1.4
	effects	MCSE	(0.87)	(0.94)	(0.95)	(1.01)	(0.48)	(0.47)	(0.48)	(0.53)	(0.26)	(0.28)	(0.36)	(0.38)
	ANCOVA	Type I	18.6	23.3	29.2	34.5	6.3	7.1	8.4	9.3	2.1	2.3	3.6	3.9
	(all)	MCSE	(0.97)	(1.06)	(1.14)	(1.19)	(0.61)	(0.64)	(0.69)	(0.73)	(0.36)	(0.38)	(0.46)	(0.49)

Table E.9: Type I error for 90%, 97.5% and 99% confidence intervals: Pattern 3

			90% confidence interval Number of time points:				**97.5% confidence interval Number of time points:**				**99% confidence interval Number of time points:**			
n	**Method**		**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**	**2**	**3**	**4**	**5**
400	ANCOVA	Type I	10.6	10.1	11.8	10.1	2.9	2.8	2.8	2.8	1.1	0.9	1.3	1.4
	(primary)	MCSE	(0.77)	(0.75)	(0.81)	(0.75)	(0.42)	(0.41)	(0.41)	(0.41)	(0.26)	(0.24)	(0.28)	(0.30)
	GEE	Type I	11.4	10.9	11.4	9.7	3.6	3.2	3.3	2.6	1.5	1.5	1.4	1.1
		MCSE	(0.80)	(0.78)	(0.79)	(0.74)	(0.47)	(0.44)	(0.44)	(0.40)	(0.30)	(0.30)	(0.30)	(0.26)
	Mixed	Type I	11.9	11.0	12.1	9.2	3.6	3.4	3.0	2.3	1.8	1.6	1.4	0.9
	effects	MCSE	(0.97)	(0.99)	(1.08)	(1.02)	(0.56)	(0.58)	(0.56)	(0.53)	(0.40)	(0.40)	(0.39)	(0.33)
	ANCOVA	Type I	18.3	24.3	30.4	34.1	5.7	7.1	8.3	9.5	2.3	3.1	3.4	4.4
	(all)	MCSE	(0.97)	(1.07)	(1.15)	(1.19)	(0.58)	(0.64)	(0.69)	(0.73)	(0.38)	(0.43)	(0.46)	(0.52)
600	ANCOVA	Type I	10.3	10.1	8.4	10.6	2.6	2.5	1.6	1.9	1.2	1.2	0.6	0.9
	(primary)	MCSE	(0.76)	(0.75)	(0.70)	(0.77)	(0.40)	(0.39)	(0.32)	(0.34)	(0.27)	(0.27)	(0.19)	(0.23)
	GEE	Type I	9.6	10.1	9.5	10.3	3.3	2.9	2.6	2.8	1.3	1.4	1.0	1.1
		MCSE	(0.73)	(0.75)	(0.73)	(0.76)	(0.44)	(0.42)	(0.40)	(0.41)	(0.28)	(0.30)	(0.25)	(0.26)
	Mixed	Type I	9.3	10.3	8.8	10.8	3.5	2.8	2.0	2.9	1.5	1.3	0.6	1.0
	effects	MCSE	(0.93)	(1.03)	(1.01)	(1.19)	(0.59)	(0.56)	(0.50)	(0.65)	(0.39)	(0.38)	(0.28)	(0.39)
	ANCOVA	Type I	16.8	23.2	27.8	33.8	4.8	6.9	7.4	9.6	2.1	3.2	2.8	3.9
	(all)	MCSE	(0.94)	(1.05)	(1.12)	(1.18)	(0.54)	(0.63)	(0.65)	(0.73)	(0.36)	(0.44)	(0.41)	(0.49)

Table E.9: Type I error for 90%, 97.5% and 99% confidence intervals: Pattern 3

n	Method		90% confidence interval Number of time points: 2	3	4	5	97.5% confidence interval Number of time points: 2	3	4	5	99% confidence interval Number of time points: 2	3	4	5
800	ANCOVA	Type I	9.6	9.1	9.1	10.4	1.9	2.3	2.4	2.6	0.7	0.9	0.7	0.9
	(primary)	MCSE	(0.73)	(0.72)	(0.72)	(0.76)	(0.34)	(0.38)	(0.39)	(0.40)	(0.21)	(0.23)	(0.21)	(0.23)
	GEE	Type I	9.5	9.1	9.1	9.7	1.8	2.6	2.4	2.5	0.4	0.9	1.0	0.8
		MCSE	(0.73)	(0.72)	(0.72)	(0.74)	(0.33)	(0.40)	(0.39)	(0.39)	(0.15)	(0.24)	(0.25)	(0.22)
	Mixed	Type I	9.2	9.2	8.8	10.1	1.3	3.0	2.1	3.4	0.0	1.5	1.2	1.2
	effects	MCSE	(0.94)	(1.06)	(1.10)	(1.24)	(0.36)	(0.62)	(0.56)	(0.74)	(0.00)	(0.44)	(0.42)	(0.44)
	ANCOVA	Type I	16.2	22.5	29.0	33.2	3.5	6.9	8.3	9.9	1.5	2.3	3.4	4.2
	(all)	MCSE	(0.92)	(1.04)	(1.13)	(1.18)	(0.46)	(0.63)	(0.69)	(0.75)	(0.30)	(0.37)	(0.46)	(0.50)

E.6 Performance measures for ANCOVA at separate time points

Table E.10: ANCOVA method performance at each time point - Power: Pattern 1

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	Power	13.0	32.1	62.6	5.5	12.9	18.9	6.1	7.2	10.8	5.3	6.5	7.3
		MCSE	(0.84)	(1.17)	(1.21)	(0.57)	(0.84)	(0.98)	(0.60)	(0.65)	(0.78)	(0.56)	(0.62)	(0.65)
	200	Power	18.9	60.1	89.6	8.5	18.4	30.4	6.9	9.2	14.6	6.7	7.5	8.2
		MCSE	(0.98)	(1.22)	(0.76)	(0.70)	(0.97)	(1.15)	(0.63)	(0.72)	(0.88)	(0.63)	(0.66)	(0.69)
	400	Power	33.1	86.3	99.1	11.0	29.2	56.1	8.2	13.8	25.4	5.3	8.9	12.2
		MCSE	(1.18)	(0.86)	(0.23)	(0.78)	(1.14)	(1.24)	(0.69)	(0.86)	(1.09)	(0.56)	(0.71)	(0.82)
	600	Power	48.3	96.4	99.9	14.9	43.2	74.9	8.4	18.2	33.8	6.7	11.4	16.9
		MCSE	(1.25)	(0.46)	(0.06)	(0.89)	(1.24)	(1.08)	(0.69)	(0.96)	(1.18)	(0.63)	(0.80)	(0.94)
	800	Power	58.1	98.8	100.0	17.8	53.1	85.6	9.8	24.2	44.6	7.3	13.2	19.4
		MCSE	(1.23)	(0.27)	(0.00)	(0.96)	(1.25)	(0.88)	(0.74)	(1.07)	(1.24)	(0.65)	(0.85)	(0.99)
2	100	Power	34.8	87.7	99.7	13.5	34.5	63.4	7.9	17.2	31.3	7.0	12.8	17.8
		MCSE	(1.19)	(0.82)	(0.14)	(0.85)	(1.19)	(1.20)	(0.67)	(0.94)	(1.16)	(0.64)	(0.84)	(0.96)
	200	Power	61.5	99.1	100.0	18.9	60.4	90.4	11.7	28.4	52.8	8.2	18.0	31.2
		MCSE	(1.22)	(0.23)	(0.00)	(0.98)	(1.22)	(0.74)	(0.80)	(1.13)	(1.25)	(0.69)	(0.96)	(1.16)
	400	Power	88.1	100.0	100.0	33.6	86.4	99.3	17.3	50.6	83.9	11.4	30.1	54.5
		MCSE	(0.81)	(0.00)	(0.00)	(1.18)	(0.86)	(0.22)	(0.95)	(1.25)	(0.92)	(0.80)	(1.15)	(1.25)
	600	Power	96.6	100.0	100.0	48.2	96.8	100.0	23.5	68.1	94.3	14.4	39.1	73.8
		MCSE	(0.45)	(0.00)	(0.00)	(1.25)	(0.44)	(0.00)	(1.06)	(1.17)	(0.58)	(0.88)	(1.22)	(1.10)
	800	Power	99.3	100.0	100.0	59.0	99.2	100.0	28.6	79.2	98.4	17.5	53.6	83.3
		MCSE	(0.22)	(0.00)	(0.00)	(1.23)	(0.22)	(0.00)	(1.13)	(1.01)	(0.31)	(0.95)	(1.25)	(0.93)

Table E.10: ANCOVA method performance at each time point - Power: Pattern 1

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	Power				37.3	88.3	99.6	19.7	54.4	87.1	12.3	33.0	60.8
		MCSE				(1.21)	(0.81)	(0.15)	(0.99)	(1.25)	(0.84)	(0.82)	(1.18)	(1.22)
	200	Power				62.1	99.8	100.0	32.3	82.9	99.6	20.8	57.4	88.9
		MCSE				(1.21)	(0.13)	(0.00)	(1.17)	(0.94)	(0.15)	(1.01)	(1.24)	(0.78)
	400	Power				89.1	100.0	100.0	56.1	98.1	100.0	33.8	87.4	99.4
		MCSE				(0.78)	(0.00)	(0.00)	(1.24)	(0.34)	(0.00)	(1.18)	(0.83)	(0.19)
	600	Power				97.3	100.0	100.0	71.7	99.8	100.0	46.3	96.5	100.0
		MCSE				(0.40)	(0.00)	(0.00)	(1.13)	(0.11)	(0.00)	(1.25)	(0.46)	(0.00)
	800	Power				99.7	100.0	100.0	84.3	100.0	100.0	60.4	99.5	100.0
		MCSE				(0.14)	(0.00)	(0.00)	(0.91)	(0.00)	(0.00)	(1.22)	(0.18)	(0.00)
4	100	Power							36.4	88.9	99.8	21.4	65.3	93.4
		MCSE							(1.20)	(0.79)	(0.11)	(1.03)	(1.19)	(0.62)
	200	Power							60.9	99.5	100.0	42.9	90.3	99.9
		MCSE							(1.22)	(0.18)	(0.00)	(1.24)	(0.74)	(0.06)
	400	Power							90.1	100.0	100.0	66.3	99.5	100.0
		MCSE							(0.75)	(0.00)	(0.00)	(1.18)	(0.18)	(0.00)
	600	Power							97.6	100.0	100.0	82.1	100.0	100.0
		MCSE							(0.38)	(0.00)	(0.00)	(0.96)	(0.00)	(0.00)
	800	Power							99.2	100.0	100.0	90.6	100.0	100.0
		MCSE							(0.22)	(0.00)	(0.00)	(0.73)	(0.00)	(0.00)

Table E.10: ANCOVA method performance at each time point - Power: Pattern 1

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	Power										36.3	88.8	99.9
		MCSE										(1.20)	(0.79)	(0.09)
	200	Power										62.3	99.5	100.0
		MCSE										(1.21)	(0.18)	(0.00)
	400	Power										90.8	100.0	100.0
		MCSE										(0.72)	(0.00)	(0.00)
	600	Power										97.4	100.0	100.0
		MCSE										(0.40)	(0.00)	(0.00)
	800	Power										99.3	100.0	100.0
		MCSE										(0.21)	(0.00)	(0.00)

Table E.11: ANCOVA method performance at each time point - Power: Pattern 2

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	Power	34.9	87.2	99.5	12.5	33.3	60.2	14.3	29.8	58.1	12.1	29.3	58.2
		MCSE	(1.19)	(0.84)	(0.18)	(0.83)	(1.18)	(1.22)	(0.87)	(1.14)	(1.23)	(0.81)	(1.14)	(1.23)
	200	Power	60.1	99.3	100.0	20.1	58.4	88.0	18.6	54.5	85.5	19.6	53.1	83.2
		MCSE	(1.22)	(0.21)	(0.00)	(1.00)	(1.23)	(0.81)	(0.97)	(1.25)	(0.88)	(0.99)	(1.25)	(0.94)
	400	Power	88.9	100.0	100.0	35.3	83.4	99.6	31.9	82.1	99.0	32.0	82.6	99.3
		MCSE	(0.78)	(0.00)	(0.00)	(1.19)	(0.93)	(0.17)	(1.17)	(0.96)	(0.25)	(1.17)	(0.95)	(0.21)
	600	Power	97.5	100.0	100.0	44.3	95.2	100.0	45.6	93.9	100.0	47.6	94.8	100.0
		MCSE	(0.39)	(0.00)	(0.00)	(1.24)	(0.54)	(0.00)	(1.25)	(0.60)	(0.00)	(1.25)	(0.55)	(0.00)
	800	Power	99.5	100.0	100.0	57.9	98.9	100.0	54.9	99.0	100.0	55.3	98.9	100.0
		MCSE	(0.18)	(0.00)	(0.00)	(1.23)	(0.26)	(0.00)	(1.24)	(0.25)	(0.00)	(1.24)	(0.26)	(0.00)
2	100	Power	34.8	87.6	99.6	35.8	87.6	99.8	37.6	86.3	99.4	34.6	87.1	99.6
		MCSE	(1.19)	(0.82)	(0.15)	(1.20)	(0.83)	(0.13)	(1.21)	(0.86)	(0.19)	(1.19)	(0.84)	(0.15)
	200	Power	60.8	99.1	100.0	61.7	99.1	100.0	61.2	99.1	100.0	60.4	99.4	100.0
		MCSE	(1.22)	(0.24)	(0.00)	(1.22)	(0.23)	(0.00)	(1.22)	(0.24)	(0.00)	(1.22)	(0.20)	(0.00)
	400	Power	90.0	100.0	100.0	86.9	100.0	100.0	88.3	100.0	100.0	87.1	100.0	100.0
		MCSE	(0.75)	(0.00)	(0.00)	(0.84)	(0.00)	(0.00)	(0.81)	(0.00)	(0.00)	(0.84)	(0.00)	(0.00)
	600	Power	97.8	100.0	100.0	96.7	100.0	100.0	97.6	100.0	100.0	96.3	100.0	100.0
		MCSE	(0.37)	(0.00)	(0.00)	(0.45)	(0.00)	(0.00)	(0.39)	(0.00)	(0.00)	(0.47)	(0.00)	(0.00)
	800	Power	99.4	100.0	100.0	99.3	100.0	100.0	99.3	100.0	100.0	99.3	100.0	100.0
		MCSE	(0.20)	(0.00)	(0.00)	(0.21)	(0.00)	(0.00)	(0.22)	(0.00)	(0.00)	(0.21)	(0.00)	(0.00)

Table E.11: ANCOVA method performance at each time point - Power: Pattern 2

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	Power				35.1	88.3	99.5	36.3	86.3	99.6	32.6	86.0	99.5
		MCSE				(1.19)	(0.81)	(0.18)	(1.20)	(0.86)	(0.15)	(1.17)	(0.87)	(0.18)
	200	Power				61.1	99.5	100.0	61.3	99.1	100.0	60.2	98.7	100.0
		MCSE				(1.22)	(0.18)	(0.00)	(1.22)	(0.23)	(0.00)	(1.22)	(0.28)	(0.00)
	400	Power				87.3	100.0	100.0	89.0	100.0	100.0	88.0	100.0	100.0
		MCSE				(0.83)	(0.00)	(0.00)	(0.78)	(0.00)	(0.00)	(0.81)	(0.00)	(0.00)
	600	Power				97.5	100.0	100.0	96.6	100.0	100.0	96.8	100.0	100.0
		MCSE				(0.39)	(0.00)	(0.00)	(0.46)	(0.00)	(0.00)	(0.44)	(0.00)	(0.00)
	800	Power				99.6	100.0	100.0	99.2	100.0	100.0	99.1	100.0	100.0
		MCSE				(0.17)	(0.00)	(0.00)	(0.22)	(0.00)	(0.00)	(0.23)	(0.00)	(0.00)
4	100	Power							36.7	88.3	99.9	34.5	86.3	99.5
		MCSE							(1.21)	(0.80)	(0.06)	(1.19)	(0.86)	(0.18)
	200	Power							62.5	99.2	100.0	60.9	99.1	100.0
		MCSE							(1.21)	(0.22)	(0.00)	(1.22)	(0.24)	(0.00)
	400	Power							88.3	100.0	100.0	87.3	100.0	100.0
		MCSE							(0.81)	(0.00)	(0.00)	(0.83)	(0.00)	(0.00)
	600	Power							96.9	100.0	100.0	96.8	100.0	100.0
		MCSE							(0.43)	(0.00)	(0.00)	(0.44)	(0.00)	(0.00)
	800	Power							99.4	100.0	100.0	99.4	100.0	100.0
		MCSE							(0.20)	(0.00)	(0.00)	(0.19)	(0.00)	(0.00)

Table E.11: ANCOVA method performance at each time point - Power: Pattern 2

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	Power										33.9	86.3	99.8
		MCSE										(1.18)	(0.86)	(0.13)
	200	Power										59.5	99.1	100.0
		MCSE										(1.23)	(0.23)	(0.00)
	400	Power										87.0	100.0	100.0
		MCSE										(0.84)	(0.00)	(0.00)
	600	Power										97.0	100.0	100.0
		MCSE										(0.43)	(0.00)	(0.00)
	800	Power										99.3	100.0	100.0
		MCSE										(0.21)	(0.00)	(0.00)

Table E.12: ANCOVA method performance at each time point - Power: Pattern 3

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	Power	37.5	90.0	99.9	37.5	88.6	99.7	37.3	87.4	99.5	36.1	89.1	99.8
		MCSE	(1.21)	(0.75)	(0.06)	(1.21)	(0.79)	(0.14)	(1.21)	(0.83)	(0.18)	(1.20)	(0.78)	(0.13)
	200	Power	61.6	99.3	100.0	60.1	99.7	100.0	59.9	99.7	100.0	62.1	99.6	100.0
		MCSE	(1.22)	(0.22)	(0.00)	(1.22)	(0.14)	(0.00)	(1.23)	(0.14)	(0.00)	(1.21)	(0.15)	(0.00)
	400	Power	89.0	100.0	100.0	90.3	100.0	100.0	90.4	100.0	100.0	87.9	100.0	100.0
		MCSE	(0.78)	(0.00)	(0.00)	(0.74)	(0.00)	(0.00)	(0.74)	(0.00)	(0.00)	(0.82)	(0.00)	(0.00)
	600	Power	97.6	100.0	100.0	97.3	100.0	100.0	97.7	100.0	100.0	97.8	100.0	100.0
		MCSE	(0.38)	(0.00)	(0.00)	(0.40)	(0.00)	(0.00)	(0.38)	(0.00)	(0.00)	(0.37)	(0.00)	(0.00)
	800	Power	99.1	100.0	100.0	99.3	100.0	100.0	99.6	100.0	100.0	99.3	100.0	100.0
		MCSE	(0.24)	(0.00)	(0.00)	(0.22)	(0.00)	(0.00)	(0.17)	(0.00)	(0.00)	(0.22)	(0.00)	(0.00)
2	100	Power	4.6	6.6	5.4	36.2	87.7	99.1	34.8	86.7	99.6	37.3	88.8	99.8
		MCSE	(0.52)	(0.62)	(0.56)	(1.20)	(0.82)	(0.23)	(1.19)	(0.85)	(0.17)	(1.21)	(0.79)	(0.13)
	200	Power	5.0	4.9	5.6	62.9	99.5	100.0	60.9	99.5	100.0	62.6	99.3	100.0
		MCSE	(0.55)	(0.54)	(0.57)	(1.21)	(0.18)	(0.00)	(1.22)	(0.18)	(0.00)	(1.21)	(0.21)	(0.00)
	400	Power	5.5	5.3	7.0	89.4	100.0	100.0	88.6	100.0	100.0	89.5	100.0	100.0
		MCSE	(0.57)	(0.56)	(0.64)	(0.77)	(0.00)	(0.00)	(0.80)	(0.00)	(0.00)	(0.77)	(0.00)	(0.00)
	600	Power	5.8	5.4	7.5	97.2	100.0	100.0	97.5	100.0	100.0	97.8	100.0	100.0
		MCSE	(0.58)	(0.56)	(0.66)	(0.41)	(0.00)	(0.00)	(0.39)	(0.00)	(0.00)	(0.37)	(0.00)	(0.00)
	800	Power	4.6	5.1	6.3	99.4	100.0	100.0	99.6	100.0	100.0	99.4	100.0	100.0
		MCSE	(0.53)	(0.55)	(0.60)	(0.20)	(0.00)	(0.00)	(0.15)	(0.00)	(0.00)	(0.20)	(0.00)	(0.00)

Table E.12: ANCOVA method performance at each time point - Power: Pattern 3

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	Power				5.0	6.4	5.6	4.4	5.6	5.9	4.6	5.6	5.6
		MCSE				(0.55)	(0.61)	(0.58)	(0.52)	(0.57)	(0.59)	(0.53)	(0.58)	(0.58)
	200	Power				5.3	4.4	6.6	4.8	4.6	6.2	5.2	5.4	6.9
		MCSE				(0.56)	(0.51)	(0.62)	(0.53)	(0.53)	(0.60)	(0.55)	(0.57)	(0.63)
	400	Power				4.1	6.1	7.3	4.1	6.5	7.3	4.9	5.4	6.9
		MCSE				(0.49)	(0.60)	(0.65)	(0.50)	(0.62)	(0.65)	(0.54)	(0.57)	(0.64)
	600	Power				4.5	6.3	10.3	4.8	6.6	9.6	5.3	5.9	8.5
		MCSE				(0.52)	(0.60)	(0.76)	(0.54)	(0.62)	(0.74)	(0.56)	(0.59)	(0.70)
	800	Power				5.3	6.1	10.2	5.5	6.6	9.1	5.1	5.9	9.6
		MCSE				(0.56)	(0.60)	(0.76)	(0.57)	(0.62)	(0.72)	(0.55)	(0.59)	(0.73)
4	100	Power							5.0	6.1	6.4	5.3	5.1	6.6
		MCSE							(0.55)	(0.60)	(0.61)	(0.56)	(0.55)	(0.62)
	200	Power							5.8	5.0	7.6	4.9	5.9	6.9
		MCSE							(0.58)	(0.55)	(0.66)	(0.54)	(0.59)	(0.64)
	400	Power							5.4	5.4	6.1	4.7	6.3	7.8
		MCSE							(0.57)	(0.57)	(0.60)	(0.53)	(0.61)	(0.67)
	600	Power							5.3	4.7	8.3	4.9	6.1	8.8
		MCSE							(0.56)	(0.53)	(0.69)	(0.54)	(0.60)	(0.71)
	800	Power							4.4	7.9	11.9	4.2	6.4	11.9
		MCSE							(0.51)	(0.68)	(0.81)	(0.50)	(0.61)	(0.81)

Table E.12: ANCOVA method performance at each time point - Power: Pattern 3

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	Power										4.9	4.6	5.1
		MCSE										(0.54)	(0.53)	(0.55)
	200	Power										5.1	6.8	7.3
		MCSE										(0.55)	(0.63)	(0.65)
	400	Power										5.1	5.6	7.8
		MCSE										(0.55)	(0.57)	(0.67)
	600	Power										4.9	6.9	8.4
		MCSE										(0.54)	(0.64)	(0.70)
	800	Power										5.8	5.7	9.4
		MCSE										(0.58)	(0.58)	(0.73)

Table E.13: ANCOVA method at each time point - Type I error: Pattern 1

			Number of time points:			
			2	**3**	**4**	**5**
Time point	**n**	**Type**				
1	100	TypeI	5.4	5.4	5.1	6.0
		MCSE	(0.57)	(0.56)	(0.55)	(0.59)
	200	TypeI	4.6	4.8	4.8	5.6
		MCSE	(0.53)	(0.54)	(0.54)	(0.58)
	400	TypeI	4.3	3.6	7.0	4.6
		MCSE	(0.51)	(0.46)	(0.64)	(0.52)
	600	TypeI	5.3	5.2	4.4	4.4
		MCSE	(0.56)	(0.55)	(0.52)	(0.51)
	800	TypeI	5.1	4.8	4.9	4.9
		MCSE	(0.55)	(0.54)	(0.54)	(0.54)
2	100	TypeI	6.8	5.1	5.9	5.8
		MCSE	(0.63)	(0.55)	(0.59)	(0.58)
	200	TypeI	5.4	5.1	5.1	5.7
		MCSE	(0.56)	(0.55)	(0.55)	(0.58)
	400	TypeI	5.5	4.6	6.0	4.6
		MCSE	(0.57)	(0.52)	(0.59)	(0.53)
	600	TypeI	4.4	5.1	4.4	4.9
		MCSE	(0.51)	(0.55)	(0.52)	(0.54)
	800	TypeI	5.1	4.7	4.0	4.8
		MCSE	(0.55)	(0.53)	(0.49)	(0.54)
3	100	TypeI		5.5	6.1	5.5
		MCSE		(0.57)	(0.60)	(0.57)
	200	TypeI		4.7	5.1	6.3
		MCSE		(0.53)	(0.55)	(0.60)
	400	TypeI		5.3	5.7	4.2
		MCSE		(0.56)	(0.58)	(0.50)
	600	TypeI		4.7	4.6	5.3
		MCSE		(0.53)	(0.52)	(0.56)
	800	TypeI		4.7	5.1	5.9
		MCSE		(0.53)	(0.55)	(0.59)

Table E.13: ANCOVA method at each time point - Type I error: Pattern 1

			Number of time points:			
			2	3	4	5
Time point	**n**	**Type**				
4	100	TypeI			5.4	4.9
		MCSE			(0.57)	(0.54)
	200	TypeI			5.6	5.8
		MCSE			(0.58)	(0.58)
	400	TypeI			5.9	5.3
		MCSE			(0.59)	(0.56)
	600	TypeI			4.4	5.5
		MCSE			(0.51)	(0.57)
	800	TypeI			4.4	4.6
		MCSE			(0.51)	(0.53)
5	100	TypeI				6.7
		MCSE				(0.63)
	200	TypeI				5.4
		MCSE				(0.56)
	400	TypeI				4.1
		MCSE				(0.49)
	600	TypeI				4.5
		MCSE				(0.52)
	800	TypeI				5.3
		MCSE				(0.56)

Table E.14: ANCOVA method at each time point - Type I error: Pattern 2

			Number of time points:			
			2	**3**	**4**	**5**
Time point	**n**	**Type**				
1	100	TypeI	5.6	6.5	6.9	5.8
		MCSE	(0.58)	(0.62)	(0.63)	(0.58)
	200	TypeI	6.2	5.4	5.6	4.7
		MCSE	(0.60)	(0.56)	(0.58)	(0.53)
	400	TypeI	5.2	5.5	5.0	5.1
		MCSE	(0.55)	(0.57)	(0.55)	(0.55)
	600	TypeI	5.6	6.1	4.3	5.1
		MCSE	(0.57)	(0.60)	(0.51)	(0.55)
	800	TypeI	6.0	5.0	5.8	5.4
		MCSE	(0.59)	(0.55)	(0.58)	(0.56)
2	100	TypeI	5.9	6.6	5.9	5.4
		MCSE	(0.59)	(0.62)	(0.59)	(0.57)
	200	TypeI	5.4	4.9	5.0	5.4
		MCSE	(0.56)	(0.54)	(0.55)	(0.56)
	400	TypeI	5.3	5.5	5.1	5.3
		MCSE	(0.56)	(0.57)	(0.55)	(0.56)
	600	TypeI	5.4	5.3	5.5	5.3
		MCSE	(0.56)	(0.56)	(0.57)	(0.56)
	800	TypeI	4.8	5.4	4.8	5.5
		MCSE	(0.53)	(0.57)	(0.53)	(0.57)
3	100	TypeI		5.9	5.1	5.4
		MCSE		(0.59)	(0.55)	(0.56)
	200	TypeI		6.3	5.5	5.9
		MCSE		(0.60)	(0.57)	(0.59)
	400	TypeI		5.1	4.9	5.4
		MCSE		(0.55)	(0.54)	(0.56)
	600	TypeI		4.5	5.1	5.6
		MCSE		(0.52)	(0.55)	(0.57)
	800	TypeI		5.3	5.3	5.3
		MCSE		(0.56)	(0.56)	(0.56)

Table E.14: ANCOVA method at each time point - Type I error: Pattern 2

			Number of time points:			
			2	3	4	5
Time point	**n**	**Type**				
4	100	TypeI			6.3	4.8
		MCSE			(0.61)	(0.54)
	200	TypeI			5.0	4.9
		MCSE			(0.55)	(0.54)
	400	TypeI			5.6	4.6
		MCSE			(0.58)	(0.52)
	600	TypeI			5.5	5.1
		MCSE			(0.57)	(0.55)
	800	TypeI			5.8	6.1
		MCSE			(0.58)	(0.60)
5	100	TypeI				5.4
		MCSE				(0.56)
	200	TypeI				5.4
		MCSE				(0.56)
	400	TypeI				5.8
		MCSE				(0.58)
	600	TypeI				5.4
		MCSE				(0.57)
	800	TypeI				5.6
		MCSE				(0.58)

Table E.15: ANCOVA method at each time point - Type I error: Pattern 3

			Number of time points:			
			2	**3**	**4**	**5**
Time point	**n**	**Type**				
1	100	TypeI	5.5	4.7	5.6	4.0
		MCSE	(0.57)	(0.53)	(0.58)	(0.49)
	200	TypeI	5.4	5.1	4.8	6.2
		MCSE	(0.56)	(0.55)	(0.53)	(0.60)
	400	TypeI	5.1	4.9	5.9	4.4
		MCSE	(0.55)	(0.54)	(0.59)	(0.51)
	600	TypeI	4.8	5.3	5.6	4.9
		MCSE	(0.54)	(0.56)	(0.57)	(0.54)
	800	TypeI	4.8	4.4	5.4	5.2
		MCSE	(0.53)	(0.51)	(0.56)	(0.55)
2	100	TypeI	6.0	6.1	5.7	5.4
		MCSE	(0.59)	(0.60)	(0.58)	(0.56)
	200	TypeI	5.7	4.7	4.3	5.4
		MCSE	(0.58)	(0.53)	(0.50)	(0.57)
	400	TypeI	6.3	5.3	6.6	5.3
		MCSE	(0.60)	(0.56)	(0.62)	(0.56)
	600	TypeI	4.8	4.9	4.1	4.8
		MCSE	(0.54)	(0.54)	(0.49)	(0.53)
	800	TypeI	4.0	4.3	5.0	4.7
		MCSE	(0.49)	(0.51)	(0.55)	(0.53)
3	100	TypeI		6.0	5.6	4.5
		MCSE		(0.59)	(0.57)	(0.52)
	200	TypeI		5.9	5.9	5.6
		MCSE		(0.59)	(0.59)	(0.58)
	400	TypeI		5.8	5.8	5.9
		MCSE		(0.58)	(0.58)	(0.59)
	600	TypeI		5.7	4.6	5.3
		MCSE		(0.58)	(0.52)	(0.56)
	800	TypeI		5.7	4.1	5.5
		MCSE		(0.58)	(0.50)	(0.57)

Table E.15: ANCOVA method at each time point - Type I error: Pattern 3

			Number of time points:			
			2	3	4	5
Time point	**n**	**Type**				
4	100	TypeI			5.0	5.4
		MCSE			(0.55)	(0.57)
	200	TypeI			6.1	5.5
		MCSE			(0.60)	(0.57)
	400	TypeI			4.3	4.4
		MCSE			(0.50)	(0.52)
	600	TypeI			5.3	4.8
		MCSE			(0.56)	(0.54)
	800	TypeI			5.2	5.6
		MCSE			(0.55)	(0.57)
5	100	TypeI				5.1
		MCSE				(0.55)
	200	TypeI				5.0
		MCSE				(0.55)
	400	TypeI				5.1
		MCSE				(0.55)
	600	TypeI				5.5
		MCSE				(0.57)
	800	TypeI				4.4
		MCSE				(0.51)

Table E.16: ANCOVA method performance at each time point - Coverage: Pattern 1

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	Cover.	94.8	94.3	93.9	95.4	94.6	94.6	94.6	94.5	93.7	95.4	94.9	94.4
		MCSE	(0.56)	(0.58)	(0.60)	(0.53)	(0.56)	(0.56)	(0.57)	(0.57)	(0.61)	(0.53)	(0.55)	(0.57)
	200	Cover.	93.9	94.2	93.6	94.8	93.2	94.3	95.3	94.8	93.8	93.9	93.5	94.1
		MCSE	(0.60)	(0.58)	(0.61)	(0.56)	(0.63)	(0.58)	(0.53)	(0.56)	(0.60)	(0.60)	(0.62)	(0.59)
	400	Cover.	95.0	94.0	92.5	94.6	94.4	92.8	94.8	95.4	92.9	95.5	93.6	92.6
		MCSE	(0.55)	(0.59)	(0.66)	(0.57)	(0.57)	(0.65)	(0.56)	(0.52)	(0.64)	(0.52)	(0.61)	(0.66)
	600	Cover.	94.9	92.9	90.3	94.8	95.1	92.6	95.6	93.6	91.7	94.9	93.6	91.5
		MCSE	(0.55)	(0.64)	(0.74)	(0.56)	(0.54)	(0.65)	(0.51)	(0.61)	(0.69)	(0.55)	(0.61)	(0.70)
	800	Cover.	95.3	92.8	89.6	94.4	94.5	90.7	95.0	93.3	91.6	95.0	93.4	89.8
		MCSE	(0.53)	(0.65)	(0.76)	(0.57)	(0.57)	(0.73)	(0.55)	(0.63)	(0.70)	(0.55)	(0.62)	(0.76)
2	100	Cover.	94.8	94.9	93.3	94.1	94.8	93.7	94.9	93.4	92.6	94.5	94.3	92.9
		MCSE	(0.55)	(0.55)	(0.63)	(0.59)	(0.55)	(0.61)	(0.55)	(0.62)	(0.65)	(0.57)	(0.58)	(0.64)
	200	Cover.	95.0	93.8	91.9	94.3	94.5	93.8	94.1	95.1	93.0	95.7	93.6	93.2
		MCSE	(0.55)	(0.60)	(0.68)	(0.58)	(0.57)	(0.60)	(0.59)	(0.54)	(0.64)	(0.51)	(0.61)	(0.63)
	400	Cover.	94.4	94.4	90.0	94.5	93.7	92.9	94.1	93.9	92.7	95.2	94.6	91.6
		MCSE	(0.58)	(0.57)	(0.75)	(0.57)	(0.61)	(0.64)	(0.59)	(0.60)	(0.65)	(0.54)	(0.56)	(0.70)
	600	Cover.	93.9	91.6	83.9	94.8	93.6	92.2	94.3	93.6	90.6	95.3	93.8	92.8
		MCSE	(0.60)	(0.70)	(0.92)	(0.55)	(0.61)	(0.67)	(0.58)	(0.61)	(0.73)	(0.53)	(0.60)	(0.65)
	800	Cover.	93.9	90.6	82.6	95.9	93.3	90.9	94.7	93.9	90.1	94.7	91.9	90.1
		MCSE	(0.60)	(0.73)	(0.95)	(0.49)	(0.63)	(0.72)	(0.56)	(0.60)	(0.75)	(0.56)	(0.68)	(0.75)

Table E.16: ANCOVA method performance at each time point - Coverage: Pattern 1

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	Cover.				94.9	93.8	93.8	94.0	94.1	93.6	94.8	95.1	94.6
		MCSE				(0.55)	(0.60)	(0.60)	(0.59)	(0.59)	(0.61)	(0.55)	(0.54)	(0.57)
	200	Cover.				95.1	93.6	92.6	95.0	93.6	93.4	94.6	94.4	93.1
		MCSE				(0.54)	(0.61)	(0.66)	(0.55)	(0.61)	(0.62)	(0.56)	(0.57)	(0.63)
	400	Cover.				94.4	93.4	88.3	94.4	93.1	91.3	93.9	92.9	92.3
		MCSE				(0.58)	(0.62)	(0.80)	(0.58)	(0.64)	(0.71)	(0.60)	(0.64)	(0.67)
	600	Cover.				94.3	92.1	86.3	94.1	92.4	88.1	94.6	92.9	91.4
		MCSE				(0.58)	(0.67)	(0.86)	(0.59)	(0.66)	(0.81)	(0.57)	(0.64)	(0.70)
	800	Cover.				94.5	91.3	81.7	94.4	92.7	87.6	94.1	93.2	88.7
		MCSE				(0.57)	(0.70)	(0.97)	(0.58)	(0.65)	(0.82)	(0.59)	(0.63)	(0.79)
4	100	Cover.							95.4	95.1	93.1	95.2	94.4	93.7
		MCSE							(0.52)	(0.54)	(0.63)	(0.54)	(0.57)	(0.61)
	200	Cover.							93.6	93.1	92.3	95.0	92.9	92.4
		MCSE							(0.61)	(0.63)	(0.67)	(0.55)	(0.64)	(0.66)
	400	Cover.							95.3	93.5	88.4	94.4	93.2	90.9
		MCSE							(0.53)	(0.62)	(0.80)	(0.57)	(0.63)	(0.72)
	600	Cover.							94.3	91.4	84.4	94.4	92.8	88.8
		MCSE							(0.58)	(0.70)	(0.91)	(0.58)	(0.65)	(0.79)
	800	Cover.							94.3	91.4	82.3	94.3	92.8	86.2
		MCSE							(0.58)	(0.70)	(0.96)	(0.58)	(0.65)	(0.86)

Table E.16: ANCOVA method performance at each time point - Coverage: Pattern 1

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	Cover.										94.5	95.4	94.7
		MCSE										(0.57)	(0.53)	(0.56)
	200	Cover.										94.6	93.9	91.0
		MCSE										(0.56)	(0.60)	(0.71)
	400	Cover.										95.2	92.3	89.5
		MCSE										(0.54)	(0.67)	(0.77)
	600	Cover.										94.3	91.5	86.1
		MCSE										(0.58)	(0.70)	(0.87)
	800	Cover.										94.5	90.8	81.1
		MCSE										(0.57)	(0.72)	(0.98)

Table E.17: ANCOVA method performance at each time point - Coverage: Pattern 2

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	Cover.	94.4	94.5	92.3	94.5	94.0	94.5	93.7	94.8	93.4	93.9	93.9	92.8
		MCSE	(0.57)	(0.57)	(0.67)	(0.57)	(0.59)	(0.57)	(0.61)	(0.55)	(0.62)	(0.60)	(0.60)	(0.65)
	200	Cover.	94.5	94.3	90.3	95.1	93.0	92.9	94.8	94.6	90.9	94.3	92.9	89.9
		MCSE	(0.57)	(0.58)	(0.74)	(0.54)	(0.64)	(0.64)	(0.55)	(0.57)	(0.72)	(0.58)	(0.64)	(0.75)
	400	Cover.	95.0	91.8	86.5	94.3	93.1	89.6	96.1	91.9	88.6	94.3	91.5	88.7
		MCSE	(0.55)	(0.69)	(0.85)	(0.58)	(0.63)	(0.76)	(0.49)	(0.68)	(0.79)	(0.58)	(0.70)	(0.79)
	600	Cover.	94.8	91.1	82.6	94.6	91.4	88.3	94.9	91.4	84.8	94.2	90.7	84.1
		MCSE	(0.55)	(0.71)	(0.95)	(0.56)	(0.70)	(0.81)	(0.55)	(0.70)	(0.90)	(0.58)	(0.73)	(0.92)
	800	Cover.	94.8	89.9	77.2	95.1	92.2	86.1	95.3	91.2	83.3	93.9	89.4	81.1
		MCSE	(0.55)	(0.75)	(1.05)	(0.54)	(0.67)	(0.86)	(0.53)	(0.71)	(0.93)	(0.60)	(0.77)	(0.98)
2	100	Cover.	94.8	94.1	92.6	94.1	94.4	92.6	92.9	94.1	92.2	94.2	93.3	92.2
		MCSE	(0.55)	(0.59)	(0.66)	(0.59)	(0.58)	(0.65)	(0.64)	(0.59)	(0.67)	(0.58)	(0.63)	(0.67)
	200	Cover.	94.4	92.9	90.1	94.7	93.6	90.4	95.1	93.8	89.8	94.5	93.1	89.3
		MCSE	(0.58)	(0.64)	(0.75)	(0.56)	(0.61)	(0.74)	(0.54)	(0.60)	(0.76)	(0.57)	(0.64)	(0.77)
	400	Cover.	94.6	92.3	86.4	94.2	93.7	86.2	94.6	92.6	84.6	93.3	91.9	85.1
		MCSE	(0.57)	(0.67)	(0.86)	(0.58)	(0.61)	(0.86)	(0.57)	(0.65)	(0.90)	(0.63)	(0.68)	(0.89)
	600	Cover.	94.9	90.8	83.4	94.0	89.2	82.9	95.6	91.0	80.3	93.3	89.8	76.5
		MCSE	(0.55)	(0.72)	(0.93)	(0.59)	(0.78)	(0.94)	(0.52)	(0.71)	(0.99)	(0.63)	(0.76)	(1.06)
	800	Cover.	93.7	89.5	78.2	93.9	89.1	77.3	94.0	87.1	74.8	93.7	86.8	71.1
		MCSE	(0.61)	(0.77)	(1.03)	(0.60)	(0.78)	(1.05)	(0.59)	(0.84)	(1.08)	(0.61)	(0.85)	(1.13)

Table E.17: ANCOVA method performance at each time point - Coverage: Pattern 2

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	Cover.				94.8	94.6	93.4	94.3	94.1	92.1	95.1	94.1	92.6
		MCSE				(0.55)	(0.57)	(0.62)	(0.58)	(0.59)	(0.67)	(0.54)	(0.59)	(0.65)
	200	Cover.				94.8	93.8	90.2	94.3	92.7	89.9	95.2	92.1	88.2
		MCSE				(0.56)	(0.60)	(0.74)	(0.58)	(0.65)	(0.75)	(0.54)	(0.68)	(0.81)
	400	Cover.				93.6	92.8	85.5	95.4	90.6	84.8	92.9	90.8	84.6
		MCSE				(0.61)	(0.65)	(0.88)	(0.52)	(0.73)	(0.90)	(0.64)	(0.72)	(0.90)
	600	Cover.				94.9	91.1	81.4	93.9	88.8	78.3	93.6	89.5	78.1
		MCSE				(0.55)	(0.71)	(0.97)	(0.60)	(0.79)	(1.03)	(0.61)	(0.77)	(1.03)
	800	Cover.				94.3	89.4	78.5	93.9	89.2	73.8	93.0	87.0	72.8
		MCSE				(0.58)	(0.77)	(1.03)	(0.60)	(0.78)	(1.10)	(0.64)	(0.84)	(1.11)
4	100	Cover.							93.4	93.5	91.8	94.8	93.6	93.1
		MCSE							(0.62)	(0.62)	(0.69)	(0.56)	(0.61)	(0.64)
	200	Cover.							94.3	93.1	90.0	95.1	92.3	89.8
		MCSE							(0.58)	(0.63)	(0.75)	(0.54)	(0.67)	(0.76)
	400	Cover.							94.4	90.3	86.9	93.5	91.9	85.5
		MCSE							(0.58)	(0.74)	(0.84)	(0.62)	(0.68)	(0.88)
	600	Cover.							93.8	90.6	80.8	93.1	90.1	79.1
		MCSE							(0.60)	(0.73)	(0.98)	(0.64)	(0.75)	(1.02)
	800	Cover.							94.2	88.3	73.6	93.7	88.9	72.8
		MCSE							(0.58)	(0.80)	(1.10)	(0.61)	(0.79)	(1.11)

Table E.17: ANCOVA method performance at each time point - Coverage: Pattern 2

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	Cover.										94.9	93.8	93.5
		MCSE										(0.55)	(0.60)	(0.62)
	200	Cover.										94.8	92.8	89.7
		MCSE										(0.56)	(0.65)	(0.76)
	400	Cover.										93.4	91.4	84.1
		MCSE										(0.62)	(0.70)	(0.91)
	600	Cover.										94.3	89.4	78.7
		MCSE										(0.58)	(0.77)	(1.02)
	800	Cover.										92.9	86.9	71.8
		MCSE										(0.64)	(0.84)	(1.13)

Table E.18: ANCOVA method performance at each time point - Coverage: Pattern 3

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	Cover.	94.6	93.9	93.3	94.7	94.1	92.1	94.6	94.4	92.4	95.1	94.8	93.6
		MCSE	(0.57)	(0.60)	(0.63)	(0.56)	(0.59)	(0.68)	(0.56)	(0.57)	(0.66)	(0.54)	(0.55)	(0.61)
	200	Cover.	95.1	93.6	91.4	95.0	94.5	90.4	94.8	94.3	91.9	94.1	94.0	91.6
		MCSE	(0.54)	(0.61)	(0.70)	(0.55)	(0.57)	(0.73)	(0.56)	(0.58)	(0.68)	(0.59)	(0.59)	(0.69)
	400	Cover.	94.1	94.0	89.5	94.9	91.8	87.1	95.5	92.4	90.1	93.8	92.1	88.4
		MCSE	(0.59)	(0.59)	(0.77)	(0.55)	(0.69)	(0.84)	(0.52)	(0.66)	(0.75)	(0.60)	(0.67)	(0.80)
	600	Cover.	94.8	91.6	84.9	94.6	91.1	84.3	94.3	92.4	83.8	95.4	92.1	85.3
		MCSE	(0.55)	(0.70)	(0.89)	(0.57)	(0.71)	(0.91)	(0.58)	(0.66)	(0.92)	(0.53)	(0.68)	(0.89)
	800	Cover.	94.8	90.8	81.6	94.2	90.4	80.1	94.3	90.9	80.5	93.4	90.6	81.9
		MCSE	(0.56)	(0.72)	(0.97)	(0.58)	(0.74)	(1.00)	(0.58)	(0.72)	(0.99)	(0.62)	(0.73)	(0.96)
2	100	Cover.	95.4	93.4	94.6	93.6	93.9	93.4	95.3	93.5	92.5	95.8	94.8	93.5
		MCSE	(0.52)	(0.62)	(0.56)	(0.61)	(0.60)	(0.62)	(0.53)	(0.62)	(0.66)	(0.50)	(0.56)	(0.62)
	200	Cover.	95.0	95.1	94.4	95.1	94.1	90.6	94.6	93.6	89.6	95.0	92.1	91.4
		MCSE	(0.55)	(0.54)	(0.57)	(0.54)	(0.59)	(0.73)	(0.56)	(0.61)	(0.76)	(0.55)	(0.67)	(0.70)
	400	Cover.	94.5	94.8	93.0	95.3	93.1	88.4	95.2	92.6	90.8	95.9	93.5	87.9
		MCSE	(0.57)	(0.56)	(0.64)	(0.53)	(0.64)	(0.80)	(0.54)	(0.66)	(0.72)	(0.49)	(0.62)	(0.82)
	600	Cover.	94.3	94.6	92.5	95.0	91.5	84.4	93.9	92.8	85.1	94.7	91.0	85.1
		MCSE	(0.58)	(0.56)	(0.66)	(0.55)	(0.70)	(0.91)	(0.60)	(0.65)	(0.89)	(0.56)	(0.71)	(0.89)
	800	Cover.	95.4	94.9	93.8	94.1	90.8	80.3	95.1	91.4	80.9	94.8	90.6	81.0
		MCSE	(0.53)	(0.55)	(0.60)	(0.59)	(0.72)	(0.99)	(0.54)	(0.70)	(0.98)	(0.55)	(0.73)	(0.98)

Table E.18: ANCOVA method performance at each time point - Coverage: Pattern 3

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	Cover.				95.0	93.6	94.4	95.6	94.4	94.1	95.4	94.4	94.4
		MCSE				(0.55)	(0.61)	(0.58)	(0.52)	(0.57)	(0.59)	(0.53)	(0.58)	(0.58)
	200	Cover.				94.7	95.6	93.4	95.3	95.4	93.8	94.8	94.6	93.1
		MCSE				(0.56)	(0.51)	(0.62)	(0.53)	(0.53)	(0.60)	(0.55)	(0.57)	(0.63)
	400	Cover.				95.9	93.9	92.8	95.9	93.5	92.7	95.1	94.6	93.1
		MCSE				(0.49)	(0.60)	(0.65)	(0.50)	(0.62)	(0.65)	(0.54)	(0.57)	(0.64)
	600	Cover.				95.5	93.8	89.8	95.2	93.4	90.4	94.7	94.1	91.5
		MCSE				(0.52)	(0.60)	(0.76)	(0.54)	(0.62)	(0.74)	(0.56)	(0.59)	(0.70)
	800	Cover.				94.7	93.9	89.8	94.5	93.4	90.9	94.9	94.1	90.4
		MCSE				(0.56)	(0.60)	(0.76)	(0.57)	(0.62)	(0.72)	(0.55)	(0.59)	(0.73)
4	100	Cover.							95.0	93.9	93.6	94.7	94.9	93.4
		MCSE							(0.55)	(0.60)	(0.61)	(0.56)	(0.55)	(0.62)
	200	Cover.							94.2	95.0	92.4	95.1	94.1	93.1
		MCSE							(0.58)	(0.55)	(0.66)	(0.54)	(0.59)	(0.64)
	400	Cover.							94.6	94.6	93.9	95.3	93.7	92.2
		MCSE							(0.57)	(0.57)	(0.60)	(0.53)	(0.61)	(0.67)
	600	Cover.							94.7	95.3	91.8	95.1	93.9	91.3
		MCSE							(0.56)	(0.53)	(0.69)	(0.54)	(0.60)	(0.71)
	800	Cover.							95.6	92.1	88.1	95.8	93.6	88.1
		MCSE							(0.51)	(0.68)	(0.81)	(0.50)	(0.61)	(0.81)

Table E.18: ANCOVA method performance at each time point - Coverage: Pattern 3

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	Cover.										95.1	95.4	94.9
		MCSE										(0.54)	(0.53)	(0.55)
	200	Cover.										94.9	93.3	92.8
		MCSE										(0.55)	(0.63)	(0.65)
	400	Cover.										94.9	94.4	92.2
		MCSE										(0.55)	(0.57)	(0.67)
	600	Cover.										95.1	93.1	91.6
		MCSE										(0.54)	(0.64)	(0.70)
	800	Cover.										94.3	94.3	90.6
		MCSE										(0.58)	(0.58)	(0.73)

Table E.19: ANCOVA method performance at each time point - Bias: Pattern 1

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	Bias	-0.14	-0.38	-0.55	-0.06	-0.24	-0.46	-0.13	-0.28	-0.42	-0.13	-0.28	-0.60
		MCSE	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias	-0.17	-0.30	-0.55	-0.13	-0.27	-0.49	-0.07	-0.28	-0.49	-0.10	-0.32	-0.54
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias	-0.19	-0.33	-0.60	-0.13	-0.31	-0.46	-0.09	-0.28	-0.47	-0.13	-0.34	-0.52
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias	-0.14	-0.37	-0.59	-0.14	-0.26	-0.44	-0.10	-0.29	-0.50	-0.14	-0.33	-0.51
		MCSE	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)
	800	Bias	-0.17	-0.32	-0.55	-0.12	-0.29	-0.47	-0.14	-0.31	-0.48	-0.11	-0.30	-0.59
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
2	100	Bias	-0.20	-0.49	-0.84	-0.11	-0.30	-0.57	-0.09	-0.32	-0.53	-0.08	-0.28	-0.50
		MCSE	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias	-0.20	-0.57	-0.86	-0.18	-0.28	-0.55	-0.11	-0.30	-0.57	-0.07	-0.33	-0.56
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias	-0.23	-0.48	-0.85	-0.15	-0.36	-0.56	-0.13	-0.31	-0.53	-0.12	-0.31	-0.53
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias	-0.20	-0.51	-0.87	-0.13	-0.33	-0.50	-0.15	-0.33	-0.53	-0.14	-0.35	-0.53
		MCSE	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
	800	Bias	-0.21	-0.47	-0.89	-0.16	-0.32	-0.54	-0.15	-0.33	-0.51	-0.15	-0.31	-0.57
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

Table E.19: ANCOVA method performance at each time point - Bias: Pattern 1

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	Bias				-0.12	-0.43	-0.78	-0.15	-0.36	-0.63	-0.14	-0.31	-0.60
		MCSE				(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias				-0.18	-0.45	-0.87	-0.18	-0.39	-0.66	-0.10	-0.40	-0.65
		MCSE				(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias				-0.19	-0.47	-0.84	-0.14	-0.40	-0.64	-0.15	-0.34	-0.62
		MCSE				(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias				-0.24	-0.45	-0.83	-0.20	-0.39	-0.67	-0.18	-0.38	-0.61
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)
	800	Bias				-0.22	-0.49	-0.84	-0.18	-0.41	-0.65	-0.14	-0.33	-0.62
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
4	100	Bias							-0.21	-0.51	-0.80	-0.18	-0.36	-0.73
		MCSE							(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias							-0.21	-0.47	-0.85	-0.08	-0.50	-0.70
		MCSE							(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias							-0.19	-0.47	-0.82	-0.15	-0.42	-0.71
		MCSE							(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias							-0.21	-0.49	-0.82	-0.19	-0.45	-0.70
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	Bias							-0.21	-0.49	-0.86	-0.17	-0.34	-0.73
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

Table E.19: ANCOVA method performance at each time point - Bias: Pattern 1

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	Bias										-0.16	-0.40	-0.82
		MCSE										(0.06)	(0.06)	(0.06)
	200	Bias										-0.15	-0.49	-0.91
		MCSE										(0.04)	(0.04)	(0.04)
	400	Bias										-0.15	-0.48	-0.88
		MCSE										(0.03)	(0.03)	(0.03)
	600	Bias										-0.23	-0.52	-0.83
		MCSE										(0.02)	(0.02)	(0.02)
	800	Bias										-0.18	-0.45	-0.90
		MCSE										(0.02)	(0.02)	(0.02)

Table E.20: ANCOVA method performance at each time point - Bias: Pattern 2

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	Bias	-0.31	-0.67	-1.08	-0.18	-0.45	-0.64	-0.18	-0.47	-0.85	-0.23	-0.72	-0.85
		MCSE	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias	-0.29	-0.57	-1.00	-0.15	-0.35	-0.73	-0.24	-0.51	-0.88	-0.20	-0.56	-1.02
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias	-0.18	-0.55	-1.01	-0.17	-0.45	-0.78	-0.23	-0.56	-0.86	-0.24	-0.58	-0.92
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias	-0.21	-0.57	-0.97	-0.18	-0.47	-0.73	-0.20	-0.52	-0.87	-0.22	-0.53	-0.95
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.02)
	800	Bias	-0.20	-0.59	-0.99	-0.17	-0.42	-0.73	-0.23	-0.46	-0.88	-0.26	-0.57	-0.93
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
2	100	Bias	-0.24	-0.57	-1.06	-0.21	-0.52	-0.93	-0.22	-0.69	-1.15	-0.27	-0.69	-1.09
		MCSE	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias	-0.25	-0.54	-1.07	-0.24	-0.50	-0.97	-0.27	-0.63	-1.09	-0.27	-0.60	-1.19
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias	-0.22	-0.58	-1.01	-0.29	-0.54	-1.02	-0.27	-0.62	-1.08	-0.31	-0.68	-1.12
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias	-0.25	-0.56	-0.98	-0.30	-0.60	-0.99	-0.26	-0.61	-1.10	-0.24	-0.62	-1.20
		MCSE	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)
	800	Bias	-0.24	-0.57	-0.99	-0.21	-0.56	-0.99	-0.28	-0.59	-1.08	-0.30	-0.67	-1.15
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

Table E.20: ANCOVA method performance at each time point - Bias: Pattern 2

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	Bias				-0.28	-0.49	-0.93	-0.24	-0.61	-1.06	-0.35	-0.69	-1.12
		MCSE				(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias				-0.21	-0.51	-0.96	-0.28	-0.62	-1.05	-0.28	-0.70	-1.16
		MCSE				(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias				-0.29	-0.56	-1.01	-0.25	-0.64	-1.10	-0.28	-0.67	-1.12
		MCSE				(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias				-0.27	-0.59	-1.01	-0.28	-0.62	-1.10	-0.26	-0.68	-1.14
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)
	800	Bias				-0.23	-0.56	-0.99	-0.28	-0.59	-1.09	-0.29	-0.68	-1.12
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
4	100	Bias							-0.29	-0.62	-1.14	-0.33	-0.72	-1.09
		MCSE							(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias							-0.27	-0.61	-1.04	-0.25	-0.67	-1.13
		MCSE							(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias							-0.24	-0.66	-1.08	-0.31	-0.70	-1.15
		MCSE							(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias							-0.25	-0.61	-1.07	-0.24	-0.67	-1.12
		MCSE							(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)
	800	Bias							-0.29	-0.59	-1.10	-0.30	-0.64	-1.16
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

Table E.20: ANCOVA method performance at each time point - Bias: Pattern 2

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	Bias										-0.30	-0.75	-1.06
		MCSE										(0.06)	(0.06)	(0.06)
	200	Bias										-0.27	-0.66	-1.22
		MCSE										(0.04)	(0.04)	(0.04)
	400	Bias										-0.30	-0.66	-1.16
		MCSE										(0.03)	(0.03)	(0.03)
	600	Bias										-0.26	-0.67	-1.14
		MCSE										(0.02)	(0.02)	(0.02)
	800	Bias										-0.30	-0.66	-1.15
		MCSE										(0.02)	(0.02)	(0.02)

Table E.21: ANCOVA method performance at each time point - Bias: Pattern 3

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	Bias	-0.09	-0.41	-0.73	-0.12	-0.46	-0.95	-0.17	-0.57	-0.94	-0.20	-0.49	-0.75
		MCSE	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias	-0.20	-0.53	-0.80	-0.23	-0.53	-0.88	-0.29	-0.47	-0.97	-0.17	-0.52	-0.87
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias	-0.21	-0.47	-0.79	-0.18	-0.53	-0.90	-0.22	-0.48	-0.83	-0.25	-0.45	-0.89
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias	-0.20	-0.44	-0.86	-0.24	-0.54	-0.92	-0.19	-0.48	-0.89	-0.19	-0.47	-0.84
		MCSE	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	Bias	-0.16	-0.46	-0.83	-0.22	-0.51	-0.92	-0.19	-0.48	-0.90	-0.19	-0.49	-0.86
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
2	100	Bias	0.01	-0.16	-0.28	-0.17	-0.51	-0.98	-0.18	-0.59	-0.89	-0.15	-0.43	-0.68
		MCSE	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias	-0.08	-0.26	-0.33	-0.14	-0.57	-0.99	-0.23	-0.45	-1.00	-0.24	-0.50	-0.83
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias	-0.10	-0.20	-0.34	-0.23	-0.52	-0.93	-0.22	-0.49	-0.80	-0.23	-0.49	-0.87
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias	-0.07	-0.19	-0.38	-0.23	-0.51	-0.94	-0.21	-0.49	-0.87	-0.17	-0.49	-0.84
		MCSE	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	Bias	-0.10	-0.16	-0.37	-0.19	-0.51	-0.93	-0.17	-0.49	-0.88	-0.18	-0.47	-0.88
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

Table E.21: ANCOVA method performance at each time point - Bias: Pattern 3

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	Bias				-0.09	-0.29	-0.73	-0.13	-0.44	-0.58	-0.10	-0.31	-0.48
		MCSE				(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias				-0.05	-0.37	-0.53	-0.14	-0.26	-0.62	-0.13	-0.30	-0.54
		MCSE				(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias				-0.12	-0.35	-0.55	-0.11	-0.24	-0.50	-0.13	-0.25	-0.56
		MCSE				(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias				-0.14	-0.32	-0.56	-0.08	-0.29	-0.54	-0.08	-0.26	-0.49
		MCSE				(0.02)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	800	Bias				-0.11	-0.29	-0.57	-0.12	-0.33	-0.53	-0.11	-0.28	-0.54
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
4	100	Bias							-0.04	-0.37	-0.62	-0.12	-0.25	-0.47
		MCSE							(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
	200	Bias							-0.23	-0.26	-0.64	-0.18	-0.29	-0.54
		MCSE							(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	400	Bias							-0.12	-0.25	-0.51	-0.16	-0.26	-0.52
		MCSE							(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	600	Bias							-0.11	-0.26	-0.56	-0.06	-0.27	-0.52
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	Bias							-0.09	-0.31	-0.55	-0.10	-0.28	-0.56
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

Table E.21: ANCOVA method performance at each time point - Bias: Pattern 3

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	Bias										-0.12	-0.30	-0.45
		MCSE										(0.06)	(0.06)	(0.06)
	200	Bias										-0.12	-0.35	-0.52
		MCSE										(0.04)	(0.04)	(0.04)
	400	Bias										-0.16	-0.27	-0.56
		MCSE										(0.03)	(0.03)	(0.03)
	600	Bias										-0.05	-0.28	-0.51
		MCSE										(0.02)	(0.03)	(0.03)
	800	Bias										-0.10	-0.26	-0.50
		MCSE										(0.02)	(0.02)	(0.02)

Table E.22: ANCOVA method performance at each time point - Empirical standard error: Pattern 1

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	EmpSE	2.43	2.37	2.41	2.29	2.46	2.40	2.39	2.35	2.45	2.36	2.45	2.40
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE	1.71	1.70	1.70	1.69	1.77	1.64	1.67	1.68	1.70	1.72	1.75	1.65
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE	1.20	1.19	1.19	1.18	1.19	1.20	1.20	1.16	1.21	1.17	1.19	1.19
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE	0.97	1.00	0.99	0.97	0.96	0.95	0.94	0.96	0.99	0.97	0.99	0.96
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE	0.82	0.85	0.84	0.83	0.83	0.84	0.84	0.84	0.83	0.83	0.85	0.84
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
2	100	EmpSE	2.35	2.39	2.40	2.44	2.42	2.40	2.36	2.49	2.49	2.40	2.44	2.44
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE	1.66	1.68	1.68	1.71	1.69	1.68	1.72	1.66	1.73	1.65	1.71	1.69
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE	1.18	1.12	1.15	1.20	1.16	1.17	1.20	1.19	1.17	1.18	1.16	1.21
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE	0.99	0.99	0.95	0.97	0.96	0.95	0.98	0.96	0.99	0.97	0.95	0.93
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE	0.84	0.84	0.81	0.79	0.83	0.84	0.85	0.83	0.85	0.85	0.88	0.84
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Table E.22: ANCOVA method performance at each time point - Empirical standard error: Pattern 1

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	EmpSE				2.40	2.41	2.33	2.45	2.38	2.38	2.41	2.37	2.34
		MCSE				(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE				1.67	1.72	1.63	1.68	1.69	1.68	1.69	1.67	1.67
		MCSE				(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE				1.20	1.18	1.20	1.20	1.19	1.21	1.20	1.19	1.18
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE				0.95	0.96	0.94	0.97	0.99	1.00	0.99	0.97	0.95
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE				0.81	0.83	0.84	0.86	0.82	0.84	0.86	0.84	0.82
		MCSE				(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
4	100	EmpSE							2.43	2.33	2.33	2.36	2.37	2.38
		MCSE							(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE							1.71	1.69	1.67	1.70	1.72	1.71
		MCSE							(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE							1.18	1.18	1.22	1.21	1.22	1.16
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE							0.97	0.98	0.96	0.97	0.96	0.97
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE							0.85	0.83	0.81	0.85	0.84	0.84
		MCSE							(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Table E.22: ANCOVA method performance at each time point - Empirical standard error: Pattern 1

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	EmpSE										2.36	2.30	2.29
		MCSE										(0.04)	(0.04)	(0.04)
	200	EmpSE										1.72	1.66	1.63
		MCSE										(0.03)	(0.03)	(0.03)
	400	EmpSE										1.16	1.20	1.15
		MCSE										(0.02)	(0.02)	(0.02)
	600	EmpSE										0.96	0.95	0.98
		MCSE										(0.02)	(0.02)	(0.02)
	800	EmpSE										0.85	0.85	0.80
		MCSE										(0.01)	(0.01)	(0.01)

Table E.23: ANCOVA method performance at each time point - Empirical standard error: Pattern 2

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	EmpSE	2.40	2.33	2.38	2.39	2.40	2.38	2.52	2.33	2.42	2.41	2.40	2.41
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE	1.67	1.67	1.65	1.65	1.77	1.67	1.69	1.69	1.69	1.72	1.70	1.70
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE	1.20	1.19	1.15	1.23	1.18	1.21	1.15	1.19	1.19	1.19	1.21	1.15
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE	0.95	0.96	0.96	0.98	0.98	0.97	0.96	0.98	0.97	0.99	0.98	0.95
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE	0.82	0.84	0.84	0.83	0.85	0.83	0.82	0.84	0.81	0.85	0.84	0.81
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
2	100	EmpSE	2.37	2.41	2.39	2.40	2.39	2.40	2.51	2.31	2.36	2.38	2.39	2.39
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE	1.70	1.69	1.65	1.64	1.67	1.69	1.66	1.66	1.69	1.68	1.67	1.68
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE	1.17	1.18	1.19	1.21	1.15	1.18	1.18	1.15	1.19	1.21	1.18	1.17
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE	0.95	0.98	0.93	0.95	0.99	0.93	0.96	0.93	0.94	0.99	0.96	0.95
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE	0.84	0.83	0.82	0.86	0.85	0.84	0.82	0.86	0.80	0.84	0.84	0.83
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Table E.23: ANCOVA method performance at each time point - Empirical standard error: Pattern 2

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	EmpSE				2.36	2.39	2.30	2.42	2.40	2.39	2.35	2.42	2.35
		MCSE				(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE				1.68	1.68	1.67	1.67	1.71	1.68	1.66	1.68	1.70
		MCSE				(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE				1.24	1.18	1.21	1.15	1.20	1.17	1.23	1.19	1.16
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE				0.98	0.96	0.96	0.97	0.99	0.95	0.98	0.97	0.96
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE				0.85	0.84	0.80	0.83	0.84	0.81	0.85	0.85	0.83
		MCSE				(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
4	100	EmpSE							2.50	2.37	2.36	2.37	2.36	2.32
		MCSE							(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE							1.67	1.70	1.69	1.65	1.71	1.66
		MCSE							(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE							1.20	1.23	1.13	1.20	1.14	1.14
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE							0.99	0.95	0.94	0.99	0.95	0.94
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE							0.82	0.86	0.81	0.82	0.82	0.82
		MCSE							(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Table E.23: ANCOVA method performance at each time point - Empirical standard error: Pattern 2

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	EmpSE										2.37	2.39	2.27
		MCSE										(0.04)	(0.04)	(0.04)
	200	EmpSE										1.70	1.68	1.65
		MCSE										(0.03)	(0.03)	(0.03)
	400	EmpSE										1.21	1.19	1.15
		MCSE										(0.02)	(0.02)	(0.02)
	600	EmpSE										0.96	0.96	0.94
		MCSE										(0.02)	(0.02)	(0.02)
	800	EmpSE										0.84	0.83	0.85
		MCSE										(0.01)	(0.01)	(0.01)

Table E.24: ANCOVA method performance at each time point - Empirical standard error: Pattern 3

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
1	100	EmpSE	2.41	2.37	2.38	2.49	2.40	2.35	2.40	2.37	2.41	2.37	2.29	2.37
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE	1.67	1.68	1.71	1.68	1.64	1.70	1.67	1.65	1.62	1.73	1.66	1.69
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE	1.20	1.14	1.18	1.20	1.19	1.20	1.15	1.19	1.13	1.22	1.22	1.17
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE	0.95	0.98	0.99	0.96	0.97	0.96	0.96	0.96	0.96	0.93	0.96	0.98
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE	0.85	0.83	0.87	0.86	0.86	0.83	0.84	0.83	0.83	0.86	0.85	0.83
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
2	100	EmpSE	2.36	2.46	2.37	2.51	2.37	2.34	2.36	2.45	2.37	2.35	2.35	2.40
		MCSE	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE	1.67	1.62	1.70	1.68	1.62	1.67	1.68	1.66	1.70	1.61	1.74	1.70
		MCSE	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE	1.19	1.18	1.21	1.16	1.17	1.17	1.17	1.18	1.14	1.13	1.19	1.19
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE	0.98	0.97	0.97	0.97	0.98	0.96	0.98	0.97	0.95	0.98	0.98	0.98
		MCSE	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE	0.84	0.83	0.84	0.86	0.83	0.82	0.82	0.84	0.81	0.82	0.85	0.83
		MCSE	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Table E.24: ANCOVA method performance at each time point - Empirical standard error: Pattern 3

			Time points: 2			**Time points: 3**			**Time points: 4**			**Time points: 5**		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
3	100	EmpSE				2.46	2.45	2.34	2.35	2.43	2.44	2.41	2.35	2.32
		MCSE				(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE				1.68	1.63	1.73	1.68	1.67	1.66	1.69	1.70	1.71
		MCSE				(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE				1.17	1.19	1.19	1.17	1.20	1.19	1.17	1.19	1.21
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE				0.95	1.00	1.00	0.96	0.99	0.98	1.00	1.00	0.99
		MCSE				(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE				0.83	0.85	0.85	0.87	0.85	0.83	0.83	0.86	0.86
		MCSE				(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
4	100	EmpSE							2.35	2.45	2.40	2.42	2.37	2.48
		MCSE							(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
	200	EmpSE							1.71	1.66	1.71	1.66	1.74	1.73
		MCSE							(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	400	EmpSE							1.17	1.19	1.16	1.15	1.19	1.21
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	600	EmpSE							0.96	0.95	0.95	0.96	0.97	0.97
		MCSE							(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	800	EmpSE							0.81	0.87	0.86	0.81	0.84	0.86
		MCSE							(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Table E.24: ANCOVA method performance at each time point - Empirical standard error: Pattern 3

			Time points: 2			Time points: 3			Time points: 4			Time points: 5		
Time point	**n**	$\beta*$: **Type**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**	**4**	**8**	**12**
5	100	EmpSE										2.39	2.35	2.35
		MCSE										(0.04)	(0.04)	(0.04)
	200	EmpSE										1.68	1.70	1.71
		MCSE										(0.03)	(0.03)	(0.03)
	400	EmpSE										1.17	1.20	1.20
		MCSE										(0.02)	(0.02)	(0.02)
	600	EmpSE										0.94	0.99	0.99
		MCSE										(0.02)	(0.02)	(0.02)
	800	EmpSE										0.84	0.84	0.85
		MCSE										(0.01)	(0.01)	(0.01)

Table E.25: ANCOVA method performance at each time point - Model-based standard error: Pattern 1

Time point	n	$\beta*$: Type	Time points: 2 4	8	12	Time points: 3 4	8	12	Time points: 4 4	8	12	Time points: 5 4	8	12
1	100	ModSE	2.41	2.39	2.40	2.41	2.40	2.40	2.40	2.41	2.40	2.40	2.40	2.40
	200	ModSE	1.69	1.69	1.69	1.69	1.69	1.69	1.69	1.69	1.69	1.68	1.68	1.69
	400	ModSE	1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19
	600	ModSE	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	800	ModSE	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
2	100	ModSE	2.40	2.39	2.38	2.40	2.39	2.40	2.40	2.40	2.39	2.40	2.40	2.40
	200	ModSE	1.69	1.68	1.67	1.69	1.68	1.68	1.69	1.69	1.69	1.69	1.68	1.68
	400	ModSE	1.19	1.19	1.18	1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19
	600	ModSE	0.97	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	800	ModSE	0.84	0.83	0.83	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
3	100	ModSE				2.40	2.39	2.38	2.39	2.40	2.38	2.39	2.39	2.39
	200	ModSE				1.68	1.68	1.67	1.69	1.68	1.68	1.68	1.68	1.68
	400	ModSE				1.19	1.18	1.18	1.19	1.19	1.19	1.19	1.19	1.19
	600	ModSE				0.97	0.97	0.96	0.97	0.97	0.97	0.97	0.97	0.97
	800	ModSE				0.84	0.84	0.83	0.84	0.84	0.84	0.84	0.84	0.84
4	100	ModSE							2.39	2.38	2.37	2.40	2.39	2.38
	200	ModSE							1.68	1.67	1.67	1.68	1.68	1.68
	400	ModSE							1.19	1.18	1.18	1.19	1.18	1.18
	600	ModSE							0.97	0.97	0.96	0.97	0.97	0.97
	800	ModSE							0.84	0.83	0.83	0.84	0.84	0.83

Table E.25: ANCOVA method performance at each time point - Model-based standard error: Pattern 1

Time point	n	$\beta*$: Type	Time points: 2 4	Time points: 2 8	Time points: 2 12	Time points: 3 4	Time points: 3 8	Time points: 3 12	Time points: 4 4	Time points: 4 8	Time points: 4 12	Time points: 5 4	Time points: 5 8	Time points: 5 12
5	100	ModSE										2.40	2.39	2.37
	200	ModSE										1.68	1.68	1.66
	400	ModSE										1.19	1.18	1.18
	600	ModSE										0.97	0.96	0.96
	800	ModSE										0.84	0.83	0.83

Table E.26: ANCOVA method performance at each time point - Model-based standard error: Pattern 2

Time point	n	$\beta*$: Type	Time points: 2 4	8	12	Time points: 3 4	8	12	Time points: 4 4	8	12	Time points: 5 4	8	12
1	100	ModSE	2.40	2.39	2.37	2.40	2.39	2.39	2.40	2.39	2.40	2.40	2.39	2.41
	200	ModSE	1.69	1.68	1.67	1.69	1.69	1.69	1.69	1.69	1.68	1.69	1.68	1.68
	400	ModSE	1.19	1.19	1.18	1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19
	600	ModSE	0.97	0.97	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	800	ModSE	0.84	0.84	0.83	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
2	100	ModSE	2.40	2.39	2.38	2.39	2.39	2.38	2.39	2.39	2.38	2.39	2.39	2.38
	200	ModSE	1.69	1.68	1.68	1.69	1.68	1.67	1.68	1.68	1.67	1.68	1.68	1.68
	400	ModSE	1.19	1.19	1.18	1.19	1.18	1.18	1.19	1.18	1.18	1.19	1.18	1.18
	600	ModSE	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.96	0.96	0.97	0.97	0.96
	800	ModSE	0.84	0.84	0.83	0.84	0.84	0.83	0.84	0.84	0.83	0.84	0.84	0.83
3	100	ModSE				2.39	2.39	2.37	2.40	2.39	2.38	2.39	2.39	2.38
	200	ModSE				1.68	1.68	1.67	1.68	1.68	1.67	1.69	1.68	1.67
	400	ModSE				1.19	1.18	1.18	1.19	1.18	1.18	1.19	1.18	1.18
	600	ModSE				0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.97	0.96
	800	ModSE				0.84	0.84	0.83	0.84	0.84	0.83	0.84	0.84	0.83
4	100	ModSE							2.40	2.38	2.38	2.39	2.39	2.38
	200	ModSE							1.68	1.68	1.67	1.69	1.68	1.67
	400	ModSE							1.19	1.18	1.18	1.19	1.18	1.18
	600	ModSE							0.97	0.97	0.96	0.97	0.97	0.96
	800	ModSE							0.84	0.83	0.83	0.84	0.84	0.83

Table E.26: ANCOVA method performance at each time point - Model-based standard error: Pattern 2

Time point	n	$\beta*$: Type	Time points: 2 4	Time points: 2 8	Time points: 2 12	Time points: 3 4	Time points: 3 8	Time points: 3 12	Time points: 4 4	Time points: 4 8	Time points: 4 12	Time points: 5 4	Time points: 5 8	Time points: 5 12
5	100	ModSE										2.39	2.39	2.38
	200	ModSE										1.68	1.68	1.67
	400	ModSE										1.19	1.18	1.18
	600	ModSE										0.97	0.96	0.96
	800	ModSE										0.84	0.84	0.83

Table E.27: ANCOVA method performance at each time point - Model-based standard error: Pattern 3

Time point	n	$\beta*$: Type	Time points: 2 4	8	12	Time points: 3 4	8	12	Time points: 4 4	8	12	Time points: 5 4	8	12
1	100	ModSE	2.40	2.39	2.37	2.41	2.39	2.37	2.39	2.38	2.37	2.38	2.38	2.37
	200	ModSE	1.69	1.68	1.66	1.69	1.68	1.67	1.68	1.68	1.67	1.69	1.68	1.66
	400	ModSE	1.19	1.18	1.18	1.19	1.18	1.18	1.19	1.18	1.18	1.19	1.18	1.18
	600	ModSE	0.97	0.97	0.96	0.97	0.97	0.96	0.97	0.96	0.96	0.97	0.96	0.96
	800	ModSE	0.84	0.84	0.83	0.84	0.84	0.83	0.84	0.83	0.83	0.84	0.83	0.83
2	100	ModSE	2.40	2.40	2.41	2.40	2.38	2.37	2.39	2.38	2.37	2.39	2.37	2.37
	200	ModSE	1.70	1.69	1.69	1.69	1.68	1.67	1.68	1.67	1.67	1.68	1.68	1.66
	400	ModSE	1.19	1.19	1.20	1.19	1.18	1.18	1.19	1.19	1.17	1.19	1.18	1.18
	600	ModSE	0.97	0.97	0.97	0.97	0.97	0.96	0.97	0.96	0.96	0.97	0.96	0.96
	800	ModSE	0.84	0.84	0.84	0.84	0.84	0.83	0.84	0.84	0.83	0.84	0.83	0.83
3	100	ModSE				2.40	2.41	2.41	2.41	2.40	2.40	2.40	2.40	2.40
	200	ModSE				1.70	1.69	1.69	1.69	1.69	1.69	1.69	1.69	1.69
	400	ModSE				1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19	1.19
	600	ModSE				0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
	800	ModSE				0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
4	100	ModSE							2.41	2.41	2.40	2.40	2.40	2.39
	200	ModSE							1.69	1.69	1.68	1.69	1.69	1.69
	400	ModSE							1.19	1.19	1.19	1.19	1.19	1.19
	600	ModSE							0.97	0.97	0.97	0.97	0.97	0.97
	800	ModSE							0.84	0.84	0.84	0.84	0.84	0.84

Table E.27: ANCOVA method performance at each time point - Model-based standard error: Pattern 3

Time point	n	$\beta*$: Type	Time points: 2			Time points: 3			Time points: 4			Time points: 5		
			4	8	12	4	8	12	4	8	12	4	8	12
5	100	ModSE										2.40	2.40	2.40
	200	ModSE										1.69	1.69	1.69
	400	ModSE										1.19	1.19	1.19
	600	ModSE										0.97	0.97	0.97
	800	ModSE										0.84	0.84	0.84

Appendix F

F.1 Publications and presentations

Publications

Chapter 2

Copsey B, Dutton S, Fitzpatrick R, Lamb SE, Cook JA: Current practice in methodology and reporting of the sample size calculation in randomised trials of hip and knee osteoarthritis: a protocol for a systematic review. Trials 2017, 18(1):466.

Copsey B, Thompson JY, Vadher K, Ali U, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA: Sample size calculations are poorly conducted and reported in many randomized trials of hip and knee osteoarthritis: results of a systematic review. Journal of Clinical Epidemiology 2018, 104:52-61.

Chapter 3

Copsey B, Thompson JY, Vadher K, Ali U, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA: Problems persist in reporting of methods and results for the WOMAC measure in hip and knee osteoarthritis trials. Quality of Life Research 2019, 28(2):335-343.

Chapter 4

Copsey B, Buchanan J, Fitzpatrick R, Lamb SE, Dutton SJ, Cook JA: Duration of treatment effect should be considered in the design and interpretation of clinical trials: Results of a discrete choice experiment. Medical Decision Making 2019, 39(4):461-473.

External oral presentations

Chapter 4

PRIFOR 2018: Copsey B, Buchanan J, Fitzpatrick R, Lamb SE, Dutton SJ, Cook JA. Development of a questionnaire for a discrete choice experiment: a case study on osteoarthritis medications. Oral presentation at Primary Healthcare Partnership Forum (St John's, Canada, June 2018).

PRIFOR 2019: Copsey B, Buchanan J, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA. What is important to osteoarthritis patients when choosing between medications?: Results of a discrete choice experiment. Oral presentation at Primary Healthcare Partnership Forum (St John's, Canada, June 2019).

Chapter 5

YSM 2019: Copsey B, Fitzpatrick R, Lamb SE, Dutton SJ, Cook JA. Should we account for duration of follow-up in the calculation of minimum clinically important differences? Oral presentation at Young Statisticians Meeting (Leeds, UK, August 2019).

External poster presentations

Chapter 2

OARSI 2018: Copsey B, Thompson JY, Vadher K, Ali U, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA. Current practice in methodology and reporting of the sample size calculation in randomised trials of hip and knee osteoarthritis: a systematic review. Poster presentation at Osteoarthritis Research Society International conference (Liverpool, UK, April 2018).

ISCB 2019: Copsey B, Thompson JY, Vadher K, Ali U, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA. Current practice in methodology of sample size calculation in randomised trials of osteoarthritis. Poster presentation at International Society for Clinical Biostatistics conference (Leuven, Belgium, July 2019).

Chapter 3

PROMS 2018: Copsey B, Thompson JY, Vadher K, Ali U, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA. Variation and poor reporting on the measurement of patient-reported outcomes can hinder the interpretation of study findings: a case study using the WOMAC measure. Poster presentation at Patient-reported Outcome Measures conference (Birmingham, UK, June 2018).

Chapter 4

YSM 2018 and RSS 2018: Copsey B, Buchanan J, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA. Using a discrete choice experiment of patient preference to inform the design and interpretation of clinical trials: a case study in osteoarthritis. Poster presentation at Young Statisticians Meeting (Oxford, UK, July 2018). Awarded best poster at YSM, where the prize was registration and presentation at Royal Statistical Society conference (Cardiff, UK, Sept 2018).

PROMS 2019: Copsey B, Buchanan J, Fitzpatrick R, Lamb SE, Dutton SJ, Cook JA. Using discrete choice experiments to test the validity of PROMs: an example using the WOMAC Index. Poster presentation at Patient-reported Outcome Measures conference (Leeds, UK, June 2019).

Chapter 6

ICTMC 2019: Copsey B, Dutton SJ, Fitzpatrick R, Lamb SE, Cook JA. A simulation study to compare longitudinal methods for the analysis of randomised trials and the implications for sample size calculation. Poster presentation at International Clinical Trials Methodology Conference (Brighton, UK, October 2019).

References

[1] A. van Walsem, S. Pandhi, R. M. Nixon, P. Guyot, A. Karabis, and R. A. Moore. Relative benefit-risk comparing diclofenac to other traditional non-steroidal anti-inflammatory drugs and cyclooxygenase-2 inhibitors in patients with osteoarthritis or rheumatoid arthritis: a network meta-analysis. *Arthritis Res Ther*, 17:66, 2015. ISSN 1478-6354. doi: 10.1186/s13075-015-0554-0.

[2] D. G. Altman. Statistics and ethics in medical research: III How large a sample? *BMJ*, 281(6251):1336–8, 1980. ISSN 0007-1447 (Print) 0007-1447.

[3] Stephen B Hulley, Steven R Cummings, Warren S Browner, Deborah G Grady, and Thomas B Newman. *Designing clinical research*. Lippincott Williams and Wilkins, 2013. ISBN 1469840545.

[4] Lawrence M Friedman, Curt Furberg, David L DeMets, David Reboussin, and Christopher B Granger. *Fundamentals of clinical trials*, volume 3. Springer, 1998.

[5] Byron W. Brown. Statistical controversies in the design of clinical trials - some personal views. *Controlled Clinical Trials*, 1(1):13–27, 1980. ISSN 0197-2456. doi: https://doi.org/10.1016/S0197-2456(80)80004-3.

[6] J. A. Cook, S. A. Julious, W. Sones, L. V. Hampson, C. Hewitt, J. A. Berlin, D. Ashby, R. Emsley, D. A. Fergusson, and et al. DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ*, 363, 2018. doi: 10.1136/bmj.k3750.

[7] E. Whitley and J. Ball. Statistics review 4: sample size calculations. *Crit Care*, 6(4):335–41, 2002. ISSN 1364-8535 (Print) 1364-8535.

[8] Stuart J Pocock. *Clinical trials: a practical approach*. John Wiley and Sons, 2013. ISBN 1118793927.

[9] J. A. Cook, J. Hislop, T. E. Adewuyi, K. Harrild, D. G. Altman, C. R. Ramsay, C. Fraser, B. Buckley, P. Fayers, I. Harvey, A. H. Briggs, J. D. Norrie, D. Fergusson, I. Ford, and L. D. Vale. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technol Assess*, 18(28):v–vi, 1–175, 2014. ISSN 1366-5278. doi: 10.3310/hta18280.

[10] R. Jaeschke, J. Singer, and G. H. Guyatt. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*, 10 (4):407–15, 1989. ISSN 0197-2456 (Print) 0197-2456.

[11] A. Wright, J. Hannon, E. J. Hegedus, and A. E. Kavchak. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). *J Man Manip Ther*, 20(3):160–6, 2012. ISSN 1066-9817 (Print) 1066-9817. doi: 10.1179/2042618612y.0000000001.

[12] C. E. Cook. Clinimetrics corner: The minimal clinically important change score (MCID): A necessary pretense. *J Man Manip Ther*, 16(4):E82–3, 2008. ISSN 1066-9817 (Print) 1066-9817. doi: 10.1179/jmt.2008.16.4.82E.

[13] A. G. Copay, B. R. Subach, S. D. Glassman, Jr. Polly, D. W., and T. C. Schuler. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J*, 7(5):541–6, 2007. ISSN 1529-9430 (Print) 1529-9430. doi: 10.1016/j.spinee.2007.01.008.

[14] H. C. de Vet and C. B. Terwee. The minimal detectable change should not replace the minimal important difference. *J Clin Epidemiol*, 63(7):804–5; author reply 806, 2010. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2009.12.015.

[15] A. F. Smelt, W. J. Assendelft, C. B. Terwee, M. D. Ferrari, and J. W. Blom. What is a clinically relevant change on the HIT-6 questionnaire? An estimation in a primary-care population of migraine patients. *Cephalalgia*, 34(1):29–36, 2014. ISSN 0333-1024. doi: 10.1177/0333102413497599.

[16] X. M. Teitsma, J. W. G. Jacobs, P. M. J. Welsing, A. Petho-Schramm, M. E. A. Borm, L. Hendriks, N. H. A. M. Denissen, J. M. van Laar, F. P. J. G. Lafeber, and J. W. J. Bijlsma. Patient-reported outcomes in newly diagnosed early rheumatoid arthritis patients treated to target with a tocilizumab- or methotrexate-based strategy. *Rheumatology (Oxford)*, 56(12):2179–2189, 2017. ISSN 1462-0324. doi: 10.1093/rheumatology/kex319.

[17] K. Deckers, K. De Smedt, B. Mitchell, D. Vivian, M. Russo, P. Georgius, M. Green, J. Vieceli, S. Eldabe, A. Gulve, J. P. van Buyten, I. Smet, V. Mehta, S. Ramaswamy, G. Baranidharan, R. Sullivan, R. Gassin, J. Rathmell, and C. Gilligan. New therapy for refractory chronic mechanical low back pain-restorative neurostimulation to activate the lumbar multifidus: One year results of a prospective multicenter clinical trial. *Neuromodulation*, 21(1):48–55, 2018. ISSN 1094-7159. doi: 10.1111/ner.12741.

[18] M. H. L. Liow, G. S. Goh, M. K. Wong, P. L. Chin, D. K. Tay, and S. J. Yeo. Robotic-assisted total knee arthroplasty may lead to improvement in quality-of-life measures: a 2-year follow-up of a prospective randomized trial. *Knee Surg Sports Traumatol Arthrosc*, 25(9):2942–2951, 2017. ISSN 0942-2056. doi: 10.1007/s00167-016-4076-3.

[19] Y. Y. Leung, B. Haaland, J. L. Huebner, S. B. S. Wong, M. Tjai,

C. Wang, B. Chowbay, J. Thumboo, B. Chakraborty, M. H. Tan, and V. B. Kraus. Colchicine lack of effectiveness in symptom and inflammation modification in knee osteoarthritis (COLKOA): a randomized controlled trial. *Osteoarthritis Cartilage*, 26(5):631–640, 2018. ISSN 1063-4584. doi: 10.1016/j.joca.2018.01.026.

[20] B. W. Carey and J. Harty. A comparison of clinical- and patient-reported outcomes of the cemented ATTUNE and PFC sigma fixed bearing cruciate sacrificing knee systems in patients who underwent total knee replacement with both prostheses in opposite knees. *J Orthop Surg Res*, 13(1):54, 2018. ISSN 1749-799x. doi: 10.1186/s13018-018-0757-6.

[21] D. Thiam, D. J. Teh, H. R. Bin Abd Razak, and A. H. Tan. Improvement in health-related quality of life after unilateral total knee arthroplasty in patients with bilateral knee osteoarthritis. *J Orthop Surg (Hong Kong)*, 24(3):294–297, 2016. ISSN 1022-5536. doi: 10.1177/1602400304.

[22] D. S. Dong, X. Yu, C. F. Wan, Y. Liu, L. Zhao, Q. Xi, W. Y. Cui, Q. S. Wang, and T. Song. Efficacy of short-term spinal cord stimulation in acute/subacute zoster-related pain: A retrospective study. *Pain Physician*, 20(5):E633–e645, 2017. ISSN 1533-3159.

[23] P. J. van der Wees, J. J. Wammes, R. P. Akkermans, J. Koetsenruijter, G. P. Westert, A. van Kampen, G. Hannink, M. de Waal-Malefijt, and B. W. Schreurs. Patient-reported health outcomes after total hip and knee surgery in a Dutch university hospital setting: results of twenty years clinical registry. *BMC Musculoskelet Disord*, 18(1):97, 2017. ISSN 1471-2474. doi: 10.1186/s12891-017-1455-y.

[24] B. G. Domb, E. O. Chaharbakhshi, I. Perets, J. P. Walsh, L. C. Yuen, and L. J. Ashberg. Patient-reported outcomes of capsular repair versus capsulotomy in patients undergoing hip arthroscopy: Minimum 5-year follow-up-a matched comparison study. *Arthroscopy*, 2018. ISSN 0749-8063. doi: 10.1016/j.arthro.2017.10.019.

[25] D. E. Hartigan, I. Perets, L. C. Yuen, and B. G. Domb. Results of hip arthroscopy in patients with MRI diagnosis of subchondral cysts - a case series. *J Hip Preserv Surg*, 4(4):324–331, 2017. ISSN 2054-8397 (Print) 2054-8397. doi: 10.1093/jhps/hnx034.

[26] A. Ruhdorfer, W. Wirth, and F. Eckstein. Relationship between isometric thigh muscle strength and minimum clinically important differences in knee function in osteoarthritis: data from the osteoarthritis initiative. *Arthritis Care Res (Hoboken)*, 67(4):509–18, 2015. ISSN 2151-464x. doi: 10.1002/acr.22488.

[27] E. R. Vina, M. J. Hannon, and C. K. Kwoh. Improvement following total knee replacement surgery: Exploring preoperative symptoms and change in preoperative symptoms. *Semin Arthritis Rheum*, 45(5):547–55, 2016. ISSN 0049-0172. doi: 10.1016/j.semarthrit.2015.10.002.

[28] J. L. Berliner, D. J. Brodke, V. Chan, N. F. SooHoo, and K. J. Bozic. Can preoperative patient-reported outcome measures be used to predict meaningful improvement in function after TKA? *Clin Orthop Relat Res*, 475(1):149–157, 2017. ISSN 0009-921x. doi: 10.1007/s11999-016-4770-y.

[29] H. R. Bin Abd Razak, C. S. Tan, Y. J. Chen, H. N. Pang, K. J. Tay, P. L. Chin, S. L. Chia, N. N. Lo, and S. J. Yeo. Age and preoperative knee society score are significant predictors of outcomes among asians following total knee arthroplasty. *J Bone Joint Surg Am*, 98(9):735–41, 2016. ISSN 0021-9355. doi: 10.2106/jbjs.15.00280.

[30] G. A. Karpouzas, T. Draper, R. Moran, E. Hernandez, P. Nicassio, M. H. Weisman, and S. Ormseth. Trends in functional disability and determinants of clinically meaningful change over time in hispanic patients with rheumatoid arthritis in the US. *Arthritis Care Res (Hoboken)*, 69(2):294–298, 2017. ISSN 2151-464x. doi: 10.1002/acr.22924.

[31] E. C. Sayre, A. Guermazi, J. M. Esdaile, J. A. Kopec, J. Singer, A. Thorne, S. Nicolaou, and J. Cibere. Associations between MRI features versus knee pain severity and progression: Data from the Vancouver Longitudinal Study of Early Knee Osteoarthritis. *PLoS One*, 12(5):e0176833, 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0176833.

[32] K. Kurosaka, S. Tsukada, H. Nakayama, T. Iseki, R. Kanto, R. Sugama, and S. Yoshiya. Periarticular injection versus femoral nerve block for pain relief after anterior cruciate ligament reconstruction: A randomized controlled trial. *Arthroscopy*, 34(1):182–188, 2018. ISSN 0749-8063. doi: 10.1016/j.arthro.2017.08.307.

[33] W. L. Dai, A. G. Zhou, H. Zhang, and J. Zhang. Efficacy of platelet-rich plasma in the treatment of knee osteoarthritis: A meta-analysis of randomized controlled trials. *Arthroscopy*, 33(3):659–670.e1, 2017. ISSN 0749-8063. doi: 10.1016/j.arthro.2016.09.024.

[34] D. M. Levy, B. D. Kuhns, J. Chahal, M. J. Philippon, B. T. Kelly, and S. J. Nho. Hip arthroscopy outcomes with respect to patient acceptable symptomatic state and minimal clinically important difference. *Arthroscopy*, 32(9):1877–86, 2016. ISSN 0749-8063. doi: 10.1016/j.arthro.2016.05.014.

[35] B. M. Saltzman, T. Leroux, M. A. Meyer, B. A. Basques, J. Chahal, Jr. Bach, B. R., A. B. Yanke, and B. J. Cole. The therapeutic effect of intra-articular normal saline injections for knee osteoarthritis: A meta-analysis of evidence level 1 studies. *Am J Sports Med*, 45(11):2647–2653, 2017. ISSN 0363-5465. doi: 10.1177/0363546516680607.

[36] Rhma Bartels, R. D. Donk, W. I. M. Verhagen, A. J. F. Hosman, and A. L. M. Verbeek. Reporting the results of meta-analyses: a plea for incorporating clinical relevance referring to an example. *Spine J*, 17(11):1625–1632, 2017. ISSN 1529-9430. doi: 10.1016/j.spinee.2017.05.019.

[37] B. U. Nwachukwu, R. S. Runyon, C. A. Kahlenberg, E. B. Gausden, W. W. Schairer, and A. A. Allen. How are we measuring clinically important outcome for operative treatments in sports medicine? *Phys Sportsmed*, 45(2):159–164, 2017. ISSN 0091-3847. doi: 10.1080/00913847.2017.1292108.

[38] M. F. Olsen, E. Bjerre, M. D. Hansen, J. Hilden, N. E. Landler, B. Tendal, and A. Hrobjartsson. Pain relief that matters to patients: systematic review of empirical studies assessing the minimum clinically important difference in acute pain. *BMC Med*, 15(1):35, 2017. ISSN 1741-7015. doi: 10.1186/s12916-016-0775-3.

[39] F. Tubach, P. Ravaud, G. Baron, B. Falissard, I. Logeart, N. Bellamy, C. Bombardier, D. Felson, M. Hochberg, D. van der Heijde, and M. Dougados. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. *Ann Rheum Dis*, 64(1):29–33, 2005. ISSN 0003-4967 (Print) 0003-4967. doi: 10.1136/ard.2004.022905.

[40] V. Corallo, M. Torre, G. Ferrara, F. Guerra, G. Nicosia, E. Romanelli, A. Lopopolo, M. P. Onesta, P. Fiore, R. Falcone, J. Bonavita, M. Molinari, and G. Scivoletto. What do spinal cord injury patients think of their improvement? A study of the minimal clinically important difference of the Spinal Cord Independence Measure III. *Eur J Phys Rehabil Med*, 53(4):508–515, 2017. ISSN 1973-9087. doi: 10.23736/s1973-9087.17.04240-x.

[41] R. Z. Tashjian, M. Hung, J. D. Keener, R. C. Bowen, J. McAllister, W. Chen, G. Ebersole, E. K. Granger, and A. M. Chamberlain. Determining the minimal clinically important difference for the american shoulder and elbow surgeons score, simple shoulder test, and visual analog scale (VAS) measuring pain after shoulder arthroplasty. *J Shoulder Elbow Surg*, 26(1):144–148, 2017. ISSN 1058-2746. doi: 10.1016/j.jse.2016.06.007.

[42] U. Bingel, V. Wanigasekera, K. Wiech, R. Ni Mhuircheartaigh, M. C. Lee, M. Ploner, and I. Tracey. The effect of treatment expectation on drug efficacy: imaging the analgesic benefit of the opioid remifentanil. *Sci Transl Med*, 3(70): 70ra14, 2011. ISSN 1946-6234. doi: 10.1126/scitranslmed.3001244.

[43] K. Wiech. Deconstructing the sensation of pain: The influence of cognitive processes on pain perception. *Science*, 354(6312):584–587, 2016. ISSN 0036-8075. doi: 10.1126/science.aaf8934.

[44] R. R. Edwards, C. Cahalan, G. Mensing, M. Smith, and J. A. Haythornthwaite. Pain, catastrophizing, and depression in the rheumatic diseases. *Nat Rev Rheumatol*, 7(4):216–24, 2011. ISSN 1759-4790. doi: 10.1038/nrrheum.2011.2.

[45] H. Radner, K. Yoshida, S. Tedeschi, P. Studenic, M. Frits, C. Iannaccone, N. A. Shadick, M. Weinblatt, D. Aletaha, J. S. Smolen, and D. H. Solomon. Different rating of global rheumatoid arthritis disease activity in rheumatoid arthritis

patients with multiple morbidities. *Arthritis Rheumatol*, 69(4):720–727, 2017. ISSN 2326-5191. doi: 10.1002/art.39988.

[46] C. Karagiannopoulos, M. Sitler, S. Michlovitz, C. Tucker, and R. Tierney. Responsiveness of the active wrist joint position sense test after distal radius fracture intervention. *J Hand Ther*, 29(4):474–482, 2016. ISSN 0894-1130. doi: 10.1016/j.jht.2016.06.009.

[47] M. Maltenfort and C. Diaz-Ledezma. Statistics in brief: Minimum clinically important difference-availability of reliable estimates. *Clin Orthop Relat Res*, 475(4):933–946, 2017. ISSN 0009-921x. doi: 10.1007/s11999-016-5204-6.

[48] S. Gupta and K. Raja. Responsiveness of Edinburgh Visual Gait Score to orthopedic surgical intervention of the lower limbs in children with cerebral palsy. *Am J Phys Med Rehabil*, 91(9):761–7, 2012. ISSN 0894-9115. doi: 10.1097/PHM.0b013e31825f1c4d.

[49] R. Simovitch, P. H. Flurin, T. Wright, J. D. Zuckerman, and C. P. Roche. Quantifying success after total shoulder arthroplasty: the minimal clinically important difference. *J Shoulder Elbow Surg*, 27(2):298–305, 2018. ISSN 1058-2746. doi: 10.1016/j.jse.2017.09.013.

[50] J. A. Singh, C. Schleck, S. Harmsen, and D. Lewallen. Clinically important improvement thresholds for Harris Hip Score and its ability to predict revision risk after primary total hip arthroplasty. *BMC Musculoskelet Disord*, 17:256, 2016. ISSN 1471-2474. doi: 10.1186/s12891-016-1106-8.

[51] H. J. Alma, C. de Jong, D. Jelusic, M. Wittmann, M. Schuler, B. J. Kollen, R. Sanderman, K. Schultz, J. W. H. Kocks, and T. Van der Molen. Assessing health status over time: impact of recall period and anchor question on the minimal clinically important difference of COPD health status tools. *Health Qual Life Outcomes*, 16(1):130, 2018. ISSN 1477-7525. doi: 10.1186/s12955-018-0950-7.

[52] M. van der Pol and J. Cairns. Estimating time preferences for health using discrete choice experiments. *Soc Sci Med*, 52(9):1459–70, 2001. ISSN 0277-9536 (Print) 0277-9536.

[53] R. Z. Tashjian, J. Deloach, C. A. Porucznik, and A. P. Powell. Minimal clinically important differences (MCID) and patient acceptable symptomatic state (PASS) for visual analog scales (VAS) measuring pain in patients treated for rotator cuff disease. *J Shoulder Elbow Surg*, 18(6):927–32, 2009. ISSN 1058-2746. doi: 10.1016/j.jse.2009.03.021.

[54] B. Speich, B. von Niederhausern, N. Schur, L. G. Hemkens, T. Furst, N. Bhatnagar, R. Alturki, A. Agarwal, B. Kasenda, C. Pauli-Magnus, M. Schwenkglenks, and M. Briel. Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data. *J Clin Epidemiol*, 96:1–11, 2018. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2017.12.018.

[55] R. C. Lawrence, D. T. Felson, C. G. Helmick, L. M. Arnold, H. Choi, R. A. Deyo, S. Gabriel, R. Hirsch, M. C. Hochberg, G. G. Hunder, J. M. Jordan, J. N. Katz, H. M. Kremers, and F. Wolfe. Estimates of the prevalence of arthritis and other rheumatic conditions in the united states. part II. *Arthritis Rheum*, 58(1):26–35, 2008. ISSN 0004-3591 (Print) 0004-3591. doi: 10.1002/art.23176.

[56] J. Menon and P. Mishra. Health care resource use, health care expenditures and absenteeism costs associated with osteoarthritis in US healthcare system. *Osteoarthritis Cartilage*, 26(4):480–484, 2018. ISSN 1063-4584. doi: 10.1016/j.joca.2017.12.007.

[57] R. Bitton. The economic burden of osteoarthritis. *Am J Manag Care*, 15(8 Suppl):S230–5, 2009. ISSN 1088-0224.

[58] R. D. Altman. Early management of osteoarthritis. *Am J Manag Care*, 16 Suppl Management:S41–7, 2010. ISSN 1088-0224.

[59] T. Vos, A. D. Flaxman, M. Naghavi, R. Lozano, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, V. Aboyans, and et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the global burden of disease study 2010. *Lancet*, 380 (9859):2163–96, 2012. ISSN 0140-6736. doi: 10.1016/s0140-6736(12)61729-2.

[60] Centre National Clinical Guideline. *National Institute for Health and Clinical Excellence: Guidance*, book section CG177, pages 1–30. National Institute for Health and Care Excellence (UK) Copyright (c) National Clinical Guideline Centre, 2014., London, 2014.

[61] S. Krasnokutsky, J. Samuels, and S. B. Abramson. Osteoarthritis in 2007. *Bull NYU Hosp Jt Dis*, 65(3):222–8, 2007. ISSN 1936-9719 (Print) 1936-9719.

[62] A. D. Woolf and B. Pfleger. Burden of major musculoskeletal conditions. *Bull World Health Organ*, 81(9):646–56, 2003. ISSN 0042-9686 (Print) 0042-9686.

[63] A. J. Carr, O. Robertsson, S. Graves, A. J. Price, N. K. Arden, A. Judge, and D. J. Beard. Knee replacement. *Lancet*, 379(9823):1331–40, 2012. ISSN 0140-6736. doi: 10.1016/s0140-6736(11)60752-6.

[64] D. J. Culliford, J. Maskell, D. J. Beard, D. W. Murray, A. J. Price, and N. K. Arden. Temporal trends in hip and knee replacement in the United Kingdom: 1991 to 2006. *J Bone Joint Surg Br*, 92(1):130–5, 2010. ISSN 0301-620x. doi: 10.1302/0301-620x.92b1.22654.

[65] J. W. Bijlsma, F. Berenbaum, and F. P. Lafeber. Osteoarthritis: an update with relevance for clinical practice. *Lancet*, 377(9783):2115–26, 2011. ISSN 0140-6736. doi: 10.1016/s0140-6736(11)60243-2.

[66] R. Altman, E. Asch, D. Bloch, G. Bole, D. Borenstein, K. Brandt, W. Christy, T. D. Cooke, R. Greenwald, M. Hochberg, and et al. Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis

of the knee. Diagnostic and Therapeutic Criteria Committee of the American Rheumatism Association. *Arthritis Rheum*, 29(8):1039–49, 1986. ISSN 0004-3591 (Print) 0004-3591.

[67] R. Altman, G. Alarcon, D. Appelrouth, D. Bloch, D. Borenstein, K. Brandt, C. Brown, T. D. Cooke, W. Daniel, D. Feldman, and et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheum*, 34(5):505–14, 1991. ISSN 0004-3591 (Print) 0004-3591.

[68] G. Jones. What's new in osteoarthritis pathogenesis? *Intern Med J*, 46(2): 229–36, 2016. ISSN 1444-0903. doi: 10.1111/imj.12763.

[69] B. Xia, Chen Di, J. Zhang, S. Hu, H. Jin, and P. Tong. Osteoarthritis pathogenesis: a review of molecular mechanisms. *Calcif Tissue Int*, 95(6):495–505, 2014. ISSN 0171-967x. doi: 10.1007/s00223-014-9917-9.

[70] T. E. McAlindon, S. Snow, C. Cooper, and P. A. Dieppe. Radiographic patterns of osteoarthritis of the knee joint in the community: the importance of the patellofemoral joint. *Ann Rheum Dis*, 51(7):844–9, 1992. ISSN 0003-4967 (Print) 0003-4967.

[71] M. Lequesne, M. Dougados, M. Abiteboul, D. Bontoux, G. Bouvenot, V. Chicheportiche, R. L. Dreiser, R. Dropsy, E. Maheu, B. Mazieres, and et al. [How to evaluate the long-term course of osteoarthritis. Tests for trials of fundamental treatments (spine excluded)]. *Rev Rhum Mal Osteoartic*, 57(9 (Pt 2)):24s–31s, 1990. ISSN 0035-2659 (Print) 0035-2659.

[72] M. G. J. Gademan, H. Putter, W. B. Van Den Hout, M. Kloppenburg, S. N. Hofstede, S. C. Cannegieter, R. G. H. H. Nelissen, and P. J. Marang-Van De Mheen. The course of pain and function in osteoarthritis and timing of arthroplasty: the CHECK cohort. *Acta Orthop*, 89(5):528–534, 2018. ISSN 1745-3674. doi: 10.1080/17453674.2018.1502533.

[73] H. S. Chua, S. L. Whitehouse, M. Lorimer, R. De Steiger, L. Guo, and R. W. Crawford. Mortality and implant survival with simultaneous and staged bilateral total knee arthroplasty experience from the Australian Orthopaedic Association National Joint Replacement Registry. *J Arthroplasty*, 33(10):3167–3173, 2018. ISSN 0883-5403. doi: 10.1016/j.arth.2018.05.019.

[74] L. Churchill, S. J. Malian, B. M. Chesworth, D. Bryant, S. J. MacDonald, J. D. Marsh, and J. R. Giffin. The development and validation of a multivariable model to predict whether patients referred for total knee replacement are suitable surgical candidates at the time of initial consultation. *Can J Surg*, 59(6): 407–414, 2016. ISSN 0008-428x. doi: 10.1503/cjs.004316.

[75] Russell V Lenth. Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3):187–193, 2001. ISSN 0003-1305.

[76] S. D. Halpern, J. H. Karlawish, and J. A. Berlin. The continuing unethical conduct of underpowered clinical trials. *Jama*, 288(3):358–62, 2002. ISSN 0098-7484 (Print) 0098-7484.

[77] D. G. Altman, D. Moher, and K. F. Schulz. Peer review of statistics in medical research. Reporting power calculations is important. *BMJ*, 325(7362):491; author reply 491, 2002. ISSN 0959-535x.

[78] P Williamson, JL Hutton, J Bliss, J Blunt, MJ Campbell, and R Nicholson. Statistical review by research ethics committees. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):5–13, 2000. ISSN 1467-985X.

[79] T. Clark, U. Berger, and U. Mansmann. Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees: review. *BMJ*, 346:f1135, 2013. ISSN 0959-535x. doi: 10.1136/bmj.f1135.

[80] S. Fernandes-Taylor, J. K. Hyun, R. N. Reeder, and A. H. Harris. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC Res Notes*, 4:304, 2011. ISSN 1756-0500. doi: 10.1186/1756-0500-4-304.

[81] P. Charles, B. Giraudeau, A. Dechartres, G. Baron, and P. Ravaud. Reporting of sample size calculation in randomised controlled trials: review. *BMJ*, 338: b1732, 2009. ISSN 0959-535x. doi: 10.1136/bmj.b1732.

[82] J. C. Rothwell, S. A. Julious, and C. L. Cooper. A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal. *Trials*, 19(1):544, 2018. ISSN 1745-6215. doi: 10.1186/s13063-018-2886-y.

[83] C. Rutterford, M. Taljaard, S. Dixon, A. Copas, and S. Eldridge. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *J Clin Epidemiol*, 68(6):716–23, 2015. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2014.10.006.

[84] S. J. Arnup, A. B. Forbes, B. C. Kahan, K. E. Morgan, and J. E. McKenzie. The quality of reporting in cluster randomised crossover trials: proposal for reporting items and an assessment of reporting quality. *Trials*, 17(1):575, 2016. ISSN 1745-6215. doi: 10.1186/s13063-016-1685-6.

[85] J. Martin, M. Taljaard, A. Girling, and K. Hemming. Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. *BMJ Open*, 6(2):e010166, 2016. ISSN 2044-6055. doi: 10.1136/bmjopen-2015-010166.

[86] A. McKeown, J. S. Gewandter, M. P. McDermott, J. R. Pawlowski, J. J. Poli, D. Rothstein, J. T. Farrar, I. Gilron, N. P. Katz, A. H. Lin, B. A. Rappaport, M. C. Rowbotham, D. C. Turk, R. H. Dworkin, and S. M. Smith. Reporting of sample size calculations in analgesic clinical trials: ACTTION

systematic review. *J Pain*, 16(3):199–206.e1–7, 2015. ISSN 1526-5900. doi: 10.1016/j.jpain.2014.11.010.

[87] M. Abdulatif, A. Mukhtar, and G. Obayah. Pitfalls in reporting sample size calculation in randomized controlled trials published in leading anaesthesia journals: a systematic review. *Br J Anaesth*, 115(5):699–707, 2015. ISSN 0007-0912. doi: 10.1093/bja/aev166.

[88] B. Olberg, M. Perleth, K. Felgentraeger, S. Schulz, and R. Busse. Quality of sample size estimation in trials of medical devices: High-risk devices for neurological conditions as example. *Int J Technol Assess Health Care*, 33(1): 103–110, 2017. ISSN 0266-4623. doi: 10.1017/s0266462317000265.

[89] E. Tavernier and B. Giraudeau. Sample size calculation: Inaccurate a priori assumptions for nuisance parameters can greatly affect the power of a randomized controlled trial. *PLoS One*, 10(7):e0132578, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0132578.

[90] A. J. Vickers. Underpowering in randomized trials reporting a sample size calculation. *J Clin Epidemiol*, 56(8):717–20, 2003. ISSN 0895-4356 (Print) 0895-4356.

[91] H. Chen, N. Zhang, X. Lu, and S. Chen. Caution regarding the choice of standard deviations to guide sample size calculations in clinical trials. *Clin Trials*, 10(4):522–9, 2013. ISSN 1740-7745. doi: 10.1177/1740774513490250.

[92] E. M. Balk, P. A. Bonis, H. Moskowitz, C. H. Schmid, J. P. Ioannidis, C. Wang, and J. Lau. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *Jama*, 287(22):2973–82, 2002. ISSN 0098-7484 (Print) 0098-7484.

[93] J. A. Cook, P. McCulloch, J. M. Blazeby, D. J. Beard, D. Marinac-Dabic, and A. Sedrakyan. Ideal framework for surgical innovation 3: randomised controlled trials in the assessment stage and evaluations in the long term study stage. *BMJ*, 346:f2820, 2013. ISSN 0959-535x. doi: 10.1136/bmj.f2820.

[94] D. M. Wenner, B. A. Brody, A. F. Jarman, J. M. Kolman, N. P. Wray, and C. M. Ashton. Do surgical trials meet the scientific standards for clinical trials? *J Am Coll Surg*, 215(5):722–30, 2012. ISSN 1072-7515. doi: 10.1016/j.jamcollsurg.2012.06.018.

[95] ICH Harmonised Tripartite Guideline. Choice of control group and related issues in clinical trials E10. *Choice*, page E10, 2000.

[96] R. Mhaskar, B. Djulbegovic, A. Magazin, H. P. Soares, and A. Kumar. Published methodological quality of randomized controlled trials does not reflect the actual quality assessed in protocols. *J Clin Epidemiol*, 65(6):602–9, 2012. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2011.10.016.

[97] A. Bafeta, A. Dechartres, L. Trinquart, A. Yavchitz, I. Boutron, and P. Ravaud.

Impact of single centre status on estimates of intervention effects in trials with continuous outcomes: meta-epidemiological study. *BMJ*, 344:e813, 2012. ISSN 0959-535x. doi: 10.1136/bmj.e813.

[98] L. H. Pengel, L. Barcena, and P. J. Morris. The quality of reporting of randomized controlled trials in solid organ transplantation. *Transpl Int*, 22 (4):377–84, 2009. ISSN 0934-0874 (Print) 0934-0874. doi: 10.1111/j.1432-2277.2008.00789.x.

[99] B. Djulbegovic, M. Lacevic, A. Cantor, K. K. Fields, C. L. Bennett, J. R. Adams, N. M. Kuderer, and G. H. Lyman. The uncertainty principle and industry-sponsored research. *Lancet*, 356(9230):635–8, 2000. ISSN 0140-6736 (Print) 0140-6736. doi: 10.1016/s0140-6736(00)02605-2.

[100] A. Lundh, S. Sismondo, J. Lexchin, O. A. Busuioc, and L. Bero. Industry sponsorship and research outcome. *Cochrane Database Syst Rev*, 12:Mr000033, 2012. ISSN 1361-6137. doi: 10.1002/14651858.MR000033.pub2.

[101] G. Schott, H. Pachl, U. Limbach, U. Gundert-Remy, K. Lieb, and W. D. Ludwig. The financing of drug trials by pharmaceutical companies and its consequences: part 2: a qualitative, systematic review of the literature on possible influences on authorship, access to trial data, and trial registration and publication. *Dtsch Arztebl Int*, 107(17):295–301, 2010. ISSN 1866-0452. doi: 10.3238/arztebl.2010.0295.

[102] R. Froud, D. Rajendran, S. Patel, P. Bright, T. Bjorkli, S. Eldridge, R. Buchbinder, and M. Underwood. The power of low back pain trials: A systematic review of power, sample size, and reporting of sample size calculations over time, in trials published between 1980 and 2012. *Spine (Phila Pa 1976)*, 42 (11):E680–e686, 2017. ISSN 0362-2436. doi: 10.1097/brs.0000000000001953.

[103] H. I. Keen, K. Pile, and C. L. Hill. The prevalence of underpowered randomized clinical trials in rheumatology. *J Rheumatol*, 32(11):2083–8, 2005. ISSN 0315-162X (Print) 0315-162x.

[104] L. Abdul Latif, J. E. Daud Amadera, D. Pimentel, T. Pimentel, and F. Fregni. Sample size calculation in physical medicine and rehabilitation: a systematic review of reporting, characteristics, and results in randomized controlled trials. *Arch Phys Med Rehabil*, 92(2):306–15, 2011. ISSN 0003-9993. doi: 10.1016/j.apmr.2010.10.003.

[105] G. Castellini, S. Gianola, S. Bonovas, and L. Moja. Improving power and sample size calculation in rehabilitation trial reports: A methodological assessment. *Arch Phys Med Rehabil*, 97(7):1195–201, 2016. ISSN 0003-9993. doi: 10.1016/j.apmr.2016.02.013.

[106] A. Rath, V. Salamon, S. Peixoto, V. Hivert, M. Laville, B. Segrestin, E. A. M. Neugebauer, M. Eikermann, V. Bertele, S. Garattini, J. Wetterslev, R. Banzi, J. C. Jakobsen, S. Djurisic, C. Kubiak, J. Demotes-Mainard, and C. Gluud.

A systematic literature review of evidence-based clinical practice for rare diseases: what are the perceived and real barriers for improving the evidence and how can they be overcome? *Trials*, 18(1):556, 2017. ISSN 1745-6215. doi: 10.1186/s13063-017-2287-7.

[107] J. M. Gierisch, E. R. Myers, K. M. Schmit, D. C. McCrory, R. R. Coeytaux, M. J. Crowley, R. Chatterjee, A. S. Kendrick, and G. D. Sanders. Prioritization of patient-centered comparative effectiveness research for osteoarthritis. *Ann Intern Med*, 160(12):836–41, 2014. ISSN 0003-4819. doi: 10.7326/m14-0318.

[108] M. Cross, E. Smith, D. Hoy, S. Nolte, I. Ackerman, M. Fransen, L. Bridgett, S. Williams, F. Guillemin, C. L. Hill, L. L. Laslett, G. Jones, F. Cicuttini, R. Osborne, T. Vos, R. Buchbinder, A. Woolf, and L. March. The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. *Ann Rheum Dis*, 73(7):1323–30, 2014. ISSN 0003-4967. doi: 10.1136/annrheumdis-2013-204763.

[109] D. Koletsi, P. S. Fleming, J. Seehra, P. G. Bagos, and N. Pandis. Are sample sizes clear and justified in RCTs published in dental journals? *PLoS One*, 9(1): e85949, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0085949.

[110] K. A. McKibbon, N. L. Wilczynski, and R. B. Haynes. Retrieving randomized controlled trials from MEDLINE: a comparison of 38 published search filters. *Health Info Libr J*, 26(3):187–202, 2009. ISSN 1471-1834 (Print) 1471-1834. doi: 10.1111/j.1471-1842.2008.00827.x.

[111] A. M. Eady, N. L. Wilczynski, and R. B. Haynes. PsycINFO search strategies identified methodologically sound therapy studies and review articles for use by clinicians and researchers. *J Clin Epidemiol*, 61(1):34–40, 2008. ISSN 0895-4356 (Print) 0895-4356. doi: 10.1016/j.jclinepi.2006.09.016.

[112] S. S. Wong, N. L. Wilczynski, and R. B. Haynes. Optimal CINAHL search strategies for identifying therapy studies and review articles. *J Nurs Scholarsh*, 38(2):194–9, 2006. ISSN 1527-6546 (Print) 1527-6546.

[113] S. S. Wong, N. L. Wilczynski, and R. B. Haynes. Developing optimal search strategies for detecting clinically sound treatment studies in EMBASE. *J Med Libr Assoc*, 94(1):41–7, 2006. ISSN 1536-5050.

[114] J. R. Lyttle, D. M. Urquhart, F. M. Cicuttini, and A. E. Wluka. Antidepressants for osteoarthritis. *Cochrane Database Syst Rev*, 2016(4), 2016. ISSN 1465-1858. doi: 10.1002/14651858.CD012157.

[115] J. S. Palmer, A. P.l Monk, S. Hopewell, L. E. Bayliss, W. Jackson, D. J. Beard, and A. J. Price. Surgical interventions for early structural knee osteoarthritis. *Cochrane Database Syst Rev*, 2016(3), 2016. ISSN 1465-1858. doi: 10.1002/14651858.CD012128.

[116] J. P. Regnaux, M. M. Lefevre-Colau, L. Trinquart, C. Nguyen, I. Boutron, L. Brosseau, and P. Ravaud. High-intensity versus low-intensity physical activ-

ity or exercise in people with hip or knee osteoarthritis. *Cochrane Database Syst Rev*, 2015(10), 2015. ISSN 1465-1858. doi: 10.1002/14651858.CD010203.pub2.

[117] J. A. Singh, S. Noorbaloochi, R. MacDonald, and L. J. Maxwell. Chondroitin for osteoarthritis. *Cochrane Database Syst Rev*, 1:Cd005614, 2015. ISSN 1361-6137. doi: 10.1002/14651858.CD005614.pub2.

[118] F. P. Kroon, L. R. van der Burg, R. Buchbinder, R. H. Osborne, R. V. Johnston, and V. Pitt. Self-management education programmes for osteoarthritis. *Cochrane Database Syst Rev*, 2014(1):Cd008963, 2014. ISSN 1361-6137. doi: 10.1002/14651858.CD008963.pub2.

[119] A. G. Witteveen, C. J. Hofstad, and G. M. Kerkhoffs. Hyaluronic acid and other conservative treatment options for osteoarthritis of the ankle. *Cochrane Database Syst Rev*, 2015(10):Cd010643, 2015. ISSN 1361-6137. doi: 10.1002/14651858.CD010643.pub2.

[120] M. Fransen, S. McConnell, G. Hernandez-Molina, and S. Reichenbach. Exercise for osteoarthritis of the hip. *Cochrane Database Syst Rev*, 2009(3):Cd007912, 2009. ISSN 1361-6137. doi: 10.1002/14651858.cd007912.

[121] Computer program. Stata IC 14. *StataCorp TX, USA.*, Accessed Apr 2019. URL `www.stata.com`.

[122] Web link. SampSize app. Accessed 05 Jul 2019. URL `https://www.epigenesys.org.uk/portfolio/sampsize//`.

[123] Joseph L Hodges Jr and Erich L Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, pages 598–611, 1963. ISSN 0003-4851.

[124] Roger Newson. Somers' D - confidence intervals for nonparametric statistics and their differences. *Stata Technical Bulletin*, 10(55), 2001.

[125] B. Copsey, J. Y. Thompson, K. Vadher, U. Ali, S. J. Dutton, R. Fitzpatrick, S. E. Lamb, and J. A. Cook. Sample size calculations are poorly conducted and reported in many randomized trials of hip and knee osteoarthritis: results of a systematic review. *J Clin Epidemiol*, 104:52–61, 2018. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2018.08.013.

[126] J. H. Kellgren and J. S. Lawrence. Radiological assessment of osteo-arthrosis. *Ann Rheum Dis*, 16(4):494–502, 1957. ISSN 0003-4967 (Print) 0003-4967.

[127] Jerzy Neyman and Egon S Pearson. The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, 29(4):492–510, 1933. ISSN 1469-8064.

[128] Steven A Julious. *Sample sizes for clinical trials*. CRC Press, 2009. ISBN 1584887400.

[129] J. D. Stamey, F. Natanegara, and Jr. Seaman, J. W. Bayesian sample size determination for a clinical trial with correlated continuous and binary outcomes. *J Biopharm Stat*, 23(4):790–803, 2013. ISSN 1054-3406. doi: 10.1080/10543406.2013.789885.

[130] M. M. Ciarleglio, C. D. Arendt, and P. N. Peduzzi. Selection of the effect size for sample size determination for a continuous response in a superiority clinical trial using a hybrid classical and Bayesian procedure. *Clin Trials*, 13(3):275–85, 2016. ISSN 1740-7745. doi: 10.1177/1740774516628825.

[131] J. Cao, J. J. Lee, and S. Alber. Comparison of Bayesian sample size criteria: ACC, ALC, and WOC. *J Stat Plan Inference*, 139(12):4111–4122, 2009. ISSN 0378-3758 (Print) 0378-3758. doi: 10.1016/j.jspi.2009.05.041.

[132] J. Parvizi, C. Restrepo, and M. G. Maltenfort. Total hip arthroplasty performed through direct anterior approach provides superior early outcome: Results of a randomized, prospective study. *Orthop Clin North Am*, 47(3):497–504, 2016. ISSN 0030-5898. doi: 10.1016/j.ocl.2016.03.003.

[133] Manasi Banerjee, Shirsendu Mondal, Rathindranath Sarkar, Hindol Mondal, and Kuntal Bhattacharya. Comparative study of efficacy and safety of tapentadol versus etoricoxib in mild to moderate grades of chronic osteorthritis of knee. *Indian Journal of Rheumatology*, 11(1):21–25, 2016. ISSN 0973-3698.

[134] F. F. Bryk, A. C. Dos Reis, D. Fingerhut, T. Araujo, M. Schutzer, P. Cury Rde, Jr. Duarte, A., and T. Y. Fukuda. Exercises with partial vascular occlusion in patients with knee osteoarthritis: a randomized clinical trial. *Knee Surg Sports Traumatol Arthrosc*, 24(5):1580–6, 2016. ISSN 0942-2056. doi: 10.1007/s00167-016-4064-7.

[135] S. N. Gopal, W. Kamal, J. George, and E. Manssor. Radiological and biochemical effects (CTX-II, MMP-3, 8, and 13) of low-level laser therapy (LLLT) in chronic osteoarthritis in Al-Kharj, Saudi Arabia. *Lasers in Medical Science*, pages 1–7, 2016.

[136] Y. Xin, L. Jianhao, S. Tiansheng, H. Yongqiang, F. Weimin, C. Ming, S. Tiezheng, Y. Jianhua, X. Liang, G. Xiaoyuan, and C. Yongping. The efficacy and safety of sodium hyaluronate injection (Adant) in treating degenerative osteoarthritis: A multi-center, randomized, double-blind, positive-drug parallel-controlled and non-inferiority clinical study. *International Journal of Rheumatic Diseases*, 19(3):271–278, 2016.

[137] M. M. Saw, T. Kruger-Jakins, N. Edries, and R. Parker. Significant improvements in pain after a six-week physiotherapist-led exercise and education intervention, in patients with osteoarthritis awaiting arthroplasty, in South Africa: A randomised controlled trial, 2016.

[138] Umit Dincer, Serkan Arbal, Hasan Saygn, Mehmet Incedayi, and Osman Rodop. The effects of closed kinetic chain exercise on articular cartilage morphology:

myth or reality? a randomized controlled clinical trial. *Turkish Journal of Physical Medicine and Rehabilitation*, 62(1):28–37, 2016. ISSN 1302-0234.

[139] A. Notarnicola, G. Maccagnano, L. Moretti, V. Pesce, S. Tafuri, A. Fiore, and B. Moretti. Methylsulfonylmethane and boswellic acids versus glucosamine sulfate in the treatment of knee arthritis: Randomized trial. *Int J Immunopathol Pharmacol*, 29(1):140–6, 2016. ISSN 0394-6320 (Print) 0394-6320. doi: 10.1177/0394632015622215.

[140] C. Beselga, F. Neto, F. Alburquerque-Sendin, T. Hall, and N. Oliveira-Campelo. Immediate effects of hip mobilization with movement in patients with hip osteoarthritis: A randomised controlled trial. *Man Ther*, 22:80–5, 2016. ISSN 1356-689x. doi: 10.1016/j.math.2015.10.007.

[141] K. D. Allen, Jr. Yancy, W. S., H. B. Bosworth, C. J. Coffman, A. S. Jeffreys, S. K. Datta, J. McDuffie, J. L. Strauss, and E. Z. Oddone. A combined patient and provider intervention for management of osteoarthritis in veterans: A randomized clinical trial. *Ann Intern Med*, 164(2):73–83, 2016. ISSN 0003-4819. doi: 10.7326/m15-0378.

[142] M. F. Schinsky, C. McCune, and J. Bonomi. Multifaceted comparison of two cryotherapy devices used after total knee arthroplasty: Cryotherapy device comparison. *Orthop Nurs*, 35(5):309–16, 2016. ISSN 0744-6020. doi: 10.1097/nor.0000000000000276.

[143] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: erlbaum, 2nd edition, 1988.

[144] A. C. Leon, L. L. Davis, and H. C. Kraemer. The role and interpretation of pilot studies in clinical research. *J Psychiatr Res*, 45(5):626–9, 2011. ISSN 0022-3956. doi: 10.1016/j.jpsychires.2010.10.008.

[145] H. P. Soares, S. Daniels, A. Kumar, M. Clarke, C. Scott, S. Swann, and B. Djulbegovic. Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. *BMJ*, 328(7430):22–4, 2004. ISSN 0959-535x. doi: 10.1136/bmj.328.7430.22.

[146] S. Ramagopalan, A. P. Skingsley, L. Handunnetthi, M. Klingel, D. Magnus, J. Pakpoor, and B. Goldacre. Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: a cross-sectional study. *F1000Res*, 3:77, 2014. ISSN 2046-1402 (Print) 2046-1402. doi: 10.12688/f1000research.3784.1.

[147] A. W. Chan, A. Hrobjartsson, M. T. Haahr, P. C. Gotzsche, and D. G. Altman. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *Jama*, 291(20):2457–65, 2004. ISSN 0098-7484. doi: 10.1001/jama.291.20.2457.

[148] T. Aagaard, H. Lund, and C. Juhl. Optimizing literature search in systematic reviews - are MEDLINE, EMBASE and CENTRAL enough for identifying effect

studies within the area of musculoskeletal disorders? *BMC Med Res Methodol*, 16(1):161, 2016. ISSN 1471-2288. doi: 10.1186/s12874-016-0264-6.

[149] S. Minozzi, V. Pistotti, and M. Forni. Searching for rehabilitation articles on MEDLINE and EMBASE. an example with cross-over design. *Arch Phys Med Rehabil*, 81(6):720–2, 2000. ISSN 0003-9993 (Print) 0003-9993.

[150] A. N. Irwin and D. Rackham. Comparison of the time-to-indexing in PubMed between biomedical journals according to impact factor, discipline, and focus. *Res Social Adm Pharm*, 13(2):389–393, 2017. ISSN 1551-7411. doi: 10.1016/j.sapharm.2016.04.006.

[151] V. Rabenda, O. Bruyere, C. Cooper, R. Rizzoli, F. Buckinx, A. Quabron, and J. Y. Reginster. Publication outcomes of the abstracts presented at the 2011 European Congress on Osteoporosis, Osteoarthritis and Musculo-Skeletal Diseases (ECCEO-IOF11): A position paper of the European Society for Clinical and Economical Aspects of Osteoporosis, Osteoarthritis and Musculo-Skeletal Diseases (ESCEO) and the International Osteoporosis and Other Skeletal Diseases Foundation (IOF). *Arch Osteoporos*, 10:11, 2015. doi: 10.1007/s11657-015-0216-5.

[152] J. Kay, M. Memon, D. de Sa, A. Duong, N. Simunovic, G. S. Athwal, and O. R. Ayeni. Five-year publication rate of clinical presentations at the open and closed American shoulder and elbow surgeons annual meeting from 2005-2010. *J Exp Orthop*, 3(1):21, 2016. ISSN 2197-1153 (Print) 2197-1153. doi: 10.1186/s40634-016-0059-z.

[153] T. Li, I. Boutron, R. Al-Shahi Salman, E. Cobo, E. Flemyng, J. M. Grimshaw, and D. G. Altman. Review and publication of protocol submissions to Trials - what have we learned in 10 years? *Trials*, 18(1):34, 2016. ISSN 1745-6215. doi: 10.1186/s13063-016-1743-0.

[154] Web link. Figshare. Accessed 22 Mar 2019. URL `https://figshare.com/`.

[155] Web link:. arXiv. Accessed 22 Mar 2019. URL `https://arxiv.org/`.

[156] S. Hopewell, G. S. Collins, I. Boutron, L. M. Yu, J. Cook, M. Shanyinde, R. Wharton, L. Shamseer, and D. G. Altman. Impact of peer review on reports of randomised trials published in open peer review journals: retrospective before and after study. *BMJ*, 349:g4145, 2014. ISSN 0959-535x. doi: 10.1136/bmj.g4145.

[157] K. F. Schulz, D. G. Altman, and D. Moher. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340:c332, 2010. ISSN 0959-535x. doi: 10.1136/bmj.c332.

[158] P. Glasziou, D. G. Altman, P. Bossuyt, I. Boutron, M. Clarke, S. Julious, S. Michie, D. Moher, and E. Wager. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*, 383(9913):267–76, 2014. ISSN 0140-6736. doi: 10.1016/s0140-6736(13)62228-x.

[159] J. P. Ioannidis, S. Greenland, M. A. Hlatky, M. J. Khoury, M. R. Macleod, D. Moher, K. F. Schulz, and R. Tibshirani. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*, 383(9912):166–75, 2014. ISSN 0140-6736. doi: 10.1016/s0140-6736(13)62227-8.

[160] M. J. Campbell. Doing clinical trials large enough to achieve adequate reductions in uncertainties about treatment effects. *J R Soc Med*, 106(2):68–71, 2013. ISSN 0141-0768. doi: 10.1177/0141076813477570.

[161] N. Black, S. van Rooyen, F. Godlee, R. Smith, and S. Evans. What makes a good reviewer and a good review for a general medical journal? *Jama*, 280(3): 231–3, 1998. ISSN 0098-7484 (Print) 0098-7484.

[162] E. Cobo, A. Selva-O'Callagham, J. M. Ribera, F. Cardellach, R. Dominguez, and M. Vilardell. Statistical reviewers improve reporting in biomedical articles: a randomized trial. *PLoS One*, 2(3):e332, 2007. ISSN 1932-6203. doi: 10.1371/journal.pone.0000332.

[163] T. A. Althunian, A. de Boer, O. H. Klungel, W. N. Insani, and R. H. Groenwold. Methods of defining the non-inferiority margin in randomized, double-blind controlled trials: a systematic review. *Trials*, 18(1):107, 2017. ISSN 1745-6215. doi: 10.1186/s13063-017-1859-x.

[164] J. P. Higgins, D. G. Altman, P. C. Gotzsche, P. Juni, D. Moher, A. D. Oxman, J. Savovic, K. F. Schulz, L. Weeks, and J. A. Sterne. The Cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343:d5928, 2011. ISSN 0959-535x. doi: 10.1136/bmj.d5928.

[165] M. L. Costa, X. L. Griffin, N. Parsons, M. Dritsaki, and D. Perry. Efficacy versus effectiveness in clinical trials. *Bone Joint J*, 99-b(4):419–420, 2017. ISSN 2049-4394. doi: 10.1302/0301-620x.99b4.bjj-2016-1247.

[166] H. C. Kraemer, J. Mintz, A. Noda, J. Tinklenberg, and J. A. Yesavage. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry*, 63(5):484–9, 2006. ISSN 0003-990X (Print) 0003-990x. doi: 10.1001/archpsyc.63.5.484.

[167] N. Bellamy, W. W. Buchanan, C. H. Goldsmith, J. Campbell, and L. W. Stitt. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol*, 15(12):1833–40, 1988. ISSN 0315-162X (Print) 0315-162x.

[168] Derek Glenn Kyte. *The methodological and ethical issues associated with patient-reported outcome measurement in clinical trials*. Thesis, School of Health and Population Sciences, University of Birmingham, 2015.

[169] M. Boers, J. R. Kirwan, G. Wells, D. Beaton, L. Gossec, M. A. d'Agostino, P. G. Conaghan, 3rd Bingham, C. O., P. Brooks, R. Landewe, L. March, L. S. Simon, J. A. Singh, V. Strand, and P. Tugwell. Developing core outcome measurement

sets for clinical trials: OMERACT filter 2.0. *J Clin Epidemiol*, 67(7):745–53, 2014. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2013.11.013.

[170] M. Bond, A. Davis, S. Lohmander, and G. Hawker. Responsiveness of the OARSI-OMERACT osteoarthritis pain and function measures. *Osteoarthritis Cartilage*, 20(6):541–7, 2012. ISSN 1063-4584. doi: 10.1016/j.joca.2012.03.001.

[171] P. Kersten, P. J. White, and A. Tennant. The visual analogue WOMAC 3.0 scale - internal validity and responsiveness of the VAS version. *BMC Musculoskeletal Disorders*, 11:80–80, 2010. ISSN 1471-2474. doi: 10.1186/1471-2474-11-80.

[172] N. J. Collins, D. Misra, D. T. Felson, K. M. Crossley, and E. M. Roos. Measures of knee function: International Knee Documentation Committee (IKDC) subjective knee evaluation form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS). *Arthritis care and research*, 63(0 11):S208–S228, 2011. ISSN 2151-464X 2151-4658. doi: 10.1002/acr.20632.

[173] S. Strange, M. R. Whitehouse, A. D. Beswick, T. Board, A. Burston, B. Burston, F. E. Carroll, P. Dieppe, K. Garfield, R. Gooberman-Hill, S. Jones, S. Kunutsor, A. Lane, E. Lenguerrand, A. MacGowan, A. Moore, S. Noble, J. Simon, I. Stockley, A. H. Taylor, A. Toms, J. Webb, J. P. Whittaker, M. Wilson, V. Wylde, and A. W. Blom. One-stage or two-stage revision surgery for prosthetic hip joint infection–the INFORM trial: a study protocol for a randomised controlled trial. *Trials*, 17:90, 2016. ISSN 1745-6215. doi: 10.1186/s13063-016-1213-8.

[174] J. N. Katz, R. H. Brophy, C. E. Chaisson, L. de Chaves, B. J. Cole, D. L. Dahm, L. A. Donnell-Fink, A. Guermazi, A. K. Haas, M. H. Jones, B. A. Levy, L. A. Mandl, S. D. Martin, R. G. Marx, A. Miniaci, M. J. Matava, J. Palmisano, E. K. Reinke, B. E. Richardson, B. N. Rome, C. E. Safran-Norton, D. J. Skoniecki, D. H. Solomon, M. V. Smith, K. P. Spindler, M. J. Stuart, J. Wright, R. W. Wright, and E. Losina. Surgery versus physical therapy for a meniscal tear and osteoarthritis. *N Engl J Med*, 368(18):1675–84, 2013. ISSN 0028-4793. doi: 10.1056/NEJMoa1301408.

[175] M. Underwood, D. Ashby, D. Carnes, E. Castelnuovo, P. Cross, G. Harding, E. Hennessy, L. Letley, J. Martin, S. Mt-Isa, S. Parsons, A. Spencer, M. Vickers, and K. Whyte. Topical or oral ibuprofen for chronic knee pain in older people. The TOIB study. *Health Technol Assess*, 12(22):iii–iv, ix–155, 2008. ISSN 1366-5278 (Print) 1366-5278.

[176] P. A. Smith. Intra-articular autologous conditioned plasma injections provide safe and efficacious treatment for knee osteoarthritis: An FDA-sanctioned, ran-

domized, double-blind, placebo-controlled clinical trial. *Am J Sports Med*, 44 (4):884–91, 2016. ISSN 0363-5465. doi: 10.1177/0363546515624678.

[177] J. Y. Reginster, S. Reiter-Niesert, O. Bruyere, F. Berenbaum, M. L. Brandi, J. Branco, J. P. Devogelaer, G. Herrero-Beaumont, J. Kanis, S. Maggi, E. Maheu, P. Richette, R. Rizzoli, and C. Cooper. Recommendations for an update of the 2010 European regulatory guideline on clinical investigation of medicinal products used in the treatment of osteoarthritis and reflections about related clinically relevant outcomes: expert consensus statement. *Osteoarthritis Cartilage*, 23(12):2086–2093, 2015. ISSN 1063-4584. doi: 10.1016/j.joca.2015.07.001.

[178] V. M. Goldberg, J. Buckwalter, M. Halpin, W. Jiranek, W. Mihalko, M. Pinzur, B. Rohan, T. Vail, P. Walker, R. Windsor, and T. Wright. Recommendations of the OARSI FDA osteoarthritis devices working group. *Osteoarthritis Cartilage*, 19(5):509–14, 2011. ISSN 1063-4584. doi: 10.1016/j.joca.2011.02.017.

[179] FDA. Clinical development programs for drugs, devices and biological products intended for the treatment of OA. *Draft guidance*, 1999. URL `https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071577.pdf`.

[180] N. Bellamy. WOMAC: a 20-year experiential review of a patient-centered self-reported health status questionnaire. *J Rheumatol*, 29(12):2473–6, 2002. ISSN 0315-162X (Print) 0315-162x.

[181] T. E. Howe, L. J. Dawson, G. Syme, L. Duncan, and J. Reid. Evaluation of outcome measures for use in clinical practice for adults with musculoskeletal conditions of the knee: a systematic review. *Man Ther*, 17(2):100–18, 2012. ISSN 1356-689x. doi: 10.1016/j.math.2011.07.002.

[182] B. Gandek. Measurement properties of the Western Ontario and McMaster Universities Osteoarthritis Index: a systematic review. *Arthritis Care Res (Hoboken)*, 67(2):216–29, 2015. ISSN 2151-464x. doi: 10.1002/acr.22415.

[183] M. A. Ahmad, F. N. Xypnitos, and P. V. Giannoudis. Measuring hip outcomes: common scales and checklists. *Injury*, 42(3):259–64, 2011. ISSN 0020-1383. doi: 10.1016/j.injury.2010.11.052.

[184] J. J. Gagnier, M. Mullins, H. Huang, D. Marinac-Dabic, A. Ghambaryan, B. Eloff, F. Mirza, and M. Bayona. A systematic review of measurement properties of patient-reported outcome measures used in patients undergoing total knee arthroplasty. *J Arthroplasty*, 32(5):1688–1697.e7, 2017. ISSN 0883-5403. doi: 10.1016/j.arth.2016.12.052.

[185] K. Harris, J. Dawson, E. Gibbons, C. R. Lim, D. J. Beard, R. Fitzpatrick, and A. J. Price. Systematic review of measurement properties of patient-reported outcome measures used in patients undergoing hip and knee arthroplasty. *Patient Relat Outcome Meas*, 7:101–8, 2016. ISSN 1179-271X (Print) 1179-271x. doi: 10.2147/prom.s97774.

[186] Frank R Noyes. *Noyes' Knee Disorders: Surgery, Rehabilitation, Clinical Outcomes E-Book*. Elsevier Health Sciences, 2016. ISBN 032342855X.

[187] Web link. WOMAC. Accessed on 29 Oct 2018. URL `www.womac.org`.

[188] J. A. Bolognese, T. J. Schnitzer, and E. W. Ehrich. Response relationship of VAS and Likert scales in osteoarthritis efficacy measurement. *Osteoarthritis Cartilage*, 11(7):499–507, 2003. ISSN 1063-4584 (Print) 1063-4584.

[189] N. F. Woolacott, M. S. Corbett, and S. J. Rice. The use and reporting of WOMAC in the assessment of the benefit of physical therapies for the pain of osteoarthritis of the knee: findings from a systematic review of clinical trials. *Rheumatology (Oxford)*, 51(8):1440–6, 2012. ISSN 1462-0324. doi: 10.1093/rheumatology/kes043.

[190] C. Sanders, M. Egger, J. Donovan, D. Tallon, and S. Frankel. Reporting on quality of life in randomised controlled trials: bibliographic study. *BMJ*, 317 (7167):1191–4, 1998. ISSN 0959-8138 (Print) 0959-535x.

[191] S. Fielding, A. Ogbuagu, S. Sivasubramaniam, G. MacLennan, and C. R. Ramsay. Reporting and dealing with missing quality of life data in RCTs: has the picture changed in the last decade? *Qual Life Res*, 25(12):2977–2983, 2016. ISSN 0962-9343. doi: 10.1007/s11136-016-1411-6.

[192] FDA. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. *Health Qual Life Outcomes*, 4:79, 2006. ISSN 1477-7525. doi: 10.1186/1477-7525-4-79.

[193] F. Angst, A. Aeschlimann, and G. Stucki. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis Rheum*, 45(4):384–91, 2001. ISSN 0004-3591 (Print) 0004-3591. doi: 10.1002/1529-0131(200108)45:4¡384::aid-art352¿3.0.co;2-0.

[194] B. Copsey, J. Y. Thompson, K. Vadher, U. Ali, S. J. Dutton, R. Fitzpatrick, S. E. Lamb, and J. A. Cook. Problems persist in reporting of methods and results for the womac measure in hip and knee osteoarthritis trials. *Qual Life Res*, 28(2):335–343, 2019. ISSN 0962-9343. doi: 10.1007/s11136-018-1978-1.

[195] N. K. Arden, S. Cro, S. Sheard, C. J. Dore, A. Bara, S. A. Tebbs, D. J. Hunter, S. James, C. Cooper, T. W. O'Neill, A. Macgregor, F. Birrell, and R. Keen. The effect of vitamin D supplementation on knee osteoarthritis, the VIDEO study: a randomised controlled trial. *Osteoarthritis Cartilage*, 24(11):1858–1866, 2016. ISSN 1063-4584. doi: 10.1016/j.joca.2016.05.020.

[196] Anna K. Nilsdotter, L. Stefan Lohmander, Maria Klssbo, and Ewa M. Roos. Hip disability and osteoarthritis outcome score (HOOS)–validity and responsiveness in total hip replacement. *BMC musculoskeletal disorders*, 4:10–10, 2003. ISSN 1471-2474. doi: 10.1186/1471-2474-4-10.

[197] Ewa M. Roos and L. Stefan Lohmander. The knee injury and osteoarthritis outcome score (KOOS): from joint injury to osteoarthritis. *Health and quality of life outcomes*, 1:64–64, 2003. ISSN 1477-7525. doi: 10.1186/1477-7525-1-64.

[198] S. Moorthy, R. Sudar Codi, R. Surendher, and K. Manimekalai. Comparison of the efficacy and safety of tramadol versus tapentadol in acute osteoarthritic knee pain: A randomized, controlled trial. *Asian Journal of Pharmaceutical and Clinical Research*, 9(3), 2016.

[199] E. Losina, J. E. Collins, J. Wright, M. E. Daigle, L. A. Donnell-Fink, D. Strnad, I. M. Usiskin, H. Y. Yang, V. Lerner, and J. N. Katz. Postoperative care navigation for total knee arthroplasty patients: A randomized controlled trial. *Arthritis Care and Research*, 68(9):1252–1259, 2016.

[200] K. D. Allen, W. S. Yancy, H. B. Bosworth, C. J. Coffman, A. S. Jeffreys, S. K. Datta, J. McDuffie, J. L. Strauss, and E. Z. Oddone. A combined patient and provider intervention for management of osteoarthritis in veterans. *Annals of Internal Medicine*, 164(2):73–83, 2016.

[201] X. Jin, G. Jones, F. Cicuttini, A. Wluka, Z. Zhu, W. Han, B. Antony, X. Wang, T. Winzenberg, L. Blizzard, and C. Ding. Effect of vitamin D supplementation on tibial cartilage volume and knee pain among patients with symptomatic knee osteoarthritis: A randomized clinical trial. *JAMA - Journal of the American Medical Association*, 315(10):1005–1013, 2016.

[202] Y. Yuan, W. Shen, Q. Han, D. Liang, L. Chen, Q. Yin, W. Zhu, and H. Xu. Clinical observation of pulsed radiofrequency in treatment of knee osteoarthritis. *International Journal of Clinical and Experimental Medicine*, 9(10):20050–20055, 2016.

[203] C. Wang, Ch. Schmid, Md. Iversen, Wf. Harvey, Ra. Fielding, Jb. Driban, Ll. Price, Jb. Wong, Kf. Reid, R. Rones, and T. McAlindon. Comparative effectiveness of Tai Chi versus physical therapy for knee osteoarthritis: A randomized trial. *Annals of internal medicine*, 165(2):77–86, 2016. doi: 10.7326/M15-2143.

[204] M. Dougados, P. Leclaire, D. van der Heijde, D. A. Bloch, N. Bellamy, and R. D. Altman. Response criteria for clinical trials on osteoarthritis of the knee and hip: a report of the Osteoarthritis Research Society International Standing Committee for Clinical Trials response criteria initiative. *Osteoarthritis Cartilage*, 8(6): 395–403, 2000. ISSN 1063-4584 (Print) 1063-4584. doi: 10.1053/joca.2000.0361.

[205] S. Bisicchia, G. Bernardi, and C. Tudisco. HYADD 4 versus methylprednisolone acetate in symptomatic knee osteoarthritis: A single-centre single blind prospective randomised controlled clinical study with a 1-year follow-up. *Clinical and Experimental Rheumatology*, 34(5):857–863, 2016.

[206] K.i D. Allen, D. Bongiorni, H. B. Bosworth, C. J. Coffman, S. K. Datta, D. Edelman, K. S. Hall, J. H. Lindquist, E. Z. Oddone, and H. Hoenig. Group versus individual physical therapy for veterans with knee osteoarthritis: Randomized

clinical trial. *Physical Therapy*, 96(5):597–608, 2016. ISSN 0031-9023. doi: 10.2522/ptj.20150194.

[207] F. Angst, A. Aeschlimann, B. A. Michel, and G. Stucki. Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities. *J Rheumatol*, 29(1):131–8, 2002. ISSN 0315-162X (Print) 0315-162x.

[208] R. S. Hinman, T. V. Wrigley, B. R. Metcalf, P. K. Campbell, K. L. Paterson, D. J. Hunter, J. Kasza, A. Forbes, and K. L. Bennell. Unloading shoes for self-management of knee osteoarthritis: A randomized trial. *Annals of Internal Medicine*, 165(6):381–389, 2016.

[209] S. Lydersen. Statistical review: frequently given comments. *Ann Rheum Dis*, 74(2):323–5, 2015. ISSN 0003-4967. doi: 10.1136/annrheumdis-2014-206186.

[210] I. Rombach, O. Rivero-Arias, A. M. Gray, C. Jenkinson, and O. Burke. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. *Qual Life Res*, 25(7):1613–23, 2016. ISSN 0962-9343. doi: 10.1007/s11136-015-1206-1.

[211] S. Adie, I. A. Harris, J. M. Naylor, and R. Mittal. CONSORT compliance in surgical randomized trials: are we there yet? A systematic review. *Ann Surg*, 258(6):872–8, 2013. ISSN 0003-4932. doi: 10.1097/SLA.0b013e31829664b9.

[212] L. Turner, L. Shamseer, D. G. Altman, L. Weeks, J. Peters, T. Kober, S. Dias, K. F. Schulz, A. C. Plint, and D. Moher. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database Syst Rev*, 11: Mr000030, 2012. ISSN 1361-6137. doi: 10.1002/14651858.MR000030.pub2.

[213] T. Devji, G. H. Guyatt, L. Lytvyn, R. Brignardello-Petersen, F. Foroutan, B. Sadeghirad, R. Buchbinder, R. W. Poolman, I. A. Harris, A. Carrasco-Labra, R. A. C. Siemieniuk, and P. O. Vandvik. Application of minimal important differences in degenerative knee disease outcomes: a systematic review and case study to inform BMJ Rapid Recommendations. *BMJ Open*, 7(5), 2017. doi: 10.1136/bmjopen-2016-015587.

[214] A. B. Hauber, N. K. Arden, A. F. Mohamed, F. R. Johnson, P. M. Peloso, D. J. Watson, P. Mavros, A. Gammaitoni, S. S. Sen, and S. D. Taylor. A discrete-choice experiment of United Kingdom patients' willingness to risk adverse events for improved function and pain control in osteoarthritis. *Osteoarthritis Cartilage*, 21(2):289–97, 2013. ISSN 1063-4584. doi: 10.1016/j.joca.2012.11.007.

[215] P. Wicks, M. Lowe, S. Gabriel, S. Sikirica, R. Sasane, and S. Arcona. Increasing patient participation in drug development. *Nat Biotechnol*, 33(2):134–5, 2015. ISSN 1087-0156. doi: 10.1038/nbt.3145.

[216] C. C. Keirns and S. D. Goold. Patient-centered care and preference-sensitive decision making. *Jama*, 302(16):1805–6, 2009. ISSN 0098-7484. doi: 10.1001/jama.2009.1550.

[217] M. Krahn and G. Naglie. The next step in guideline development: incorporating patient preferences. *Jama*, 300(4):436–8, 2008. ISSN 0098-7484. doi: 10.1001/jama.300.4.436.

[218] C. Charles, A. Gafni, and T. Whelan. Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). *Soc Sci Med*, 44(5):681–92, 1997. ISSN 0277-9536 (Print) 0277-9536.

[219] F. Legare, D. Stacey, S. Turcotte, M. J. Cossi, J. Kryworuchko, I. D. Graham, A. Lyddiatt, M. C. Politi, R. Thomson, G. Elwyn, and N. Donner-Banzhoff. Interventions for improving the adoption of shared decision making by healthcare professionals. *Cochrane Database Syst Rev*, 2014(9):Cd006732, 2014. ISSN 1361-6137. doi: 10.1002/14651858.CD006732.pub3.

[220] A. M. Stiggelbout, T. Van der Weijden, M. P. De Wit, D. Frosch, F. Legare, V. M. Montori, L. Trevena, and G. Elwyn. Shared decision making: really putting patients at the centre of healthcare. *BMJ*, 344:e256, 2012. ISSN 0959-535x. doi: 10.1136/bmj.e256.

[221] S. F. Terry. Clinical trial result reporting: Time to move into the 21st century. *Clin Trials*, 13(6):597–598, 2016. ISSN 1740-7745. doi: 10.1177/1740774516665599.

[222] A. Kearney, P. Williamson, B. Young, H. Bagley, C. Gamble, S. Denegri, D. Muir, N. A. Simon, S. Thomas, J. T. Elliot, H. Bulbeck, J. C. Crocker, C. Planner, C. Vale, M. Clarke, T. Sprosen, and K. Woolfall. Priorities for methodological research on patient and public involvement in clinical trials: A modified Delphi process. *Health Expect*, 20(6):1401–1410, 2017. ISSN 1369-6513. doi: 10.1111/hex.12583.

[223] E. W. de Bekker-Grob, C. Berlin, B. Levitan, K. Raza, K. Christoforidi, I. Cleemput, J. Pelouchova, H. Enzmann, N. Cook, and M. G. Hansson. Giving patients' preferences a voice in medical treatment life cycle: The PREFER public-private project. *Patient*, 10(3):263–266, 2017. ISSN 1178-1653. doi: 10.1007/s40271-017-0222-3.

[224] CBER CDRH. Patient preference information - Voluntary submission, review in premarket approval applications, humanitarian device exemption applications, and de novo requests, and inclusion in decision summaries and device labeling. Draft guidance for industry. *Food and Drug Administration Staff, and Other Stakeholders*, 2016.

[225] J. Cordero-Ampuero, A. Darder, J. Santillana, M. T. Caloto, and G. Nocea. Evaluation of patients' and physicians' expectations and attributes of osteoarthritis treatment using Kano methodology. *Qual Life Res*, 21(8):1391–404, 2012. ISSN 0962-9343. doi: 10.1007/s11136-011-0058-6.

[226] C. Heneghan, B. Goldacre, and K. R. Mahtani. Why clinical trial outcomes fail

to translate into benefits for patients. *Trials*, 18(1):122, 2017. ISSN 1745-6215. doi: 10.1186/s13063-017-1870-2.

[227] T. Pincus. Rheumatoid arthritis: disappointing long-term outcomes despite successful short-term clinical trials. *J Clin Epidemiol*, 41(11):1037–41, 1988. ISSN 0895-4356 (Print) 0895-4356.

[228] S. M. McCurry, S. M. Shortreed, M. Von Korff, B. H. Balderson, L. D. Baker, B. D. Rybarczyk, and M. V. Vitiello. Who benefits from CBT for insomnia in primary care? Important patient selection and trial design lessons from longitudinal results of the Lifestyles trial. *Sleep*, 37(2):299–308, 2014. ISSN 0161-8105. doi: 10.5665/sleep.3402.

[229] Institute of Medicine (US) Committee for the Substance Abuse Coverage Study. *Treating Drug Problems: Volume 1: A Study of the Evolution, Effectiveness, and Financing of Public and Private Drug Treatment Systems*. National Academies Press (US), Washington (DC), 1990. doi: 10.17226/1551.

[230] J. Askling, M. Holmqvist, and L. Ljung. Editorial: Is rheumatoid arthritis a mortal disease? *Arthritis Rheumatol*, 69(8):1509–1511, 2017. ISSN 2326-5191. doi: 10.1002/art.40145.

[231] S. Chan, A. Jonsson, and M. Bhandari. Planning a clinical research study. *Indian J Orthop*, 41(1):16–22, 2007. ISSN 0019-5413 (Print) 0019-5413. doi: 10.4103/0019-5413.30520.

[232] T. Pincus and T. Sokka. Should contemporary rheumatoid arthritis clinical trials be more like standard patient care and vice versa? *Ann Rheum Dis*, 63 Suppl 2:ii32–ii39, 2004. ISSN 0003-4967 (Print) 0003-4967. doi: 10.1136/ard.2004.028415.

[233] Martin E Backhouse and Paul Fenn. The use of discrete choice analysis in the design of randomised controlled trials. Report, University of Nottingham, 2006.

[234] D. Stacey, M. Taljaard, G. Dervin, P. Tugwell, A. M. O'Connor, M. P. Pomey, L. Boland, S. Beach, D. Meltzer, and G. Hawker. Impact of patient decision aids on appropriate and timely access to hip or knee arthroplasty for osteoarthritis: a randomized controlled trial. *Osteoarthritis Cartilage*, 24(1):99–107, 2016. ISSN 1063-4584. doi: 10.1016/j.joca.2015.07.024.

[235] G. S. Hazlewood, C. Bombardier, G. Tomlinson, and D. Marshall. A Bayesian model that jointly considers comparative effectiveness research and patients' preferences may help inform GRADE recommendations: an application to rheumatoid arthritis treatment recommendations. *J Clin Epidemiol*, 93:56–65, 2018. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2017.10.003.

[236] E. Stamuli, D. Torgerson, M. Northgraves, S. Ronaldson, and L. Cherry. Identifying the primary outcome for a randomised controlled trial in rheumatoid arthritis: the role of a discrete choice experiment. *J Foot Ankle Res*, 10(1):57, 2017. ISSN 1757-1146.

[237] P. Blinman, M. King, R. Norman, R. Viney, and M. R. Stockler. Preferences for cancer treatments: an overview of methods and applications in oncology. *Ann Oncol*, 23(5):1104–10, 2012. ISSN 0923-7534. doi: 10.1093/annonc/mdr559.

[238] O. Ethgen, A. Tancredi, E. Lejeune, A. Kvasz, B. Zegels, and J. Y. Reginster. Do utility values and willingness to pay suitably reflect health outcome in hip and knee osteoarthritis? A comparative analysis with the WOMAC index. *J Rheumatol*, 30(11):2452–9, 2003. ISSN 0315-162X (Print) 0315-162x.

[239] T. Maddala, K. A. Phillips, and F. Reed Johnson. An experiment on simplifying conjoint analysis designs for measuring preferences. *Health Econ*, 12(12):1035–47, 2003. ISSN 1057-9230 (Print) 1057-9230. doi: 10.1002/hec.798.

[240] J. Ratcliffe, L. Couzner, T. Flynn, M. Sawyer, K. Stevens, J. Brazier, and L. Burgess. Valuing Child Health Utility 9D health states with a young adolescent sample: a feasibility study to compare best-worst scaling discrete-choice experiment, standard gamble and time trade-off methods. *Appl Health Econ Health Policy*, 9(1):15–27, 2011. ISSN 1175-5652. doi: 10.2165/11536960-000000000-00000.

[241] D. Bijlenga, E. Birnie, and G. J. Bonsel. Feasibility, reliability, and validity of three health-state valuation methods using multiple-outcome vignettes on moderate-risk pregnancy at term. *Value Health*, 12(5):821–7, 2009. ISSN 1098-3015. doi: 10.1111/j.1524-4733.2009.00503.x.

[242] A. Robinson, A. E. Spencer, J. L. Pinto-Prades, and J. A. Covey. Exploring differences between TTO and DCE in the valuation of health states. *Med Decis Making*, 37(3):273–284, 2017. ISSN 0272-989x. doi: 10.1177/0272989x16668343.

[243] L. A. Augestad, K. Stavem, I. S. Kristiansen, C. H. Samuelsen, and K. Rand-Hendriksen. Influenced from the start: anchoring bias in time trade-off valuations. *Qual Life Res*, 25(9):2179–91, 2016. ISSN 0962-9343. doi: 10.1007/s11136-016-1266-x.

[244] N. Bansback, A. R. Hole, B. Mulhern, and A. Tsuchiya. Testing a discrete choice experiment including duration to value health states for large descriptive systems: addressing design and sampling issues. *Soc Sci Med*, 114:38–48, 2014. ISSN 0277-9536. doi: 10.1016/j.socscimed.2014.05.026.

[245] A. Tsuchiya and P. Dolan. The QALY model and individual preferences for health states and health profiles over time: a systematic review of the literature. *Med Decis Making*, 25(4):460–7, 2005. ISSN 0272-989X (Print) 0272-989x. doi: 10.1177/0272989x05276854.

[246] B. Barrett, R. Brown, M. Mundt, L. Dye, J. Alt, N. Safdar, and R. Maberry. Using benefit harm tradeoffs to estimate sufficiently important difference: the case of the common cold. *Med Decis Making*, 25(1):47–55, 2005. ISSN 0272-989X (Print) 0272-989x. doi: 10.1177/0272989x04273147.

[247] B. Barrett, B. Harahan, D. Brown, Z. Zhang, and R. Brown. Sufficiently important difference for common cold: severity reduction. *Ann Fam Med*, 5(3): 216–23, 2007. ISSN 1544-1709. doi: 10.1370/afm.698.

[248] M. L. Schaarschmidt, A. Schmieder, N. Umar, D. Terris, M. Goebeler, S. Goerdt, and W. K. Peitsch. Patient preferences for psoriasis treatments: process characteristics can outweigh outcome attributes. *Arch Dermatol*, 147(11):1285–94, 2011. ISSN 0003-987x. doi: 10.1001/archdermatol.2011.309.

[249] J. M. Gonzalez. Evaluating risk tolerance from a systematic review of preferences: The case of patients with psoriasis. *Patient*, 11(3):285–300, 2018. ISSN 1178-1653. doi: 10.1007/s40271-017-0295-z.

[250] L. Fraenkel, Jr. Bogardus, S. T., J. Concato, and D. R. Wittink. Treatment options in knee osteoarthritis: the patient's perspective. *Arch Intern Med*, 164 (12):1299–304, 2004. ISSN 0003-9926 (Print) 0003-9926. doi: 10.1001/archinte.164.12.1299.

[251] J. Posnett, S. Dixit, B. Oppenheimer, S. Kili, and N. Mehin. Patient preference and willingness to pay for knee osteoarthritis treatments. *Patient Prefer Adherence*, 9:733–44, 2015. ISSN 1177-889X (Print) 1177-889x. doi: 10.2147/ppa.s84251.

[252] S. Groenewoud, N. J. Van Exel, A. Bobinac, M. Berg, R. Huijsman, and E. A. Stolk. What influences patients' decisions when choosing a health care provider? Measuring preferences of patients with knee arthrosis, chronic depression, or alzheimer's disease, using discrete choice experiments. *Health Serv Res*, 50(6): 1941–72, 2015. ISSN 0017-9124. doi: 10.1111/1475-6773.12306.

[253] N. N. O'Hara, G. P. Slobogean, T. Mohammadi, C. A. Marra, M. R. Vicente, A. Khakban, and M. D. McKee. Are patients willing to pay for total shoulder arthroplasty? Evidence from a discrete choice experiment. *Can J Surg*, 59(2): 107–12, 2016. ISSN 0008-428x.

[254] T. K. Kim, J. Choi, K. S. Shin, C. B. Chang, and S. C. Seong. Patients' perspective on controversial issues in total knee arthroplasty. *Knee Surg Sports Traumatol Arthrosc*, 16(3):297–304, 2008. ISSN 0942-2056 (Print) 0942-2056. doi: 10.1007/s00167-007-0468-8.

[255] C. K. Kwoh, E. R. Vina, Y. K. Cloonan, M. J. Hannon, R. M. Boudreau, and S. A. Ibrahim. Determinants of patient preferences for total knee replacement: African-Americans and whites. *Arthritis Res Ther*, 17:348, 2015. ISSN 1478-6354. doi: 10.1186/s13075-015-0864-2.

[256] E. R. Vina, D. Richardson, E. Medvedeva, C. Kent Kwoh, A. Collier, and S. A. Ibrahim. Does a patient-centered educational intervention affect african-american access to knee replacement? A randomized trial. *Clin Orthop Relat Res*, 474(8):1755–64, 2016. ISSN 0009-921x. doi: 10.1007/s11999-016-4834-z.

[257] C. S. Li, R. W. Poolman, and M. Bhandari. Treatment preferences of patients with early knee osteoarthritis: a decision board analysis assessing high tibial osteotomy versus the KineSpring(R) Knee Implant System. *J Long Term Eff Med Implants*, 23(2-3):175–88, 2013. ISSN 1050-6934 (Print) 1050-6934.

[258] M. M. Byrne, K. O'Malley, and M. E. Suarez-Almazor. Willingness to pay per quality-adjusted life year in a study of knee osteoarthritis. *Med Decis Making*, 25(6):655–66, 2005. ISSN 0272-989X (Print) 0272-989x. doi: 10.1177/0272989x05282638.

[259] K. J. Bozic and V. Chiu. Emerging ideas: Shared decision making in patients with osteoarthritis of the hip and knee. *Clin Orthop Relat Res*, 469(7):2081–5, 2011. ISSN 0009-921x. doi: 10.1007/s11999-010-1740-7.

[260] J. F. Bridges, A. B. Hauber, D. Marshall, A. Lloyd, L. A. Prosser, D. A. Regier, F. R. Johnson, and J. Mauskopf. Conjoint analysis applications in health–a checklist: a report of the ISPOR good research practices for conjoint analysis task force. *Value Health*, 14(4):403–13, 2011. ISSN 1098-3015. doi: 10.1016/j.jval.2010.11.013.

[261] Nicholas Bellamy. *WOMAC osteoarthritis index: user guide IX*. Nicholas Bellamy, 2008.

[262] P. McGettigan and D. Henry. Cardiovascular risk with non-steroidal anti-inflammatory drugs: systematic review of population-based controlled observational studies. *PLoS Med*, 8(9):e1001098, 2011. ISSN 1549-1277. doi: 10.1371/journal.pmed.1001098.

[263] S. Hernandez-Diaz, C. Varas-Lorenzo, and L. A. Garcia Rodriguez. Non-steroidal antiinflammatory drugs and the risk of acute myocardial infarction. *Basic Clin Pharmacol Toxicol*, 98(3):266–74, 2006. ISSN 1742-7835 (Print) 1742-7835.

[264] J. M. Scheiman and A. M. Fendrick. Practical approaches to minimizing gastrointestinal and cardiovascular safety concerns with COX-2 inhibitors and NSAIDs. *Arthritis Res Ther*, 7 Suppl 4:S23–9, 2005. ISSN 1478-6354. doi: 10.1186/ar1795.

[265] C. P. Cannon, S. P. Curtis, G. A. FitzGerald, H. Krum, A. Kaur, J. A. Bolognese, A. S. Reicin, C. Bombardier, M. E. Weinblatt, D. van der Heijde, E. Erdmann, and L. Laine. Cardiovascular outcomes with etoricoxib and diclofenac in patients with osteoarthritis and rheumatoid arthritis in the multinational etoricoxib and diclofenac arthritis long-term (MEDAL) programme: a randomised comparison. *Lancet*, 368(9549):1771–81, 2006. ISSN 0140-6736. doi: 10.1016/s0140-6736(06)69666-9.

[266] H. E. Paulus. FDA arthritis advisory committee meeting: serious gastrointestinal toxicity of nonsteroidal antiinflammatory drugs; drugcontaining renal

and biliary stones; diclofenac and carprofen approved. *Arthritis Rheum*, 31(11): 1450–1451, 1988. ISSN 1529-0131.

[267] P. Guyot, S. Pandhi, R. M. Nixon, A. Iqbal, R. L. Chaves, and R. Andrew Moore. Efficacy and safety of diclofenac in osteoarthritis: Results of a network meta-analysis of unpublished legacy studies. *Scand J Pain*, 16:74–88, 2017. ISSN 1877-8860. doi: 10.1016/j.sjpain.2017.03.006.

[268] N. Latimer, J. Lord, R. L. Grant, R. O'Mahony, J. Dickson, and P. G. Conaghan. Value of information in the osteoarthritis setting: cost effectiveness of COX-2 selective inhibitors, traditional NSAIDs and proton pump inhibitors. *Pharmacoeconomics*, 29(3):225–37, 2011. ISSN 1170-7690. doi: 10.2165/11584200-000000000-00000.

[269] S. R. Smith, B. R. Deshpande, J. E. Collins, J. N. Katz, and E. Losina. Comparative pain reduction of oral non-steroidal anti-inflammatory drugs and opioids for knee osteoarthritis: systematic analytic review. *Osteoarthritis Cartilage*, 24 (6):962–72, 2016. ISSN 1063-4584. doi: 10.1016/j.joca.2016.01.135.

[270] M. C. Hochberg, J. Martel-Pelletier, J. Monfort, I. Moller, J. R. Castillo, N. Arden, F. Berenbaum, F. J. Blanco, P. G. Conaghan, G. Domenech, Y. Henrotin, T. Pap, P. Richette, A. Sawitzke, P. du Souich, and J. P. Pelletier. Combined chondroitin sulfate and glucosamine for painful knee osteoarthritis: a multicentre, randomised, double-blind, non-inferiority trial versus celecoxib. *Ann Rheum Dis*, 75(1):37–44, 2016. ISSN 0003-4967. doi: 10.1136/annrheumdis-2014-206792.

[271] B. Copsey, J. Buchanan, R. Fitzpatrick, S. E. Lamb, S. J. Dutton, and J. A. Cook. Duration of treatment effect should be considered in the design and interpretation of clinical trials: Results of a discrete choice experiment. *Med Decis Making*, 39(4):461–473, 2019. ISSN 0272-989x. doi: 10.1177/0272989x19841877.

[272] L. E. Dahlberg, I. Holme, K. Hoye, and B. Ringertz. A randomized, multicentre, double-blind, parallel-group study to assess the adverse event-related discontinuation rate with celecoxib and diclofenac in elderly patients with osteoarthritis. *Scand J Rheumatol*, 38(2):133–43, 2009. ISSN 0300-9742. doi: 10.1080/03009740802419065.

[273] F. E. Silverstein, G. Faich, J. L. Goldstein, L. S. Simon, T. Pincus, A. Whelton, R. Makuch, G. Eisen, N. M. Agrawal, W. F. Stenson, A. M. Burr, W. W. Zhao, J. D. Kent, J. B. Lefkowith, K. M. Verburg, and G. S. Geis. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: the CLASS study: A randomized controlled trial. Celecoxib Long-term Arthritis Safety Study. *Jama*, 284(10):1247–55, 2000. ISSN 0098-7484 (Print) 0098-7484.

[274] M. E. Farkouh, H. Kirshner, R. A. Harrington, S. Ruland, F. W. Verheugt, T. J. Schnitzer, G. R. Burmester, E. Mysler, M. C. Hochberg, M. Doherty, E. Ehrsam, X. Gitton, G. Krammer, B. Mellein, A. Gimona, P. Matchaba,

C. J. Hawkey, and J. H. Chesebro. Comparison of lumiracoxib with naproxen and ibuprofen in the Therapeutic Arthritis Research and Gastrointestinal Event Trial (TARGET), cardiovascular outcomes: randomised controlled trial. *Lancet*, 364(9435):675–84, 2004. ISSN 0140-6736. doi: 10.1016/s0140-6736(04)16894-3.

[275] L. S. Simon, H. T. Hatoum, R. M. Bittman, W. T. Archambault, and R. P. Polisson. Risk factors for serious nonsteroidal-induced gastrointestinal complications: regression analysis of the MUCOSA trial. *Fam Med*, 28(3):204–10, 1996. ISSN 0742-3225 (Print) 0742-3225.

[276] D. R. Ramey, D. J. Watson, C. Yu, J. A. Bolognese, S. P. Curtis, and A. S. Reicin. The incidence of upper gastrointestinal adverse events in clinical trials of etoricoxib vs. non-selective NSAIDs: an updated combined analysis. *Curr Med Res Opin*, 21(5):715–22, 2005. ISSN 0300-7995 (Print) 0300-7995.

[277] W. E. Smalley, W. A. Ray, J. R. Daugherty, and M. R. Griffin. Nonsteroidal anti-inflammatory drugs and the incidence of hospitalizations for peptic ulcer disease in elderly persons. *Am J Epidemiol*, 141(6):539–45, 1995. ISSN 0002-9262 (Print) 0002-9262.

[278] S. P. Gutthann, L. A. Garcia Rodriguez, and D. S. Raiford. Individual nonsteroidal antiinflammatory drugs and other risk factors for upper gastrointestinal bleeding and perforation. *Epidemiology*, 8(1):18–24, 1997. ISSN 1044-3983 (Print) 1044-3983.

[279] L. A. Garcia Rodriguez and S. Hernandez-Diaz. The risk of upper gastrointestinal complications associated with nonsteroidal anti-inflammatory drugs, glucocorticoids, acetaminophen, and combinations of these agents. *Arthritis Res*, 3(2):98–101, 2001. ISSN 1465-9905 (Print) 1465-9905. doi: 10.1186/ar146.

[280] J. L. Goldstein, F. E. Silverstein, N. M. Agrawal, R. C. Hubbard, J. Kaiser, C. J. Maurath, K. M. Verburg, and G. S. Geis. Reduced risk of upper gastrointestinal ulcer complications with celecoxib, a novel COX-2 inhibitor. *Am J Gastroenterol*, 95(7):1681–90, 2000. ISSN 0002-9270 (Print) 0002-9270. doi: 10.1111/j.1572-0241.2000.02194.x.

[281] T. Turajane, R. Wongbunnak, T. Patcharatrakul, K. Ratansumawong, Y. Poigampetch, and T. Songpatanasilp. Gastrointestinal and cardiovascular risk of non-selective NSAIDs and COX-2 inhibitors in elderly patients with knee osteoarthritis. *J Med Assoc Thai*, 92 Suppl 6:S19–26, 2009. ISSN 0125-2208 (Print) 0125-2208.

[282] J. Kongtharvonskul, T. Anothaisintawee, M. McEvoy, J. Attia, P. Woratanarat, and A. Thakkinstian. Efficacy and safety of glucosamine, diacerein, and NSAIDs in osteoarthritis knee: a systematic review and network meta-analysis. *Eur J Med Res*, 20:24, 2015. ISSN 0949-2321. doi: 10.1186/s40001-015-0115-7.

[283] J. A Singh, T. Wilt, and R. MacDonald. Chondroitin for osteoarthritis. *The Cochrane Library*, 2006.

[284] L. Puljak, A. Marin, D. Vrdoljak, F. Markotic, A. Utrobicic, and P. Tugwell. Celecoxib for osteoarthritis. *Cochrane Database Syst Rev*, 5:Cd009865, 2017. ISSN 1361-6137. doi: 10.1002/14651858.CD009865.pub2.

[285] D. Rochon, J. M. Eberth, L. Fraenkel, R. J. Volk, and S. N. Whitney. Elderly patients' experiences using adaptive conjoint analysis software as a decision aid for osteoarthritis of the knee. *Health Expect*, 17(6):840–51, 2014. ISSN 1369-6513. doi: 10.1111/j.1369-7625.2012.00811.x.

[286] M. Bech, T. Kjaer, and J. Lauridsen. Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. *Health Econ*, 20(3):273–86, 2011. ISSN 1057-9230. doi: 10.1002/hec.1587.

[287] F. Reed Johnson, E. Lancsar, D. Marshall, V. Kilambi, A. Muhlbacher, D. A. Regier, B. W. Bresnahan, B. Kanninen, and J. F. Bridges. Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task force. *Value Health*, 16(1):3–13, 2013. ISSN 1098-3015. doi: 10.1016/j.jval.2012.08.2223.

[288] M. Czajkowski, M. Giergiczny, and W. H. Greene. Learning and fatigue effects revisited. the impact of accounting for unobservable preference and scale heterogeneity on perceived ordering effects in multiple choice task discrete choice experiments. Report, University of Warsaw, 2012.

[289] L. J. Mangham, K. Hanson, and B. McPake. How to do (or not to do) ... designing a discrete choice experiment for application in a low-income country. *Health Policy Plan*, 24(2):151–8, 2009. ISSN 0268-1080 (Print) 0268-1080. doi: 10.1093/heapol/czn047.

[290] Choice Metrics. Ngene 1.1. 1 user manual and reference guide. *Sydney, Australia: ChoiceMetrics*, 2012.

[291] N. K. Arden, A. B. Hauber, A. F. Mohamed, F. R. Johnson, P. M. Peloso, D. J. Watson, P. Mavros, A. Gammaitoni, S. S. Sen, and S. D. Taylor. How do physicians weigh benefits and risks associated with treatments in patients with osteoarthritis in the United Kingdom? *J Rheumatol*, 39(5):1056–63, 2012. ISSN 0315-162X (Print) 0315-162x. doi: 10.3899/jrheum.111066.

[292] Limesurvey GmbH. Limesurvey: An open source survey tool. *LimeSurvey Project Hamburg, Germany.*, 2012. URL `http://www.limesurvey.org`.

[293] E. W. de Bekker-Grob, B. Donkers, M. F. Jonker, and E. A. Stolk. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. *Patient*, 8(5):373–84, 2015. ISSN 1178-1653. doi: 10.1007/s40271-015-0118-z.

[294] D. Marshall, J. F. Bridges, B. Hauber, R. Cameron, L. Donnalley, K. Fyie, and F. R. Johnson. Conjoint analysis applications in health - How are studies being designed and reported?: An update on current practice in the published literature between 2005 and 2008. *Patient*, 3(4):249–56, 2010. ISSN 1178-1653 (Print) 1178-1653. doi: 10.2165/11539650-000000000-00000.

[295] Bryan Orme. Sample size issues for conjoint analysis studies. *Sawthooth Software Research paper Series. Squim, WA, USA: Sawthooth Software Inc*, 1998.

[296] M. Bech and D. Gyrd-Hansen. Effects coding in discrete choice experiments. *Health Econ*, 14(10):1079–83, 2005. ISSN 1057-9230 (Print) 1057-9230. doi: 10.1002/hec.984.

[297] Computer program. Stata IC 15. *StataCorp TX, USA.*, Accessed Apr 2019. URL `www.stata.com`.

[298] Daniel McFadden. *Conditional logit analysis of qualitative choice behavior*. Frontiers In Econometrics. Academic Press, New York, 1973.

[299] K. Train. Halton sequences for mixed logit. *Econometrics*, 2001(0012002):1–18, 2001. URL `https://ideas.repec.org/p/wpa/wuwpem/0012002.html`.

[300] Revelt David and Train Kenneth. Customer-specific taste parameters and mixed logit: Households' choice of electricity supplier. Report, University of California at Berkeley, 2000. URL `https://EconPapers.repec.org/RePEc:ucb:calbwp:e00-274`.

[301] F. S. Miguel, M. Ryan, and M. Amaya-Amaya. 'Irrational' stated preferences: a quantitative and qualitative investigation. *Health Econ*, 14(3):307–22, 2005. ISSN 1057-9230 (Print) 1057-9230. doi: 10.1002/hec.912.

[302] J. Swait and W. Adamowicz. Choice environment, market complexity, and consumer behavior: a theoretical and empirical approach for incorporating decision complexity into models of consumer choice. *Organ Behav Hum Decis Process*, 86(2):141–167, 2001. ISSN 0749-5978.

[303] J. Veldwijk, D. Determann, M. S. Lambooij, J. A. van Til, I. J. Korfage, E. W. de Bekker-Grob, and G. A. de Wit. Exploring how individuals complete the choice tasks in a discrete choice experiment: an interview study. *BMC Med Res Methodol*, 16:45, 2016. ISSN 1471-2288. doi: 10.1186/s12874-016-0140-4.

[304] D. Yu, K. P. Jordan, J. Bedson, M. Englund, F. Blyth, A. Turkiewicz, D. Prieto-Alhambra, and G. Peat. Population trends in the incidence and initial management of osteoarthritis: age-period-cohort analysis of the Clinical Practice Research Datalink, 1992-2013. *Rheumatology (Oxford)*, 56(11):1902–1917, 2017. ISSN 1462-0324. doi: 10.1093/rheumatology/kex270.

[305] H. Jackson, L. A. Barnett, K. P. Jordan, K. S. Dziedzic, E. Cottrell, A. G. Finney, Z. Paskins, and J. J. Edwards. Patterns of routine primary care for osteoarthritis in the UK: a cross-sectional electronic health records study. *BMJ Open*, 7(12):e019694, 2017. ISSN 2044-6055. doi: 10.1136/bmjopen-2017-019694.

[306] D. Culliford, J. Maskell, A. Judge, C. Cooper, D. Prieto-Alhambra, and N. K. Arden. Future projections of total hip and knee arthroplasty in the UK: results

from the UK Clinical Practice Research Datalink. *Osteoarthritis Cartilage*, 23 (4):594–600, 2015. ISSN 1063-4584. doi: 10.1016/j.joca.2014.12.022.

[307] R. Duncan, G. Peat, E. Thomas, E. M. Hay, and P. Croft. Incidence, progression and sequence of development of radiographic knee osteoarthritis in a symptomatic population. *Ann Rheum Dis*, 70(11):1944–8, 2011. ISSN 0003-4967. doi: 10.1136/ard.2011.151050.

[308] J. Broadbent, S. Maisey, R. Holland, and N. Steel. Recorded quality of primary care for osteoarthritis: an observational study. *Br J Gen Pract*, 58(557):839–43, 2008. ISSN 0960-1643. doi: 10.3399/bjgp08X376177.

[309] M. Grotle, K. B. Hagen, B. Natvig, F. A. Dahl, and T. K. Kvien. Obesity and osteoarthritis in knee, hip and/or hand: an epidemiological study in the general population with 10 years follow-up. *BMC Musculoskelet Disord*, 9:132, 2008. ISSN 1471-2474. doi: 10.1186/1471-2474-9-132.

[310] P. G. Conaghan, P. M. Peloso, S. V. Everett, S. Rajagopalan, C. M. Black, P. Mavros, N. K. Arden, C. J. Phillips, F. Rannou, and et al. Inadequate pain relief and large functional loss among patients with knee osteoarthritis: evidence from a prospective multinational longitudinal study of osteoarthritis real-world therapies. *Rheumatology (Oxford)*, 54(2):270–7, 2015. ISSN 1462-0324. doi: 10.1093/rheumatology/keu332.

[311] K. L. Paterson, J. Kasza, D. J. Hunter, R. S. Hinman, H. B. Menz, G. Peat, and K. L. Bennell. Longitudinal association between foot and ankle symptoms and worsening of symptomatic radiographic knee osteoarthritis: data from the osteoarthritis initiative. *Osteoarthritis Cartilage*, 25(9):1407–1413, 2017. ISSN 1063-4584. doi: 10.1016/j.joca.2017.05.002.

[312] A. Shelbaya, C. T. Solem, C. Walker, Y. Wan, C. Johnson, and J. C. Cappelleri. The economic and clinical burden of early versus late initiation of celecoxib among patients with osteoarthritis. *Clinicoecon Outcomes Res*, 10:213–222, 2018. ISSN 1178-6981 (Print) 1178-6981. doi: 10.2147/ceor.s140208.

[313] S. E. Nissen. Cardiovascular safety of celecoxib, naproxen, or ibuprofen for arthritis. *N Engl J Med*, 376(14):1390, 2017. ISSN 0028-4793. doi: 10.1056/NEJMc1702534.

[314] G. Vanderstraeten, T. M. Lejeune, H. Piessevaux, D. De Bacquer, C. Walker, and B. De Beleyr. Gastrointestinal risk assessment in patients requiring non-steroidal anti-inflammatory drugs for osteoarthritis: The GIRANO study. *J Rehabil Med*, 48(8):705–710, 2016. ISSN 1650-1977. doi: 10.2340/16501977-2119.

[315] M. J. Parkes, M. J. Callaghan, L. Tive, M. Lunt, and D. T. Felson. Responsiveness of single versus composite measures of pain in knee osteoarthritis. *J Rheumatol*, 2018. ISSN 0315-162X (Print) 0315-162x. doi: 10.3899/jrheum.170928.

[316] J. Lin, W. Zhang, A. Jones, and M. Doherty. Efficacy of topical non-steroidal anti-inflammatory drugs in the treatment of osteoarthritis: meta-analysis of randomised controlled trials. *BMJ*, 329(7461):324, 2004. ISSN 0959-8138. doi: 10.1136/bmj.38159.639028.7C.

[317] A. Arfe, L. Scotti, C. Varas-Lorenzo, F. Nicotra, A. Zambon, B. Kollhorst, T. Schink, E. Garbe, R. Herings, H. Straatman, and et al. Non-steroidal anti-inflammatory drugs and risk of heart failure in four European countries: nested case-control study. *BMJ*, 354:i4857, 2016. ISSN 0959-8138. doi: 10.1136/bmj.i4857.

[318] R. Raja, B. Dube, E. M. Hensor, S. F. Hogg, P. G. Conaghan, and S. R. Kingsbury. The clinical characteristics of older people with chronic multiple-site joint pains and their utilisation of therapeutic interventions: data from a prospective cohort study. *BMC Musculoskelet Disord*, 17:194, 2016. ISSN 1471-2474. doi: 10.1186/s12891-016-1049-0.

[319] J. Martel-Pelletier, A. J. Barr, F. M. Cicuttini, P. G. Conaghan, C. Cooper, M. B. Goldring, S. R. Goldring, G. Jones, A. J. Teichtahl, and J. P. Pelletier. Osteoarthritis. *Nat Rev Dis Primers*, 2:16072, 2016. ISSN 2056-676x. doi: 10.1038/nrdp.2016.72.

[320] S. A. Abdalbary. Ultrasound with mineral water or aqua gel to reduce pain and improve the WOMAC of knee osteoarthritis. *Future Sci OA*, 2(1):Fso110, 2016. ISSN 2056-5623 (Print) 2056-5623. doi: 10.4155/fsoa-2016-0003.

[321] D. O. Clegg, D. J. Reda, C. L. Harris, M. A. Klein, J. R. O'Dell, M. M. Hooper, J. D. Bradley, C. O. Bingham (3rd), M. H. Weisman, C. G. Jackson, and et al. Glucosamine, chondroitin sulfate, and the two in combination for painful knee osteoarthritis. *N Engl J Med*, 354(8):795–808, 2006. ISSN 0028-4793. doi: 10.1056/NEJMoa052771.

[322] A. Kahan, D. Uebelhart, F. De Vathaire, P. D. Delmas, and J. Y. Reginster. Long-term effects of chondroitins 4 and 6 sulfate on knee osteoarthritis: the study on osteoarthritis progression prevention, a two-year, randomized, double-blind, placebo-controlled trial. *Arthritis Rheum*, 60(2):524–33, 2009. ISSN 0004-3591 (Print) 0004-3591. doi: 10.1002/art.24255.

[323] R. Whittle, K. P. Jordan, E. Thomas, and G. Peat. Average symptom trajectories following incident radiographic knee osteoarthritis: data from the Osteoarthritis Initiative. *RMD Open*, 2(2):e000281, 2016. ISSN 2056-5933 (Print) 2056-5933. doi: 10.1136/rmdopen-2016-000281.

[324] B. Al-Omari, J. Sim, P. Croft, and M. Frisher. Generating individual patient preferences for the treatment of osteoarthritis using adaptive choice-based conjoint (ACBC) analysis. *Rheumatol Ther*, 4(1):167–182, 2017. ISSN 2198-6576 (Print) 2198-6576. doi: 10.1007/s40744-017-0056-4.

[325] J. Chang, T. L. Kauf, S. Mahajan, J. M. Jordan, V. B. Kraus, T. P. Vail, S. D.

Reed, M. A. Omar, K. H. Kahler, and K. A. Schulman. Impact of disease severity and gastrointestinal side effects on the health state preferences of patients with osteoarthritis. *Arthritis Rheum*, 52(8):2366–75, 2005. ISSN 0004-3591 (Print) 0004-3591. doi: 10.1002/art.21227.

[326] M. Underwood, D. Ashby, P. Cross, E. Hennessy, L. Letley, J. Martin, S. Mt-Isa, S. Parsons, M. Vickers, and K. Whyte. Advice to use topical or oral ibuprofen for chronic knee pain in older people: randomised controlled trial and patient preference study. *BMJ*, 336(7636):138–42, 2008. ISSN 0959-8138. doi: 10.1136/bmj.39399.656331.25.

[327] J. Ratcliffe, M. Buxton, T. McGarry, R. Sheldon, and J. Chancellor. Patients' preferences for characteristics associated with treatments for osteoarthritis. *Rheumatology (Oxford)*, 43(3):337–45, 2004. ISSN 1462-0324 (Print) 1462-0324. doi: 10.1093/rheumatology/keh038.

[328] C. Berchi, P. Degieux, H. Halhol, B. Danel, M. Bennani, and C. Philippe. Impact of falling reimbursement rates on physician preferences regarding drug therapy for osteoarthritis using a discrete choice experiment. *Int J Pharm Pract*, 24(2):114–22, 2016. ISSN 0961-7671. doi: 10.1111/ijpp.12220.

[329] J. A. Kopec, C. G. Richardson, H. Llewellyn-Thomas, A. Klinkhoff, A. Carswell, and A. Chalmers. Probabilistic threshold technique showed that patients' preferences for specific trade-offs between pain relief and each side effect of treatment in osteoarthritis varied. *J Clin Epidemiol*, 60(9):929–38, 2007. ISSN 0895-4356 (Print) 0895-4356. doi: 10.1016/j.jclinepi.2007.01.001.

[330] W. Louthrenoo, S. Nilganuwong, S. Aksaranugraha, P. Asavatanabodee, and S. Saengnipanthkul. The efficacy, safety and carry-over effect of diacerein in the treatment of painful knee osteoarthritis: a randomised, double-blind, NSAID-controlled study. *Osteoarthritis Cartilage*, 15(6):605–14, 2007. ISSN 1063-4584 (Print) 1063-4584. doi: 10.1016/j.joca.2007.02.021.

[331] W. J. Zheng, F. L. Tang, J. Li, F. C. Zhang, Z. G. Li, Y. Su, D. H. Wu, L. Ma, H. Q. Zhou, F. Huang, J. L. Zhang, D. F. Liang, Y. X. Zhou, and H. Xu. Evaluation of efficacy and safety of diacerein in knee osteoarthritis in Chinese patients. *Chin Med Sci J*, 21(2):75–80, 2006. ISSN 1001-9294 (Print) 1001-9294.

[332] C. Bombardier, L. Laine, A. Reicin, D. Shapiro, R. Burgos-Vargas, B. Davis, R. Day, M. B. Ferraz, C. J. Hawkey, M. C. Hochberg, T. K. Kvien, and T. J. Schnitzer. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR study group. *N Engl J Med*, 343(21):1520–8, 2 p following 1528, 2000. ISSN 0028-4793 (Print) 0028-4793. doi: 10.1056/nejm200011233432103.

[333] M. Burnier. The safety of rofecoxib. *Expert Opin Drug Saf*, 4(3):491–9, 2005. ISSN 1474-0338. doi: 10.1517/14740338.4.3.491.

[334] P. A. Dieppe, S. Ebrahim, R. M. Martin, and P. Juni. Lessons from the

withdrawal of rofecoxib. *BMJ*, 329(7471):867–8, 2004. ISSN 0959-8138. doi: 10.1136/bmj.329.7471.867.

[335] M. R. Weir, R. S. Sperling, A. Reicin, and B. J. Gertz. Selective COX-2 inhibition and cardiovascular effects: a review of the rofecoxib development program. *Am Heart J*, 146(4):591–604, 2003. ISSN 0002-8703. doi: 10.1016/s0002-8703(03)00398-3.

[336] Z. Varga, S. R. A. Sabzwari, and V. Vargova. Cardiovascular risk of nonsteroidal anti-inflammatory drugs: An under-recognized public health issue. *Cureus*, 9 (4):e1144, 2017. ISSN 2168-8184 (Print) 2168-8184. doi: 10.7759/cureus.1144.

[337] DB Adekanmbi. Comparison of probit and logit models for binary response variable with applications to birth data in South-Western Nigeria. *American Journal of Mathematics and Statistics*, 7(5):199–208, 2017. ISSN 2162-8475.

[338] K. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2003. ISBN 9780521017152.

[339] M. Ryan, N. Krucien, and F. Hermens. The eyes have it: Using eye tracking to inform information processing strategies in multi-attributes choices. *Health Econ*, 2017. ISSN 1057-9230. doi: 10.1002/hec.3626.

[340] C. Vass, D. Rigby, K. Tate, A. Stewart, and K. Payne. An exploratory application of eye-tracking methods in a discrete choice experiment. *Med Decis Making*, 38(6):658–672, 2018. ISSN 0272-989x. doi: 10.1177/0272989x18782197.

[341] R. Norman, M. T. King, D. Clarke, R. Viney, P. Cronin, and D. Street. Does mode of administration matter? Comparison of online and face-to-face administration of a time trade-off task. *Qual Life Res*, 19(4):499–508, 2010. ISSN 0962-9343. doi: 10.1007/s11136-010-9609-5.

[342] M. M. Byrne, J. Souchek, M. Richardson, and M. Suarez-Almazor. Racial/ethnic differences in preferences for total knee replacement surgery. *J Clin Epidemiol*, 59(10):1078–86, 2006. ISSN 0895-4356 (Print) 0895-4356. doi: 10.1016/j.jclinepi.2006.01.010.

[343] M. K. Figaro, P. W. Russo, and J. P. Allegrante. Preferences for arthritis care among urban African Americans: "I don't want to be cut". *Health Psychol*, 23(3):324–9, 2004. ISSN 0278-6133 (Print) 0278-6133. doi: 10.1037/0278-6133.23.3.324.

[344] L. R. Hausmann, M. Mor, B. H. Hanusa, S. Zickmund, P. Z. Cohen, R. Grant, D. M. Kresevic, H. S. Gordon, B. S. Ling, C. K. Kwoh, and S. A. Ibrahim. The effect of patient race on total joint replacement recommendations and utilization in the orthopedic setting. *J Gen Intern Med*, 25(9):982–8, 2010. ISSN 0884-8734. doi: 10.1007/s11606-010-1399-5.

[345] W. F. Huang, F. Y. Hsiao, Y. W. Wen, and Y. W. Tsai. Cardiovascular events associated with the use of four nonselective NSAIDs (etodolac, nabume-

tone, ibuprofen, or naproxen) versus a cyclooxygenase-2 inhibitor (celecoxib): a population-based analysis in Taiwanese adults. *Clin Ther*, 28(11):1827–36, 2006. ISSN 0149-2918 (Print) 0149-2918. doi: 10.1016/j.clinthera.2006.11.009.

[346] T. Vanniyasingam, C. E. Cunningham, G. Foster, and L. Thabane. Simulation study to determine the impact of different design features on design efficiency in discrete choice experiments. *BMJ Open*, 6(7):e011985, 2016. ISSN 2044-6055. doi: 10.1136/bmjopen-2016-011985.

[347] N. Bansback, M. Hudson, C. Koehn, T.-L. Laba, and M. Harrison. FRI0607 how to design clinical trials to be more patient oriented: an example from preventative treatments for rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 77(Suppl 2):827–827, 2018. doi: 10.1136/annrheumdis-2018-eular.5712.

[348] C. Cedraschi, S. Delezay, M. Marty, F. Berenbaum, D. Bouhassira, Y. Henrotin, F. Laroche, and S. Perrot. "Let's talk about OA pain": a qualitative analysis of the perceptions of people suffering from OA. Towards the development of a specific pain OA-related questionnaire, the Osteoarthritis Symptom Inventory Scale (OASIS). *PLoS One*, 8(11):e79988, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0079988.

[349] M. G. Rabbani, S. A. Haq, N. Bellamy, M. N. Islam, M. R. Choudhury, A. Naheed, S. Ahmed, and A. Shahin. Development, linguistic and clinimetric validation of the WOMAC VA3.01 Bangla for Bangladesh Index. *Rheumatol Int*, 35(6):997–1003, 2015. ISSN 0172-8172. doi: 10.1007/s00296-014-3192-y.

[350] A. M. Davis, E. M. Badley, D. E. Beaton, J. Kopec, J. G. Wright, N. L. Young, and J. I. Williams. Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. *J Clin Epidemiol*, 56(11):1076–83, 2003. ISSN 0895-4356 (Print) 0895-4356.

[351] L. A. Garcia Rodriguez, A. Gonzalez-Perez, H. Bueno, and J. Hwa. Nsaid use selectively increases the risk of non-fatal myocardial infarction: a systematic review of randomised trials and observational studies. *PLoS One*, 6(2):e16780, 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0016780.

[352] A. Edwards, G. Elwyn, and A. Mulley. Explaining risks: turning numerical data into meaningful pictures. *BMJ*, 324(7341):827–830, 2002. ISSN 0959-8138 1468-5833.

[353] N. Bansback, M. Harrison, and C. Marra. Does introducing imprecision around probabilities for benefit and harm influence the way people value treatments? *Med Decis Making*, 36(4):490–502, 2016. ISSN 0272-989x. doi: 10.1177/0272989x15600708.

[354] M. Harrison, C. A. Marra, and N. Bansback. Preferences for 'new' treatments diminish in the face of ambiguity. *Health Econ*, 26(6):743–752, 2017. ISSN 1057-9230. doi: 10.1002/hec.3353.

[355] D. Hughes, J. Charles, D. Dawoud, R. T. Edwards, E. Holmes, C. Jones, P. Parham, C. Plumpton, C. Ridyard, H. Lloyd-Williams, E. Wood, and S. T. Yeo. Conducting economic evaluations alongside randomised trials: Current methodological issues and novel approaches. *Pharmacoeconomics*, 34(5):447–61, 2016. ISSN 1170-7690. doi: 10.1007/s40273-015-0371-y.

[356] Mandy Ryan, Karen Gerard, and Mabel Amaya-Amaya. *Using discrete choice experiments to value health and health care*, volume 11. Springer Science and Business Media, 2007. ISBN 1402057539.

[357] M. J. Hancock, C. G. Maher, J. Latimer, A. J. McLachlan, C. W. Cooper, R. O. Day, M. F. Spindler, and J. H. McAuley. Assessment of diclofenac or spinal manipulative therapy, or both, in addition to recommended first-line treatment for acute low back pain: a randomised controlled trial. *Lancet*, 370(9599): 1638–43, 2007. ISSN 0140-6736. doi: 10.1016/s0140-6736(07)61686-9.

[358] D. L. Mason, V. A. Dickens, and A. Vail. Rehabilitation for hamstring injuries. *Cochrane Database Syst Rev*, 12:Cd004575, 2012. ISSN 1361-6137. doi: 10.1002/14651858.CD004575.pub3.

[359] C. M. Williams, C. G. Maher, J. Latimer, A. J. McLachlan, M. J. Hancock, R. O. Day, and C. W. Lin. Efficacy of paracetamol for acute low-back pain: a double-blind, randomised controlled trial. *Lancet*, 384(9954):1586–96, 2014. ISSN 0140-6736. doi: 10.1016/s0140-6736(14)60805-9.

[360] National Clinical Guideline Centre UK. Osteoarthritis: care and management in adults. *NICE*, 2014.

[361] L. A. Deveza, D. J. Hunter, and W. E. Van Spil. Too much opioid, too much harm. *Osteoarthritis Cartilage*, 26(3):293–295, 2018. ISSN 1063-4584. doi: 10.1016/j.joca.2017.12.003.

[362] R. Llewellyn-Bennett, D. Edwards, N. Roberts, A. H. Hainsworth, R. Bulbulia, and L. Bowman. Post-trial follow-up methodology in large randomised controlled trials: a systematic review. *Trials*, 19(1):298, 2018. ISSN 1745-6215. doi: 10.1186/s13063-018-2653-0.

[363] D. Brixner, E. O. Meltzer, K. Morland, C. A. Carroll, U. Munzel, and B. J. Lipworth. Implication of alternative minimal clinically important difference threshold estimation methods on technology assessment. *Int J Technol Assess Health Care*, 32(6):371–375, 2016. ISSN 0266-4623. doi: 10.1017/s0266462316000593.

[364] J. E. Ang, H. R. Bin Abd Razak, T. S. Howe, B. K. Tay, and S. J. Yeo. Obesity does not affect outcomes in hybrid versus cemented total knee arthroplasty in Asians. *J Arthroplasty*, 32(12):3643–3646, 2017. ISSN 0883-5403. doi: 10.1016/j.arth.2017.06.043.

[365] W. L. Dai, Z. M. Lin, D. H. Guo, Z. J. Shi, and J. Wang. Efficacy and safety of hylan versus hyaluronic acid in the treatment of knee osteoarthritis. *J Knee Surg*, 2018. ISSN 1538-8506. doi: 10.1055/s-0038-1641142.

[366] K. M. Dunn, P. Campbell, and K. P. Jordan. Long-term trajectories of back pain: cohort study with 7-year follow-up. *BMJ Open*, 3(12):e003838, 2013. ISSN 2044-6055 (Print) 2044-6055. doi: 10.1136/bmjopen-2013-003838.

[367] B. Doganay Erdogan, Y. Y. Leung, C. Pohl, A. Tennant, and P. G. Conaghan. Minimal clinically important difference as applied in rheumatology: An OMERACT Rasch working group systematic review and critique. *J Rheumatol*, 43(1):194–202, 2016. ISSN 0315-162X (Print) 0315-162x. doi: 10.3899/jrheum.141150.

[368] A. Escobar, J. M. Quintana, A. Bilbao, I. Arostegui, I. Lafuente, and I. Vidaurreta. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. *Osteoarthritis Cartilage*, 15(3):273–80, 2007. ISSN 1063-4584 (Print) 1063-4584. doi: 10.1016/j.joca.2006.09.001.

[369] C. B. Terwee, L. D. Roorda, J. Dekker, S. M. Bierma-Zeinstra, G. Peat, K. P. Jordan, P. Croft, and H. C. de Vet. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol*, 63(5):524–34, 2010. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2009.08.010.

[370] Web link. OAI. Accessed 11 Dec 2018. URL `https://oai.epi-ucsf.org/datarelease/`.

[371] C. G. Peterfy, E. Schneider, and M. Nevitt. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis Cartilage*, 16(12):1433–41, 2008. ISSN 1063-4584. doi: 10.1016/j.joca.2008.06.016.

[372] Web link. OAI (updated). Accessed 22 Mar 2019. URL `https://data-archive.nimh.nih.gov/oai`.

[373] D. Hui, O. Shamieh, C. E. Paiva, O. Khamash, P. E. Perez-Cruz, J. H. Kwon, M. A. Muckaden, M. Park, J. Arthur, and E. Bruera. Minimal clinically important difference in the physical, emotional, and total symptom distress scores of the Edmonton symptom assessment system. *J Pain Symptom Manage*, 51(2): 262–9, 2016. ISSN 0885-3924. doi: 10.1016/j.jpainsymman.2015.10.004.

[374] D. E. Beaton, M. Boers, and G. A. Wells. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol*, 14(2):109–14, 2002. ISSN 1040-8711 (Print) 1040-8711.

[375] F. Angst, A. Aeschlimann, and J. Angst. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. *J Clin Epidemiol*, 82:128–136, 2017. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2016.11.016.

[376] R. Froud and G. Abel. Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: the forgotten lesson of pythagoras. theoretical considerations and an example application of

change in health status. *PLoS One*, 9(12):e114468, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0114468.

[377] A. J. Vickers and D. G. Altman. Statistics notes: Analysing controlled trials with baseline and follow up measurements. *BMJ*, 323(7321):1123–4, 2001. ISSN 0959-8138 (Print) 0959-8138. doi: 10.1136/bmj.323.7321.1123.

[378] G. A. Miller and J. P. Chapman. Misunderstanding analysis of covariance. *J Abnorm Psychol*, 110(1):40–8, 2001. ISSN 0021-843X (Print) 0021-843x.

[379] P. M. Fayers and R. D. Hays. Don't middle your MIDs: regression to the mean shrinks estimates of minimally important differences. *Qual Life Res*, 23(1):1–4, 2014. ISSN 0962-9343. doi: 10.1007/s11136-013-0443-4.

[380] M. M. Glymour, J. Weuve, L. F. Berkman, I. Kawachi, and J. M. Robins. When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *Am J Epidemiol*, 162(3):267–78, 2005. ISSN 0002-9262 (Print) 0002-9262. doi: 10.1093/aje/kwi187.

[381] Sophia Rabe-Hesketh and Anders Skrondal. *Multilevel and longitudinal modeling using Stata*. Stata Press, College Station, Texas, 2nd edition, 2008. ISBN 9781597180405.

[382] J. N. Matthews, D. G. Altman, M. J. Campbell, and P. Royston. Analysis of serial measurements in medical research. *BMJ*, 300(6719):230–5, 1990. ISSN 0959-8138 (Print) 0959-8138.

[383] Stephen John Walters. *Quality of life outcomes in clinical trials and health-care evaluation: a practical guide to analysis and interpretation*. Statistics in practice. John Wiley and Sons, Chichester, West Sussex, U.K., 2009.

[384] V. J. Williams, S. R. Piva, J. J. Irrgang, C. Crossley, and G. K. Fitzgerald. Comparison of reliability and responsiveness of patient-reported clinical outcome measures in knee osteoarthritis rehabilitation. *J Orthop Sports Phys Ther*, 42(8):716–23, 2012. ISSN 0190-6011. doi: 10.2519/jospt.2012.4038.

[385] D. L. McCreary, B. C. Sandberg, D. C. Bohn, H. R. Parikh, and B. P. Cunningham. Interpreting patient-reported outcome results: Is one minimum clinically important difference really enough? *Hand (N Y)*, page 1558944718812180, 2018. ISSN 1558-9447. doi: 10.1177/1558944718812180.

[386] M. Hung, J. F. Baumhauer, F. W. Licari, M. W. Voss, J. Bounsanga, and C. L. Saltzman. PROMIS and FAAM minimal clinically important differences in foot and ankle orthopedics. *Foot Ankle Int*, page 1071100718800304, 2018. ISSN 1071-1007. doi: 10.1177/1071100718800304.

[387] T. Ogura, J. Ackermann, A. Barbieri Mestriner, G. Merkely, and A. H. Gomoll. Minimal clinically important differences and substantial clinical benefit in patient-reported outcome measures after autologous chondrocyte im-

plantation. *Cartilage*, page 1947603518799839, 2018. ISSN 1947-6035. doi: 10.1177/1947603518799839.

[388] A. G. Copay, S. D. Glassman, B. R. Subach, S. Berven, T. C. Schuler, and L. Y. Carreon. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. *Spine J*, 8(6):968–74, 2008. ISSN 1529-9430 (Print) 1529-9430. doi: 10.1016/j.spinee.2007.11.006.

[389] K. Horvath, Z. Aschermann, M. Kovacs, A. Makkos, M. Harmat, J. Janszky, S. Komoly, K. Karadi, and N. Kovacs. Changes in quality of life in Parkinson's disease: How large must they be to be relevant? *Neuroepidemiology*, 48(1-2): 1–8, 2017. ISSN 0251-5350. doi: 10.1159/000455863.

[390] M. J. Diaz-Arribas, M. Fernandez-Serrano, A. Royuela, F. M. Kovacs, T. Gallego-Izquierdo, M. Ramos-Sanchez, R. Llorca-Palomera, P. Pardo-Hervas, and O. S. Martin-Pariente. Minimal clinically important difference in quality of life for patients with low back pain. *Spine (Phila Pa 1976)*, 42(24):1908–1916, 2017. ISSN 0362-2436. doi: 10.1097/brs.0000000000002298.

[391] G. Singla, M. Singh, A. Singh, I. Kaur, K. Harsh, and K. Jasmeen. Is sinonasal outcome test-22 reliable for guiding chronic rhinosinusitis patients for endoscopic sinus surgery? *Niger J Clin Pract*, 21(9):1228–1233, 2018. ISSN 1119-3077 (Print).

[392] A. G. Copay, B. Eyberg, A. S. Chung, K. S. Zurcher, N. Chutkan, and M. J. Spangehl. Minimum clinically important difference: Current trends in the orthopaedic literature, part II: Lower extremity: A systematic review. *JBJS Rev*, 6(9):e2, 2018. ISSN 2329-9185. doi: 10.2106/jbjs.rvw.17.00160.

[393] J. D. Maratt, Y. Y. Lee, S. Lyman, and G. H. Westrich. Predictors of satisfaction following total knee arthroplasty. *J Arthroplasty*, 30(7):1142–5, 2015. ISSN 0883-5403. doi: 10.1016/j.arth.2015.01.039.

[394] I. Hmamouchi, F. Allali, L. Tahiri, H. Khazzani, L. E. Mansouri, S. Ali Ou Alla, R. Abouqal, and N. Hajjaj-Hassouni. Clinically important improvement in the WOMAC and predictor factors for response to non-specific non-steroidal anti-inflammatory drugs in osteoarthritic patients: a prospective study. *BMC Res Notes*, 5:58, 2012. ISSN 1756-0500. doi: 10.1186/1756-0500-5-58.

[395] W. C. Lee, Y. H. Kwan, H. C. Chong, and S. J. Yeo. The minimal clinically important difference for Knee Society Clinical Rating System after total knee arthroplasty for primary osteoarthritis. *Knee Surg Sports Traumatol Arthrosc*, 25(11):3354–3359, 2017. ISSN 0942-2056. doi: 10.1007/s00167-016-4208-9.

[396] C. J. Hwang, R. Ellis, R. M. Davis, and S. Tolleson-Rinehart. Determination of the minimal clinically important difference of the University of North Carolina Dry Eye Management Scale. *Cornea*, 36(9):1054–1060, 2017. ISSN 0277-3740. doi: 10.1097/ico.0000000000001287.

[397] F. Angst, T. Benz, S. Lehmann, A. Aeschlimann, and J. Angst. Multidimensional minimal clinically important differences in knee osteoarthritis after comprehensive rehabilitation: a prospective evaluation from the Bad Zurzach Osteoarthritis Study. *RMD Open*, 4(2):e000685, 2018. ISSN 2056-5933 (Print) 2056-5933. doi: 10.1136/rmdopen-2018-000685.

[398] M. E. O'Connell, B. Gould, J. Ursenbach, J. Enright, and D. G. Morgan. Reliable change and minimum clinically important difference (MCID) of the repeatable battery for the assessment of neuropsychology status (RBANS) in a heterogeneous dementia sample: Support for reliable change methods but not the MCID. *Appl Neuropsychol Adult*, pages 1–7, 2018. ISSN 2327-9095. doi: 10.1080/23279095.2017.1413575.

[399] D. Revicki, R. D. Hays, D. Cella, and J. Sloan. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*, 61(2):102–9, 2008. ISSN 0895-4356 (Print) 0895-4356. doi: 10.1016/j.jclinepi.2007.03.012.

[400] A. Akaberi, F. A. Klok, D. M. Cohn, A. Hirsch, J. Granton, and S. R. Kahn. Determining the minimal clinically important difference for the PEmbQoL questionnaire, a measure of pulmonary embolism-specific quality of life. *J Thromb Haemost*, 2018. ISSN 1538-7836. doi: 10.1111/jth.14302.

[401] G. Kjellsson, P. Clarke, and U. G. Gerdtham. Forgetting to remember or remembering to forget: a study of the recall period length in health care survey questions. *J Health Econ*, 35:34–46, 2014. ISSN 0167-6296. doi: 10.1016/j.jhealeco.2014.01.007.

[402] J. S. Schmitt and J. H. Abbott. Patient global ratings of change did not adequately reflect change over time: a clinical cohort study. *Phys Ther*, 94(4): 534–42, 2014. ISSN 0031-9023. doi: 10.2522/ptj.20130162.

[403] M. L. Ferreira, P. H. Ferreira, R. D. Herbert, and J. Latimer. People with low back pain typically need to feel 'much better' to consider intervention worthwhile: an observational study. *Aust J Physiother*, 55(2):123–7, 2009. ISSN 0004-9514 (Print) 0004-9514.

[404] W. Wang, Y. Ma, Y. Huang, and H. Chen. Generalizability analysis for clinical trials: a simulation study. *Stat Med*, 36(10):1523–1531, 2017. ISSN 0277-6715. doi: 10.1002/sim.7238.

[405] S. W. Stirman, R. J. DeRubeis, P. Crits-Christoph, and P. E. Brody. Are samples in randomized controlled trials of psychotherapy representative of community outpatients? A new methodology and initial findings. *J Consult Clin Psychol*, 71(6):963–72, 2003. ISSN 0022-006X (Print) 0022-006x. doi: 10.1037/0022-006x.71.6.963.

[406] M. L. Bell, M. T. King, and D. L. Fairclough. Bias in area under the curve for longitudinal clinical trials with missing patient reported outcome data: sum-

mary measures versus summary statistics. *SAGE Open*, 4(2):2158244014534858, 2014. ISSN 2158-2440.

[407] M. L. Costa, J. Achten, S. Hennings, N. Boota, J. Griffin, S. Petrou, M. Maredza, M. Dritsaki, T. Wood, J. Masters, I. Pallister, S. E. Lamb, and N. R. Parsons. Intramedullary nail fixation versus locking plate fixation for adults with a fracture of the distal tibia: the UK FixDT RCT. *Health Technol Assess*, 22(25):1–148, 2018. ISSN 1366-5278. doi: 10.3310/hta22250.

[408] M. F. Olsen, E. Bjerre, M. D. Hansen, B. Tendal, J. Hilden, and A. Hrobjartsson. Minimum clinically important differences in chronic pain vary considerably by baseline pain and methodological factors: systematic review of empirical studies. *J Clin Epidemiol*, 101:87–106.e2, 2018. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2018.05.007.

[409] M. van Middelkoop, K. S. Dziedzic, M. Doherty, W. Zhang, J. W. Bijlsma, T. E. McAlindon, S. L. Lohmander, and S. M. Bierma-Zeinstra. Individual patient data meta-analysis of trials investigating the effectiveness of intra-articular glucocorticoid injections in patients with knee or hip osteoarthritis: an OA Trial Bank protocol for a systematic review. *Syst Rev*, 2:54, 2013. ISSN 2046-4053. doi: 10.1186/2046-4053-2-54.

[410] Web link. OA trial bank. Accessed 11 Dec 2018. URL `http://www.oatrialbank.com/`.

[411] A. Ousmen, T. Conroy, F. Guillemin, M. Velten, D. Jolly, M. Mercier, S. Causeret, J. Cuisenier, O. Graesslin, Z. Hamidou, F. Bonnetain, and A. Anota. Impact of the occurrence of a response shift on the determination of the minimal important difference in a health-related quality of life score over time. *Health Qual Life Outcomes*, 14(1):167, 2016. ISSN 1477-7525. doi: 10.1186/s12955-016-0569-5.

[412] A. Kiran, D. J. Hunter, A. Judge, R. E. Field, M. K. Javaid, C. Cooper, and N. K. Arden. A novel methodological approach for measuring symptomatic change following total joint arthroplasty. *J Arthroplasty*, 29(11):2140–5, 2014. ISSN 0883-5403. doi: 10.1016/j.arth.2014.06.008.

[413] J. E. Pope, D. Khanna, D. Norrie, and J. M. Ouimet. The minimally important difference for the health assessment questionnaire in rheumatoid arthritis clinical practice is smaller than in randomized controlled trials. *J Rheumatol*, 36(2): 254–9, 2009. ISSN 0315-162X (Print) 0315-162x. doi: 10.3899/jrheum.080479.

[414] Y. Guo, H. L. Logan, D. H. Glueck, and K. E. Muller. Selecting a sample size for studies with repeated measures. *BMC Med Res Methodol*, 13:100, 2013. ISSN 1471-2288. doi: 10.1186/1471-2288-13-100.

[415] T. Freitag, M. A. Hein, D. Wernerus, H. Reichel, and R. Bieger. Bone remodelling after femoral short stem implantation in total hip arthroplasty: 1-year

results from a randomized DEXA study. *Archives of Orthopaedic & Trauma Surgery*, 136(1):125–30, 2016.

[416] M. de Rooij, M. van der Leeden, J. Cheung, M. van der Esch, A. Hakkinen, D. Haverkamp, L. D. Roorda, J. Twisk, J. Vollebregt, W. F. Lems, and J. Dekker. Efficacy of tailored exercise therapy on physical functioning in patients with knee osteoarthritis and comorbidity: a randomized controlled trial [with consumer summary]. *Arthritis Care & Research 2016 Aug 26:Epub ahead of print*, 2016.

[417] T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Stat Med*, 38(11):2074–2102, 2019. ISSN 0277-6715. doi: 10.1002/sim.8086.

[418] A. Richards. University of Oxford Advanced Research Computing. 2015. URL `Zenodo.10.5281/zenodo.22558`.

[419] C. R. Madan. Multiple statistical tests: Lessons from a d20. *F1000Res*, 5:1129, 2016. ISSN 2046-1402 (Print) 2046-1402. doi: 10.12688/f1000research.8834.2.

[420] A. M. De Livera, S. Zaloumis, and J. A. Simpson. Models for the analysis of repeated continuous outcome measures in clinical trials. *Respirology*, 19(2): 155–161, 2014. ISSN 1323-7799. doi: 10.1111/resp.12217.

[421] C. H. Mallinckrod, P. W. Lane, D. Schnell, Y. Peng, and J. P. Mancuso. Recommendations for the primary analysis of continuous endpoints in longitudinal clinical trials. *Drug Information Journal*, 42(4):303–319, 2008. doi: 10.1177/009286150804200402.

[422] A. Prakash, R. C. Risser, and C. H. Mallinckrodt. The impact of analytic method on interpretation of outcomes in longitudinal clinical trials. *Int J Clin Pract*, 62(8):1147–58, 2008. ISSN 1368-5031. doi: 10.1111/j.1742-1241.2008.01808.x.

[423] R. Chu, L. Thabane, J. Ma, A. Holbrook, E. Pullenayegum, and P. J. Devereaux. Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study. *BMC Med Res Methodol*, 11:21, 2011. ISSN 1471-2288. doi: 10.1186/1471-2288-11-21.

[424] G. Zhang and J. J. Chen. Adaptive fitting of linear mixed-effects models with correlated random-effects. *J Stat Comput Simul*, 83(12), 2013. ISSN 0094-9655. doi: 10.1080/00949655.2012.690763.

[425] B. A. Evans, Z. Feng, and A. V. Peterson. A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials. *Stat Med*, 20(22):3353–73, 2001. ISSN 0277-6715 (Print) 0277-6715.

[426] N. Mayer-Hamblett and R. A. Kronmal. Improving the estimation of change from baseline in a continuous outcome measure in the clinical trial setting.

Contemp Clin Trials, 26(1):2–16, 2005. ISSN 1551-7144 (Print) 1551-7144. doi: 10.1016/j.cct.2004.08.008.

[427] K. A. Hallgren, D. C. Atkins, and K. Witkiewitz. Aggregating and analyzing daily drinking data in clinical trials: A comparison of type I errors, power, and bias. *J Stud Alcohol Drugs*, 77(6):986–991, 2016. ISSN 1937-1888.

[428] Y. Ma, M. Mazumdar, and S. G. Memtsoudis. Beyond repeated-measures analysis of variance: advanced statistical methods for the analysis of longitudinal data in anesthesia research. *Reg Anesth Pain Med*, 37(1):99–105, 2012. ISSN 1098-7339. doi: 10.1097/AAP.0b013e31823ebc74.

[429] J. C. Gardiner, Z. Luo, and L. A. Roman. Fixed effects, random effects and GEE: what are the differences? *Stat Med*, 28(2):221–39, 2009. ISSN 0277-6715 (Print) 0277-6715. doi: 10.1002/sim.3478.

[430] P. Schober and T. R. Vetter. Repeated measures designs and analysis of longitudinal data: If at first you do not succeed-try, try again. *Anesth Analg*, 127 (2):569–575, 2018. ISSN 0003-2999. doi: 10.1213/ane.0000000000003511.

[431] G. F. Borm, J. Fransen, and W. A. Lemmens. A simple sample size formula for analysis of covariance in randomized clinical trials. *J Clin Epidemiol*, 60(12):1234–8, 2007. ISSN 0895-4356 (Print) 0895-4356. doi: 10.1016/j.jclinepi.2007.02.006.

[432] S. Zhang and C. Ahn. Sample size calculation for time-averaged differences in the presence of missing data. *Contemp Clin Trials*, 33(3):550–6, 2012. ISSN 1551-7144.

[433] Keith E. Muller and Paul Wilder Stewart. *Linear model theory : univariate, multivariate, and mixed models*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J., 2006. ISBN 9780471214885.

[434] I. C. Olsen, T. K. Kvien, and T. Uhlig. Consequences of handling missing data for treatment response in osteoarthritis: a simulation study. *Osteoarthritis Cartilage*, 20(8):822–8, 2012. ISSN 1063-4584. doi: 10.1016/j.joca.2012.03.005.

[435] H. M. Ghomrawi, L. A. Mandl, J. Rutledge, M. M. Alexiades, and M. Mazumdar. Is there a role for expectation maximization imputation in addressing missing data in research using WOMAC questionnaire? Comparison to the standard mean approach and a tutorial. *BMC Musculoskelet Disord*, 12:109, 2011. ISSN 1471-2474. doi: 10.1186/1471-2474-12-109.

[436] J. Twisk, L. Bosman, T. Hoekstra, J. Rijnhart, M. Welten, and M. Heymans. Different ways to estimate treatment effects in randomised controlled trials. *Contemp Clin Trials Commun*, 10:80–85, 2018. ISSN 2451-8654. doi: 10.1016/j.conctc.2018.03.008.

[437] M. L. Costa, J. Achten, J. Griffin, S. Petrou, I. Pallister, S. E. Lamb, and N. R. Parsons. Effect of locking plate fixation vs intramedullary nail fixation on 6-

month disability among adults with displaced fracture of the distal tibia: The UK FixDT randomized clinical trial. *Jama*, 318(18):1767–1776, 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.16429.

[438] J. Kasza, K. Hemming, R. Hooper, J. Matthews, and A. B. Forbes. Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Stat Methods Med Res*, 28(3):703–716, 2019. ISSN 0962-2802. doi: 10.1177/0962280217734981.

[439] Melanie L. Bell. New guidance to improve sample size calculations for trials: eliciting the target difference. *Trials*, 19(1):605–605, 2018. ISSN 1745-6215. doi: 10.1186/s13063-018-2894-y.

[440] R. H. Dworkin, S. Peirce-Sandner, D. C. Turk, M. P. McDermott, A. Gibofsky, L. S. Simon, J. T. Farrar, and N. P. Katz. Outcome measures in placebo-controlled trials of osteoarthritis: responsiveness to treatment effects in the REPORT database. *Osteoarthritis Cartilage*, 19(5):483–92, 2011. ISSN 1063-4584. doi: 10.1016/j.joca.2011.02.020.

[441] N. Bellamy, C. Wilson, J. Hendrikz, S. L. Whitehouse, B. Patel, S. Dennison, and T. Davis. Osteoarthritis index delivered by mobile phone (m-WOMAC) is valid, reliable, and responsive. *J Clin Epidemiol*, 64(2):182–90, 2011. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2010.03.013.

[442] J. O. Jansen, P. Pallmann, G. MacLennan, and M. K. Campbell. Bayesian clinical trial designs: Another option for trauma trials? *J Trauma Acute Care Surg*, 83(4):736–741, 2017. ISSN 2163-0755. doi: 10.1097/ta.0000000000001638.

[443] V. Prasad, C. Kim, M. Burotto, and A. Vandross. The strength of association between surrogate end points and survival in oncology: A systematic review of trial-level meta-analyses. *JAMA Intern Med*, 175(8):1389–98, 2015. ISSN 2168-6106. doi: 10.1001/jamainternmed.2015.2829.

[444] M. K. Wilson, K. Karakasis, and A. M. Oza. Outcomes and endpoints in trials of cancer treatment: the past, present, and future. *Lancet Oncol*, 16(1):e32–42, 2015. ISSN 1470-2045. doi: 10.1016/s1470-2045(14)70375-4.

[445] W. Sones, S. A. Julious, J. C. Rothwell, C. R. Ramsay, L. V. Hampson, R. Emsley, S. J. Walters, C. Hewitt, M. Bland, D. A. Fergusson, J. A. Berlin, D. Altman, L. D. Vale, and J. A. Cook. Choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial - the development of the DELTA2 guidance. *Trials*, 19(1):542, 2018. ISSN 1745-6215. doi: 10.1186/s13063-018-2887-x.

[446] R. Hooper, S. Teerenstra, E. de Hoop, and S. Eldridge. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Stat Med*, 35(26):4718–4728, 2016. ISSN 0277-6715. doi: 10.1002/sim.7028.

[447] K. Hemming, S. Eldridge, G. Forbes, C. Weijer, and M. Taljaard. How to design

efficient cluster randomised trials. *BMJ*, 358:j3064, 2017. ISSN 0959-8138. doi: 10.1136/bmj.j3064.

[448] J. A. Cook, S. A. Julious, W. Sones, L. V. Hampson, C. Hewitt, J. A. Berlin, D. Ashby, R. Emsley, D. A. Fergusson, S. J. Walters, E. C. F. Wilson, G. Maclennan, N. Stallard, J. C. Rothwell, M. Bland, L. Brown, C. R. Ramsay, A. Cook, D. Armstrong, D. Altman, and L. D. Vale. Delta(2) guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *Trials*, 19(1):606, 2018. ISSN 1745-6215. doi: 10.1186/s13063-018-2884-0.

[449] M. Boers, P. Brooks, J. F. Fries, L. S. Simon, V. Strand, and P. Tugwell. A first step to assess harm and benefit in clinical trials in one scale. *J Clin Epidemiol*, 63(6):627–32, 2010. ISSN 0895-4356. doi: 10.1016/j.jclinepi.2009.07.002.

[450] M. Boers, J. A. Singh, S. S. Cofield, Jr. Bridges, S. L., L. W. Moreland, J. R. O'Dell, H. Wu, S. Leatherman, and J. R. Curtis. A novel method to combine assessment of benefit and harm: Outcome measures in rheumatology 3x3 methodology applied to two active comparator trials. *Arthritis Care Res (Hoboken)*, 71(2):319–322, 2019. ISSN 2151-464x. doi: 10.1002/acr.23590.

[451] P. A. Shaw. Use of composite outcomes to assess risk-benefit in clinical trials. *Clin Trials*, 15(4):352–358, 2018. ISSN 1740-7745. doi: 10.1177/1740774518784010.

[452] T. W. LeBlanc and A. P. Abernethy. Patient-reported outcomes in cancer care - hearing the patient voice at greater volume. *Nat Rev Clin Oncol*, 14(12): 763–772, 2017. ISSN 1759-4774. doi: 10.1038/nrclinonc.2017.153.

[453] K. Gustafsson, O. Rolfson, M. Eriksson, L. Dahlberg, and J. Kvist. Study protocol for an observational register-based study on health and risk factors in patients with hip and knee osteoarthritis. *BMJ Open*, 8(10):e022812, 2018. ISSN 2044-6055. doi: 10.1136/bmjopen-2018-022812.

[454] O. Rolfson, E. Bohm, P. Franklin, S. Lyman, G. Denissen, J. Dawson, J. Dunn, K. Eresian Chenok, M. Dunbar, S. Overgaard, G. Garellick, and A. Lubbeke. Patient-reported outcome measures in arthroplasty registries report of the Patient-Reported Outcome Measures Working Group of the International Society of Arthroplasty Registries Part II. Recommendations for selection, administration, and analysis. *Acta Orthop*, 87 Suppl 1:9–23, 2016. ISSN 1745-3674. doi: 10.1080/17453674.2016.1181816.

[455] Web link. National Joint Registry. Accessed 19 Apr 2019. URL `http://www.njrcentre.org.uk/njrcentre/default.aspx`.

[456] R. W. Bohannon. Minimal clinically important difference for grip strength: a systematic review. *J Phys Ther Sci*, 31(1):75–78, 2019. ISSN 0915-5287 (Print) 0915-5287. doi: 10.1589/jpts.31.75.

[457] M. Harden and T. Friede. Sample size calculation in multi-centre clinical trials. *BMC Med Res Methodol*, 18(1):156, 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0602-y.

[458] Web link. Shiny CRT calculator: Power and sample size for cluster randomised trials. Accessed 18 Apr 2019. URL `https://clusterrcts.shinyapps.io/rshinyapp/`.

[459] L. E. Bothwell, J. Avorn, N. F. Khan, and A. S. Kesselheim. Adaptive design clinical trials: a review of the literature and clinicaltrials.gov. *BMJ Open*, 8(2): e018320, 2018. ISSN 2044-6055. doi: 10.1136/bmjopen-2017-018320.

[460] J. J. Lee and C. T. Chu. Bayesian clinical trials in action. *Stat Med*, 31(25): 2955–72, 2012. ISSN 0277-6715. doi: 10.1002/sim.5404.