Open cohort designs for cluster-randomised trials in institutional settings

Laura Elizabeth Marsden

University of Leeds

School of Medicine

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

March 2022

The candidate confirms that the work submitted is their own and that appropriate credit has been given where reference has been made to the work of others.

OPIS-CRTs

OPen cohorts in Institutional Settings: designs for CRTs

User engagement work

This thesis was part of a larger MRC-funded project as a bolt-on to the DCM-EPIC trial, named OPIS-CRTs. The project included a user engagement work package which involved my supervisors Dr Rebecca Walwyn and Professor Amanda Farrin, as well as Professor Claire Surr (CS) and Dr Alys Griffiths (AG) of Leeds Beckett University. As part of the user engagement work, two workshops and an online survey were conducted.

The in-person user engagement workshops took place on 1st May 2019 and 16th October 2019. The first workshop included trialists working in care home settings, and the second workshop included experts working in the field of cluster-randomised trials across a range of settings. Both workshops and the online survey included participants working in a range of roles. I was heavily involved in the planning of all of the user engagement work, and also delivered sessions and facilitated discussions in both workshops. The online survey was set up by CS and AG but all of the classification system questions in the survey (see Chapter 4) were written and provided by myself. The first response was received on 13th February 2020 and the last on 24th November 2020.

Dedicated to the memory of my grandparents:

Alice, Bill, Dyliss and Ronnie.

Acknowledgements

First and foremost I would like to thank my supervisor Rebecca Walwyn, whose enthusiasm and dedication to my development has been a constant throughout the whole process. I am very grateful for your insights and our many thought-provoking conversations. I am also greatly indebted to my co-supervisors Andrew Copas and Amanda Farrin for their continued support and guidance over the last three years. Many thanks also to Stuart Barber for stepping in over the last year, and for helping me to see the bigger picture.

I would also like to thank the Medical Research Council for funding the OPIS-CRTs project, and Leeds CTRU for the opportunity to pursue a PhD. A big thank you also to Claire Surr and Alys Griffiths for your collaboration with the user engagement work.

The simulation studies within this thesis would not have been possible without the use of the High Performance Computing facilities at the University of Leeds and their extremely helpful staff - I very much appreciate your time. Thanks also to the Information Specialists in LIHS at the University of Leeds who assisted me with my electronic searches right in the beginning.

I would also like to thank:

My fellow researchers and PhD students at Leeds CTRU. Ultimately, nobody understands the pain and struggles of a PhD quite like those who already have one, or are doing one too. Our Friday catch-ups were invaluable to me.

My friends outside of work, who helped me to keep things in perspective, and for carefully treading around the “P word” in the final few months.

My work colleagues who have become close friends; there were times where you genuinely saved me, and I will be forever glad for our talks.

My family - especially my lovely mam and dad - who, right from the beginning, always instilled in me that I can do anything I put my mind to. That, and “fresh fruit!” and in times of crisis, the importance of “a nice bath”. In all seriousness, without your love and sacrifices I may not have even begun this journey in the first place.

And lastly, to Richard, my counsellor and cheerleader, who has been behind me every step of the way. Thank you for the encouragement, the reassurance, for pulling me away from my computer when I needed it, and for believing in me way more than I believed in myself a lot of the time. I genuinely could not have done this without you.

Abstract

Cluster-randomised trials (CRTs) are often conducted in institutions such as care homes, schools and hospitals. These institutions can be defined as 'open cohorts', as individuals join and leave the clusters over time. I propose that open-cohort designs allow recruitment of participants after cluster-randomisation and have the ability to link participants' repeated measurements. This thesis aims to develop methods surrounding open-cohort designs for parallel-group CRTs.

A scoping review was conducted, which suggested care homes, palliative care and prisons could have greater need for open-cohort designs than other institutional settings. Previous definitions of an open-cohort design were inconsistent, and it was often ambiguous whether the population, design or analysis was 'open'.

Identifying CRT designs in the scoping review was challenging. To overcome this, I developed a CRT classification system which allows trialists to examine particular design components and assess whether a design falls into one of six proposed sub-types. Timescales were highlighted as a crucial design aspect.

A simulation study was used to compare established closed-cohort and cross-sectional designs to open-cohort designs over a range of study parameters and complications inevitable in practice. Analysis models, all variations of a mixed effects model, were fixed. Results were highly dependent on the scenario and estimand of interest.

Alternative analysis models were compared in a second simulation study, including a joint model for longitudinal and survival outcomes, a cluster-weighted model, and heteroscedastic models which partition individual variance over time. Results again varied depending on the estimand and complications, but overall the heteroscedastic models exhibited more improvements in terms of bias, precision and convergence.

CRT designs can be complex with many unique features. I therefore consider throughout how to fully specify CRT designs. The thesis concludes with suggestions on how to improve reporting of CRT designs, including specific recommendations on how established frameworks could be updated.

Total word count: 72,540

Contents

List of Figures

List of Tables

List of boxes

List of abbreviations

ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
CAC	Cluster autocorrelation
CC	Closed cohort
CC-D	Closed cohort estimand with drop-out
CC-I	Immortal closed cohort estimand
CMAI	Cohen-Mansfield Agitation Inventory
CR	Cluster-randomisation
CRSE	Continuous recruitment, short exposure
CRT	Cluster randomised trial
CRXO	Cluster-randomised cross-over
CS-OC-D	Cross-sectional or open cohort estimand with drop-out
CW	Cluster-weighted model
DCM	Dementia Care Mapping
DGM	Data generating model (or mechanism)
EHR	Electronic health record
GEE	Generalized estimating equations
HETW	Heteroscedastic model with cluster-period random effect
HETWO	Heteroscedastic model without cluster-period random effect

IAC	Individual autocorrelation
ICC	Intra-cluster correlation coefficient
JM	Joint model
LTFU	Loss to follow-up
MAR	Missing at random
MCAR	Missing completely at random
MCSE	Monte Carlo standard error
MNAR	Missing not at random
NACR	New admission continuous recruitment
NMR	Neonatal mortality rate
OC	Open cohort
OC-26-I	Immortal open cohort estimand with 26 weeks of individual exposure
OC-52-I	Immortal open cohort estimand with 52 weeks of individual exposure
PMM	Pattern mixture model
R-CS	(Repeated) cross-sections
REML	Restricted maximum likelihood
SE	Standard error
SM	Selection model
SPM	Shared parameter model
SW-CRT	Stepped-wedge CRT
UC	Usual care

Chapter 1

Introduction

This chapter will introduce relevant concepts and theory related to this research project that existed prior to its commencement. Whilst this project is rooted in applied medical statistics and will draw largely from the clinical trials literature, relevant aspects of the related fields of surveys and epidemiology will also be discussed. The motivating example will be described in detail, highlighting the gaps in knowledge that exist as motivation for this research. Finally, this chapter concludes with the thesis aims.

The structure of this chapter will be based on a clear distinction between population, desifgn and analysis, as this will be a key theme throughout the thesis.

1.1 Population

1.1.1 Entry into and exit from clusters

Groups of individuals, such as schools, hospitals, prisons or even whole communities, are referred to as clusters. Groups of individuals treated by the same clinician can also be defined as clusters. In many cluster-randomised trials (CRTs), individuals can move in and out of clusters over time. Entry into and exit from clusters occurs continuously, in the sense that it could happen at any time. Individuals may move in and out of clusters at completely unique times to others, such as residents moving into a new care home or individuals exiting a prison facility dependent on their sentence. Despite the continuous movement of individuals, groups of individuals may also enter or exit clusters at the same time; for example, all first year students in a secondary school could enter the cluster on the first day of term.

Loss to follow-up (LTFU) occurs when an outcome for a randomised participant is not obtained, and is a general clinical trials term covering multiple eventualities. LTFU can occur when participants remain in the cluster, such as withdrawal of consent to data collection in CRTs or administrative errors[1]. LTFU also includes situations where participants no longer remain in the cluster; this could be due to relocation or death. Some trials classify death as a type of LTFU, whereas others treat it separately. Throughout this thesis, LTFU will denote missing outcomes when participants remain in the cluster, as well as missing outcomes when participants are no longer present in the cluster, including death. This was the same approach taken in the DCM-EPIC trial (see Section 1.4). I will use drop-out to describe missing outcomes when participants are no longer present in the cluster.

The concept of participants being unobservable versus observable is discussed later in Section 2.3.2.1. In some trials, individuals who move out of a cluster are still seen as observable and attempts are made to follow them up in their new location; in this work it is assumed that once an individual has left the cluster they are unobservable, and they are treated the same as individuals who have died.

Entire clusters may also be LTFU in a CRT. This could be due to challenges of recruitment, retraction of cluster-level consent, or in extreme cases the loss of individuals may be so high that the entire cluster is also lost [5].

1.1.2 Distinctions between populations and cohorts

A distinction will firstly be made between populations and cohorts following the epidemiological literature. A *population* is defined in the Dictionary of Epidemiology as "all the inhabitants of a given country or area, considered together," and is usually described in one of three ways [6]. A *dynamic* or *open population* gains and loses members over time; natural populations such as cities or countries are examples of these. In contrast to this definition, a *closed* population does not gain members over time and loses members only to death, as opposed to other reasons such as LTFU. Rothman further adds that any study population with LTFU is then an open population, because members can be lost for reasons other than death [7].

[1] The possibility of withdrawal of consent to randomisation is ignored here because the requirement of participant consent to randomisation in RCTs is not always feasible in CRTs; if consent can be gained from all members of a cluster prior to cluster-randomisation (CR), then this would generally be preferable, but in some CRTs consent to cluster-randomisation is made at a higher level on behalf of the whole cluster [5].

Whether a population is open or closed can be a subtle distinction, and depends both on how the membership of the population is defined and the time axis involved [7, 8]. Rothman provides an example of a group of individuals who use a particular treatment [7]. These individuals could be viewed as a closed population if membership begins with treatment initiation, and the time axis is anchored to each individual's "time zero". If time is instead viewed as calendar time, the same population could be defined as open because over time new individuals can join this population. A population can therefore be closed and open, depending on which time axis is of interest.

A population that does not appear to be discussed in the literature is a cross-sectional population. A cross-sectional population could be thought of as a population at a specific time point.

In contrast to a population, membership of a *cohort* is permanent, by Rothman and Vandenbroucke's definitions, either defined by a time representing time zero or an event experienced by all members [7, 8]. One example is a birth cohort consisting of individuals born within a certain time period. A birth cohort will not grow over time, and if somebody is lost to follow-up they are still defined as a member of the cohort. Szklo provides a more general definition of a cohort as "any designated and defined group of individuals who are followed or traced over a given time period" [9].

Whilst Rothman and Vandenbroucke take 'cohort' to imply fixed membership, others do not, resulting in further definitions for clarity. '*Fixed cohort*' can be used to explicitly describe a cohort which cannot gain members over time; the birth cohort described above is an example of this. In contrast, in *open* or *dynamic* cohorts, members are able to join and leave at any time. Rothman adds that this terminology could be seen as contradictory if a cohort is interpreted as an unchanging group of individuals, and suggests using open or dynamic population instead [7]. Finally, a '*closed cohort*' (CC) can be used to describe a cohort where membership ends "only through occurrence of the study outcome or the end of eligibility for membership" [6]. Using previous definitions, if a fixed cohort does not lose any members, the definition of a closed population is satisfied, and so can be called a closed cohort.

Throughout this thesis an underlying open population is assumed. An open population in the context of CRTs could be seen as consisting of all individuals present in the cluster between baseline and final follow-up. When an "open cohort" (OC) design or analysis is referred to, "open" acknowledges an underlying open population and "cohort" refers to the use of longitudinal data, as opposed to a fixed group of individuals.

1.1.3 Institutional populations

An institutional population is commonly defined in surveys and census literature as the population of individuals who do not belong to a household [10]. Examples of such institutions include care homes, prisons, hospitals, religious institutions and more. These institutional populations could be open or closed depending on the circumstances, but the focus in this thesis is on those that have open populations. For example, a prison population could be open if prisoners come and go over time, but if the prison is for those with life sentences only, and if the time origin for each prisoner is the start of their sentence, then it is a closed population.

In many cases the participants will be patients, but as this does not apply in all cases (for example in prisons), individual members of the institutions will be called *participants* throughout the thesis. To differentiate between individual and cluster levels, participant will sometimes be used interchangeably with individual.

1.1.4 Steady state assumption

In a steady state population, people who exit due to death or moving are continually replaced by people with similar characteristics [8]. As a closed population does not allow newcomers, a steady state is only possible in open/dynamic populations [7]. How reasonable the assumption of a steady state is depends on the population and the time period under study. For example, assuming a steady state for a country over a ten year period is much less realistic than for a period of one year. The steady state assumption is used by epidemiologists to calculate incidence rates in dynamic populations. Under this assumption, because the average population size is approximately the same at any time point, any chosen time interval can provide valid estimates of incidence rates [7].

Whilst this project focuses on continuous outcomes, and incidence rates are therefore not of interest, the idea of a steady state could be a useful simplification tool for simulation or analytical purposes. In institutions such as care homes, for example, there can be an upper limit on cluster size, and when someone leaves they will be replaced by someone who is likely to be suffering from similar conditions and be of a similar age. A steady state assumption could be a fair assumption to make in these circumstances.

1.2 Design

The design of a CRT includes several components, and should be dictated primarily by the research question of interest. The type of design, recruitment process and data collection schedule will be the main focus of the design choices. Simple randomisation is assumed throughout and methods such as matching and stratification are beyond the scope of this project and will not be elaborated on any further.

1.2.1 Types of CRT design

Throughout this thesis the focus is on parallel-group randomisation, but this work could equally apply to factorial designs as in both cases, cluster is nested in treatment (described later in Section 1.3.7). This work does not apply to stepped-wedge or cluster cross-over designs because in these cases, treatment and cluster are crossed. However, the latter two designs are still introduced in this section to highlight the use of terminology used in these related literatures to date.

1.2.1.1 Traditional parallel-group designs

Parallel-group CRTs randomise clusters to different arms of a clinical trial as opposed to individuals [11]. A CRT may be chosen over an individually randomised trial for several reasons [12]. Most commonly, CRTs are used to reduce contamination; if control participants and intervention participants exist in the same hospital, for example, the control participants may become aware of the intervention and weaken its effect, introducing bias. CRTs may also be chosen over individually randomised trials when the intervention is delivered at cluster-level, in an attempt to change the culture, environment or general practices of a cluster rather than targeting individuals directly. There are currently two accepted, *named* designs for parallel-group CRTs. Other designs such as the new admission continuous recruitment (NACR) design (see Chapter 4) are also commonly used in practice, but do not appear to be named as a ‘type’ in the literature.

1.2.1.1.1 Closed-cohort designs

A closed-cohort (CC) design, sometimes shortened to just cohort and also known as a longitudinal design, samples a group of individuals at baseline and collects one or more repeated measurements from these individuals at subsequent measurement time points [13].

A closed-cohort design is appropriate for an underlying closed population. To reflect a closed population, the closed-cohort design is therefore optimal when there are low migration rates in and out of the cluster [14, 15]. If a significant number of individuals leave the clusters over the trial period, those that remain in the trial would not necessarily represent the characteristics of those in the original cohort, introducing bias and loss of statistical power. Although some state that short intervention periods are preferable when using a closed-cohort design [14], ultimately the turnover rate is most influential, as a closed-cohort design could still provide unbiased results in a trial with a long intervention period if there is a relatively low turnover of individuals. However, with longer intervention periods, the age of the closed-cohorts will change, possibly along with other related characteristics, potentially confounding changes in outcome. Cluster size is not necessarily an issue for the closed-cohort design; all eligible individuals in the clusters could be measured if feasible, or if the clusters are inherently large, such as in communities, smaller samples could be taken.

1.2.1.1.2 (Repeated) cross-sectional designs

Before introducing a cross-sectional design for parallel-group CRTs, cross-sectional studies, one of the main types of observational, non-experimental studies in epidemiology [7, 16], will be summarised. A cross-sectional study involves measuring a population, or a representative sample from the population, at a specific point in time; they therefore provide only a 'snapshot' of information. Cross-sectional studies often estimate prevalence or the association between two or more variables, but crucially inference can only be made on the prevalence or associations at the specific time of measurement. Cross-sectional studies are prone to several types of bias, but one which is relevant to this project is length-biased sampling [17]. This occurs when individuals with longer durations, say of a disease, are over-represented in the sample, and those with shorter durations are under-represented [7].

Using the simplest definition, a cross-sectional design samples and measures different participants at each time point in a CRT. The same method of sampling at a specific time point as in observational cross-sectional studies applies, but there is now the opportunity to make one or more repeated assessments of the same cluster, accounted for by the (repeated) prefix. The cross-sectional design in the context of CRTs has also been called an independent-sample design [14], a nested cross-sectional design [13], a repeated cross-sectional design [11] or a design with repeated cross-sections [18].

Most definitions of a cross-sectional design in the trials literature appear to be consistent, insofar as they all describe how samples are taken from different individuals at each time point. When sampling within a finite population, however, it is possible that the same person may be measured at multiple time points - these will be referred to as 'overlaps' - and this is a grey area. Some definitions of a cross-sectional design do not explicitly explain the possibility of, or what to do with, such overlapping measurements. The nested cross-sectional design definition [19] implies that there is no overlap and that individuals are measured at most once, because individuals are nested within measurement occasions and so only contribute at a single time point (nesting is described fully in Section 1.3.7). Similarly, the independent-samples definition suggests that each sample is independent of the rest with no overlaps [14]. There is a subtle distinction here between independent sampling, which implies that the sampling process is independent at different time points and so overlaps must be permitted, and independent samples which implies that the samples themselves are independent of each other, with no overlaps. Throughout this thesis a distinction is made between the (repeated) cross-sectional (R-CS) design without overlaps, where samples consist of *completely* different people at each time point, and the R-CS design with overlaps, where overlaps are possible.

In both of these R-CS designs, high turnover rates are not an issue as in the closed-cohort design because samples will still be representative of the cross-sectional populations at fixed measurement points. However, the turnover rate, time between measurements and cluster size are determining factors of whether a R-CS design with or without overlaps should be chosen. Populations with higher turnovers, longer intervals between measurements and larger cluster sizes are more suited to the R-CS design without overlaps, because the possibility of overlaps is minimised. In some situations it may not even be possible to obtain more than one measurement from individuals; in this case the R-CS design with no overlaps is the only available option, even if the underlying population is open. If cluster sizes are small it may not be possible, or desirable, to obtain the required number of measurements from completely unique individuals at each time point for a R-CS design without overlaps, and the version with overlaps is more appropriate. The R-CS design without overlaps overcomes the issue of re-examination bias prevalent in the closed-cohort design [20], whereas the R-CS design with overlaps does not control for this.

The combination of small cluster sizes in the motivating example of DCM-EPIC (see

Section 1.4) with short intervals between measurement points means that using the R-CS design without overlaps in this thesis would force the exclusion of any individuals that are re-sampled, resulting in samples that are unrepresentative of the population at specific time points. The existence of the design without overlaps is therefore acknowledged in Chapter 3 and an example given in Chapter 4, but in the simulation studies of Chapters 5 and 6 the R-CS design will refer to the version with overlaps unless otherwise specified.

A further disadvantage of the cross-sectional design is that individuals with very short periods of exposure to the intervention may be sampled [21]. This is the opposing idea to length-biased sampling, where instead of those with longer durations being over-represented, those with short durations and potentially very little exposure are given as much weighting in the sample as individuals with long durations. Possible strategies to overcome this include the use of length of stay as a covariate in the cross-sectional model, or only including individuals who have had a minimum length of exposure to the intervention at a particular follow-up point [21, 22].

Although overlaps have been discussed here as though their existence would always be transparent to trialists, it is important to note that in many situations it may not be possible to determine whether or not overlaps exist in the data. If data does not include an individual identification variable or is aggregated at a cluster-level it is impossible to link repeated measurements over time that actually originated from the same individual. In these cases, trialists are forced into using a R-CS design, even though the underlying population may be open.

1.2.1.1.3 Comparison of the two established designs

The strengths and weaknesses of closed-cohort and cross-sectional CRT designs have been thoroughly discussed [14, 15, 19–24]. Notably, Feldman and McKinlay [14] compared the two designs via a unifying analysis model which could be applied to either design; this will be discussed in Section 1.4 and later in Section 5.4.

The closed-cohort design has been shown to have superior precision and power to the cross-sectional design in theory [19, 21, 22]. However, Feldman and McKinlay stress that such gains in precision can be minimal unless there is the ability to link individuals' repeated measurements over time, and there is a high correlation between the repeated measurements [14]. McKinlay also concluded that the closed-cohort design is more cost-efficient for shorter trial durations and high correlation between repeated measures [25].

An important difference between the two designs is the type of inference each permits [12, 19, 21, 23, 26]. As the closed-cohort design involves taking repeated measures from individuals, it is best suited to assessing the effect of the intervention on an individual's change in outcome over time. Despite measuring at the individual level, the R-CS design in contrast can only provide inference on cluster-level changes over time. The choice between the two designs can depend on various considerations, but arguably the objectives of the trial and the corresponding requirements for inference should be the most important [12, 19].

1.2.1.1.4 Use of both closed-cohort and cross-sectional designs

Whilst the majority of published trials opt for either a closed-cohort or cross-sectional design, some non-randomised studies have utilised both designs in the same trial. The Minnesota Heart Health Program used cross-sectional data for inference on the community, and closed-cohort data for inference on individuals, for example [27]. In contrast, the Stanford Five-City Project and Pawtucket Heart Health Program collected both closed-cohort and cross-sectional data, but ultimately emphasised analyses based on the latter due to extremely high LTFU of 55% [26, 28]. In another approach, Wagner conducted surveys on a variety of groups (adults, adolescents, grocery stores and so on), and provided both practical and statistical reasons to choose either closed-cohort, cross-sectional or both for each group [29]. Different again was the method by Blair, who performed analysis of the primary endpoint using both cross-sectional and closed-cohort samples over three time points [30]. By combining cross-sectional and closed-cohort samples in the same trial, Diehr *et al.* were able to compare closed-cohort individuals to those who were sampled only once [15].

Although the trials above used both closed-cohort and cross-sectional designs, the designs and subsequent analyses were kept separate, and a single design which unites the two approaches was not seen.

1.2.1.2 Stepped-wedge designs

A relatively modern design for CRTs is the stepped-wedge design (SW-CRT). In a SW-CRT, all clusters are initially allocated to the control condition before crossing over to the intervention at different times, depending on the sequence they are randomised to [31]. Copas *et al.* categorised SW-CRTs as having one of three designs: closed cohort, open cohort or continuous recruitment short exposure [32]. As in the parallel-group case, closed-cohort SW-CRTs recruit individuals at the start of the trial and measure these same individuals over time. An open-cohort SW-CRT design allows individuals to enter and leave the trial throughout the trial period. A continuous recruitment short exposure design involves only a short exposure to the intervention, and recruitment happens on a continuous basis as individuals become eligible.

1.2.1.3 Cluster cross-over designs

A third type of CRT is the cluster randomised cross-over (CRXO) design; this is the cluster randomised equivalent of individually randomised cross-over trials, where experimental units act as their own control. In CRXO trials with two treatments, clusters are often assigned to both control and intervention conditions, and undergo two separate exposure periods, one to each condition with a "wash-out period" in between; in this case there are two 'cluster-periods' [33]. The order in which a cluster is exposed to the different conditions is determined by randomisation.

Previously, CRXO designs have been referred to as cluster cross-over or individual cross-over CRTs, respectively, to distinguish whether cross-over applies to the whole cluster or to individuals [34]. More recently, as with SW-CRT designs, 'cross-sectional' and 'cohort' terminology has also been used [33]. In a cross-sectional CRXO design, different individuals are measured in each cluster-period; conversely, the cohort CRXO measures the same individuals in each cluster-period.

1.2.1.4 Open-cohort designs

Whilst there has been discussion over the choice between closed-cohort and cross-sectional designs, they are usually viewed as the only two possible, mutually exclusive options for parallel-group CRTs, forcing trialists to choose between them. A design which falls between the two is briefly alluded to by Feldman and McKinlay as a design with "randomly overlapping samples" [14]. Koepsell also hints at this design, adding that replacement of those LTFU with new entrants during the trial would ensure that samples remained representative of clusters, but that this approach would come with added complexities [35]. The existing designs are relevant for closed populations and cross-sectional populations, but a design suitable for open populations is only currently emerging, principally within SW-CRTs [36]. This hybrid design has been referred to as an open cohort design. The individuals recruited before CR will be referred to as the *original cohort* and those recruited after the *additional cohort*.

1.2.1.5 Kasza's open cohort sampling schemes

More recently, Kasza presented three broad types of open cohort sampling scheme for use in open cohort CRT designs, focusing on schemes where the number of participants sampled at each time point in each cluster is constant over the trial period, but individuals can contribute different numbers of measurements overall [36].

The "core group" scheme involves following a fixed group of individuals over the trial period and additionally sampling others not in this group at most once at different time points. It is the same concept used in split panel surveys (see Section 1.2.2.3) and is analogous to following a closed cohort over time as well as sampling non-overlapping cross-sections at given time points. Parallels can also be seen with studies that used both closed-cohort and cross-sectional designs in Section 1.2.1.1.4. The "closed population" scheme initially defines a population before taking repeated random samples from the same closed group of individuals, with no new joiners permitted. The third suggested sampling scheme is rotation sampling, which is also known as rotating panels in the surveys literature (see Section 1.2.2.2). For each individual, a maximum number of consecutive measurements is imposed and the proportion of individuals to be replaced at each time point is specified. The main motivating factor of this scheme is the reduction of measurement burden for participants. These schemes will be discussed in more detail in Chapter 3.

1.2.2 Types of design in the surveys literature

1.2.2.1 Panel, longitudinal and repeated surveys

A sampling frame is a list used to select a sample from; it need not list every element in the population, but each element must have some chance of being selected for the sample [16]. For example, for the population of students in English schools, a suitable sampling frame could contain all English schools.

Within the surveys literature, the definition of repeated cross-sectional and longitudinal data is similar to that in observational studies and trials. Repeated cross-sectional data, however, is described by the UK data service as consisting of a "new sample of interviewees at successive time points" [37]. Data from consecutive years obtained using repeated cross-sectional sampling in surveys is assumed independent, and there may be a "small probability" of resampling the same individual, and the samples not being entirely independent. It is perhaps due to the large expected sample size of surveys that this statement is made, as in settings where the population size is small and the turnover rate low, the probability of resampling could actually be large which would violate the assumption of independence.

Longitudinal surveys are further split into either cohort surveys or panel surveys [37, 38]. Whilst both types of longitudinal survey aim to assess change over time, they target different populations. A cohort study in the context of surveys is defined as following "a specific population that is defined by geography and time" [38]. The previously mentioned birth cohort in Section 1.1.2 from epidemiology is an example of this population. Some cohort studies will measure a single cohort only, for example those born in 2000, and as such are better suited to answering research questions about changes over time in specific cohorts of individuals, where questions can be adapted over time to suit the changing age of participants if desired. If a cohort consists of those born in 2000, a cohort effect is a change attributable to those specific individuals only, whereas a period effect affects all cohorts and ages. If a single cohort is used, the effects of age can be investigated, however, a cohort effect cannot be distinguished from a period effect.

In contrast, a panel survey, where the group of individuals is called the "panel", aims to target the whole population. A panel survey is sometimes thought of as an amalgamation of several cohort studies [38]. Because surveys are taken from individuals across all ages, panel surveys are best used in answering research questions on all of society, though will be inferior to cohort studies if interest is in a specific subgroup, for example 30-40 year olds, as generally less data will be collected per age group. As panel surveys cover the entire population, questions have to be more general than in cohort studies as questions must be relevant to all age groups.

The differences between the two longitudinal surveys also dictates whether "refreshment samples", samples of new respondents, are taken over time to replace those lost [38]. Panel surveys take into account the underlying dynamic population mentioned previously, by removing those not present at a particular wave from the sampling frame, making them impossible to sample, and by including any newcomers. Refreshment samples have been used to simply top-up the sample size, replacing those lost so as to maintain a steady state, but they can also be used to diagnose and/or correct for potential attrition bias [39]. In contrast, cohort surveys do not generally recruit new individuals and in theory lose members only to death.

These two types of longitudinal survey can be likened to the idea of closed cohort versus open cohort designs in CRTs. A cohort survey is defined using a particular point in time, as is the closed cohort design whose members are defined by their presence at a specific time point. In contrast, panel surveys with multiple waves of recruitment are analogous to open cohort designs, where refreshment samples of new individuals are taken.

Longitudinal surveys can also provide aggregate estimates of change but it is important to note that, in the case of a cohort study, these aggregate estimates refer only to the population who were initially sampled. If longitudinal surveys are to be used to provide aggregate estimates of change for the current population, then the population frame would need to be updated at each time point to remove any individuals who are no longer present and to include any newcomers [40].

One of the disadvantages of a panel survey is that with frequent data collection, repeated questioning of the same individual can influence their behaviour and, as a result, their responses; in this context it has been called panel conditioning, or time-in-sample bias [10], whereas in trials it has been called re-examination bias [20]. To overcome this, a potential solution is to use a *rotating panel design*.

Panels are usually very large, often consisting of thousands of individuals, which lends itself to taking sub-samples. Analogously, taking sub-samples is useful in CRTs with large cluster sizes, but for small cluster sizes this may not be an option and the whole cluster may be attempted to be sampled so as not to decrease the precision of estimates.

1.2.2.2 Rotating panels

In a rotating panel survey design, or rotation sampling, individuals are limited to the number of measurements they are able to provide before being removed from the sample and replaced by someone else, whilst still remaining in the population. They are often used in labour force studies, where investigators intend to ask frequent questions on the same topic [38]. This author further notes that a rotating panel design can therefore be thought of as a hybrid between the repeated cross-sectional and longitudinal designs. In the context of CRTs, large cluster sizes make it possible to implement a rotating panel design in an attempt to reduce the risk of re-examination bias, but with smaller cluster sizes this may not be an option.

1.2.2.3 Split panels

A split panel survey involves a panel survey with additional independent samples at each time point [40]. This type of design permits both cross-sectional inference (by using the panel plus independent samples at each time point) as well as inference on individual change over time (using just the panel), without the restrictions of a rotating panel survey. Kasza refers to this design within CRTs as a “core group” scheme, again suggesting it as a possible open cohort sampling scheme (see Section 5.4).

1.2.2.4 Cluster sampling

Another concept in the surveys literature that translates to CRTs is the idea of cluster sampling. A cluster in this context can be defined as a sampling unit containing at least one lower-level unit [16]. For simple one-stage cluster sampling, the sampling frame of all clusters in the population is compiled, and a simple random sample of clusters is taken. All of the lower-level members of the selected clusters are taken as the final sample.

If individuals in a city are the population of interest in a survey, it may not be possible to identify the sampling frame of all individuals in the city due to the vast size of this list. A solution to this problem is to use two-stage or multi-stage cluster sampling. The first stage of cluster sampling would be the same as in the one-stage case; this could be postcodes in the city, for example. For the second stage, individuals in the selected clusters could then be listed in the sampling frame, with the reduction in size making this process more feasible. Individuals in each postcode would then be sampled to complete the process.

These methods are analogous to the way clusters, and potentially individuals in clusters, may be randomly sampled from a larger pool of clusters in a CRT. One-stage cluster sampling is likely to be used when cluster sizes are small, and taking a further sample within the clusters would not be possible; these are defined as ‘full-samples’ in Chapter 3. In this case, the individuals belonging to the selected clusters are chosen by proxy rather than by random sampling, as in the two or multi-stage case. In Chapter 3, two-stage sampling is also defined as taking ‘sub-samples’.

1.2.3 Measurement and recruitment

The measurement scheme involves the choice between discrete or continuous measurement, the number of measurement points, and the timing and spacing of these points (if discrete). Similarly, the recruitment process can also occur in a discrete or continuous fashion. Both the measurement and recruitment details have implications for the design and will be discussed further in Section 2.3.1.

Measurement of participants throughout a trial can either be time-structured, where measurement points are fixed in advance at specific times and are identical across participants, or allowed to vary for each participant [41]. The fixed times could be based on time at a cluster-level, for example three months after cluster-randomisation, or on an individual-level, for example three months after recruitment, but in both of these cases the schedule is consistent across individuals. The schedule is also common to both arms. Throughout this thesis the focus is on time-structured designs.

1.3 Analysis

1.3.1 Longitudinal and clustered data

Longitudinal data consists of repeated measurements from multiple individuals over time. Measurements from the same individual are likely to be more similar than two measurements from different individuals, and this correlation is accounted for in a longitudinal analysis. If a dataset includes repeated measurements from the same individuals over time, one option is to conduct separate analyses for each time point. A longitudinal analysis, in contrast, uses the data from all time points in one overall model.

Longitudinal data could be described as clustered, where the clusters are the individuals and measurements at different occasions are members of the cluster. More commonly, clustered data refers to data collected from different individuals in the same cluster in a CRT. In both cases, dependence exists within clusters and must be accounted for in the analysis. Throughout this thesis, the datasets are both longitudinal and clustered in that there are repeated measurements from individuals who are also members of a cluster.

1.3.2 Two analysis methods

Two major approaches exist for the analysis of clustered, normally distributed data; random effects (RE) models and generalised estimating equation (GEE) models. The latter were primarily introduced for longitudinal data, but are equally applicable to clustered data found in CRTs.

RE models are also known as mixed effects models (due to the inclusion of both fixed and random effects) [5], multilevel models [1] and hierarchical models [42]. The latter two names are useful because they convey the structures involved in their many applications; there may be repeated observations within an individual, students in a school, or even three-level or higher models such as workers in cities in countries. Briefly, when RE models are used for the analysis of CRTs, random effects for each cluster are used to account for between-cluster variation. The cluster random effects are assumed to follow the same normal distribution with zero mean. Similarly, if the data is longitudinal rather than from a CRT, random effects would be included for individuals. Further random effects can also be included for other levels of dependence in the data. As a full probability model can be obtained with a RE model, estimates are calculated using maximum likelihood estimation, in particular restricted maximum likelihood (REML) [43]. As each cluster (or individual) has its own random effect, RE models have cluster- (or individual)-specific interpretations.

GEE models were first introduced by Liang and Zeger as a method of analysing longitudinal data when the outcome is not normally distributed (though they can also be used when the outcome is normally distributed) [44]. Unlike RE models, GEE models do not use random effects to explicitly account for dependency in the data. Instead, GEE models make assumptions about the correlation structure, and in estimating the correlation parameters associated with this structure are able to estimate regression coefficients [5]. As a result, GEE model estimates have only a population-average or marginal interpretation. As GEE models are unable to obtain a full probability model, estimation requires iterative generalised least squares methods, and sandwich estimators to estimate robust standard errors.

As explicit modelling of cluster and individual correlation is of interest, RE models will be used throughout this thesis rather than GEE.

1.3.3 Individual versus cluster level analysis

Analysis of longitudinal (or clustered) data must always take into account the similarity between measurements from the same individual (or measurements in the same cluster), and such methods fall into two categories; individual-level regression methods or cluster-level summary methods.

In the 1950s, when statisticians debated whether the correct unit of analysis in CRTs should be the cluster or the individual, the cluster-level summary method, or “means analysis”, was proposed as a solution [13]. Cluster-level summary methods involve aggregating individual-level outcomes from clusters, either using a mean or some other measure, and analysing these summaries, ignoring the individual-level data they originated from. These methods can be thought of as two-stage methods, where the first stage consists of the aggregation process, and the second uses an appropriate analysis method to compare cluster-level summaries across arms [5]. As the analysis occurs at the cluster level, there are no correlations to account for within the summary data, so random effects other than residual error are not required and a simple two-sample t-test or equivalent can be used.

From the design of experiments literature [45], individuals are the observational units as they are the smallest units which are measured, and clusters are the experimental units as they are the smallest unit that treatment is assigned to. From this perspective, analysis of cluster-level summary data makes sense. Moreover, whilst individual-level methods like RE or GEE models can be susceptible to type I error rate inflation with small numbers of clusters, the simpler cluster-level summary methods are more robust [46].

Individual-level methods are, however, more efficient, as more detailed information from clusters is used to weight clusters appropriately in the analysis when cluster sizes are variable. Random effects can also be used to improve the precision of the intervention effect estimate by removing other sources of variation. Cluster-level summaries obscure information on change over time for individuals within clusters which may be of key interest, and are therefore not usually the chosen analysis method for closed-cohort designs. For the R-CS designs, an individual or cluster-level analysis could be used. In this thesis, individual-level regression analysis using individual-level data will be the focus because individual change over time is of interest.

1.3.4 Handling of baseline measurements

A baseline in clinical trials is a measurement taken from participants before the intervention takes place [47]. There are several ways to analyse a randomised trial with baseline and follow-up measurements. Coffman explains that the approaches can be categorised into conditional or unconditional models; although this is specifically for individually randomised trials, the same principles still apply [48]. Conditional models include the analysis of covariance (ANCOVA) method, which involves treating the baseline measurement as a covariate in the model for the outcome; the outcome is therefore conditional on this covariate. Alternatively, unconditional methods do not include the baseline measurement as a covariate in the model, and instead treat the baseline measurement as an outcome in its own right. The latter can be called longitudinal or repeated measures approaches.

1.3.4.1 Longitudinal ANCOVA

Assuming a baseline measurement and one follow-up, using the longitudinal ANCOVA approach each trial participant would have just one response, as the baseline outcome is used as a covariate in the follow-up model.

The most intuitive method for a closed cohort CRT would be to adjust individual outcomes for individual baselines. If a baseline is missing, however, individuals would either have to be dropped from the analysis, losing information in what is called a complete case analysis, or have their baselines imputed; this is a disadvantage of this approach.

As discussed previously, individual level outcomes can be aggregated at the cluster level, and subsequently used to perform a cluster-level analysis. Cluster-level aggregation could also be used within the ANCOVA method to adjust for cluster-averaged baselines. Cluster-averaged baselines are primarily used for R-CS designs, as the population is expected to change at each time point, but could also be used in CC designs, either on their own or in addition to individual level baselines [46].

If the only timescale of interest is that of individual exposure, baselines could easily be defined as the first measurement before the intervention is delivered and could be adjusted for at an individual level. However, in situations where the timescale of cluster-level exposure is as important as the timescale of individual exposure, when some individuals are recruited before and others are recruited after cluster-randomisation, the idea of a ‘true’ baseline to adjust for is more complex. Whilst those recruited post-CR could have a baseline measured on their individual timescale, they would be missing a baseline measurement at $t = 0$ on the cluster-level timescale. As with the R-CS design, adjusting for individual baselines would not make sense for this situation, because the additional cohort would have missing baselines. Cluster-averaged baselines could be an option as they can be used whether individuals are in the original or additional cohort, however, they have not been included in this thesis.

1.3.4.2 Difference of differences

The first unconditional approach includes fixed effects of arm, time and the arm by time interaction, where the interaction term is the estimand of interest to test whether there is a difference between baseline and follow-up between arms; this is called the 'difference of differences' (DOD) approach. For two time points, the first difference is between follow-up and baseline outcomes, and the second finds the difference of the differences between arms.

The DOD approach is the least desirable way to control for baseline outcomes in CRTs, because it does not take into account the fact that before randomisation the arms should have approximately the same outcomes, and does not control for regression to the mean, which can give way to misleading conclusions [46].

1.3.4.3 Constrained baseline

The second unconditional approach is so-called because in contrast to the DOD approach, baseline means are constrained to be equal between the randomised arms [46]. To constrain the baselines, the fixed effect for arm is dropped from the terms included in the DOD approach, leaving just time and the arm by time interaction.

The constrained baseline approach has the advantage of using all available data unlike longitudinal ANCOVA, with no cluster-level aggregation required, whether this is applied to the repeated cross-sectional, closed cohort or as an extension the open cohort design.

1.3.5 Discrete versus continuous time

Usually, when time intervals are discrete, time is treated as a categorical variable in the analysis, but when there are two or more measurement points there is also the option to assume time is a continuous variable; whether or not this assumption is valid then depends on the timing and spacing of the measurements [49].

Due to its cross-sectional nature, use of discrete time in an R-CS analysis seems logical. Depending on the outcome, a CC analysis could treat time as either discrete or continuous. If an open population is to be captured as accurately as possible, continuous time in an open cohort analysis would be optimal, however, a high number of discrete time points may offer similar benefits without the burden of continuous measurement in practice.

1.3.6 Outcome type

Attention will be restricted throughout to continuous outcomes only, due to the motivating example (see Section 1.4).

1.3.7 Hierarchical and non-hierarchical structures

In a hierarchical structure there is pure nesting of factors within higher level factors. In a simple example, students are nested within schools if each student belongs to only one school (Figure 1.1). Non-hierarchical structures occur when pure nesting is not present, and can be categorised as having either cross-classified or multiple membership relationships. These will be outlined here as both cross-classified and multiple membership relationships will be referred to throughout the thesis.

Classification diagrams and network diagrams are provided for illustration. In a classification diagram, the arrow moves from a lower level unit to a higher level unit. A single arrow from a lower level unit implies nesting; multiple arrows to different higher level units imply cross-classification, and multiple arrows to the same higher level unit implies multiple-membership.

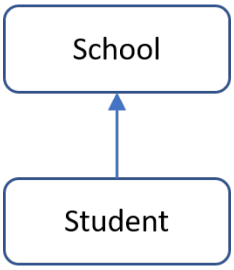


Figure 1.1: Classification diagram for students nested within schools (adapted from [1]).

1.3.7.1 Nested models

Extending the example of students in schools, students and both the primary school and secondary school they belong(ed) to could be of interest. Taking 5 primary schools as P1 to P5, and two secondary schools, S1 and S2, a purely nested relationship of students within primaries within secondaries would assume that all students who attended say P1, P2 and P3 also attended S1. The remaining students attending P4 and P5 also attended S2 (Figure 1.2). This hierarchical relationship occurs because each primary school appears in only a single secondary school, and each student in a single primary school.

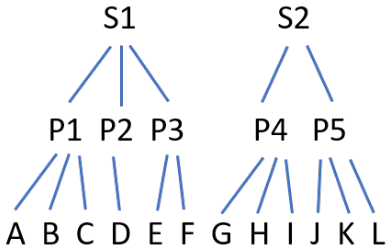


Figure 1.2: Network diagram for students nested within primary schools and primary schools nested within secondary schools (adapted from [2]). S = secondary schools; P = primary schools; letters denote students.

1.3.7.2 Cross-classified models

The models proposed by Hooper and Kasza in Section 1.3.8.4.2 and Section 1.3.8.4.3 are examples of cross-classified models, so they are introduced here.

The nested models in the previous section are very simplistic and in reality more complex relationships exist. If primary schools occur in combination with multiple secondary schools, the relationship is cross-classified. The student is said to be nested in the cross-classification between primary schools and secondary schools, meaning they can be members of any combination of primary and secondary school (Figures 1.3 and 1.4) [1].

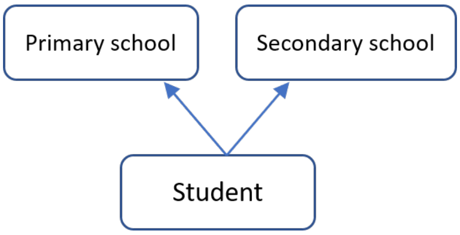


Figure 1.3: Classification diagram for students nested within the cross-classification of primary and secondary schools (adapted from [1]).

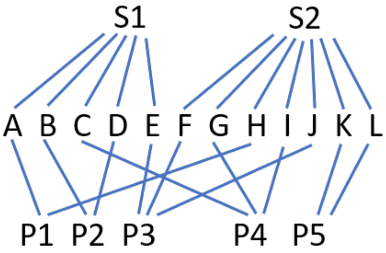


Figure 1.4: Network diagram for students nested within the cross-classification of primary and secondary schools (adapted from [2]). S = secondary schools; P = primary schools; letters denote students.

1.3.7.3 Multiple membership models

The concept of multiple membership is of interest in this thesis because it could potentially be used as an analysis method for settings with open populations (see Chapter 6). Inspiration originated from similar analogies by Goldstein and Walwyn [1, 50], and the basic premise is introduced here.

Extending previous ideas once more, it may be too simplistic to assume that students attend only a single primary or secondary school. Multiple membership relationships allow lower level units to be a member of multiple higher level units of the same type (Figure 1.5). For example, a student may have attended S1 for 4 years and S2 for 1 year. Multiple membership and cross-classified relationships can also occur in combination with each other.

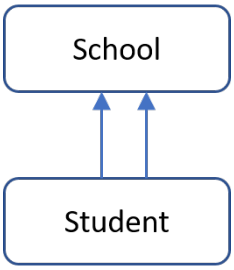


Figure 1.5: Classification diagram for the multiple membership relationship between students and schools (adapted from [1]).

1.3.7.4 Relationships within the designs

There are four factors to consider in the motivating example of this thesis; treatment arm, clusters, individuals and time [19]. Clusters are nested in treatment arms and individuals are nested in clusters and therefore treatment arms. Time is crossed with both cluster and treatment arm because “the same measurement schedule applies in both treatment [arms] and all clusters" [19]. The last pair of factors, individuals and times, differs between the cross-sectional (without overlaps) and closed cohort designs: in the former, individuals are nested within times because they supply just one measurement, but in the latter individuals are fully crossed with time because each individual is expected to be measured at each time point. For an open cohort design, the crossed relationship rather than nested is used because some participants will have repeated measurements, but to reflect the changing open population over time full crossing is not expected, so it is said to be partially crossed.

1.3.8 Traditional analysis of the three designs

1.3.8.1 Notation

The notation to be used throughout the thesis is defined here. Subscripts will start with the lowest level in the model where possible and move upwards. Random effects are differentiated from fixed effects using boldface as in [49].

1.3.8.2 Cross-sectional analysis

A traditional mixed model analysis of variance (ANOVA) is a common approach for a cross-sectional analysis of CRTs [49]. A continuous outcome Y_{tijk} is given by

$$\begin{aligned} &Y_{tijk} = \mu + G_k + T_t + GT_{tk} + \mathbf{C}_{jk} + \mathbf{CT}_{tjk} + \mathbf{E}_{tijk} \\ &\mathbf{C}_{jk} \sim N(0, \sigma_C^2), \quad \mathbf{CT}_{tjk} \sim N(0, \sigma_{CT}^2), \quad \mathbf{E}_{tijk} \sim N(0, \sigma_E^2), \end{aligned} \tag{1.1}$$

for individual i nested in time t, cluster j and arm k. This can be applied to any number of time intervals $t = 1, \ldots, T$. Fixed effects include grand mean μ, the treatment arm effect G_k, effect of time T_t and the arm by time interaction GT_{tk}. $\mathbf{C}_{jk}$ represents the random effect for cluster, $\mathbf{CT}_{tjk}$ the interaction between time and cluster which is also a random effect, and $\mathbf{E}_{tijk}$ is the residual error.

The total variance of the outcome is given by $\sigma^2 = \sigma_C^2 + \sigma_{CT}^2 + \sigma_E^2$, and the intra-cluster correlation coefficient (ICC), ρ, is the correlation between measurements of two individuals in the same cluster at the same time [51]. The ICC is assumed the same over all time points. It can also be described as the proportion of total variance of the outcome that is attributable to clustering [49]:

$$\rho = \frac{\sigma_C^2 + \sigma_{CT}^2}{\sigma^2}.$$

The cluster autocorrelation (CAC), π, is the correlation between population means in the same cluster measured at two different times [52]:

$$\pi = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_{CT}^2}.$$

This model allows a random intercept for cluster by inclusion of $\mathbf{C}_{jk}$, and a random coefficient for each cluster-time combination given by $\mathbf{CT}_{tjk}$. This means there is correlation between members of the same cluster and correlation between members of the same cluster-time combination. Note that there are no random effects relating to individual here because individuals are assumed to contribute just one measurement. It is also possible to adjust for further covariates to reduce confounding but these are omitted in this model and models throughout this thesis.

1.3.8.3 Closed cohort analysis

The most common approach for closed cohort analysis of CRTs is similar to the cross-sectional equivalent, but with additional random effects for individual to remove variation due to repeated measures from the same individual [49]. For continuous outcome Y_{tijk},

$$Y_{tijk} = \mu + G_k + T_t + GT_{tk} + \mathbf{C}_{jk} + \mathbf{CT}_{tjk} + \mathbf{I}_{ijk} + \mathbf{IT}_{tijk} + \mathbf{E}_{tijk} \tag{1.2}$$

$$\mathbf{C}_{jk} \sim N(0, \sigma_C^2), \quad \mathbf{CT}_{tjk} \sim N(0, \sigma_{CT}^2), \quad \mathbf{I}_{ijk} \sim N(0, \sigma_I^2), \mathbf{IT}_{tijk} \sim N(0, \sigma_{IT}^2), \tag{1.3}$$

$$\mathbf{E}_{tijk} \sim N(0, \sigma_E^2),$$

where individual i is no longer nested in time t but crossed with time, whilst still being nested in cluster j and arm k. Fixed and random effects of model (1.1) are also included here, and the additional $\mathbf{I}_{ijk}$ and $\mathbf{IT}_{tijk}$ terms represent random effects for individual and the interaction between time and individual, respectively. If there are multiple measurements per individual at the same time point, the $\mathbf{IT}_{tijk}$ term can be estimated separately from the residual error and provides information on correlation between the multiple measurements; if, as was the case in DCM-EPIC, individuals are measured just once at each time point, this term can be dropped from the model.

In addition to the ICC and CAC, the individual autocorrelation (IAC), τ, can also now be defined as the correlation between two measurements from the same individual taken at different times in a given cluster, where $\sigma_{IT}^2 = 0$ if the individual by time interaction is dropped:

$$\tau = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_{IT}^2 + \sigma_E^2}.$$

1.3.8.4 Specific models

1.3.8.4.1 Feldman and McKinlay's unifying model

Feldman and McKinlay call their proposed model a 'unifying model' as it can be applied to both R-CS and CC designs, as well as OC designs which are only alluded to [14]. There is an assumption of a common estimand across the designs, which is the DOD as described previously. The arm by time interaction is used to assess this treatment effect.

Let Y_{mtijk} be the continuous outcome for replicate measurement $m = 1, \ldots, M$, at time $t = 1, \ldots, T$, for individual $i = 1, \ldots, I$, nested in cluster $j = 1, \ldots, J$, nested in treatment arm $k = 1, \ldots, K$. The response is then given by

$$\begin{aligned} &Y_{mtijk} = \mu + G_k + T_t + GT_{tk} + \mathbf{C}_{jk} + \mathbf{CT}_{tjk} + \mathbf{I}_{ijk} + \mathbf{IT}_{tijk} + \mathbf{E}_{mtijk} \qquad (1.4)\\ &\mathbf{C}_{jk} \sim N(0, \sigma_C^2), \quad \mathbf{CT}_{tjk} \sim N(0, \sigma_{CT}^2), \quad \mathbf{I}_{ijk} \sim N(0, \sigma_I^2), \quad \mathbf{IT}_{tijk} \sim N(0, \sigma_{IT}^2),\\ &\mathbf{E}_{mtijk} \sim N(0, \sigma_E^2). \end{aligned}$$

Fixed effects consist of the grand mean μ, treatment arm G_k, time T_t and the treatment arm by time interaction GT_{tk}. Random effects include the random effect for cluster $\mathbf{C}_{jk}$, the cluster-time interaction $\mathbf{CT}_{tjk}$, the random effect for individual $\mathbf{I}_{ijk}$, the individual-time interaction $\mathbf{IT}_{tijk}$ and the residual measurement error $\mathbf{E}_{mtijk}$. This model assumes that all random effects are Gaussian, homoscedastic and independent. As a result,

$$\sigma^2 = \sigma_C^2 + \sigma_{CT}^2 + \sigma_I^2 + \sigma_{IT}^2 + \sigma_E^2,$$
$$\rho = \frac{\sigma_C^2 + \sigma_{CT}^2}{\sigma^2}, \quad \pi = \frac{\sigma_C^2}{\sigma_C^2 + \sigma_{CT}^2}, \quad \tau = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_{IT}^2 + \sigma_E^2},$$

where σ^2 is the total variance of the outcome, ρ is the ICC, π is the CAC and τ is the IAC. This ANOVA model assumes mutual independence amongst random effects and so no covariance terms are present.

In this model G_k is used to denote arm, the single subscript implying that allocations are fixed over the entire duration of the trial. As a result, it appears to be applicable to parallel-group CRTs only, because SW-CRTs and CRXO trials involve clusters switching between treatment conditions, though this could be easily amended.

The ability of this model to be applied to all three designs is due to the CAC and IAC terms; in models that are specific to the R-CS or CC designs, these parameters will be set to 0 or 1. The CAC is defined using σ_C^2 and σ_{CT}^2, the time-invariant and time-varying components of the cluster-level variance, respectively. The same partitions occur at the individual level with σ_I^2 and σ_{IT}^2. As a result of including these additive random effects for cluster and individual, constant correlation (compound symmetry) among repeated measures on a given unit (cluster or individual) is also assumed. If $\tau = 0$ there is no individual autocorrelation and so this describes the situation where measurements over time are independent and each individual is measured just once (previously described as the R-CS design without overlaps) [52]. If $0 < \tau \leq 1$, there is correlation between measurements over time. If τ is “quite high”, but not necessarily equal to 1, measurements over time are likely to originate from the same individual and the design is CC [14]. In an OC design τ would be expected to be positive but smaller.

1.3.8.4.2 Hooper’s parameterisation of the closed cohort model

Hooper provides separate models for the CC and R-CS designs [51], but no reference is made to the OC design. The CC model has an alternative parameterisation as such:

$$\begin{aligned} &Y_{tijk} = \mu + T_t + A_{tk}\delta + \mathbf{C}_{jk} + \mathbf{CT}_{tjk} + \mathbf{I}_{ijk} + \mathbf{E}_{tijk} \qquad (1.5) \\ &\mathbf{C}_{jk} \sim N(0, \sigma_C^2), \quad \mathbf{CT}_{tjk} \sim N(0, \sigma_{CT}^2), \quad \mathbf{I}_{ijk} \sim N(0, \sigma_I^2), \quad \mathbf{E}_{tijk} \sim N(0, \sigma_E^2), \end{aligned}$$

where Y_{tijk} is the continuous outcome for individual i, at time t, nested in cluster j, nested in arm k. Fixed effects include the grand mean μ, time T_t, and a treatment indicator for arm at time t, A_{tk}, with δ as the treatment effect.

Model (1.5) uses a constrained baseline approach due to its inclusion of time and arm by time interaction terms, but omitting the fixed effect for arm. Another difference between models (1.5) and (1.4) in terms of fixed effects is due to the difference between the terms

$$GT_{tk} = \begin{cases} t & \text{if arm } k \text{ receives the intervention treatment at time } t \\ 0 & \text{if arm } k \text{ receives the control treatment at time } t \end{cases}$$

$$A_{tk} = \begin{cases} 1 & \text{if arm } k \text{ receives the intervention treatment at time } t \\ 0 & \text{if arm } k \text{ receives the control treatment at time } t. \end{cases}$$

As a result, model (1.4) allows a linear change in treatment effect over time, whereas (1.5) assumes a constant treatment effect.

In Feldman and McKinlay's model (1.4), the only relationships described are cluster as nested in treatment arm, and individual as nested in cluster. Given that their model can be applied to both the R-CS and CC designs, any other nesting, such as individuals being nested in times in the R-CS design, is unclear. However, Hooper provides extra information on this by explaining that the closed cohort model (1.5) is not (purely) hierarchical because individuals and times do not have a nested relationship, therefore the model is cross-classified.

1.3.8.4.3 Kasza's open cohort model

In recent work by Kasza, a similar model to (1.5) is proposed specifically for open cohort CRTs [36]. The model for continuous outcome Y_{tij} is:

$$\begin{aligned} & Y_{tij} = T_t + \theta X_{tj} + \mathbf{C}_j + \mathbf{CT}_{tj} + \mathbf{I}_{ij} + \mathbf{E}_{tij} \\ & \mathbf{C}_j \sim N(0, \sigma_C^2), \quad \mathbf{CT}_{tj} \sim N(0, \sigma_{CT}^2), \quad \mathbf{I}_{ij} \sim N(0, \sigma_I^2), \quad \mathbf{E}_{tij} \sim N(0, \sigma_E^2) \end{aligned} \tag{1.6}$$

for participant i at time t in cluster j, where arm k has been dropped in contrast to the Feldman and McKinlay and Hooper models. Again, fixed effects include T_t for time and X_{tj} which is assumed to be a treatment indicator for cluster j at time t with corresponding treatment effect θ.

This model is similar to the model (1.5) in that it allows baseline means to be constrained, but the arm by time interaction A_{tk} in (1.5) is replaced by a cluster by time interaction X_{tj}; this allows the crossing from control to treatment to occur in all clusters, and as such there are no longer clearly defined treatment arms but instead clusters are defined by the time they switch from the control to intervention condition. This can be seen in model (1) and (2) in Hussey and Hughes [53]. Kasza's model is therefore applicable to all types of CRTs, including SW-CRTs and CRXO designs, and the parallel-group CRT with constrained baselines is a special case of this general model. Though it is not described as such, this is again a cross-classified model. Although there are small differences between the Hooper and Kasza models, they are very similar in essence, and the Hooper model (1.5) would equally apply to SW-CRTs and CRXO designs if arm were replaced with sequence.

1.3.8.5 Open cohort analysis

In epidemiology, underlying open populations are acknowledged in analyses by calculating incidence rates for binary or time-to-event outcomes using person-time. As the population is dynamically changing, it cannot be assumed that all individuals have the same time at risk as in a closed cohort design. Average risks can't be calculated for an open population, because to calculate average risks everyone must have been at risk for the whole of the follow-up period [7]. Use of person-time involves dividing the number of events by the sum of each individuals' unique follow-up time. However, this calculation assumes that the event rate over this period of time is constant, so the implications of using person-time in any context should be considered and used with caution.

In CRTs, the definition of an open cohort analysis is unclear. The only two known analysis models applicable to open cohort designs with continuous outcomes are that of Feldman and McKinlay and Kasza [14, 36]. In both of these models, repeated measurements from individuals are linked using random effects. However, unlike the person-time approach, the only timescale used is the timescale from cluster-randomisation; neither of these models explicitly account for individuals who enter the cluster after cluster-randomisation and have differing individual exposure times to those in the original cohort. Discussing a repeated cross-sectional design, Hayes and Moulton [5] add that:

> "it is also possible, if desired, to include in the sample subjects who have entered the cluster during the study; for example, in-migrants to the study population. When interpreting the results for such individuals, however, it may be necessary to take into account their degree of exposure to the intervention, particularly for interventions of long duration."

If in-migrants are also allowed in an open cohort design, the same accounting of individual exposure should be made, possibly along with the linking of repeated measurements in the Feldman and McKinlay and Kasza models. This quote also highlights the lack of consideration most R-CS analyses have for individuals' length of exposure, if individuals are permitted to be recruited after cluster-randomisation.

Finally, to reflect the changing open population, an open cohort analysis could instead only viewing missingness as missing values within an individual's length of stay in the cluster; that is, missing values when they were present. If there are missing values which correspond to times where an individual is not present in the cluster, these values could be treated as unobservable rather than missing.

1.4 Motivating example

1.4.1 The DCM-EPIC trial

The DCM-EPIC trial aimed to evaluate the clinical- and cost-effectiveness of Dementia Care Mapping (DCM), a quality improvement tool used by care home staff to improve care outcomes for people with dementia [54]. A parallel-group CRT design was used, with care homes randomised to DCM and usual care (UC) or UC only. Follow-ups were conducted at 6 and 16 months following cluster-randomisation. The primary continuous outcome was agitation assessed at the resident-level at 16 months, using the total Cohen-Mansfield Agitation Inventory (CMAI) score.

Recruitment of 750 residents from 50 care homes, assigned to intervention and control in the ratio 3:2 respectively, was required to provide 90% power to detect a standardised effect size of 0.4 for the between arm difference in mean scores at final follow-up with a two-sided 5% significance level. This calculation assumed an ICC of 0.1, which was based on the ICC for CMAI in a similar nursing home trial. An assumption of 25% LTFU at 16 months was also made. A sensitivity analysis confirmed a maximum tolerable LTFU of 35% which would reduce the power to no less than 85%. Each care home was expected to recruit an average of 15 participants, based on the number of beds available in a care home setting.

The original DCM-EPIC trial had a closed-cohort design and the primary endpoint was to be analysed at 16 months. The 726 residents initially recruited between October 2014 and December 2015 were the closed-cohort sample to be followed over 16 months. In November 2015, monitoring indicated that 32-48% of residents would be lost to follow-up by 16 months. This meant that the trial would have been underpowered if a design change had not been implemented. It would have also caused issues with external validity as the sample of residents may have no longer been representative of the general care home population due to attrition bias. Relating this back to Section 1.3, the closed cohort estimand and estimates would have been questionable as the underlying population was clearly not closed.

Following this, a design change was approved between baseline and final follow-up to an “open-cohort” design. An additional 261 residents were recruited at 16 months to increase the size of the cross-sectional sample at 16 months, and a cross-sectional analysis was the primary analysis of the primary endpoint. Although repeated measurements were available for some residents - that is a baseline, 6 month and 16 month measurement - these were not linked at an individual level in the primary analysis. Cluster-averaged baselines only were included as covariates in the model.

Figure 1.6 shows four possible scenarios for residents in the care homes during the trial. Residents A and B were part of the original cohort, whereas residents C and D joined a care home post-randomisation. Only residents A and C remained in their care home at final follow-up; residents B and D either moved, withdrew or died before this point. Resident C was present at 16 months and so was available for recruitment into the cross-sectional sample. Although resident D was present for a significant period of time and therefore received the intervention, they were not recruited due to the timing and spacing of recruitment and measurement points, so their data was not collected.

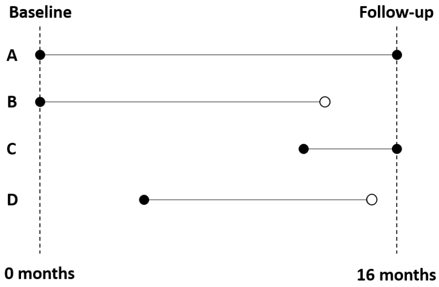


Figure 1.6: Illustration of the four different scenarios for residents in DCM-EPIC care homes. Black circles denote a resident's presence; white circles denote the time where a resident moved, withdrew from the trial or died. Resident D was not recruited into the trial.

If the trial were analysed as originally intended, the closed-cohort analysis would have used data from residents A and B only. Resident A has complete data whereas resident B has one or more observations missing. With approximately 43% of residents in the closed-cohort having a scenario like resident B, this created a huge missing data problem. The cross-sectional sample at 16 months allowed new entrants to the trial to be measured (resident C) but this design was unable to capture data from residents who entered the trial after randomisation but left before the 16 month time point (resident D). Thus, number of and timing of measurement and recruitment points is important.

The DCM-EPIC trial exemplifies how neither of the original CC or new cross-sectional designs were entirely suited to the trial's objectives. Neither approach made full use of the data collected from residents, which has implications for statistical power, cost and interpretation. An ideal design would collect both closed-cohort (A) and cross-sectional (C) data, as well as data from closed-cohort participants lost to follow-up (B) and from those present in between baseline and final follow-up (D). Although the final DCM-EPIC design was called “open-cohort” due to the allowance for recruitment after cluster-randomisation and the ability to link repeated measurements, it was not optimal in terms of measurement and recruitment time points in ensuring that all residents were captured.

The DCM-EPIC trial, and more generally care home trials, have several unique complexities that make the need for a new design greater. Care home trials have an unavoidably high turnover of residents due to the age and health of the care home population. They also have inherently small cluster sizes as there is often an upper limit on the number of beds available. For a closed-cohort design, this means that more clusters would need to be recruited to overcome expected attrition, increasing costs. Finally, the most common reasons for LTFU, moving out of the care home and death, require special consideration and cannot be handled using standard methods for missing data (see Section 2.3.2.1).

1.5 Aims and objectives of the thesis

The overarching aim of this thesis is to develop the design and analysis of novel open cohort parallel-group CRTs when the underlying population is open. The motivational example of DCM-EPIC was carried out in care homes, so this setting will be the primary focus, but as other institutional settings could also benefit from this research, extensions to other settings and how implementation may differ in these settings will be made throughout.

In Chapter 2, a scoping review will be conducted to investigate the use of open cohort designs in epidemiology and CRTs. Further details on the institutional settings will be provided here which will inform the remainder of the thesis. Complications which are inherent in institutional settings will also be explained. Unforeseen difficulties encountered in the searches relating to poor reporting will also be discussed, and how these obstacles led to the development of a classification system to improve reporting and design identification in CRTs, presented in Chapter 4.

Following Chapter 2 and before delving into more detail in Chapters 4-6, it will be necessary to set out clearly what I mean by a design in this thesis, and provide more detail on sampling schemes which have been introduced in this chapter.

Comparison of designs will be the focus of Chapter 5, which details methods and results from a large simulation study. Data will be generated in a base case and then with different combinations of the complications discussed in Chapter 2 to reflect other settings. The resulting CC, R-CS and OC designs will be compared in terms of bias and precision using variations of mixed effects analysis models. Consideration of a range of estimands of interest will also be given here.

In Chapter 6, alternative analysis models will be explored to determine whether improved convergence, bias and precision can be obtained compared to the mixed effects models of Chapter 5. This will also be carried out using a simulation study, using the complications scenarios generated in Chapter 5.

The thesis will then conclude with a final chapter on general conclusions in which I aim to highlight the key themes of this work, implications, general limitations and directions for future research.

Chapter 2

Use of open cohort designs in CRTs and epidemiology: a methodological scoping review

2.1 Introduction

In Chapter 1, the concept of a novel open (or dynamic) cohort design was introduced, in the context of parallel-group CRTs. This design was motivated by the DCM-EPIC trial, set in care homes. The specific complications of DCM-EPIC suggest that an open cohort design could be a useful third design option for parallel-group CRTs in care homes; one aim of this chapter is to determine whether this is true, and whether this new design would be equally beneficial across other institutional settings. Six other institutional settings will be investigated in addition to care homes: prisons, palliative care, schools, primary care, hospitals and communities. To do this, in the first part of this chapter electronic searches will be carried out to obtain samples of CRTs in care homes and other institutional settings. The range of designs used will be assessed, as well as characteristics of the trials that could indicate issues that an open cohort design could potentially solve; this includes LTFU rates, number and size of clusters and length of follow-up.

Preliminary searches indicated a lack of explicit use of an open (or dynamic) cohort design within parallel-group CRTs, but explicit usage is known to exist within epidemiology. A scoping review is therefore also presented of the explicit usage of open (or dynamic) cohort designs within CRTs and epidemiology.

This chapter begins with detailed aims and is followed by a description of some key design components of DCM-EPIC that motivated the data extracted in the following reviews. Methods for the searches are then described in detail, followed by two separate results and discussion sections for the different aims, and a final section with recommendations on how reporting could be improved in CRTs.

2.2 Aims

The key aims of this review were to:

1. a) Update a recent systematic review of CRTs in care homes [55] by extending it from 2011 onwards. Assess whether open cohort designs have been used in this context, explicitly or implicitly, or how they *could* be used if no usage is found.

   b) Repeat the above in other institutional settings.

2. Investigate the *explicit* usage of open/dynamic cohort designs within CRTs and epidemiological studies, summarising their characteristics and selecting a subset for review.

The main distinction between aims 1 and 2 is whether a trial or study uses an open/-dynamic design explicitly or not, by naming the design. For explicit usage in aim 2, the open/dynamic keywords are included in the search, whereas in aim 1 they are not.

2.3 DCM-EPIC components considered in the searches

2.3.1 Recruitment and measurement

2.3.1.1 Continuous versus discrete recruitment

Recruitment of individuals in CRTs can occur discretely at fixed time points or in a continuous process. Choice of recruitment could be due to the setting, the condition under investigation, type of outcome, participants' eligibility criteria, who is recruiting the participants and potentially other factors. Trials investigating acute events may be more suited to continuous recruitment, whereas those studying chronic conditions would be able to take advantage of discrete recruitment [56]. For example, a hospital trial recruiting participants who have recently suffered a stroke would be best suited to continuous recruitment, because at the start of the trial eligible participants are not identifiable due to not yet having suffered the event of interest. In contrast, participants suffering from chronic conditions like diabetes could be recruited at one discrete time point, because recruiters would be able to use records or routine data to compile a list of eligible participants in advance.

In a closed cohort design, participants are recruited at a single discrete time point before randomisation. In a (repeated) cross-sectional design, recruitment is still discrete but can occur at multiple time points. The open cohort design could have either discrete or continuous recruitment.

The original DCM-EPIC closed cohort design included a single recruitment point before cluster-randomisation. After the change in design, an additional discrete recruitment point at 16 months was introduced in order to recruit participants present at this time point.

2.3.1.2 Continuous versus discrete measurement of participant data

Discrete measurement refers to the collection of data at fixed time points throughout a trial. Trialists may wish to discretely collect data in situations where the measurement process is time or labour intensive, such as in-depth interviewing or biological testing, or in an effort to minimise costs. If collecting data discretely with a limited number of time points, however, participants with a limited time in the cluster could potentially be missed. DCM-EPIC measured participants at 3 discrete time points planned in advance; baseline, 6 months and 16 months.

Continuous measurement usually occurs ‘during’ a trial period rather than at exact fixed points, or is done in such a way that no participant is omitted. Continuous measurement is sometimes known as continuous ‘surveillance’, and Murray adds that data is collected “as endpoints occur” [49]. Thus, some endpoints lend themselves better to continuous surveillance, such as time-to-event or binary endpoints. Routinely collected data or electronic health records (EHR) can be used for such purposes, which also reduces the measurement burden for staff and in some cases avoids the need for participant consent. Several discrete measurements could be used to mimic continuous measurement, but this could increase costs and resources needed dramatically. Whether several discrete measurements could be seen as equivalent to continuous would depend on whether the frequency is often enough for a particular endpoint; for example, weekly questionnaires could be sufficient for a continuous outcome such as depression, whereas measurement may have to occur every day or even hour for others.

2.3.1.3 Timing of measurement points

When a trial collects data discretely, measurements should be timed so that the expected intervention effects are captured, if this knowledge is available a priori [13]. The first consideration is how quickly the intervention can be expected to take effect. If the intervention effect is immediate, the whole trial period need not be extremely long, whereas more complicated interventions that take a longer time to embed may need extended trial periods. The second consideration is the expected shape of the intervention effect. For a constant or linear intervention effect over time, timing of measurements is not as important, so measurements that are equidistant would be a sensible approach, for example baseline, 3 and 6 months. If an intervention effect is non-constant, however, and more change is expected either at the start or the end of a trial period, then measurement timing should be chosen accordingly.

2.3.1.4 Number of measurement points (if discrete measurement)

Firstly, designs may have a single measurement occasion, usually after an intervention; this is called the “posttest only control group” design by Murray [49]. If one wishes to investigate individual change over time, designs with a single measurement are insufficient, because only a single snapshot of information is known. Lack of baseline measurements also means that there is no way of knowing whether arms were balanced before the intervention, opening up the possibility of selection bias [49].

Two measurements offer an improvement over a single measurement, and indeed can be effective if intervention effects can be described in a single ‘step’. However, if individual change over time is the goal, two measurements still do not provide information on the shape of change, and true change can be confounded with measurement error [41]. With two measurements, these will usually take place before randomisation and at the trial end, and can be called a ‘pretest-posttest’ or ‘before-after’ design.

There may also be more than two discrete measurement points. Singer and Willett add that the more measurement points in a design, the more flexible your analysis can be without having to make strong assumptions [41]. However, more measurement points will usually lead to more complicated trial logistics and increased costs, and there are more possibilities for the covariance matrix of time points [49].

2.3.2 Complications

2.3.2.1 Missing data

2.3.2.1.1 Missing data mechanisms

A complete dataset Y can be partitioned into observed and missing components [57, 58]. If the probability of missingness is independent of both the observed and missing components of Y, the data are said to be missing completely at random (MCAR). Simply ignoring the missing data under a MCAR mechanism can still provide unbiased estimates as the observed data are a random subsample of the full data. Data are missing at random (MAR) if the probability of missingness is dependent on observed data, but not on the unobserved data. In this case unbiased estimates can be obtained by using information contained within the observed data [59]. However, if the probability of being missing depends on values that cannot be observed, use of the observed information still does not provide enough information to obtain unbiased estimates. In this case, the data are said to be missing not at random (MNAR).

2.3.2.1.2 Data truncated due to death or moving out of a cluster

When data is missing due to death, careful thought is required about how the analysis method deals with this data specifically and the subsequent implications for inference. Considering both the longitudinal and survival responses from participants, all statistical models can be divided into either unconditional, partly conditional or fully conditional, where the conditioning is with respect to the survival outcome. Unconditional models describe the longitudinal outcome without any consideration of survival outcomes, and as such are suitable if either deaths do not occur, if death is independent of the longitudinal response or if the longitudinal response can be given a value after death that makes sense [60]. In unconditional models, such as mixed effects models, missing values due to death can be implicitly imputed when this is not desirable; this is a key result that is potentially not widely known by researchers [59]. This happens as a result of the covariance structure that is applied to the matrix of individuals' responses, meaning that missing values due to death are still 'filled in' according to this structure.

The idea of ‘mortal’ and ‘immortal’ cohorts is also relevant here [60]. The unconditional model provides inference for an immortal cohort, because those that die are implicitly included after death, and subsequently this inference targets individuals present at the start of the trial and assumes that they do not die. Mortal cohort inference can be obtained using the partly conditional or fully conditional models, and targets only the individuals that are alive at each time point throughout the trial.

In the same way that Seaman *et al.* partition missing data into that which is “unobservable” and “potentially observable”, data missing due to death could be thought of as ‘not applicable’ rather than ‘missing’ [61]. If residents have died then their data is not missing in the same way as somebody that has withdrawn consent to further data collection. As the open-cohort design focuses on the change in individual outcome as a result of interventions delivered at cluster-level, resident data when they are not present in the cluster is not of interest. In other words, our inference only concerns the change in individual outcome whilst a resident is present in the cluster. Therefore, unobservable data due to moving out of a cluster should be treated in the same way as unobservable data due to death. The majority of LTFU, specifically unobservable data, was due to moving out of the cluster and death in DCM-EPIC.

2.4 Methods

This chapter is described as a scoping review as the aims are to explore the level of potential and actual use of open cohort designs in CRTs and epidemiology [62, 63]. For this reason, and due to poor reporting encountered, searches were not intended to be fully systematic or exhaustive but are fully documented.

Several searches were carried out to fulfil the aims of this review. These will be referred to as:

- Care homes and other institutional settings searches (Aim 1)
- CRTs search, not restricted by setting (explicit usage) (Aim 2)
- Epidemiology search, not restricted by setting (explicit usage) (Aim 2)

2.4.1 Search strategies and data sources

Care homes and other institutional settings

The care homes and institutional settings totalled 7 individual searches which will be treated as a set. MEDLINE and EMBASE were searched in March 2019 for the care homes search. The other institutional settings searches were carried out in May 2019 and also used Criminal Justice Abstracts for the prisons setting.

The search strategy initially included the same content as the previous systematic review in care homes [55]. Extra keywords were then included in the search in order to reproduce their results exactly. When searching for CRTs, the same strategy was always used (Appendix, Table A.1). This involved using the subject heading for randomised controlled trials and many ways of describing cluster-randomised trials. Although the CONSORT extension for CRTs [64] advises to include community, field and group randomised trials when searching for CRTs, a decision was made to include only group randomised as in the previous review, because field and community trials broadened the search significantly. The remainder of the search strategies were specific to each setting, using a mixture of MeSH headings and keywords (Appendix, Tables A.2-A.8).

The care home search was limited to 2011 onwards as the previous systematic review included articles published up to the end of 2010. For prisons and palliative care no restrictions were made by year, and after screening only 4 and 8 trials remained in these settings, respectively. Searches in schools, primary care, hospitals and communities returned significantly more results and so were limited to the year 2019 only. Due to the scope of the project, once screening had taken place and a list of eligible articles were obtained in each setting, random samples of eligible articles were taken using a random number generator in Microsoft Excel. A larger random sample of 50 was taken in the care homes articles, and smaller samples of 5 were taken from all other settings, except for prisons where only 4 results were found.

CRTs (explicit usage)

MEDLINE and EMBASE databases were initially searched in April 2019 (Appendix, Table A.9). No limits were made on publication date. Two relevant results were returned; one for ‘open cohort’ and one for ‘dynamic cohort’. As ‘dynamic population’ returned only irrelevant results, these keywords were not pursued any further within the CRT searches, just epidemiology.

A Google Scholar search was carried out to complement the initial database searches because other results were known to exist from previous Google Scholar searches. Google Scholar is not recommended to be used as the only method due to its changing content and inability for users to store searches and use advanced searches [65], but has been recommended to complement established search methods such as those tried initially [66]. Search techniques are not as advanced in Google Scholar and so the search terms were simply "open cohort" AND ("cluster randomised" OR "cluster randomized"), and "dynamic cohort" AND ("cluster randomised" OR "cluster randomized"). It has previously been recommended that only the first 300 results in Google Scholar should be used to avoid grey literature [66]. For the open cohort search, 512 results were found in total with just one relevant article between records 187 and 227, so only the first 227 results were used. 131 results were returned for the dynamic cohort search. Search results were updated in September 2021.

Epidemiology (explicit usage)

This search was also carried out in MEDLINE and EMBASE in April 2019. Keywords included 'open' and 'dynamic' cohorts as well as 'dynamic population', along with MeSH headings for epidemiological studies (Appendix, Table A.10). Again, no limits were made on publication date.

2.4.2 Inclusion and exclusion criteria

After removing duplicates, search results were screened by title and abstract initially using the eligibility criteria discussed in the following sections. Articles included after screening were read in full. Screening and full-text reading was carried out by a single reviewer.

Care homes and other institutional settings

The inclusion and exclusion criteria from the previous systematic review [55] were used as a starting point. This involved excluding articles where outcomes are not reported such as study protocols, designs, pilot and feasibility studies and cost-effectiveness reports. Extra exclusion criteria were also added, which included conference abstracts, reviews, meta-analyses, secondary analyses, process evaluations and substudies, as well as results where outcomes were not focused on participants. Articles which did not involve randomisation by clusters were also excluded, and other designs including split plot and stepped wedge. Results were excluded if the trial did not take place in the care home setting, or if a mixture of settings were used. The inclusion criteria were results papers from parallel-group or factorial CRTs, written in English, and carried out in a single setting of interest. Full detail is provided in the PRISMA [67] flowchart (Appendix, Figure A.1).

PRISMA flowcharts for the other six settings are omitted but the same approach was used. Many of the trials found in the communities search were linked to an institution, such as a community clinic or community pharmacy. These trials were excluded on the basis that they were too similar to those in primary care and instead the search focused on community clusters defined by a geographical area.

CRTs (explicit usage)

Both protocols and results papers written in English were retained for parallel-group or factorial CRTs. Any article not reporting on a CRT was excluded; this consisted of observational studies, cross-sectional studies and individually randomised trials, amongst others. Purely methodological or statistical results were excluded but retained in a separate folder for future reference. Results relating to DCM-EPIC were also excluded. A decision was made to not exclude protocols as in the institutional settings search due to the low number of results and a majority being study protocols. Limits on setting were not imposed in this search. Full detail is provided in the PRISMA flowchart (Appendix, Figure A.2).

Epidemiology (explicit usage)

Some of the exclusion criteria from the previous two searches were used for the epidemiological searches but this was amended throughout the selection process. Articles that were written in English and reported results of studies were retained. To be included a study had to provide details on the setting, and the setting had to have multiple institutions (at least 2) of the same type. Registries that were 'population-based' with no detail on setting were excluded, whereas studies using large databases and routine data were retained as long as they specified the setting. Full detail is provided in the PRISMA flowchart (Appendix, Figure A.3).

2.4.3 Data extraction

Descriptive data extracted that was common to all searches included the first author, year of publication and a description of the setting. The rest of the data extracted was specific to each search. All data extraction was carried out by a single reviewer.

Care homes and other institutional settings

The six institutional settings after care homes were decided on by the project team; they were deemed to be settings in which CRTs are commonly conducted where individuals could be expected to move in and out of clusters over time, for different reasons.

Whether the design used was named was extracted first, to assess whether there was any explicit usage of open or dynamic cohort designs. Whether measurement was discrete or continuous and the number of measurement points was also extracted, where applicable. The total length of follow-up from randomisation and the overall LTFU rate of residents for the primary outcome only was also recorded, as well as the total number of clusters and average cluster size. The full data extraction form is given in Appendix, Table A.11.

CRTs and epidemiology (explicit usage)

The data extraction form for the CRT and epidemiology searches is provided in Appendix, Table A.12. Components that are specific to CRTs were not extracted from the epidemiological studies: this included whether a design was parallel/factorial, the number of arms, whether it was a protocol or results paper and the primary outcome type. For both CRTs and epidemiological studies, specific mentions of ‘open’ or ‘dynamic cohort’ and whether this was describing the population, design, study, analysis or other was recorded. An attempt was also made to categorise whether an open cohort analysis had been used. Information on measurement and recruitment was extracted along with specific parts of the analysis relating to missing data, length of stay, time and the steady state assumption. Finally, any mention of an open cohort estimand or rationale for an open cohort design was recorded.

2.5 Results - care homes and institutional settings

Tables of included trials for the care homes and other institutional settings are provided in Appendix, Tables A.13 and A.14 respectively.

2.5.1 Poor reporting

It was not possible to address the first part of aim 1 of the scoping review, to assess which designs had been used in the care home search. The designs used in the care home CRTs was very poorly reported, with just 2 of 50 (4%) explicitly mentioning this. Chenoweth *et al.* [68] used a “cohort design” and Weintraub *et al.* [69] described the use of repeated cross-sectional samples, referring to residents who were randomly selected in both samples as a “cohort” and the “longitudinal cohort”. Some used the word ‘cohort’, though this appeared to be referring to the group of individuals under study which is potentially misleading. Underwood *et al.* described how their study used different analyses, including a “cohort analysis” and a “cross-sectional analysis”, though a design was not specifically mentioned [70].

Given the poor reporting of designs in the larger care home search, I did not attempt to extract the designs used in the other institutional settings searches. The following results therefore focused on the second part of aim 1, to assess how an open cohort design *could* be used, instead.

2.5.2 Care homes

The majority of trials (92%) took place in long-term care facilities for older people where nursing and/or health care is provided 24 hours a day; some of these facilities provided specialist care, for example to those with dementia. Three trials were carried out in retirement villages or communities, where nursing and/or social care is not provided, and residents live independently in multiple dwellings. These settings can still be viewed as clusters as the villages are often 'gated' with certain amenities or common-rooms shared by residents. One trial described its setting as 'assisted living' for those with dementia. These different settings were retained to assess whether differences existed.

The majority of trials commented on the difficulties encountered with research in care home settings; participants are usually frail and vulnerable with a high risk of LTFU due to death, moving care home or hospitalisation. Some trials classified death separately to LTFU, suggesting that losses due to death were viewed as unavoidable and inevitable in this population [70–72]. Of the 32 trials that provided reasons for LTFU, an average of 60% of these were due to death. In one case, losses were so great that 63% of clusters were also LTFU because all of the residents in the cluster died [71]. There were three occasions where deaths did not feature as a reason for LTFU: one trial with a short two week follow-up, where there were no losses at all [73], and two trials in retirement villages where, although medical reasons were given as some of the reasons for LTFU, no deaths occurred [74, 75]. Although their LTFU rates were still an issue (9% and 27%), this could suggest that the retirement village population does differ slightly to that of residential care homes.

Many trials anticipated high levels of LTFU and made various efforts to alleviate it. Some trials limited trial inclusion to residents with a particular life expectancy: 4 weeks (4 months follow-up) [76]; 180 days (between 12-32 months follow-up) [77]; 6 months (9 months follow-up) [78] and 12 months (12 months follow-up) [79]. None of these minimum life expectancies are greater than the time to follow-up. Others commented on the difficulty of choosing a suitable follow-up period due to expected attrition, identifying a trade-off between the trial being long enough to assess intervention sustainability, but short enough to minimise LTFU [76, 80–82]. In some cases, long-term effects were purposely not investigated because the attrition rates were expected to be too high [83]. In contrast, one trial analysed a closed-cohort of residents who had been present in the home for 24 months, with only 29% of those at baseline remaining at the end [69]. The authors state themselves that this subset of individuals did not represent the general nursing home population who at that time nationally had an average stay of 6 months.

2.5.3 Prisons

Amongst the four trials identified in prison and related settings, each was notably different. Three took place in institutions where residents permanently reside: large prisons in Ethiopia, mother and baby units in UK prisons, and dormitories of a large correctional facility for youths in New York [84–86]. The fourth involved a probation setting for 13-17 year olds, where the individuals are linked to the organisation but do not live there permanently [87].

Of the three trials that reported LTFU rates, the probation setting suffered the lowest attrition rate of 14.8% over 6 months, followed by the youth correctional facility with 26.5% over 21 weeks. The MBU attrition rates were significantly higher, with 29% of dyads lost at 5 weeks and 83% lost at 13 weeks, leaving the investigators unable to perform analyses for the second time point. This was reportedly due to "very rapid turnover" of female prisoners, the majority of whose sentences are less than 6 months [85]. Within the prison setting, LTFU rates vary greatly depending on the exact context, specifically whether prisoners have short term or long term sentences which may be linked to the sex of the prisoner or the type of setting (e.g. incarcerated versus probation).

Reasons for LTFU in prisons were not due to death but predominantly due to release or transferring to other facilities.

Investigators are usually aware of a prisoner's release date, and this can sometimes form part of the inclusion criteria; for example, limiting to those that are expected to remain in the facility for follow-up. However, Sleed *et al.* add that even with this precaution in place, many prisoners will be waiting to be sentenced at the time of enrolment and then will no longer be available at follow-up [85]. As well as using prisoners' release date in the inclusion criteria, one trial excluded prisons (clusters) where the incarceration period was “too short”, retaining only the prisons with longer periods of imprisonment, of “many years” [84]. Reasons for this are not given, but the decision to focus on prisoners with a longer length of stay could have been to reduce the amount of missing outcome data.

2.5.4 Palliative care

The settings also varied widely in the palliative care trials. Three took place in hospitals where the participants were inpatients; two in cancer clinics and one in an acute geriatric ward [88–90]. One randomised GP practices and involved community-based palliative care, and the last randomised offices of a home nursing organisation; in these cases, participants lived at home but were provided with regular care by staff in the community [91, 92]. Despite the range in cluster type, all participants involved were approaching end of life.

One trial assessed participant outcomes by nurse and carer proxy after death, so LTFU rates were not well-defined. The four remaining trials had a median LTFU rate of 37.5%, the highest in all of the settings. Of the three trials that reported reasons for these, a median of 25% of losses were due to death. In all cases there were other reasons that were more common than death, mainly participant withdrawal, possibly due to the deterioration of these participants in such difficult circumstances.

Some of the palliative care trials also limited their study population using life expectancy: specifically, participants needed to be expected to live for more than 48 hours (8 weeks follow-up) [91], more than 2 months (3-4 weeks follow-up) [92], or between 6-24 months (3 months follow-up) [90]. In the latter two cases, in contrast to care homes the minimum life expectancies were greater than the time to follow-up. In one case these strict criteria led to poor recruitment rates and failure to reach target sample sizes [92].

Attempts to minimise missing outcome data were also made in this setting. One trial explained having to balance the need for a minimum trial duration to implement the intervention and for participants to see benefit, whilst reducing an expected high amount of attrition [91]. In another case an original 4 month endpoint was reduced to 3 months before trial initiation, to ensure that losses were minimised and to maintain high power whilst still having sufficient time for participant improvement [90].

Due to the nature of the palliative care population, ethical considerations also complicate the design of trials in this setting. One trial assessed outcomes using proxies after a participant’s death to eliminate the burden for dying participants [88]. Another trial purposely did not collect all outcome measures at baseline, again to minimise the burden for participants, stating this as a trade-off for conducting a trial with “highly distressed patients” [89].

Finally, the only trial in this setting using a factorial design discussed how the design, essentially allowing the conduct of three separate trials, was more economical, reduced participant burden and reduced the overall time frame [91]. In particular, they voiced concerns that funding for one of the individual trials would not have been as generous, resulting from “unique research challenges” in this type of setting.

2.5.5 Schools

Starting with the youngest, the school trials included 5-7 year olds in Spain [93], 5-9 year olds in the US [94], 8-12 year olds in Norway [95], 10-16 year olds in Uganda [96] and 16-17 year old high school students in Norway [97].

Of the three schools that had LTFU information, the median rate was 12.7%. Whilst two were below 15%, one had a very large 55.8% due to a large number of non-responders to surveys, for which reasons are not given [97]. This particular trial involved the older students, who were emailed a link to the surveys to complete outside of school hours. This method of data collection may have been seen as efficient or cost-effective, but placing the responsibility on the students could have drastically reduced participation in comparison to trials where data collection occurred in school. Only one trial provided reasons for LTFU, which included changing school and health problems [93].

All of the school trials appear to have chosen age groups such that the students are contained in the same establishment over the study period, rather than looking at students transitioning from primary to secondary school for example, possibly to avoid attrition.

2.5.6 Primary care

Three of the five trials described the setting as general practices [98–100], one as primary care clinics [101] and the other as healthcare centres [102]. Eligible participants were recruited across most age groups, unlike care homes or schools, with eligibility criteria defined by the specific study aims.

One of the primary care trials was unclear about LTFU, and another defined losses in terms of cytology results rather than participants; both of these were excluded from calculations. Of the three remaining, LTFU rates ranged from 0.9% to 27.3% with a median of 13.9%. Some provided reasons for losses, including refusal to participate and relocation, but others were vague, simply stating "lost to follow-up". The median proportion of losses due to death was 14.3%.

In one trial where the length of follow-up was 24 months, participants with a life expectancy of less than 6 months were excluded [100]. Another trial with 6 month follow-up excluded participants who were planning to move away from the area within a year [101]. Besides this, any inclusion criteria appeared to be directly related to the research question.

2.5.7 Hospitals

The five trials in this setting were described as either hospitals or wards/units of medical centres, but each study population was unique. These trials involved a mixture of inpatient or outpatient stays. Inpatients are defined by the NHS as patients who stay in hospital for at least one night, whereas outpatients may have a bed whilst operations or tests are performed but will not stay overnight [103]. Four of the five hospital trials involved inpatients; one studied participants with drug-drug interactions across various units in a hospital [104], and another involved participants over 65 after gastrointestinal surgery [105]. The third studied those hospitalised with acute myocardial infarction, but followed them up for a year after discharge to assess their medication persistence [106], and the fourth included women giving birth [107]. The outpatients trial focused on men diagnosed with prostate cancer, not residing in the hospital [108].

For one trial, the primary outcome was the frequency of time intervals where potassium levels were not monitored [104]. Although 36 participants died during the study period, the authors did not define LTFU or mention how deaths were treated in the analysis, so this trial was omitted from calculations. For the remaining trials, the median LTFU rate was 5.8%. The trial involving women giving birth had no losses, which is worth noting as a special case; because the participants were enrolled either at second-trimester ultrasound or upon admission to the labour ward, there was potentially not enough opportunity for the women to be lost.

Only one trial provided reasons for LTFU; in the trial involving older participants after GI surgery, 10.8% of the losses to follow-up were due to death. Whilst this may seem low in comparison to the care home trials, it should be noted that the intervention period in this trial was just 7 days on average.

Participants with a life expectancy of less than a year were excluded in one trial with one year follow-up [106]. In the GI surgery trial, participants with an expected length of stay less than 6 days were excluded; the length of follow-up is not explicitly reported but the intervention was 7 days long on average [105]. These criteria seem reasonable as they correspond with the length of follow-up.

2.5.8 Communities

Of the five trials in the communities setting, three specifically defined the clusters as villages [109–111], one had clusters which were areas comprising several neighbouring villages [112], and another's clusters were neighbourhoods covered by a health worker [113]. The villages in one trial were also spread across multiple countries [111]. In all of these trials, the setting was described in some way as "rural". Some study populations included everyone in the cluster (e.g. treating for malaria), and others focused on specific groups (eg. women affected by conflict).

Excluding the trial that admitted newcomers, making LTFU difficult to measure, LTFU in the remaining trials had a median rate of 4%, though one of these is calculated based on pregnancies not participants [112]. In these settings not enough information was provided on reasons for LTFU.

Compromises in length of follow-up or inclusion criteria for the trials were not found in the community setting.

The communities setting is the only one where a trial took a random sample of eligible inhabitants as opposed to assessing everyone eligible in the cluster [113]. This suggests that they could have increased their cluster size if desired, and that larger clusters may be more common in the community setting without an upper limit, as is the case in others.

2.5.9 Summary statistics across the different settings

It is important to note in this section that even if all samples are included, a sample size of 5 is insufficient to make generalisations about these populations, and the samples should be used only to get a flavour of the different settings. Due to the uniqueness of each trial, some calculations have excluded trials because a characteristic is not well-defined. LTFU rates were standardised by dividing the LTFU rate (%) by the length of follow-up in weeks across each setting to enable comparison.

LTFU was found to be higher in the care home, prison and palliative care studies included, with communities having the lowest LTFU rates. The shortest lengths of follow-up and lowest number of clusters were also found in care homes, prisons and palliative care in comparison to the other settings.

Within this sample, care home, prison and palliative care trials were most restricted in their cluster sizes, and the range of cluster sizes was also smaller. Whilst schools, primary care and hospital trials fall in the middle, the community trials had hundreds of participants on average in their clusters.

Characteristic	Care homes	Prisons	Palliative care	Schools	Primary care	Hospitals	Communities
LTFU rate (%/week)	0.8 (0.6-1.0) n = 38	1.3 (0.9-3.8) n = 3	2.9 (2.9-4.7) n = 4	0.3 (0.2-0.8) n = 3	0.3 (0.1-0.3) n = 3	0.2 (0.1-0.3) n = 4	0.1 (0.0-0.3) n = 4
Length of follow-up (weeks)	26.1 (20.4-50.0) n = 50	23.5 (19-32.5) n = 4	10.5 (6.9-13) n = 4	52 (43.3-104) n = 5	62.8 (45.5-81.3) n = 4	52 (52-52) n = 2	45.5 (32.5-69.3) n = 4
Total number of clusters	21 (12-38) n = 50	11.5 (9.3-13.8) n = 2	18 (10-24) n = 5	36 (30-48) n = 5	52 (20-63) n = 5	29 (18-38) n = 5	24 (20-34) n = 5
Average cluster size	11.9 (6.3-21.2) n = 42	23.3* n = 1	19.2 (4.4-28.2) n = 5	76.6 (29.4-81.5) n = 5	39 (35.8-94.4) n = 5	28.9 (18.4-139.5) n = 4	344.5 (196.3-388.3) n = 5

Table 2.1: Summary statistics of characteristics across different settings. Median (IQR). The number of data points used in each calculation is given by n. *Only one measurement available.

2.5.10 Discussion

In this sample of 50 care home CRTs from 2011 onwards, there was no explicit usage of an open cohort design; dynamic cohort designs were also not found. In the two trials that reported a design, one used a closed cohort design and another appeared to use both a repeated cross-sectional and a closed cohort design. It was difficult to determine whether an open cohort design had been used implicitly in the remaining 48 care home CRTs. This was exacerbated by some of the design characteristics also being reported poorly. Instead, the focus has shifted to how an open cohort design *could* be used in institutional settings. Given that literature could not be found to aid the identification of designs given particular features for parallel-group CRTs, this motivated the classification system of Chapter 4.

Findings from this set of searches suggest that care homes, prisons and palliative care settings have the greatest need for a new open cohort design option. These trials made the most compromises in terms of trial design and ultimately their research question in order to alleviate their higher attrition rates and to overcome the difficulties encountered in these settings. With shorter trial durations, they would also be less able to answer questions on longer term intervention effects in comparison to other settings. Attempts to reduce missing outcome data also resulted in narrowing down the target population, for example to healthier, younger participants or to prisoners with long sentences only, which may not always be the population of interest.

Trials in care homes, palliative care, primary care and hospitals used exclusion criteria relating to participants' life expectancy. In hospitals these life expectancies aligned with the length of follow-up, but in the other settings a discrepancy can be seen. In care homes, for example, one trial excluded those with a life expectancy of less than 4 weeks for a follow-up of 4 months. This approach could have improved LTFU to some degree by excluding the sickest participants, but perhaps the investigators were hesitant to exclude all those with a life expectancy of less than 4 months in case this led to a severely reduced sample size.

It is more difficult to say in a general sense whether the school, primary care and hospital settings would benefit from an open cohort design, as the study populations and characteristics in this review were more varied. A ward for psychiatric participants in a hospital studying chronic mental health conditions would have very different characteristics to a stroke rehabilitation ward, for example. Further investigation into sub-types of ward or condition is therefore needed.

Information on average cluster size suggests that there could be upper limits for capacity for clusters in some settings. In settings where individuals are linked to a central hub and visit the institution occasionally, in primary care for example or large geographical areas, there may not be an upper limit on cluster size. The smaller ranges in care homes and palliative care also suggest that cluster sizes in these trials may have been more similar to each other within settings, whereas clusters in the other settings have a much wider range, potentially due to the heterogeneous study populations across trials.

The number of clusters could be an indication of availability of these institutions in practice; for example, an abundance of primary care clusters in contrast to a limited pool of prisons. The exception is where there are a relatively low number of clusters but the average cluster size is very large, as in communities. In settings where there are fewer clusters available *and* the average cluster size is small, as in care homes, prisons and palliative care, it is even more important to use data efficiently from the clusters that exist.

The results also suggest that communities, as they have been defined in this context as geographical regions, are the setting least in need of a new CRT design. The low LTFU rate, abundance of clusters and participants and lack of compromises in trial design mean that an alternative to account for high migration is not necessarily required. For these reasons, the community setting will not receive as much consideration in the remainder of the thesis as the other settings.

Special considerations may have to be made in palliative care and care home settings when designing an open cohort trial, specifically to the number of measurements and participant burden due to the vulnerable population under study.

There are several limitations to this scoping review. The small sample sizes of the other six institutional settings besides care homes mean that general statements cannot be made about these settings. Whilst the sample size for care home trials was larger and results are therefore more generalisable, this still consisted of a random sample from a much larger pool of eligible articles. However, given that poor reporting prevented determination of the designs used in the institutional settings, it could be argued that full systematic searches with larger sample sizes would have provided little benefit over what is reported here.

2.6 Results - CRTs and epidemiological studies (explicit usage)

For the CRTs search, 15 records were identified, with a further 358 found using Google Scholar. Of the 356 unique records, 318 were excluded after screening just titles and abstracts, and 38 were assessed by reading the full text. A further 18 were excluded after reading the full text, leaving 20 eligible CRTs for inclusion. Of the 20 eligible CRTs, only 4 were identified using MEDLINE and EMBASE alone.

For the epidemiology search, 148 records were identified after removing duplicates. A further 127 records were excluded after screening, leaving 21 eligible articles.

Tables of included trials for the CRTs and epidemiology searches are provided in Appendix, Tables A.15 and A.16 respectively.

2.6.1 Characteristics of included trials and studies

2.6.1.1 Settings

The CRTs and epidemiological studies included generally fit into one of the seven institutional settings from the previous search. Of the 20 CRTs, 9 were based in a rural community setting (with one being shared water points in communities [114]), 3 were in care homes (or in one case, social housing for over 55's), 3 were in primary care, 3 were in schools and 1 was in a hospital. One unique setting was food service operations within school districts, where any worker with at least one month of work was included in the sample [115].

Of the 21 epidemiological studies, 1 was in schools, 8 took place in primary care, and 10 in secondary care, which encompassed hospitals, specialised clinics such as infectious disease or dialysis units, and more. Two settings not falling into any of these categories were refugee sites and football clubs. Refugee camps could be a plausible setting in which open cohorts exist; in this example, a "steady number of refugees was maintained during the study, with weekly arrivals and departures towards mainland or other refugee camps" [116]. Similarly, in football clubs it is common for players to move between clubs over time [117].

2.6.1.2 Type of outcome

Within the CRT articles, for the primary outcome there were 2 binary outcomes, 5 count outcomes, 3 continuous outcomes, 9 event outcomes and 1 time-to-event outcome. A distinction is made here between event outcomes where only the occurrence of the event is recorded and not the specific time as in a time-to-event outcome, and binary outcomes which do not relate to events. One of the CRTs reporting use of a continuous outcome actually collected count data but treated the data as continuous in the analysis [118].

The outcome type may in some cases enable trialists to collect data for the primary outcome using EHR or secondary data sources. For example, whether or not an individual has been tested for HIV can be present in EHRs that trialists can use without having to carry out tests themselves [119]. The primary outcome in another trial was the number of hospitalizations, obtained from insurance data [120]. For a continuous outcome which needs to be administered to participants in person and is potentially time-consuming or requires specialists to administer, these outcomes may be less likely to be collected routinely and may also be a reason why only three of these were seen.

Complications related to missing data also differ according to the outcome type (see Section 2.6.2.7.1).

Primary outcome was not extracted from the epidemiological studies as this is not reported as clearly as in trials, with many studies assessing a range of associations, prevalences and risk factors.

2.6.1.3 Recruitment and measurement

Another aim was to determine whether recruitment and data collection happened on either a discrete or continuous basis, and if discrete how many measurement and recruitment points were chosen. Recruitment was more difficult to extract than measurement as it was often not explicitly given; this is not reported as it was unclear in over half of the CRTs, though two cases are provided as good examples. One CRT used helpful terminology of "continuously enrolled" [121], and another enrolled new children "in each follow-up round" [114], which clearly explains that recruitment of new children occurred discretely at the same time as the measurement. The measurement and recruitment processes may have aligned like this in many of the other CRTs but this was only explained clearly in one case. Within the epidemiological studies, the ICONA study also mentioned "continuous enrolment" [122].

Five trials used a surveillance system to recruit and/or collect data, and these were all within the communities setting [119, 123–126]. In these cases, it was implied that the surveillance occurred prospectively in time. However, it was not clear whether surveillance is synonymous with continuous data collection/recruitment, because one of those reporting surveillance also said visits occurred every 3 months in the follow-up period [124], and another two said births or deaths were recorded within 6 weeks [123, 125].

Regarding measurement, five of the CRTs [115, 120, 121, 127, 128] used external data sources such as insurance claims data to assess the primary outcome and were categorised separately given that the trialists had no input in these design choices. Of the remaining 15 CRTs, 12 measured at discrete points and 3 measured continuously, though some of the discrete examples could also be classed as continuous because of how frequent the measurements were. In increasing order, five trials had 2 measurement points, and the rest were unique with 5-7, 9, 12, 25, 28, 37 and 105 measurement points. The latter trials measured very frequently, for example: daily over 28 days; monthly over 3 years; or weekly over 2 years. The 3 CRTs assumed to have continuous measurement mentioned surveillance without referring to specific visiting times, and one of these used the helpful phrase "continuous data collection" [119]. Two of these measured the neonatal mortality rate (NMR) within 6 weeks of a birth [123, 125], and the other measured HIV testing uptake [119]; all used event outcomes and occurred in community settings. The CRTs with continuous outcomes all collected data discretely.

2.6.2 Usage

2.6.2.1 Distinction between population, design and analysis

Four of the CRTs used the phrase 'dynamic cohort', and interestingly were all describing the study population as such rather than the design or analysis. The remaining 16 used 'open cohort'; 6 described the study population, 5 described the design, 2 described the study or trial, 2 described the design *and* study, and one article was unclear. This illustrates that there is some recognition in the literature of the separation between open cohort designs and populations, but that there was no mention of an open cohort analysis.

Despite this terminology being more established in epidemiology, the situation in the epidemiological studies was more complicated; 5 studies described the study population as open/dynamic cohort, 8 described the study, 4 described the study *and* population, 3 described the study *and* design, and one described the study, design *and* population all as open cohort.

Whilst 5 of the CRTs used the phrase 'open/dynamic cohort' without explanation, 15 described their usage in some way. The majority of these definitions involved the allowance of new participants to join the trial throughout the trial period, or stating that anyone who had been present in a cluster at some point during the trial period would be included, with a few also mentioning out-migration. Chaboyer *et al.* further added that because of this, each participant's length of follow-up was allowed to vary [129], and similarly Bell *et al.* that workers "contributed differing amounts of working-months" [115]. It is surprising that so many of the definitions relate to the allowance of participants to enter the study after cluster-randomisation, because this would also be the case for a (repeated) cross-sectional design.

Some of the other definitions warrant special attention. Baiocchi *et al.* described their study as "an open-cohort study, so individual student participants were not tracked between baseline and followup surveys" [130]. In this particular study, student measurements were anonymised due to their sensitive nature, but this explanation implies that an open-cohort study does not attempt to link repeated measurements over time, which contradicts the linkage that was seen in the open cohort analysis models of Feldman and McKinlay and Kasza in Chapter 1.

The unclear result came from Pape *et al.*, who said "the analysis allowed for any new patients joining the practice during the 2 years to be included in the results (open cohort)" [121]. Whilst it is not clear what is being described as 'open cohort' here, there is some reference to the analysis. This highlights potential confusion over terminology. In another case, Ivers *et al.* also linked the definition of an open-cohort design to who is included in the analysis: "we will use an open cohort design; all eligible residents present in the home at any observation time will be included in the analysis" [118].

Piotrowski *et al.* used different designs according to the study population targeted. The closed cohort design included "all insured residents who live in participating NHs [nursing homes] at the start of the study", whereas an open cohort design, used as a sensitivity analysis, included "all insured residents who live in participating NHs at the start of the study *or* during the study" [120].

One misleading definition was given by Agarwal *et al.* [131], who said that "all individuals residing within the buildings at the start of the intervention period are included (intention to treat, open cohort)" - this appears to fit the description of a closed cohort design, rather than open cohort. No further explanation about whether new residents are included is provided in the text and no response was received after seeking clarification from the author.

Lippman described the registry of data as an open cohort, adding that "linkage of health facility and census data allow[ed] us to establish an open cohort of approximately 34,000 18-49-year-old residents" [119]. Pickering *et al.* [114] also discussed the concept of children younger than 5 years migrating into the cluster at birth or upon moving into the area, and 'aging out' of the cluster when they reached 5 years; in this case, the inclusion criteria of the trial dictates out-migration rather than physically moving out of the cluster.

As well as describing a dynamic cohort of individuals, Staedke described the recruitment itself as dynamic [132]. Another interesting point was made by Bavarian, who said that the dynamic cohort occurred at the student level; this specification of a level implies that the dynamic nature could occur at other levels, for example clusters moving in and out of the study over time [133].

Finally, Tripathy *et al.* and Azad *et al.* both used the same definition: "the study population was an open cohort - ie, women could enter the study at any time during the trial period if they had given birth" [123, 126]. Whilst the underlying study population is acknowledged here, this definition could benefit from being rephrased. As the study population is an open cohort, women are expected to join and leave the study population over the trial period. The actual *allowance* of women to enter the study throughout the trial period is not related to the study population, however, but the design.

There was a distinct lack of explanation of the open/dynamic cohort terminology within the epidemiological studies, perhaps as researchers felt that the concepts are already established and do not warrant explanation. In one case an open cohort study design was defined as a design where "each patient's time at risk commenced at a different time point, and some exited prior to the end of the study period," providing a good reminder that a participant's time at risk is a crucial consideration in epidemiological studies [134]. The ICONA study [122] described an open cohort as a "data set extract" including all participants who were seen for care in *either* 2004 or 2014, in comparison to the closed cohort which included all participants seen in both 2004 and 2014, which is similar to Piotrowski *et al.*'s previous definition from CRTs. The most recent definition from Sawicki stated that "all eligible patients of a given year were considered, so that the composition of the cohorts changed over the years and repeated measurement per subject was possible"; they also divided the patients into one or more of four "open cohorts" depending on health conditions, for example if they were elderly or had diabetes amongst others [135].

2.6.2.2 Open cohort design

Given the lack of definitions, I propose for now that an open cohort design allows recruitment of participants after cluster-randomisation and also has the *ability* to link repeated measurements from participants. Seven of the 20 CRTs reported the use of an open cohort design. Both Agarwal and Baiocchi presumably used this terminology because new participants were allowed to enter the study after cluster-randomisation [130, 136]. In the former, surveys were anonymous so with no way to link measurements from the same student, measurements were aggregated at the cluster level. Given that linkage was not possible in either case, from the definition proposed here both of these designs should be defined as repeated cross-sectional designs rather than open cohort.

Clasen and Pickering [114, 124] combined repeated outcomes without linking at the level of the individual, but whether the trialists had the ability to link repeated measurements over time is unclear. It is possible that they used OC designs but did not take advantage of linking repeated measurements in the analysis.

Piotrowski linked recurrent hospitalisations of individuals using generalized linear mixed models [120]. Although the OC design was used to carry out a sensitivity analysis rather than primary, it would appear that in this case it does meet the proposed definition of an open cohort design. However, analysis was carried out at the cluster level to calculate hospitalisation rates, so it would appear that whilst the ability to link was possible, this was not necessarily done. The work of Greiver was similar in that the number of prescriptions prescribed for participants were linked, but the analysis for this took place at the individual level [128].

Finally, Ivers collected monthly number of medications for participants in care homes [118]. Though it is not explicitly reported, linkage must have been possible because in the statistical analysis random intercepts and slopes are included in mixed effects models for repeated measurements. This trial would therefore meet the definition of having both an open cohort design and analysis.

Only the CRTs that claim to have used an open cohort design have been mentioned in this section, but there may have been others who used an open cohort design without reporting it. A full classification of designs is carried out in Section 4.5.

2.6.2.3 Open cohort analysis

Whether the included trials and studies used an open cohort analysis was also of interest. As a definition of an open cohort analysis does not currently exist, an open cohort analysis was defined as one that involves recruitment after cluster-randomisation and links repeated measures from individuals. An open cohort analysis can therefore only be used when the design is also open cohort. To determine whether an open cohort analysis was used, it was necessary to know whether outcomes in a trial consisted of a single measurement or repeated measurements. Just over half (11/20) of the CRTs had a single measurement or event, with the remainder using repeated measurements.

If an event outcome involves a terminal event such as death, it is only measured once and is therefore classed as a single measurement. For example, in four CRTs [123, 125, 126, 137], the primary outcome was the neonatal mortality rate (NMR), an event outcome at the level of the newborn. Another trial measured whether sexual assault had occurred [130]. Similarly, for binary outcomes, one trial measured whether participants met their cholesterol or not [121], and another measured presence of anaemia [132]. As these outcomes included only a single measurement, they will be classed as not applicable in terms of the analysis. One time-to-event outcome was seen where acquiring a pressure ulcer was a single event and is also be classed as not applicable [129]. There were also two cases of count measurements used over a single period that again could be classed as not applicable [120, 128], and one where a single continuous measurement was assessed [136].

Five cases were seen where the primary outcome was event or count, and repeated measurements over time were combined without explicitly acknowledging that repeated measurements came from the same individual. Two similar trials, both studying diarrhoea in children, collected repeated event outcomes but combined the data from all time points in the analysis to calculate prevalence [114, 124]. Three trials used repeated count outcomes, with the first of these collecting cluster-level monthly hospital visits in aggregate form that cannot be used to link individuals [131]. The second of these measured the number of episodes/days absent due to diarrhoea at the individual-level, but combined repeated counts to calculate absence rates across clusters [138]. The third trial collected data on the number of antibiotics dispensed per year using health insurance claims data, but again combined information across years using person-time to calculate dispensing rates [127]. Even though some of these outcomes were collected at the individual level, they were ultimately analysed and reported at the cluster level. There were no examples found where repeated continuous outcomes were combined in a simple way without acknowledging the repeated measurements.

There were 4 cases in the CRTs where repeated measurements were collected and linked at the individual-level, meeting the proposed definition of an open cohort analysis; 2 had event outcomes and 2 were continuous. GEE were used to do this in both of the event outcome examples [115, 119], and mixed effects models with random effects at the individual-level were used for the trials with continuous outcomes [118, 133]. Similarly, in an observational school study with continuous outcomes, Cairney used a mixed effects model with seven time points and a random intercept for each child, again meaning that repeated measurements from the same child were linked over time [139]. The missing data implications of these examples using mixed effects models will be discussed further in Section 2.6.2.7.1.

An open cohort analysis should not necessarily be defined as one that includes all of the individuals present at any point during the trial/study, because when recruitment is discrete and there are a small number of measurement points, it is likely that many individuals who enter the cluster will be 'missed' by recruitment and therefore not included in the analysis. With continuous recruitment/surveillance this definition would hold but it is preferable to keep this more general by saying an open cohort analysis can only be applied to an open cohort design, where there is the ability to link repeated measurements from individuals and new recruits are allowed after CR. Moreover, if an open cohort analysis were defined as simply including all individuals present throughout the trial, the five previously mentioned examples of trials combining repeated event or count measurements over time without linking individuals would be classed as having an open cohort analysis. I propose that this is not sufficient, and that linkage is an important part of an open cohort analysis.

From a practical point of view, in some settings it may be more difficult and/or costly to obtain linkage information which would be a barrier to using an open cohort analysis.

2.6.2.4 Populations and cohorts

There is also confusion in both the trials and epidemiological literature regarding the cohort versus population terminology, potentially due to the contradictory nature of these definitions as highlighted by Rothman in Chapter 1. The Dictionary of Epidemiology gives an example of a closed cohort as women in labour where the outcome is the vital status of their child [6]. In this case, a woman becomes a member of the closed cohort when she is in labour, with membership ending after she has given birth and the study outcome is known. In contrast, in a CRT within the same setting, Azad *et al.* describe women who give birth in a particular region over a specific study period as an 'open cohort study population', merging the two concepts of cohort and population into one [123]. In the latter, this population/cohort is viewed as open because women who become pregnant over the study period can become eligible and join the cohort as time progresses.

2.6.2.5 Rationale for an open cohort design

Only one of the CRTs provided reasoning for using an open cohort design, saying it was due to "high urban migration rates" [114]. More rationale was provided in the epidemiological studies.

Cairney *et al.* [139] conducted a study in public schools, choosing an open cohort design because

> “excluding [new students] would have served no obvious purpose, and would have required the provision of alternative activities during assessments.”

Another study [140] provided rationale for using an open cohort design, stating that it

> “allow[ed] patients to enter the population throughout the whole study period rather than requiring registration on 1 January 1993, thus better reflecting the realities of routine general practice.”

This may well translate to other settings where the influx of participants occurs continuously over time, and demonstrates that this design could offer a more realistic or pragmatic approach than that of closed cohort designs.

2.6.2.6 Estimands

Only one of the CRTs [130] commented on how the estimand changes when using a different design:

> “open-cohort estimands are useful for answering the question, “What will happen to the rate of rape within a school a year after introducing the intervention, assuming natural turnover in the enrollment of the school?” ... a closed-cohort design would be better at answering the question, “What will happen to the probability of rape for a particular girl who received the intervention?” ”

In other words, the authors believe that use of an open cohort design means that the expected turnover in that population should be considered in the research question.

In the ICONA observational study [122] discussed previously, use of both a closed cohort and open cohort design allowed the authors to “address different scenarios”, or in other words, answer different research questions within the same study. The closed cohort participants would age over this period and therefore the prevalence of NCDs would be expected to increase. These analyses should be interpreted with caution, however, as the closed cohort includes people that have survived for at least ten years, and may not be representative of the target population of those currently living with HIV. The open cohort design was used to assess the “net effect” of NCD prevalence between 2004 and 2014 as opposed to the change in a specific group of people. It also allowed the authors to look at how demographics changed at population-level over time; for example, people at the later time point were more likely to be older and male.

2.6.2.7 Analysis methods

2.6.2.7.1 Missing data

For missing data, specific methods that dealt with missing values outside of an individual's length of stay in a cluster or institution were sought, as opposed to missing values within a length of stay. No examples were found of the former. Multiple imputation was used in some of the trials and studies, along with single imputation methods (specifically last observation carried forward), inverse probability weighting and complete case analyses, but these only pertained to missing values within a length of stay.

In Section 2.6.2.3, the analyses used by Bavarian, Ivers and Cairney with continuous outcome variables were described as open cohort because they linked repeated measurements from individuals using mixed effects models. Mixed effects models are unconditional models and so no methods were found that are conditional on the time an individual spent in the cluster or institution.

Cairney states that they left "missing data as missing" because "attrition typically resulted from movement between schools" [139]. Imputation was not carried out for children if they were missing *during* their length of stay in the school. Moreover, no comments were made about missing observations *outside* of a child's length of stay, but implicit imputation will have occurred here for children when they were not present in the cluster due to the use of a mixed effects model. Mixed effects models assume that missing data is MAR, which could be acceptable in the case of missingness within a length of stay (in other words, if a child was observable but missing). It could also be plausible to impute for unobservable data in a school setting, where reasons for missingness are more likely to be MAR, and moving schools in this case may not have been related to the child's fitness levels. However, mixed models were also used in the CRT by Ivers in care homes [118], where reasons for missingness could potentially be MNAR and use of a mixed effects model could be questionable.

2.6.2.7.2 Time

Time is an important consideration of the open-cohort design because with migration in and out of the cluster over time there are potentially multiple timescales at play. Within the CRTs and epidemiological studies there were examples of time-varying covariates and random slopes in mixed effects models, neither of which related specifically to dealing with multiple time scales.

2.6.2.7.3 Use of person-time

As introduced in Chapter 1, incidence rates are often calculated in epidemiological studies for binary or time-to-event outcomes using person-time. Of the 21 epidemiological studies, 11 calculated incidence rates using person-time, or some variation [116, 117, 134, 140–147]. Examples include the incidence rate ratio of fractures per 1000 person-years [141] or the incidence rate of infections per every 1000 CVC-days (days spent with a catheter) [144]. Person-time was also used in 5 of the CRTs [115, 120, 127, 129, 138].

As an example, one epidemiological study using person-time looked at participants over a 15 year period registered in eligible GP practices, recording their entry date to the cohort and using electronic records to see whether any cardiovascular disease events were recorded over this period [140]. The amount of time from cohort entry to event, or censoring, is recorded as each individuals' amount of person-time, and the sum of everyone's person-time is used in the denominator of the incidence rate calculations. Use of person-time is effective for accounting for differing lengths of stay when the outcome is binary or time-to-event, but it is unclear how this could be adapted for continuous outcomes.

2.6.2.7.4 Length of stay

Whether length of stay in a cluster or institution was adjusted for in any of the analyses was also of interest. In an observational study, Hechter *et al.* adjusted for length of follow-up in mixed-effects Poisson regression models, though whether this was through the inclusion of length of follow-up as a covariate or as the exposure is not explained [148]. Whilst a reason is not given, this could have been in an attempt to reduce possible bias caused by participants being followed up for differing periods of time, with those being followed for a longer time possibly having a greater chance of initiating the HPV4 vaccination. In a similar way, another epidemiological study used log person-years of follow-up as an offset term in a negative binomial mixed effects model [146].

2.6.2.7.5 Examples of the steady state assumption

There were no mentions or links to the steady state assumption in the CRT articles.

In one observational study in dialysis centres, participants were followed up for a maximum of 3 years or until they were lost to follow-up when they were censored, due to either death or transplantation [149]. The initial population was sampled randomly. Those lost to follow-up in the initial population were replaced by new incident participants, which ensured a constant number of participants in the study at any one time. Power calculations were based on a fixed number of 5,000 participants to be followed over the 3 years, providing 15,000 person-years of follow-up, thus the steady state method used here appears to have facilitated both the calculation of required sample size and calculation of incidence rates.

2.6.3 Discussion

The main aim of this chapter was to assess the use of open and dynamic cohort designs and summarise their characteristics. Given that 'dynamic cohort' was only mentioned in reference to study populations, for the rest of the thesis only open cohort designs will be referred to. Use of 'open cohort' varied widely, but there was some evidence of differentiation between an open cohort study population and an open cohort design. Definitions of an open population, open cohort design and open cohort analysis need to be restated with improved clarity. Mention of an open cohort population does not imply that an open cohort design has been used; authors may simply be acknowledging that the underlying population is dynamically changing. Furthermore, use of an open cohort study or design does not imply that a corresponding open cohort analysis has been performed.

There appeared to be no consensus amongst the results as to what an open cohort design is, with a wide range of definitions from those who reported to have used one. An open cohort design is proposed here to be one that allows recruitment of participants post-CR and also has the ability to link repeated measurements from participants. In other words, it is a property of the design that allows new participants to enter the trial after cluster-randomisation, as opposed to the study population or the analysis, and linkage is important. Of the 7 CRTs that reported to have used an open cohort design, only 3 clearly met the proposed definition; it was difficult to determine in some of the cases whether there was an ability to link measurements. In one case, an open cohort design was said to be used with repeated cross-sectional samples taken. This was the only example where a type of sampling was reported under the label of an open cohort design. Clearly there is confusion over whether the sampling is synonymous with design, a component of the design or a separate entity altogether.

In this chapter I also defined an open cohort analysis as only being able to be applied to an open cohort design, with repeated measurements linked in some way. Amongst the CRTs, four trials met this definition; two with repeated event outcomes performed linkage using GEE, and two with repeated continuous outcomes used mixed effects models with random effects for individuals. This approach for continuous outcomes is the same method used in the Feldman and McKinlay and Kasza models discussed in Chapter 1 and is the only method seen so far.

The epidemiological literature was also searched to address the second aim of this scoping review on ‘explicit use’ of open cohort designs. The steady state assumption was used in one of the epidemiological studies found, but this appeared to be for calculation of sample size and to aid calculation of incidence rates, which are not of interest in this thesis as they do not apply to continuous outcomes. Having said this, the steady state assumption does feature later in the thesis as it is useful for simulation of an open population (see Chapter 5). The use of person-time was also seen in the epidemiological studies as well as the CRTs, but it was unclear how this method could translate to continuous outcomes. Overall, there were no methods found unique to epidemiology regarding an open cohort design; a mixed effects model for continuous outcomes was seen in an observational study, used in the same way as it would in CRTs. The lack of methodologies, specifically for continuous outcomes, meant that the epidemiological literature could not guide the remainder of the thesis, with the mixed effects model being the only method to take forward. This will be used as a fixed analysis method in Chapter 5, with alternative analysis models attempted in Chapter 6.

DCM-EPIC used discrete recruitment and measurement, so one aim was to investigate the use of discrete or continuous processes within open cohort designs. Within the CRTs, three examples of continuous measurement were found, and were used within community settings for event outcomes. There were however cases of discrete measurement with large numbers of time points which could be classed as continuous. Investigation into the recruitment process was not particularly illuminating as this was reported on poorly.

A further complexity in this review, which also occurred for the first search, was that when designs were not named, it was difficult to determine the design due to a lack of guidance on the components of an open cohort design, or other parallel-group CRT designs. A framework for classifying designs in this area is required, and this gap in the literature motivates Chapter 4.

The use of Google Scholar for the explicit usage in CRTs search is a limitation of this work, but without this a large proportion of the articles would not have been found using database searching alone. It provided a useful supplement in this case because the ‘open cohort’ or ‘dynamic cohort’ keywords were often only cited in the main body of articles, not searched by databases. However, as Google Scholar is not as robust and reproducible as database searching, this was a cumbersome task.

It is perhaps not entirely surprising that the ‘open cohort’ or ‘dynamic cohort’ keywords were often not found in the title and abstract of articles, given that previous reviews found that, between 2000 and 2007, at least 35% of articles failed to specify the much broader category of ‘cluster randomised trial’ in their title or abstract [150]. The search strategy used in all of the searches in this chapter, namely using the ‘randomised controlled trial’ subject heading with a combination of cluster-specific keywords, was also used by authors of this article, though the exact keywords used differ as those used in this work were based on a previous systematic review of care homes [55]. However, the use of the ‘AND’ Boolean operator to combine these, as opposed to ‘OR’, is the same strategy advocated by the authors to increase precision and reduce the reading time for reviewers [150].

2.7 Guidance for improving reporting

When reporting a parallel-group CRT, trialists should aim to clearly describe the measurement and recruitment processes within a trial. Use of ‘continuous’ and ‘discrete’ in the description is encouraged; there were three examples of this in the review which provided greater clarity on the trial design. The CONSORT statement extension to cluster randomised trials [64] item 10c requires trialists to report “whether consent was sought before or after randomisation”, and whilst this is vital in understanding the trial schedule, this could be improved on even further. CONSORT diagrams can sometimes be used to determine the order of recruitment and randomisation, but there is potential for confusion as they are not necessarily temporal in nature, and sometimes imply that individuals are recruited at the same time when this is not the case. A SPIRIT schedule [151] is recommended as it can be clearly marked to show multiple recruitment points and timing of assessments, as well as relative timing of screening, informed consent and cluster-randomisation. One trial used ‘X’s to denote discrete measurements at baseline, and then monthly up to the last visit on the SPIRIT schedule [131], but a horizontal line could be used for continuous processes. If data collection and/or recruitment occur through the use of EHR or other external sources, this should be stated clearly, and whether this is a retrospective or prospective process. Staedke provided a clear study timeline which distinguished between cross-sectional and cohort processes, and initial periods versus ‘dynamic’ periods of recruitment [132]. Greiver provided a schedule which identified enrolment, intervention and measurement periods [128].

Though not ideal for determining the order of processes, CONSORT flow diagrams [152] of participants are useful to report numbers lost to follow-up, but they could also be used to show in-migration when this is expected. This could be used by trials with continuous recruitment, by summarising all new participants after cluster-randomisation, but also when individuals are recruited at multiple discrete points. Bell included the number of workers present in baseline and follow-up periods but no information on who was common to both and who was new [115]. A number of the trials did distinguish who was new [121, 124, 130], but the most information was provided by Pickering *et al.* [114] who summarised how many children had left the population (due to migration as well as ‘aging out’), how many were present at both start and end, and how many had ‘aged in’ or moved into the study area during the trial.

Whether repeated measurements from the same individual were linked over time was at times difficult to determine in this review. If linkage is a key part of trial design, trialists are encouraged to mention this linkage explicitly if it has occurred or will occur, or to provide reasons for why it could not happen or was chosen not to be done. One trial [130] made this clear by saying: “this level of anonymity meant that there was no way to link baseline surveys to outcome surveys for any particular adolescent.”

Finally, when reporting an open/dynamic cohort design, study population or analysis, to aid future research it is recommended that this should be stated in the abstract where possible. In the explicit usage in CRTs search, Google Scholar had to be used to find mentions of these phrases as many mentioned them only in the article’s main text. The CONSORT specific guidance for abstracts currently only includes the same information as the CRT specific statement, with “description of the trial design (e.g. parallel, cluster, non-inferiority)” given under the “trial design” heading; this does not seem comprehensive enough to capture the developing level of detail of trial design in this field. Reporting could also be greatly improved by specifying exactly what element or elements are open/dynamic, specifically, rather than reporting an ‘open cohort trial/study’, trialists should explicitly mention the population, design and analysis.

Chapter 3

Designs and sampling schemes

3.1 Introduction

The introductory chapter to this thesis was split into the concepts of population, design and analysis, and CRT designs already established in the literature were summarised. In the scoping review, electronic searches were conducted to assess the characteristics of and methods used in open and dynamic cohort designs in both epidemiology and CRTs. It was apparent over the course of writing these chapters, however, that the definition of a design can be ambiguous, and if insufficient information is provided, incorrect assumptions can be made by the reader. Designs can range from simple to extremely complex but even the most simple designs consist of several components that often go unsaid or are assumed. An aim of this chapter is therefore to set out the definition of a design, in the context of CRTs, for the remainder of the thesis for absolute transparency.

Chapter 1 also introduced elements of the surveys literature relevant to this thesis, such as the concept of a sampling frame. The second aim of this chapter is to extend these concepts to the field of CRTs. Different types of sampling scheme for CRTs will be discussed, along with practical considerations for their implementation. Some of these sampling schemes are already established and commonly used, and some have recently been proposed by Kasza, specifically for open cohorts [36]. In addition to this, I propose two novel sampling schemes. Consideration of how some sampling schemes may be more suited to particular situations, such as large cluster sample sizes, is given, along with a summary of whether, within a sampling scheme, there is independence between successive samples. This chapter concludes with guidance on how reporting of CRT designs could be improved.

3.2 Designs of CRTs

As this thesis focuses on designs for CRTs, it is important to be clear on what a design consists of exactly. Murray [49] proposed that:

> "The design will specify the number of conditions, the number of levels of each condition, and whether those conditions are completely crossed or not. It will specify whether the data are from discrete time intervals or represent continuous surveillance. It will specify whether each group and member was measured only once or more than once. It will specify whether the groups were matched or stratified *a priori*. Finally, it will specify the selection and allocation schemes employed for the groups and members."

Firstly, terminology can vary between experimental design and clinical trial literatures, but also within clinical trial literatures; for example, for some public health trials, group is used to refer to clusters, and conditions are used to refer to arms or treatment groups, as Murray uses above. The terminology of treatment group or arm is commonly used in parallel-group CRTs, but cannot be used in SW or CRXO CRTs as all clusters experience both intervention and control *conditions* and treatment groups are not clearly defined. To avoid ambiguity for the remainder of the thesis, I refer to clusters as the groups of individuals which serve as the unit of randomisation. As I will focus on parallel-group CRTs only, I will specifically use arms to describe the distinct collections of clusters who are exposed to differing treatments; in a two-arm trial, assumed for this thesis, this will specifically be an intervention and control arm. I also do not adopt Murray's use of the term 'levels', which appears to refer to the number of clusters within an arm or condition, as this will predominantly be used in the thesis instead to refer to the level at which an intervention is delivered (individual- or cluster-level).

In addition to Murray's specifications, if discrete measurement takes place, further information could also be provided on whether the measurement timings are linked to timescales pertaining to the individual, the cluster, or other. For example, measurements could be anchored to an individual's time of recruitment or exposure or to the time of cluster recruitment (see Chapter 4). It could also be argued that in addition to discrete or continuous measurement that discrete or continuous recruitment is an important part of the design. From the scoping review of Chapter 2, in many cases the term 'OC design' was used to mean new participants were allowed to join post-CR; this is again a factor that I believe is part of the design. Lastly, assumptions are often made that sub-samples are taken from larger cluster population sizes. This is not always desirable depending on the setting, as in the case of DCM-EPIC for example the limit on cluster sizes meant that full samples of all available participants had to be taken. I propose that full- versus sub-samples should also be specified as part of the design.

A design for a CRT should therefore specify the following:

1. Whether the CRT type is parallel-group, factorial, SW or CRXO[1]
2. Number of arms
3. Discrete or continuous measurement. If discrete, specify whether measurements are linked to individual or cluster timescales
4. Whether a single measurement is taken per participant or repeated measurements. This could also apply to clusters, but is outside the scope of this thesis
5. Discrete or continuous recruitment
6. Matching or stratification
7. Whether full- or sub-samples of clusters are taken
8. The sampling scheme used for participants, if sub-samples are taken. This could also apply to clusters
9. Whether new participants are able to join post-CR
10. The randomisation scheme for clusters. This could also apply to participants, but is outside the scope of this thesis

Now that this is set out, it is clear to see that referring to a "R-CS design" could be misleading and does not provide enough information. In most cases, this means a sub-sample is assumed, and a sampling scheme of repeated random samples is used. However, the design consists of much more than just the sampling method, so it is recommended that when describing a design, all of these components are clearly broken down. Similarly, whilst an "open cohort design" in Chapter 2 most commonly referred to the ability to include new recruits after CR, it is crucial that the other elements of this design are laid out clearly.

Given that there are a lot of elements to specify in a CRT design, from this point onwards

[1] An equivalent statement using experimental design language, as alluded to by Murray [49], is whether clusters are nested in arms (parallel, factorial) or crossed with arms (SW, CRXO), and for the former, whether arms are completely crossed with each other (factorial) or not (parallel-group). However, this could be seen as overly complex for those not familiar with this literature.

I will make some assumptions. I assume a two-arm, parallel design, though this could be extended to more arms and a factorial design. Matching and stratification are beyond the scope of this thesis, as are single versus repeated measurements of clusters, sampling schemes for clusters, and randomisation schemes for participants within clusters.

R-CS design for sub-samples means that a random sampling scheme is used at each time point. The OC design for sub-samples similarly means that an open cohort sampling scheme is used at each time point. The CC design for sub-samples means that a closed cohort sampling scheme is used.

I then extend these definitions to their full-sample versions. As full-samples of clusters are taken, a sampling scheme is not required. The CC design for full-samples would measure all those available in the cluster but not allow new recruits post-CR. Given that the R-CS and OC designs for full-samples both allow new recruits post-CR, another factor is needed to differentiate between them. I propose that the linkage of repeated measurements is part of the OC design for full-samples, and not for the R-CS design for full-samples. Therefore, throughout the thesis, "R-CS design", "OC design" and "CC design" will be used as shorthand to refer to these designs, but it is important to note that they differ depending on whether they are full- or sub-samples. As previously described in Chapter 1, the R-CS design could also refer to either the type with overlaps or without. The specific sampling schemes are described in the following section.

Design	**Full/sub sample**	**Sampling scheme**	**Ability for new recruits post-CR**	**Ability to link repeated measurements**
CC (full)	Full	NA	✗	✓
CC (sub)	Sub	CC	✗	✓
R-CS (full)	Full	NA	✓	✗
R-CS (sub)	Sub	R-CS	✓	✗
OC (full)	Full	NA	✓	✓
OC (sub)	Sub	OC (beds)	✓	✓

Table 3.1: How the CC, R-CS and OC designs will be defined throughout the thesis. Sampling schemes are not required for full-samples and are given as NA. R-CS could refer to the version with or without overlaps.

3.3 Sampling schemes

In this section a range of sampling schemes will be described, some of which are well-established in the literature, some which are entirely novel that I propose, and some recently proposed by Kasza which are inspired by the surveys literature (see Chapter 1). Sampling schemes are only used within a design that takes sub-samples from clusters. The sampling scheme is linked closely to whether new recruits are allowed post-CR.

I define the 'cluster sample size', m, as the number of individuals sampled at each measurement point in each cluster, and the 'cluster population size', M, as the full *available* size of the clusters at each measurement point.

3.3.1 Established sampling schemes

3.3.1.1 Closed cohort

This sampling scheme typically involves taking a sample at the start of the trial and measuring those selected until trial end, unless they are LTFU. This is the simplest of the sampling schemes as sampling only occurs once. In practice this would also be simple to implement in settings such as care homes, as the same individuals are followed. Simple random sampling could be used to select the initial sample, but alternatives such as stratified, purposive or systematic sampling could also be used [153].

To reduce the possibility of selection bias, the random sample should be made by a third party [64].

3.3.1.2 Repeated cross-sections (with overlaps)

Perhaps the most widely known sampling scheme, the R-CS sampling scheme takes random samples at each measurement point. This version permits overlaps, that is, the same individual can be chosen more than once. The scheme is illustrated in Figure 3.1 alongside the OC (beds) sampling scheme. The 15 participants are randomly sampled at $t = 0$ and $t = 1$, with those that happen to be re-sampled ('overlaps') given in pink. Unlike the closed cohort scheme, sampling has to be conducted at each time point, and at each time point an update would have to be made of the full eligible population, or the sampling frame, which would require more work, particularly with high turnovers. However, the actual method of sampling from eligible individuals is simple.

Simple random sampling is likely to be the most common choice of sampling in this scheme due to its unbiasedness and simplicity in implementation, especially when repeated sampling occurs, but other sampling methods could also be used as described above.

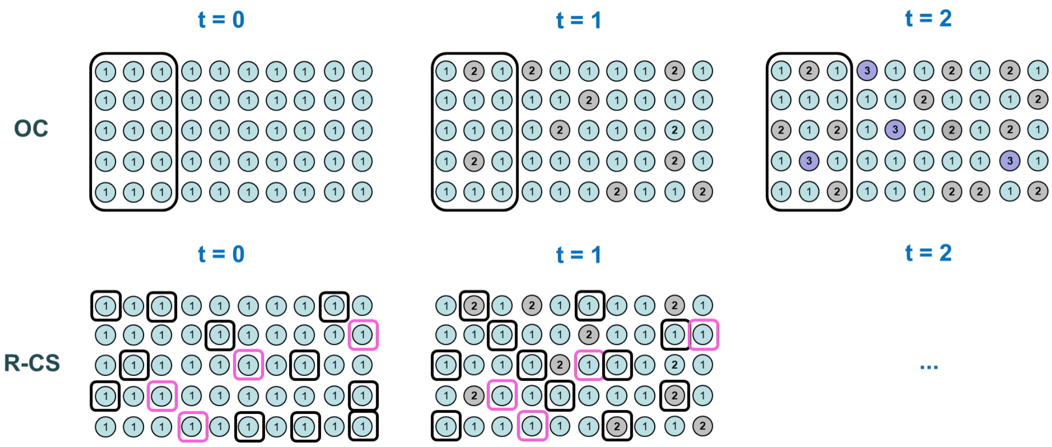


Figure 3.1: Illustration of how sampling differs between OC and R-CS designs in sub-samples with a cluster population size (M) of 50 and cluster sample size (m) of 15. A number 1, 2 or 3 in a circle represents the first, second or third person to be present in this bed since time zero. Pink selections for R-CS denote overlaps between $t = 0$ and $t = 1$.

3.3.1.3 Repeated cross-sections (without overlaps)

This scheme also has random samples at each measurement point, but overlaps are not permitted, so individuals previously chosen are omitted from future samples. This would therefore be slightly more cumbersome in practice than the previous scheme.

3.3.2 Novel sampling schemes

3.3.2.1 Open cohort (beds)

In this novel open cohort sampling scheme, a sample of size m is taken at $t = 0$, and any drop-outs are replaced directly by the person entering that same bed (as in the case of the full-samples). Participants are retained in the sample for as long as they remain in the cluster, similar to the CC sampling scheme. This sampling scheme could therefore be thought of as sampling beds rather than individuals, as the same beds are measured over the course of the trial as people move in and out of them, assuming individuals do not move between beds. This scheme is illustrated in Figure 3.1. In a setting such as a care home, if a certain number of beds or rooms in a cluster are known to be included in the sample, researchers know who will be replacing those that drop-out and they can be more easily identified, making this scheme relatively simple to implement. However, if replacement individuals are not eligible, a replacement would have to be sought from elsewhere in the cluster, or if a replacement is not found there may be a period of time where the bed is empty and there is a gap between participants and the steady state assumption does not hold. There may also be situations where multiple individuals leave and enter the cluster at the same time. In this case, there would be a choice of who replaces who, so to reduce bias a simple random sample of those entering could be taken.

3.3.2.2 Open cohort (wider cluster)

This sampling scheme, also novel, is similar to OC (beds) in that participants are retained in the sample for as long as they remain in the cluster. However, when someone drops out, instead of sampling the replacement individual who enters the same bed as in the OC (beds) sampling scheme, the replacement individual is sampled from the wider cluster population at that time. This is illustrated in Figure 3.2. For example, if m = 15 and M = 50 and someone from the sample of 15 drops out after the $t = 0$ measurement (orange), a replacement (blue) enters that bed. There are also four other drop-outs after $t = 0$ who were not in the sample of 15 (green). The sample at $t = 1$ will then consist of the 14 members of the original cohort (grey), and the 15th individual is sampled from those that remain (contained in dashed lines). For this sampling scheme, the probability of sampling an individual from the additional cohort is just 5/36, so there is a high chance of sampling original cohort individuals as replacements. In general the probability of sampling an individual from the additional cohort is $\frac{1+D_{ns}}{M-m+1}$, where D_{ns} is the number of drop-outs in those that are not sampled, equal to 4 in this example. In practice this scheme would likely be more difficult to conduct than the previous OC scheme, as random samples are instead taken from the wider cluster for replacements, and like the R-CS schemes this would involve updating the sampling frame at each time point.

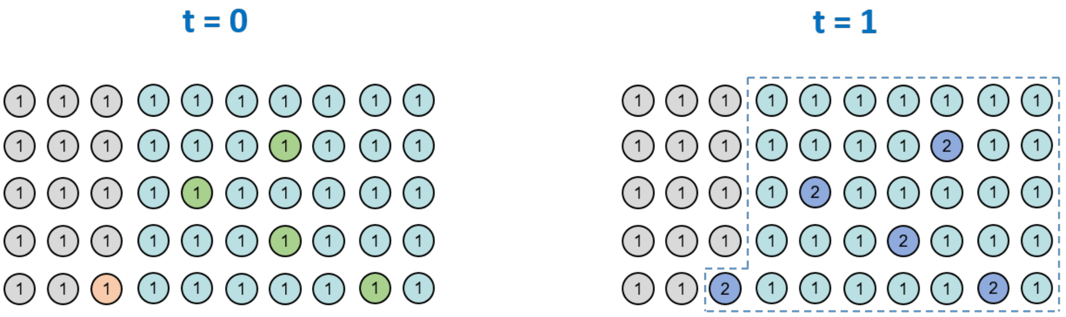


Figure 3.2: Illustration of the alternative OC sampling method for sub-samples with M = 50 and m = 15. A number 1 or 2 in a circle represents the first or second person to be present in this bed since time zero. The single drop-out after measurement at $t = 0$ from the sample of 15 is given in orange; four more drop-outs not in the sample of 15 who drop-out are given in green. Replacement for drop-outs at $t = 1$ are in blue. The sampling frame from which the replacement is sampled at $t = 1$ is contained within the dashed lines.

3.3.3 Recent sampling schemes proposed by Kasza

3.3.3.1 Core group

The “core group” sampling scheme collects measurements from a fixed group of individuals over the trial period as well as also measuring others outside of this group once at most at different time points [36]. This is a step up in terms of complexity in practice as there are two sampling processes occurring simultaneously. Kasza adds that this design would be most appropriate when participants are expected to remain in the trial either for the entire duration or for a single trial period. The expectation of some individuals to remain in a trial for the entire duration is a luxury that may be feasible in some settings, such as communities, but would not be realistic in settings considered in this thesis such as care homes where missing data is inevitable. In R-CS sampling schemes without overlaps, one measurement per person may be expected due to the design, for example if measurements are very far apart, but to expect two extreme types of individual may not be realistic in practice. With larger cluster sizes it would be feasible to collect non-overlapping samples, but with small cluster sizes the pool of individuals to sample from would be limited and the chance of overlaps increases. As this thesis is motivated by DCM-EPIC where small cluster sizes were an issue, this sampling scheme will not be pursued further.

3.3.3.2 Closed population

This sampling scheme identifies the population at the start, and repeatedly samples from this same closed population, as defined in Chapter 1, not allowing new recruits [36]. This scheme would be fairly simple to implement as once the sampling frame is determined in the beginning, random samples are taken from it over time, without needing to update. This scheme does not recognise the underlying open population assumed in this thesis, and is therefore unsuitable, but could be suited to situations where the underlying population is closed. It could also be considered as an alternative in settings where the population under study cannot be fully identified, but for the settings considered in this thesis it is assumed that identifiability of the sampling frame is not an issue. This sampling scheme was proposed by Kasza as a possible *open cohort* sampling scheme, which appears to be contradictory given it concerns a closed population, however, this scheme may have been considered as its simple ‘churn rate’ can be used with the formulae provided by Kasza to calculate a sample size for an open cohort design.

3.3.3.3 Rotation sampling

The final sampling scheme discussed by Kasza is rotation sampling, which involves imposing a maximum number of times each individual can be measured consecutively and the fraction of individuals to be replaced at each time point [36]. In practice this would be more complicated to conduct and keep track of the participants in rotation in comparison to simple random sampling. This method is said to be suitable if reduction of measurement burden for participants is a priority, and given that this is not of key importance for this thesis will not be pursued further.

3.3.4 Independence of successive samples

In Chapter 1, I outlined that the R-CS design with overlaps has no dependence between past and future samples. This is also the case for the closed population scheme. All of the other sampling schemes involve dependence between successive samples in some way. In the R-CS without overlaps, dependence exists if individuals have to be manually removed from the sample. The closed cohort, core group, and OC schemes similarly have dependence as future samples depend on who was in previous samples. This dependence could be seen as a positive; for example, in the OC and CC schemes, retaining individuals for as long as they remain maximises the amount of linkage that can occur for repeated measurements. However, although the initial sample is taken randomly and should be representative of the cluster, any ‘strangeness’ in this initial sample is present for as long as these individuals remain, which could be throughout the whole trial period.

3.4 An example

The R-CS design is defined by Copas *et al.* in the context of SW-CRTs as “open cohort with repeated cross-sectional sampling” [32]. It is unclear whether ‘open cohort’ in this context means that individuals move in and out of clusters over time, or the ability to be recruited post-CR. If the former is true and the underlying population is open, this definition would benefit from separating the open population from the design applied. However, if the latter is true, this serves as a good example of stating some of the individual components of a design. This definition explains that the overall design is called R-CS, but within this individuals are able to be recruited post-CR, sub-samples are taken (implied), and the sampling scheme is repeated cross-sections.

3.5 Discussion

This chapter outlines how the umbrella term of 'design' can include many complex components and how, without full elaboration, designs can be unclear.

The R-CS sampling scheme *with* linkage is not investigated in this thesis, which is a limitation. Typically the nature of R-CS sampling implies that linkage is not possible, or desirable, but it could be that in certain circumstances repeated cross-sectional sampling with linkage is beneficial; this is an avenue for future research. Trialists may be tempted to report this design with R-CS sampling and linkage as a "R-CS design", which is different to what I have defined in this chapter and a potential issue going forward. However, stating a "R-CS design" *in addition* to a full description of its components would help clarify exactly what was meant by this phrase, and a further reason to provide as much detail as possible.

3.6 Guidance for improving reporting

The existing CONSORT statement extension for CRTs provides trialists with specific guidance on information that should be reported in a CRT results article [64]. The standard CONSORT item 3a requires trialists to report "description of trial design (such as parallel, factorial)," and to extend to CRTs, "definition of cluster and description of how design features apply to clusters". Understandably this is provided in a vague way, due to the many components of designs, but this item could benefit from more specific prompts such as those given herein so that trialists can describe a design fully. The statement does not specifically instruct to include CRT types such as parallel-group, factorial, SW or CRXO. A further element that is missing completely from the CONSORT examples is whether or not participants can be recruited after CR. Whether full- or sub-samples are taken, and the sampling schemes used if sub-samples are taken, are included in the CONSORT extension somewhat under Item 10b: "mechanism by which individual participants were included in clusters for the purposes of the trial (such as complete enumeration, random sampling)". Here, complete enumeration is assumed to be equivalent to full-samples, and random sampling is a vague type of sampling scheme. This item could benefit from adopting this terminology of full- versus sub-samples and sampling schemes, and making it clear that many variations of sampling scheme exist.

Chapter 4

A classification system for parallel-group CRTs

4.1 Introduction

One of the aims of the scoping review in Chapter 2 was to investigate the use of open and dynamic cohort designs within CRTs and epidemiological studies when the clusters have underlying open populations. The first main challenge of Chapter 2 was poor reporting; only 2/50 (4%) of the care home trials explicitly reported the type of trial design used. Attempts were made to identify designs based on their components, but this led to the second challenge: a lack of guidance in the literature linking design components to specific named designs. Moreover, in the explicit usage search where the "open cohort" phrase was used, the definition of an open cohort design varied widely. In this chapter, some guidance is provided for improving reporting to address the first challenge in addition to that already provided in Section 2.7, but the main focus is to address the second challenge via the development of a framework to enable trialists to classify their parallel-group CRTs based on individual components.

Building on a previous framework proposed for classifying stepped-wedge CRTs [32], I developed a classification system for parallel-group CRTs which breaks down the important components of six sub-designs, enabling trialists to identify their full trial design and assign it a name. Whilst the classification system can be used to retrospectively assess CRTs if a design is unclear, it could also be used by trialists at the planning stage to view possible design options, increase transparency of the processes involved and to prompt consideration of the statistical issues that may arise. To accompany the classification system, I also produced diagrams of the designs to allow clear visualisation of each design, and a flowchart to quickly identify the design.

As this thesis is focused on institutional settings, the classification system applies only to clusters that are institutions: schools, hospitals and so on. It does not apply to CRTs where the unit of randomisation is the individual delivering treatment (the ARTIST trial, for example, where rheumatologists were randomised [154]).

The six proposed sub-designs include the 'closed-cohort' and 'repeated cross-sectional' designs previously discussed, as well as the 'new admission continuous recruitment' (NACR) design and two proposed 'open-cohort' designs. The NACR design involves continuously recruiting participants as they become eligible over time, therefore focusing on participants that are newly admitted only, and measurement at fixed time points relative to a participant's date of recruitment or other personal milestone. In both open-cohort designs, participants are recruited before and after cluster-randomisation. They differ by the type of recruitment; discrete or continuous. The crucial components of an open-cohort design are identification and recruitment of participants both before and after cluster-randomisation, the entry of new participants throughout the trial period resulting in variable exposure periods and, to distinguish it from the repeated cross-sectional design, ability to link repeated measurements over time from the same participant. Within this framework I have further differentiated between two types of the closed cohort design; a 'standard' type as well as a 'non-standard' variant.

These six sub-designs can be thought of as broad combinations of different elements that I previously listed in Chapter 3 as components of a design. Many of these elements can be seen in the classification system (ability to link repeated measurements, ability for new recruits post-CR, discrete versus continuous recruitment), but there are also further components in this chapter. Specifically, timescales are introduced in this chapter as well as consideration of the different levels of intervention effect that can occur in CRTs, when these exposures begin and how long they last. In this chapter, the general shorthand version of CC, R-CS and OC designs is used as described in Chapter 3, without reference to full- or sub-samples.

Some of the same terminology from Copas' work is used here, such as "continuous recruitment", which describes how participants can be recruited in continuous time as opposed to discrete intervals [32]. Whilst the closed cohort and open cohort designs presented here are the same as those described by Copas (albeit for SW-CRTs), the continuous recruitment design in their work is specifically 'continuous recruitment *short* exposure' (CRSE). In this CRSE design, the very short exposure period relative to the trial duration means that follow-up measurements may be collected from participants much longer after they have stopped being exposed to the intervention. The 'short exposure' element in Copas' work is important in the context of SW-CRTs because it leads to participants experiencing just one of the intervention or control conditions, not both as is the case in the closed cohort and open cohort designs.

In this chapter I will describe how the classification system was developed using user engagement work, giving details on specific revisions that were made. A flowchart for quick identification of designs is then given, followed by the full classification system and a description of each item. Illustrative examples are then provided for each of the designs alongside diagrams. The chapter is concluded with a final discussion of this work in context of the wider literature, as well as limitations and future avenues for research. A summary of the trials provided by respondents in the user engagement work that I classified are also included in Section 4.5.

4.2 Methods

4.2.1 Overview

The initial classification system was first reviewed by my supervisory team and then tested using a sample of trials from my scoping review of Chapter 2 to ensure it was as widely applicable as possible. In the second user engagement workshop, trialists from a range of settings who work on CRTs were asked for their feedback on the next iteration of the classification system; they used it as a tool to classify some of their own trials. From this, one suggestion was made that staggering of processes can occur individually as well as all together or in batches; this is now option 10c ("timing of intervention delivery and measurement is unique for each cluster"). The final stage of testing was through the user engagement survey. As described at the very beginning of this thesis, the online survey was created by other members of the user engagement team, whereby items from my classification system were transcribed into questions. Other questions relating to respondents' characteristics and questions about implementation and barriers for an open cohort design were also included in the survey, but these are beyond the scope of this thesis and will be presented elsewhere. Following each item of the classification system, respondents were given an option of 'none of the above (please specify any problems/options you feel are missing)' in a free text field. The majority of amendments were made following this survey.

In this final revision, the aim was to focus on what is important statistically rather than on logistical problems, as this classification system is intended as a tool for statisticians when designing CRTs and assessing risk of bias.

4.2.2 Revisions made following the workshop and survey

Prior to the survey, the classification system had just one open-cohort design option, but this was subsequently split into the discrete and continuous recruitment versions for improved clarity.

Item 1 was initially about the timing of participants 'joining' clusters. Workshop attendees commented that this phrasing did not necessarily apply to primary care and the term 'joining' could potentially be ambiguous. For all of the settings I considered, except for primary care, there is a clear interpretation of what it means to join a cluster. In care homes, hospitals, communities, prison and palliative care, this could be thought of as physically moving into and residing in the cluster; in schools this could be defined as enrolling at the school. For primary care, however, many people register at birth or when they move to a new area, so can be a "silent" member of the cluster for a long time without necessarily having any interaction with the cluster: "it's only when you come into contact with them or they contact you because you're on a list of potentially eligible patients that you've actually entered the cluster. So it's almost like a two-stage process." In light of this, the concept of joining the cluster was removed from the classification system and replaced with the statement: 'whether all eligible trial participants are identified before cluster randomisation or not'.

Two of the previous responses to Item 1 report whether presence in the cluster is a co-intervention. Two survey respondents commented that they were unsure what this meant. In light of this, and due to presence in the cluster prior to randomisation being a separate issue, these options were removed from Item 1. However, whether length of stay in a cluster before cluster-randomisation may affect outcomes in the intervention arm remains something that should be considered in terms of bias, so this issue was moved to a section on 'additional design considerations' at the end of the classification system.

Under 'measurement', the classification system initially included options which conflated the outcome type, the timing of measurements and ability to link repeated measurements. The final version separates these into components in Items 2 and 3 to emphasise the statistical reasoning behind each. After these components had been separated, the outcome type (eg. continuous, binary, time-to-event) was deemed no longer integral to this aspect of the design and was removed. The number of measurements was also removed for the same reason, with only a distinction now made between a single measurement and more than one measurement. Both of these aspects would be integral to the calculation of sample size but do not aid the differentiation between designs.

Some of the survey responses indicated that the system did not necessarily cater for CRTs where individual data is available but where there is no recruitment or consent at the patient level. In the final version, I clearly differentiate which items can only be answered for trials which include participant recruitment, and reduce the number of references to recruitment when not necessary. Initially, the 'timing of participant recruitment' item contained a mix of information on whether recruitment of participants was discrete or continuous, and whether they were recruited before or after cluster randomisation or both. To simplify this, and to attribute each component clearly to a statistical issue, I split up when participants are recruited and the type of recruitment (discrete or continuous). However, as Item 1 now asks about knowledge of eligible participants before cluster randomisation, when participants are recruited does not provide any further information and so this component has been removed. For the type of recruitment, the closed cohort and cross-sectional designs will always have discrete recruitment under Item 4, and the NACR design continuous recruitment, so the main purpose of this item is now to distinguish between discrete and continuous versions of the open-cohort design.

Initially in Item 4 regarding recruitment, one of the options was for participants to be recruited at 'discrete time points'; this has been changed to 'discrete windows' because it would be impossible in many cases to recruit all participants on a certain day or week, and in reality there would be a tolerance of a few weeks. The initial options for this item also implied that all participants must be recruited before cluster randomisation in a closed cohort design, but this is not always the case.

For Item 5, a survey respondent commented that the start of exposure to the intervention might occur for all participants simultaneously but this does not necessarily mean that exposure occurs immediately after cluster randomisation, so this has been incorporated into the final version. For example, there may be a lag to accommodate a staff training period.

For Item 6, 'planned duration of exposure', I felt more elaboration was required in the case where exposure to the intervention occurs until the end of the data collection period, after comments from some survey respondents. I assume that even if formal training elements of the intervention finish after a short period of time, that such training would lead to a change in the practices, culture or environment of the cluster and so the cluster-level intervention continues until the end of the data collection period *and* possibly beyond.

The 'type of intervention' is now referred to as 'level of intervention delivery' in Item 7. At the time of the survey, this item only included options of a cluster-level or individual-level intervention. One survey respondent commented that it was possible to have both, so the options were revised to include individual-level only, and cluster-level only or both cluster-level and individual-level. One statistical implication for this item is whether the timescale of interest in the trial is the timescale of individuals or both the timescale of the cluster and of individuals. Furthermore, at the design stage, knowing the level of intervention delivery prompts discussion about the level at which outcomes should be measured, which outcomes are most relevant to the intervention itself, and which levels should be randomised. Looking retrospectively at a completed trial, if the level of intervention delivery is clear, this allows the reader to understand the reasons the outcomes were chosen, the level at which they were collected, and the choice of units of randomisation.

Items 8 and 9 were initially a single item but are now separated for further clarity. Cluster-level learning and cluster-level drift are additional considerations but do not necessarily dictate this aspect of the trial design. Staff training was initially referred to as the reason for learning and/or drift, but this reason has been removed to generalise this point after a survey respondent commented that staff training may not necessarily be a part of an intervention.

At the time of the survey, Item 10 was referred to as 'timing of cluster *recruitment*', but this is not necessarily what is of interest here. Nor was I interested in the timing of cluster *randomisation*, because even if clusters are randomised at different times, their interventions and measurement schedules may start at the same times; also conversely, clusters randomised at the same time still may start the intervention and measurements at different times. This item has therefore been revised to make it clear that interest is specifically in intervention delivery and data collection.

Items 8-10 in this system only apply to CRTs with cluster-level interventions. Individual learning can also apply, and literature exists for example on learning effects in surgical trials [155, 156]. This issue is therefore not specific to CRTs. Similarly, intervention delivery and measurement schedules are an issue for all randomised trials, so the focus here is on the implications for cluster-level intervention effects only. The final consideration under

Item 11, ‘presence in the cluster before cluster-randomisation is a co-intervention’, could also be an issue for CRTs with an individual-level intervention only.

Initial versions of the classification system included an item on missing data. As with type of outcome and number of measurements, this is important for the sample size calculation but has been removed so this tool can focus on the design only. Moreover, missing data considerations were specific to individual trials, and a number of respondents indicated that they found it challenging to answer.

4.3 Results

4.3.1 Design flowchart

The design flowchart (Figure 4.1) is intended as a shortcut to identify a trial design using a minimal number of key questions. The following section provides more elaboration on each item and may be required for complex cases.

4.3.2 Overview

Generally within each item, I begin with the simplest option, becoming progressively more complex. Options are not necessarily exhaustive. Some combinations across items are also not possible. While this system has been designed to encompass a range of settings, multiple primary outcomes, discrete and continuous measurements, and simple and complex interventions, it may not cover all eventualities.

There are 6 core ‘design indicators’ which are properties of trials which I propose directly influence this aspect of the design, and 5 ‘additional design considerations’ which I suggest trialists consider when designing parallel-group CRTs (see Figures 4.2-4.4).

4.3.3 Core design indicators

Item 1: Identification of eligible participants before cluster randomisation

This item not only helps to classify a trial design; it is a partial indicator of whether or not the design is susceptible to recruitment and/or identification bias, both types of selection bias. I refer to this as ‘partial’ because, as well as knowing when identification of eligible participants takes place, when eligible participants are recruited would also need

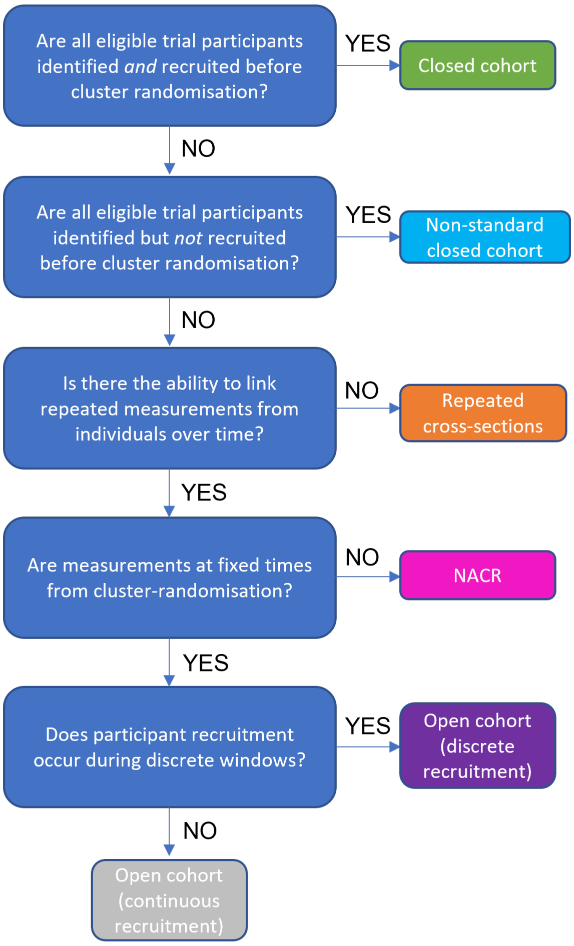

Are all eligible trial participants identified *and* recruited before cluster randomisation?
YES
Closed cohort
NO
Are all eligible trial participants identified but *not* recruited before cluster randomisation?
YES
Non-standard closed cohort
NO
Is there the ability to link repeated measurements from individuals over time?
NO
Repeated cross-sections
YES
Are measurements at fixed times from cluster-randomisation?
NO
NACR
YES
Does participant recruitment occur during discrete windows?
YES
Open cohort (discrete recruitment)
NO
Open cohort (continuous recruitment)

Figure 4.1: Flowchart to quickly identify a trial design.

to be known. If participants are recruited before cluster-randomisation (option a), the design has no risk of identification or recruitment bias. However, option a) paired with recruitment post-randomisation eliminates risk of identification bias but not the risk of recruitment bias; even though the list of eligible participants has been compiled, finding out their allocation may cause participants to refuse consent or influence recruiters, resulting in differential recruitment rates over the trial arms [157]. If no eligible trial participants are identified before cluster randomisation (option b), or some are and some are not (option c) , there is also a risk of recruitment bias. Even so, there are ways to reduce or eliminate the risk of recruitment bias when recruitment occurs post-randomisation, such as blinding those performing identification and recruitment, using broad, objective eligibility criteria for individuals or recruiting in settings outside of the clusters [3, 56, 157, 158].

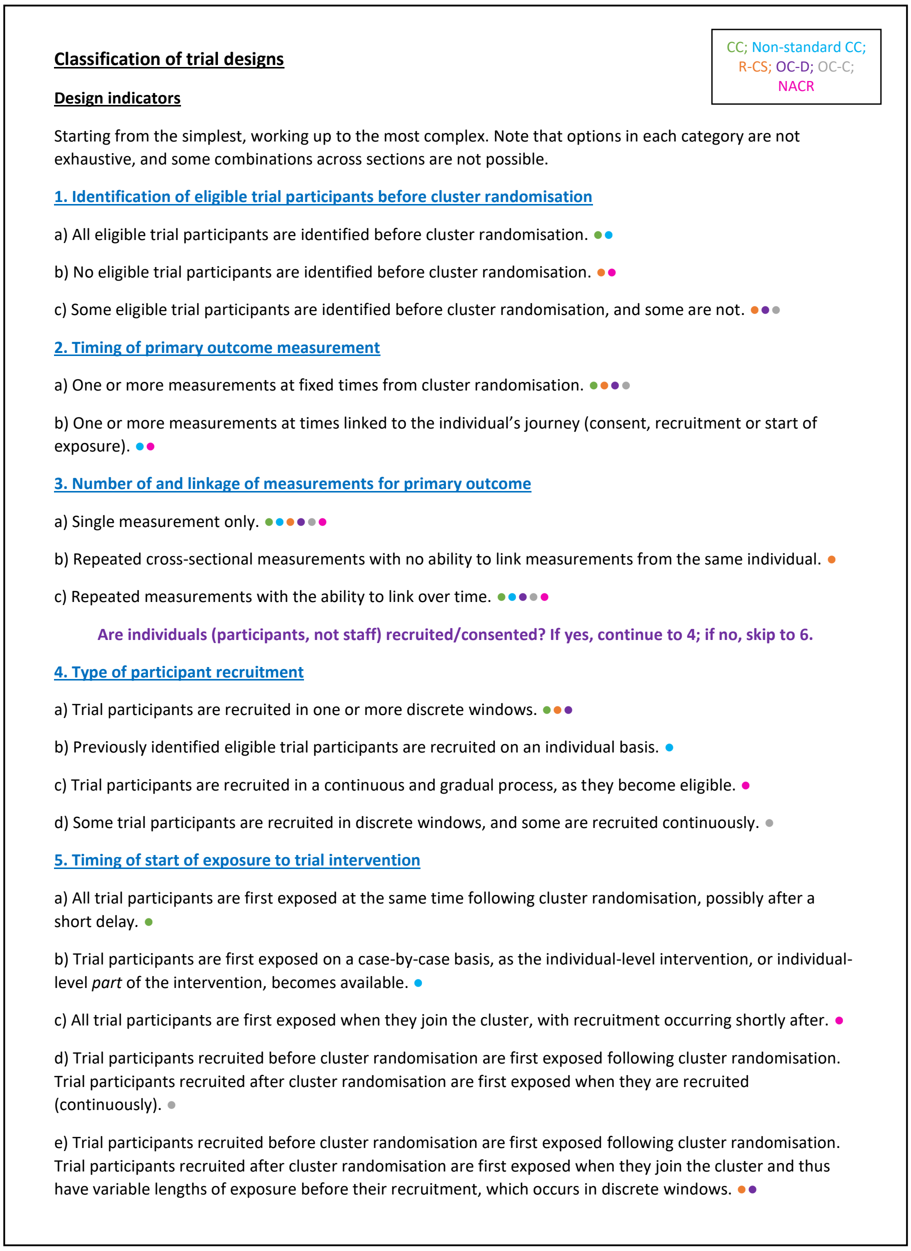
In trials where participants are not recruited or consented, identification bias can still occur if participants are identified after cluster randomisation [159]. In this item, I therefore do not refer to recruitment or consent processes. For example, in a trial where participants are not aware they are part of a research study and are not contacted, identification bias can still occur if the recruiter is not blinded to allocation [157]. If blinding of the identifier is not possible, identification bias could still occur if the intervention improves the identification skills of staff post-randomisation, or participants are attracted to the intervention site because they are seeking a treatment and know it is being offered at a particular site [159].

Item 2: Timing of primary outcome measurement

I suggest that the first important part of the measurement process is when outcomes are measured, as this links to whether cluster or individual level timescales are at play. If the measurement schedule is anchored to the timing of cluster randomisation, a cluster-level timescale is of interest, whereas if the timings are specific to a participant the focus shifts to individual-level timescales. Use of a cluster-level timescale implies a closed, open-cohort or repeated cross-sectional design, whereas individual-level timescales are most commonly used in a NACR design. In a 'non-standard' closed-cohort design, measurements could also be linked to individual times. There may be cases where some trial outcomes are measured using the individual timescale, and others on the cluster-randomisation timescale, so I suggest looking at the primary outcome only for this item. If there are multiple primary outcomes both linked to the same timescale this is not an issue, but if multiple primary outcomes span both timescales this would require looking at other items more clearly to classify the design.

In a 'standard' closed-cohort design where all participants are recruited before cluster-randomisation, usually outcomes will be measured at a fixed time from cluster randomisation as everyone receives the intervention at the same time. In a 'non-standard' closed-cohort design, where participants are recruited after cluster randomisation from a closed list of eligible individuals, participants may be measured at a fixed time after their own recruitment. In summary, measurements based on a fixed time from cluster randomisation do not guarantee a closed-cohort design, because it depends on timing of participant recruitment. However, cross-sectional and open cohort designs will usually have measurements at fixed times from cluster randomisation.

For retrospectively assessed outcomes, measurement items should be used as if outcomes had been assessed prospectively.



CC; Non-standard CC; R-CS; OC-D; OC-C; NACR

Classification of trial designs

Design indicators

Starting from the simplest, working up to the most complex. Note that options in each category are not exhaustive, and some combinations across sections are not possible.

1. Identification of eligible trial participants before cluster randomisation

a) All eligible trial participants are identified before cluster randomisation. ●●

b) No eligible trial participants are identified before cluster randomisation. ●●

c) Some eligible trial participants are identified before cluster randomisation, and some are not. ●●●

2. Timing of primary outcome measurement

a) One or more measurements at fixed times from cluster randomisation. ●●●●

b) One or more measurements at times linked to the individual's journey (consent, recruitment or start of exposure). ●●

3. Number of and linkage of measurements for primary outcome

a) Single measurement only. ●●●●●●

b) Repeated cross-sectional measurements with no ability to link measurements from the same individual. ●

c) Repeated measurements with the ability to link over time. ●●●●●

Are individuals (participants, not staff) recruited/consented? If yes, continue to 4; if no, skip to 6.

4. Type of participant recruitment

a) Trial participants are recruited in one or more discrete windows. ●●●

b) Previously identified eligible trial participants are recruited on an individual basis. ●

c) Trial participants are recruited in a continuous and gradual process, as they become eligible. ●

d) Some trial participants are recruited in discrete windows, and some are recruited continuously. ●

5. Timing of start of exposure to trial intervention

a) All trial participants are first exposed at the same time following cluster randomisation, possibly after a short delay. ●

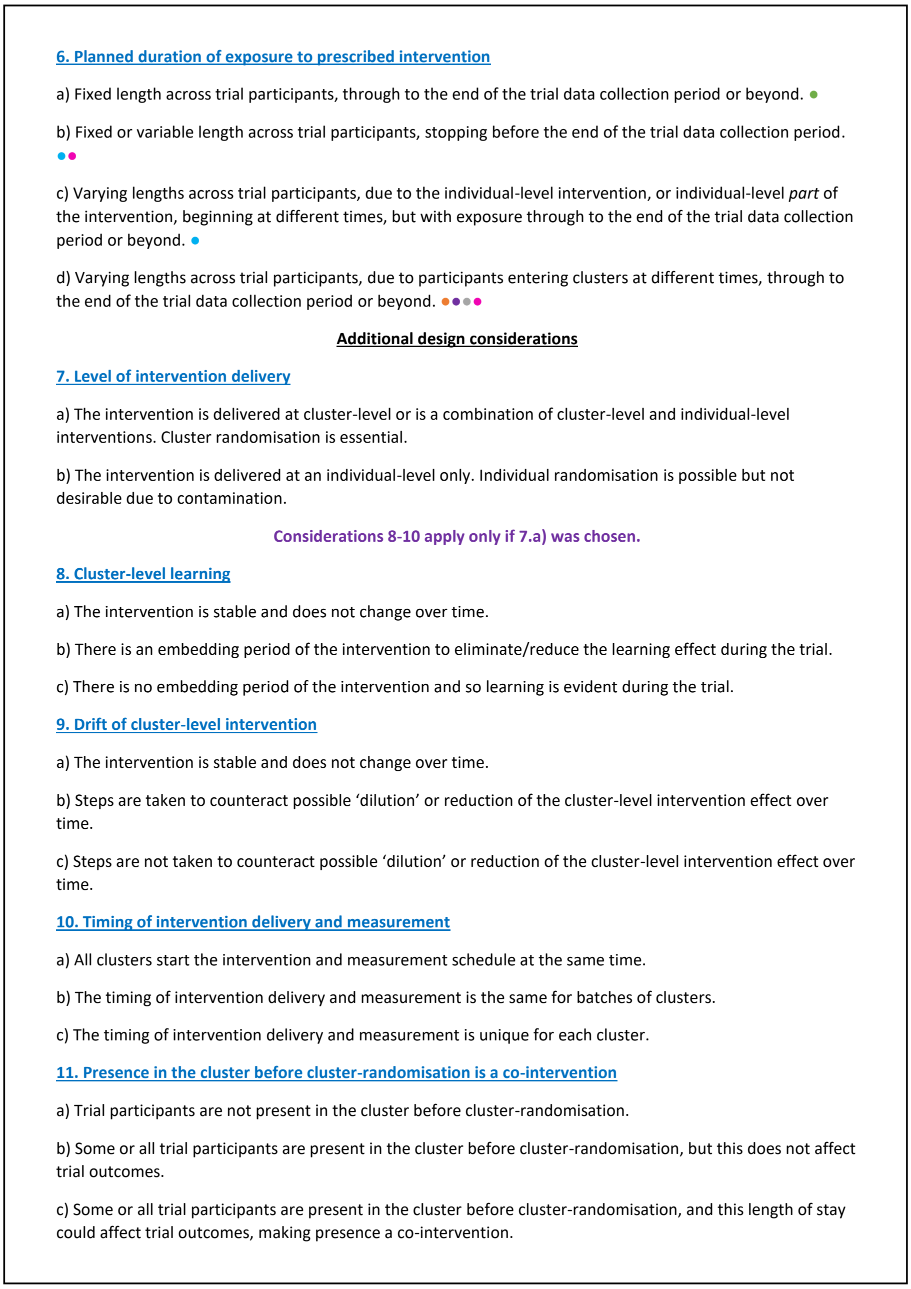
b) Trial participants are first exposed on a case-by-case basis, as the individual-level intervention, or individual-level *part* of the intervention, becomes available. ●

c) All trial participants are first exposed when they join the cluster, with recruitment occurring shortly after. ●

d) Trial participants recruited before cluster randomisation are first exposed following cluster randomisation. Trial participants recruited after cluster randomisation are first exposed when they are recruited (continuously). ●

e) Trial participants recruited before cluster randomisation are first exposed following cluster randomisation. Trial participants recruited after cluster randomisation are first exposed when they join the cluster and thus have variable lengths of exposure before their recruitment, which occurs in discrete windows. ●●

Figure 4.2: Classification system (page 1).



6. Planned duration of exposure to prescribed intervention

a) Fixed length across trial participants, through to the end of the trial data collection period or beyond. ●

b) Fixed or variable length across trial participants, stopping before the end of the trial data collection period. ●●

c) Varying lengths across trial participants, due to the individual-level intervention, or individual-level *part* of the intervention, beginning at different times, but with exposure through to the end of the trial data collection period or beyond. ●

d) Varying lengths across trial participants, due to participants entering clusters at different times, through to the end of the trial data collection period or beyond. ●●●●

Additional design considerations

7. Level of intervention delivery

a) The intervention is delivered at cluster-level or is a combination of cluster-level and individual-level interventions. Cluster randomisation is essential.

b) The intervention is delivered at an individual-level only. Individual randomisation is possible but not desirable due to contamination.

Considerations 8-10 apply only if 7.a) was chosen.

8. Cluster-level learning

a) The intervention is stable and does not change over time.

b) There is an embedding period of the intervention to eliminate/reduce the learning effect during the trial.

c) There is no embedding period of the intervention and so learning is evident during the trial.

9. Drift of cluster-level intervention

a) The intervention is stable and does not change over time.

b) Steps are taken to counteract possible ‘dilution’ or reduction of the cluster-level intervention effect over time.

c) Steps are not taken to counteract possible ‘dilution’ or reduction of the cluster-level intervention effect over time.

10. Timing of intervention delivery and measurement

a) All clusters start the intervention and measurement schedule at the same time.

b) The timing of intervention delivery and measurement is the same for batches of clusters.

c) The timing of intervention delivery and measurement is unique for each cluster.

11. Presence in the cluster before cluster-randomisation is a co-intervention

a) Trial participants are not present in the cluster before cluster-randomisation.

b) Some or all trial participants are present in the cluster before cluster-randomisation, but this does not affect trial outcomes.

c) Some or all trial participants are present in the cluster before cluster-randomisation, and this length of stay could affect trial outcomes, making presence a co-intervention.

Figure 4.3: Classification system (page 2).

Item		Option	Design					
			CC	CC v2	R-CS	OC-D	OC-C	NACR
1. Identification of eligible trial participants before cluster randomisation	a	All eligible trial participants are identified before CR.	●	●				
	b	No eligible trial participants are identified before CR.			●			●
	c	Some eligible trial participants are identified before CR, and some are not.			●	●	●	
2. Timing of primary outcome measurement	a	One or more measurements at fixed times from CR.	●		●	●	●	
	b	One or more measurements at times linked to the individual's journey (consent, recruitment or start of exposure).		●				●
3. Number of and linkage of measurements for primary outcome	a	Single measurement only.	●	●	●	●	●	●
	b	Repeated cross-sectional measurements with no ability to link measurements from the same individual.			●			
	c	Repeated measurements with the ability to link over time.	●	●		●	●	●
4. Type of participant recruitment	a	Trial participants are recruited in one or more discrete windows.	●		●	●		
	b	Previously identified eligible trial participants are recruited on an individual basis.		●				
	c	Trial participants are recruited in a continuous and gradual process, as they become eligible.						●
	d	Some trial participants are recruited in discrete windows, and some are recruited continuously.					●	
5. Timing of start of exposure to trial intervention	a	All trial participants are first exposed at the same time following CR, possibly after a short delay.	●					
	b	Trial participants are first exposed on a case-by-case basis, as the individual-level intervention, or individual-level part of the intervention, becomes available.		●				
	c	All trial participants are first exposed when they join the cluster, with recruitment occurring shortly after.						●
	d	Trial participants recruited before CR are first exposed following CR. Trial participants recruited after CR are first exposed when they are recruited (continuously).					●	
	e	Trial participants recruited before CR are first exposed following CR. Trial participants recruited after CR are first exposed when they join the cluster and thus have variable lengths of exposure before their recruitment, which occurs in discrete windows.			●	●		
6. Planned duration of exposure to prescribed intervention	a	Fixed length across trial participants, through to the end of the trial DC period or beyond.	●					
	b	Fixed or variable length across trial participants, stopping before the end of the trial DC period.		●				●
	c	Varying lengths across trial participants, due to the individual-level intervention, or individual-level part of the intervention, beginning at different times, but with exposure through to the end of the trial DC period or beyond.		●				
	d	Varying lengths across trial participants, due to participants entering clusters at different times, through to the end of the trial DC period or beyond.			●	●	●	●

CR = cluster randomisation, DC = data collection

Figure 4.4: Classification system (page 3).

Item 3: Number of and linkage of measurements for primary outcome

The second important element I propose that relates to measurement is the ability to link repeated measurements from the same individuals over time. Whilst the actual linking of repeated measurements from the same individual is part of the analysis, the trial may collect data in such a way that linking will be possible or not at the analysis stage. In a cross-sectional design, for example, measurements may be taken where individual ID is not available. This item helps to distinguish between the cross-sectional design and the other designs. The phrasing of 3b) and 3c) with "ability to link over time" makes it clear that by choosing a cross-sectional design, linkage will not be possible at the analysis stage, but by choosing one of the other designs that linkage will be possible. However, even if linkage is possible, it may not be done; an open-cohort design could have a cross-sectional analysis if the information is available to link but is not made use of.

Item 4: Type of participant recruitment

By recruitment, I refer here to consent for data collection as is often the case in CRTs, when consent to randomisation occurs at a higher level for the whole cluster [5].

By determining whether recruitment of participants (if applicable) is discrete or continuous, distinction can be made between the two open-cohort designs and the purely continuous NACR design. Whilst continuous recruitment may occur during a fixed window, it is the way that participants become eligible that informs whether the recruitment process is continuous. For example, participants may be recruited over a fixed period but become eligible as they join a cluster. Although this item is unreportable for trials without participant recruitment or consent, this is not to say that individuals' data are not included in such trials. For example, although not consented, individual level outcomes may be accessed using anonymised routine data retrospectively.

Item 5: Timing of start of exposure to trial intervention

The start of exposure refers to the time when trial participants first experience the trial intervention. This item highlights possible bias that could arise with a cluster-level intervention and a discrepancy between the start of an individual's exposure to the intervention and the time of their recruitment, and only applies to trials with participant recruitment or consent. If the intervention is directed at an individual level only, there is no risk of this type of bias. If participants are exposed to a cluster-level intervention for longer periods than is accounted for, this could lead to the intervention effect being estimated too strongly; however, this only applies when 'time exposed' is used in the analysis. In a standard closed-cohort design (option a) where all participants are recruited before cluster-randomisation, this is not an issue. With the non-standard closed-cohort design (option b), if a cluster-level intervention is rolled out once the individual-level intervention is available, following participant recruitment, then again this is not an issue. With continuous recruitment (options c and d), this gap is likely to be negligible, for example if participants are recruited within hours of admission to a ward. However, in an open-cohort or cross-sectional design with discrete recruitment (option e), individuals could potentially be exposed to a cluster-level intervention for a long period of time before being formally recruited and measured, and as such are at higher risk of this type of bias. When the risk of bias is high, increasing the number of recruitment points could be helpful.

Item 6: Planned duration of exposure to prescribed intervention

This item classifies whether all participants are exposed for the same fixed duration or for varying lengths of time. This highlights possible bias that could arise if all participants are assumed to have undergone the same exposure length when in reality the exposure is variable across participants; it is a suggestion that in these cases, some form of adjustment could be considered in the analysis. A distinction is also made between whether the exposure lasts until the end of the data collection period or ends before this time. In cases where the exposure is expected to last through to the end of the trial data collection period, there is an expectation that participants will remain in the cluster and therefore remain exposed. However, for the case of 6.b), this includes cases where there is an expectation that participants will leave the cluster at some point; an example is expectation of discharge from a ward, as opposed to an indefinite stay in a care home.

Due to planning an intention-to-treat analysis, there is interest in the *planned* duration; in other words, assuming there is no withdrawal from treatment. In a hospital CRT, if individuals are exposed to the intervention for 3 months or until early discharge, I assume that 3 months is planned whereas early discharge is not, and so focus on the 3 month time frame. Reference to ‘varying lengths’ of exposure periods means that due to the design, the varying lengths are expected; I am not referring to exposure lengths that vary unintentionally. The two reasons for varying lengths in parts c) and d) are due to either an individual-level part of the intervention being staggered, or participants entering the cluster at different times. The NACR design has different sub-types and so appears both in part b) and part d) (see Section 4.4 for examples).

4.3.4 Additional design considerations

The following design considerations are not indicative of a particular design, but could be potential sources of bias.

Item 7: Level of intervention delivery

I argue that this item emphasises not only the rationale for using a cluster-randomised design but also propose it has implications for the timescales used. If the intervention occurs only at the individual-level, individual randomisation may be possible but undesirable due to contamination. Here, the intervention can be thought of as independent of the cluster as presence in the cluster does not necessarily lead to exposure to the intervention. An example of this could be a one-to-one intervention between a participant and a therapist in a care home. Alternatively, if part of the intervention occurs at the level of the cluster, cluster-randomisation is essential. Cluster-level interventions involve a change to the environment, culture or general practices of the institution, and unlike with individual-level interventions, being present in the cluster results in exposure to the intervention.

For CRTs with individual-level interventions only, a NACR design may be considered. If an intervention component targets the individual, an individual's start of exposure to that component can be used as a starting point, and the individual's start of exposure becomes a timescale of interest. In contrast, if at least part of the intervention is delivered at cluster-level, and the intervention starts following cluster-randomisation, the timescale of the cluster as opposed to the timescale of the individual may be of interest. There are grey areas where interventions have multiple components operating at multiple levels and a parallel-group CRT design is used.

However, this is not to say that NACR designs always consist of an individual-level intervention only. In the same way, closed cohort designs do not necessarily have to include cluster-level interventions. For this reason, this is an additional design consideration rather than a dictator of design.

Item 8: Cluster-level learning

A cluster-level intervention may be stable over time or it may vary, either becoming stronger ('learning') or weaker ('drift'). One way in which an intervention may strengthen over time is via staff experience, whereby staff gradually increase their level of expertise as they practice their skills more. The difference between Item 8.b) and c) is whether or not the design allows for an 'embedding period' of the intervention before participants are exposed to it. An embedding period allows the intervention effect to reach a plateau before outcomes are measured; in contrast, without an embedding period, learning may occur during the trial. Whether or not there is an embedding period has implications relating to the dose of the intervention provided as well as the timing of participants entering. For example, a participant exposed at the start would be provided with a different dose of the intervention to a participant starting after the embedding period is completed. If an embedding period is implemented, all participants should in theory be provided with the same dose. An example of an intervention effect which is constant over time could be a leaflet placed in a GP surgery, as it is not influenced by other factors such as staff knowledge and/or turnover. For this item, interest lies only in the initial period following cluster-randomisation to determine whether or not there is an embedding period.

Item 9: Drift of cluster-level intervention

A cluster-level intervention may also weaken over time. This could occur if trained staff leave an institution and there is some level of dilution of the intervention effect participants are provided with. Even with no staff turnover, drift can occur if those trained at the start are trained just once or infrequently, and elements of their training are forgotten or become less effective. Lack of supervision of staff and monitoring of their effectiveness can also exacerbate drift. If any of these are expected, there may be subsequent training for new staff or refresher training for existing staff to mitigate drift when there is an expectation of staff turnover.

Item 10: Timing of intervention delivery and measurement

Here I consider how bias could arise from either the intervention beginning or data being collected at different calendar times across clusters. An example is if some clusters start the intervention or data collection after a change in policy has been introduced, and others start before. In the case of 10.a) when these schedules are the same across all clusters, there is no risk of bias. However, if schedules are different for batches of clusters, or even individual clusters, this could require further thought. With a NACR design, even if participant recruitment begins at the same time across all clusters, measurements are taken at varying calendar times across individuals which could cause bias in a similar way to in an IRT.

Item 11: Presence in the cluster before cluster-randomisation is a co-intervention

In settings where participants can be present in the cluster before randomisation, this period of time in the cluster before the trial begins could affect trial outcomes. Presence in the cluster prior to cluster-randomisation could therefore be seen as a co-intervention, which is not a trial intervention but recognised as an intervention in its own right. For example, the amount of time residents are present in a care home before a trial begins could affect their agitation because this may stabilise over time.

4.4 Illustrative examples

In this section, one example CRT is provided per design, taken from those given by trialists in the online survey. Each design example is supplemented with a diagram to highlight some of the classification system items. Purple lines represent a participant's total time in cluster; when presence is possible before CR (all designs except NACR and 'non-standard' CC), there is a possibility for this to influence outcomes (Item 11).

For the OC design with discrete recruitment and the R-CS design, assuming a cluster-level intervention is in progress when participants join, there may be a delay between first exposure to the intervention and individual recruitment (Item 5); this discrepancy is indicated by the blue arrows. The orange arrows show the time from cluster-randomisation when new joiners enter, which could be seen as time zero on these individuals' unique timescale.

An example of the 'standard' closed cohort design is given in Box 1 and Figure 4.5.

Box 1: Closed cohort - Falls in Care Homes (FinCH)

FinCH was a CRT based in care homes aimed at assessing the effectiveness of a fall prevention programme [160]. All eligible trial participants were identified and recruited before cluster-randomisation (1a, 4a). Participants were assessed for the primary outcome measure, the number of falls, between three and six months post-randomisation (2a, 3c). The Guide to Action Care Homes (GtACH) Intervention Tool was a cluster-level intervention (7a); this included training of care home staff, provision of manuals and a poster displayed in care homes. All participants were exposed to this intervention at approximately the same time following cluster-randomisation (5a) through to the end of the data collection period of the trial (6a). Whilst the provision of manuals and posters in care homes were stable interventions that did not change over time (8a, 9a), the staff training element was variable. The staff training intervention had an embedding period of three months post-randomisation before outcomes were assessed (8b), which reduced the risk of learning during the outcome assessment period, and refresher training was also provided to counteract possible dilution of this intervention effect (9b). No information on the timing of intervention delivery and measurement across clusters was found. All participants were present before cluster-randomisation, so there is a possibility that their length of stay before cluster-randomisation could have affected trial outcomes (11b/c).

A R-CS example is given in Box 2 and Figure 4.6. AFFINITIE is an example of the R-CS design with no overlaps, because different patients were audited at baseline and follow-up.

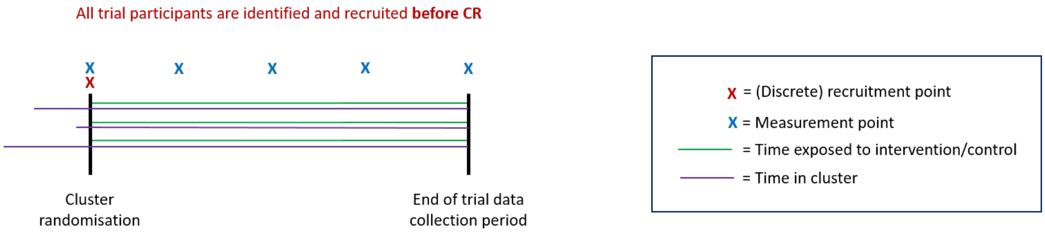

All trial participants are identified and recruited before CR
X
X
X
X
X
X
Cluster randomisation
End of trial data collection period
X = (Discrete) recruitment point
X = Measurement point
= Time exposed to intervention/control
= Time in cluster

Figure 4.5: Diagram of the 'standard' closed cohort design in the FinCH trial.

This occurred because it was very unlikely that patients would have surgery more than once within 12 months.

Box 2: Repeated cross-sectional - Audit and Feedback INterventions to Increase evidence-based Transfusion practIcE (AFFINITIE)

AFFINITIE consisted of two 2x2 factorial CRTs which used routinely collected audit data to assess whether two feedback interventions could reduce the proportion of unnecessary blood transfusions [161]. The first trial comprised surgery patients, and the second haematology patients; I will focus solely on the surgery trial for simplicity. Two interventions were studied: enhanced content in the audit, and enhanced follow-on support. The unit of randomisation was hospital trusts. Participants presenting at baseline and 12 months follow-up were likely to be different, so some eligible trial participants may have been identified before cluster randomisation, but many were not (1c). The primary outcome was assessed both at baseline and 12 months following cluster-randomisation (2a). Because data was anonymised, it was not possible to link any repeated measurements from the same individual, if this occurred (3b). The interventions targeted individuals who compiled the audits and staff involved in transfusion at the hospital, and were therefore cluster-level interventions (7a). Although participant-level data was used in AFFINITIE, participants were not recruited, so Items 4-5 on the classification system are skipped. Participants were exposed to the intervention during their hospital stay, so the exposure across participants was variable (6d). No embedding period of the intervention was implemented so cluster-level learning would have occurred throughout the trial (8c). The enhanced follow-on support intervention in AFFINITIE could be seen as a method for counteracting possible dilution of the cluster-level intervention effect, although there was only one round of audit and feedback (9b/c). The enhanced content intervention was delivered within the same 24 to 48 hour period for all clusters, and the support intervention was available from randomisation across all clusters (10a). Some participants were present before cluster-randomisation, so there is a possibility that their length of stay before cluster-randomisation could affect trial outcomes (11b/c). At the analysis stage, the proportion of unnecessary transfusions at 12 months was adjusted for the proportion of unnecessary transfusions at baseline as a cluster-level covariate. As no individual-level linkage occurred, the analysis was cross-sectional.

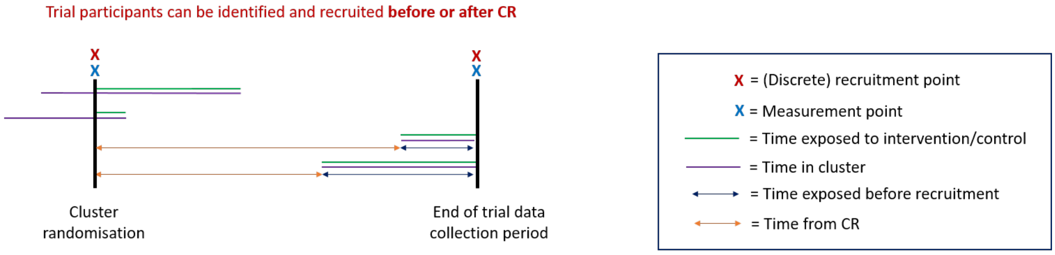

Trial participants can be identified and recruited before or after CR
X
X
X
X
Cluster randomisation
End of trial data collection period
X = (Discrete) recruitment point
X = Measurement point
= Time exposed to intervention/control
= Time in cluster
= Time exposed before recruitment
= Time from CR

Figure 4.6: Diagram of the R-CS design with no overlaps in the AFFINITIE trial.

The SEHER trial is an example of the open cohort design with discrete recruitment (Box 3, Figure 4.7). This example highlights the similarity and subtle difference between the repeated cross-sectional and open-cohort (with discrete recruitment) designs. For each item in the classification system, this design fulfils the open-cohort (with discrete recruitment) design, but also fulfils the cross-sectional design on all but Item 3, which concerns linkage of measurements. I believe that, because this design had the ability to link repeated measurements from individuals, that it should be classed as an open-cohort design. This changes, however, at the analysis stage, where the trialists conduct a repeated cross-sectional analysis, not requiring any linkage. The authors describe this design as repeated cross-sectional, again highlighting more distinction is needed between the design and analysis of CRTs, and that a repeated cross-sectional analysis does not imply that the design was also cross-sectional. In contrast, AFFINITIE had a repeated cross-sectional design which is described as such because there is an inability to link measurements from individuals.

The SEHER trial does not match Figure 4.7 exactly as SEHER only had two measurement and recruitment points. The extra measurement and recruitment points have been included to show the possibilities for this design. With only two recruitment points in the SEHER trial, the time exposed before recruitment represented by blue arrows would be considerably larger.

The POD trial is an example of the NACR design (Box 4, Figure 4.8). Two other examples have also been provided in diagram form only for the NACR design to illustrate its different sub-types. POD and PEARL both have variable exposure periods, but in POD discharge from the ward is expected and so the exposure periods are not expected to endure until the end of the trial (6b). In PEARL, participants are residents in care homes and so are not expected to leave the cluster with the exposure therefore lasting until the end of the trial (6d). MINT is different again in that the exposure period is fixed, but as it is based in an emergency department, discharge is again expected and so the exposure stops before the end of the trial for all participants (6b). MINT also differs from POD and PEARL as it has a particularly short exposure period, as identified by Copas as the CRSE design for SW-CRTs [32]. These designs differ only by Item 6.

The final example of PROSPER (Box 5, Figure 4.11) could be seen as having a 'non-standard' closed cohort design, because use of the classification system gives a mix of closed cohort and NACR components. Although the 'standard' closed cohort design requires participants to be both identified and recruited before cluster-randomisation, in

Box 3: Open cohort with discrete recruitment - Strengthening Evidence base on scHool-based intErventions for pRomoting adolescent health (SEHER)

SEHER was a secondary school CRT which aimed to assess the effectiveness of an intervention to improve school climate and health outcomes [162]. Grade 9 students present on the day of baseline and final follow-up assessments (or both) were eligible to participate, so some trial participants were identified before cluster randomisation and some not (1c). The primary outcome of school climate was assessed at the individual-level before and 8 months after the intervention, at fixed calendar times, though it could not be determined when cluster-randomisation occurred in relation to the baseline survey (2a). Some students were present for both baseline and endline so trialists had the ability to link repeated measurements (3c) based on their design, but chose not to in the analysis; only cluster-level baselines were adjusted for and random effects for students were not included, even for the analysis model including only 'closed-cohort' students. The intervention was multicomponent and consisted of both cluster and individual-level interventions (7a); in fact, it was delivered in three levels to the whole school, individuals and groups of individuals. Students were consented into the study at baseline and endline in two discrete windows (4a). Students present in school at baseline were exposed to the intervention following cluster-randomisation, whereas those enrolling in the school later were exposed when they joined and were not then recruited until 8 months later (5e). This means there would have been variable exposure lengths across students (6d). A pilot test of the intervention was carried out in schools involved in the trial before the main trial, which ensured that the intervention was embedded to some degree before the main trial began (8b). Training and supervision of the counsellors and teachers heavily involved in the intervention was provided throughout the intervention period, to counteract possible drift of the intervention effect (9b). Information on the timing of intervention delivery and measurement across clusters was not found. Some students were present before cluster-randomisation, so there is a possibility that their length of stay before cluster-randomisation could affect trial outcomes (11b/c).

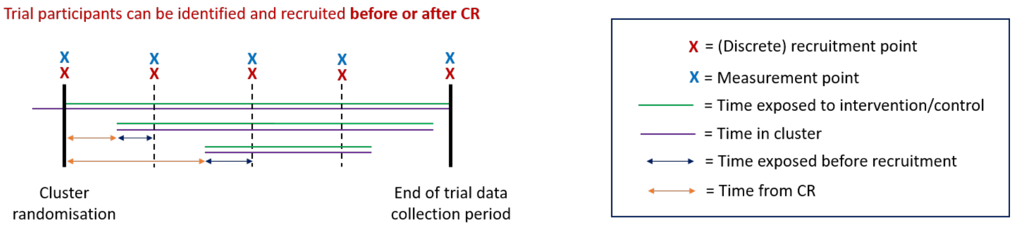

Trial participants can be identified and recruited **before or after CR**
X
X
Cluster randomisation
End of trial data collection period
X = (Discrete) recruitment point
X = Measurement point
= Time exposed to intervention/control
= Time in cluster
= Time exposed before recruitment
= Time from CR

Figure 4.7: Diagram of the open cohort design with discrete recruitment.

some cases this is not feasible, and so instead a list of eligible participants are identified before CR which is then used for recruitment, with no further recruitment after CR from individuals outside of this list. This pragmatic approach still avoids the risk of selection bias as in the standard closed cohort design, but has extra flexibility in patient recruitment timings.

This example also portrays how the decision to use individual- or cluster-randomisation is closely linked to the design. Initially, cluster randomisation was chosen to reduce contamination and because the intervention was thought to be cluster-level in its delivery. However, following the feasibility trial, what was previously thought to be a multidisciplinary

Box 4: New admission continuous recruitment - Prevention Of Delirium (POD)

POD was a feasibility CRT set in elderly care medicine and orthopaedic trauma wards [163]. Since participants became eligible when they were admitted to a ward, no eligible trial participants were identified before cluster randomisation (1b). All assessments were made at times post-admission, unique to each individual (2b). A primary objective was to obtain an estimate of the intended primary outcome in the definitive trial, incidence of new-onset delirium, within 10 days of a participant being admitted to hospital, so repeated measurements were obtained with the ability to link over time (3c). The POD programme was a ward-based intervention involving staff and volunteers, to change practices and the environment experienced by participants; a cluster-level intervention (7a). Trial participants were recruited in a continuous fashion, as they became eligible (4c). As a result of the cluster-level intervention, they were first exposed when they joined the ward, and if they met other eligibility criteria were asked for consent to data collection within 48 hours of admission (5c). With a maximum of 48 hours between exposure and recruitment, risk of bias from unaccounted exposure to the cluster is low. Participants were then exposed until they were discharged (6b). Post-randomisation, wards assigned to the intervention underwent a 6-month embedding period before participants were recruited (8b). No information could be found about refresher training for staff in the ward or training for new staff, but this could be because the intervention period was relatively short at 6 months. All wards began participant recruitment at the same time (10a), however, as measurements are based on individual timescales in a NACR design there would still have been a chance that measurements at varying calendar times could introduce bias. Participants were not present in the cluster before cluster-randomisation so presence as a co-intervention was not an issue (11a).

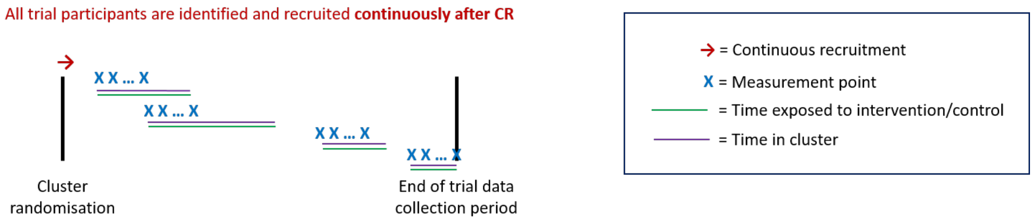

All trial participants are identified and recruited continuously after CR
X X … X
X X … X
X X … X
X X … X
Cluster randomisation
End of trial data collection period
→ = Continuous recruitment
X = Measurement point
= Time exposed to intervention/control
= Time in cluster

Figure 4.8: Diagram of one variation of the NACR design as in the POD trial. The X's with ellipses in between represent the 10 daily measurements.

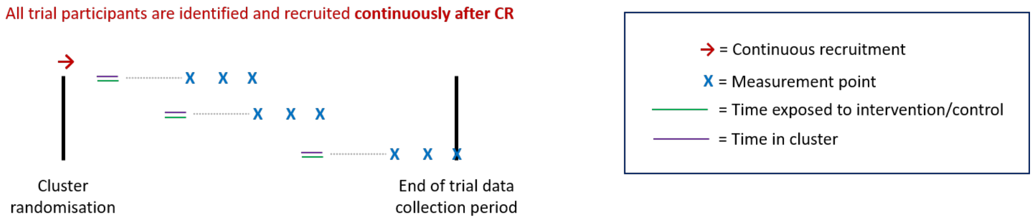

All trial participants are identified and recruited continuously after CR
X X X
X X X
X X X
Cluster randomisation
End of trial data collection period
→ = Continuous recruitment
X = Measurement point
= Time exposed to intervention/control
= Time in cluster

Figure 4.9: Diagram of one variation of the NACR design as in the MINT trial. The grey dots represent the measurements taking place long after the exposure period.

intervention at the cluster-level was actually more of an individual-focused intervention focused on social rather than medical care, with almost zero input from general practice staff. Given this, use of individual randomisation in the main trial would have a low risk of

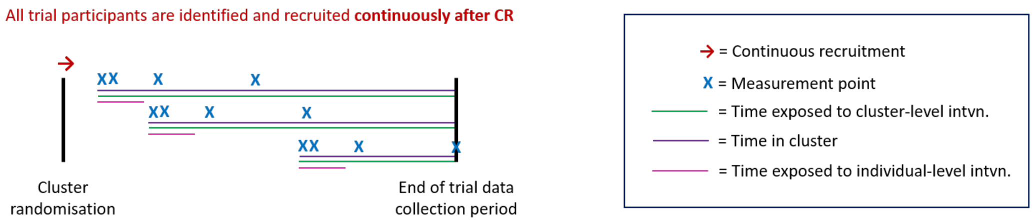

All trial participants are identified and recruited continuously after CR
XX X X
XX X X
XX X X
Cluster randomisation
End of trial data collection period
= Continuous recruitment
X = Measurement point
= Time exposed to cluster-level intvn.
= Time in cluster
= Time exposed to individual-level intvn.

Figure 4.10: Diagram of one variation of the NACR design as in the PEARL trial.

contamination and so the need for a cluster-randomised design was significantly reduced, with the disadvantages associated with randomising clusters [164, 165] not necessarily being outweighed by a minor reduction in contamination. As a result, an IRT was chosen over a CRT for the definitive main trial; this makes sense because the feasibility trial had some NACR design components, such as the use of individual timescales for measurement schedules.

Similarities can be seen between the 'non-standard' CC and NACR designs by studying the diagrams. The main difference is that continuous recruitment in NACR is inherently different to the recruitment process in the former. In NACR, trial participants are unknown before CR and the participants become eligible as they enter the cluster; in contrast, the CC trial participants are identified before CR. Their recruitment points are not fixed in advance, rather they are staggered depending on logistical or implementation constraints.

Box 5: 'Non-standard' closed cohort - PeRsOnaliSed Care Planning for OldER People (PROSPER)

The PROSPER feasibility CRT, where the clusters were general practices, aimed to assess whether a personalised care planning intervention could improve frailty in older people [166]. All eligible trial participants were identified before cluster randomisation (1a). SF-12 and SF-36 Physical Component Summary (PCS), and Mental Component Summary (MCS) were assessed at baseline and 12 months post-registration (2b, 3c). The team-based intervention was to be delivered at the cluster-level of the general practice (7a). Participants who had previously been identified before cluster randomisation were recruited after cluster randomisation on an individual basis for logistical reasons (4b). Participants were exposed on an individual basis after they had been recruited (5b). The intervention was a fixed duration of 12 weeks (6b).

None of the trials provided had an open cohort design with continuous recruitment, but a diagram is provided in Figure 4.12 to depict this design.

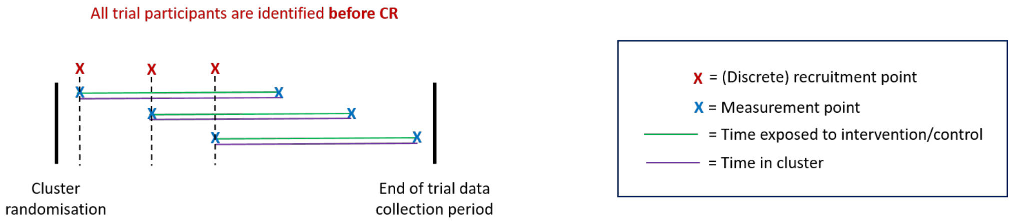

All trial participants are identified before CR
X
X
X
X
X
X
X
X
X
Cluster randomisation
End of trial data collection period
X = (Discrete) recruitment point
X = Measurement point
= Time exposed to intervention/control
= Time in cluster

Figure 4.11: Diagram of the ‘non-standard’ closed cohort design as in the PROSPER trial.

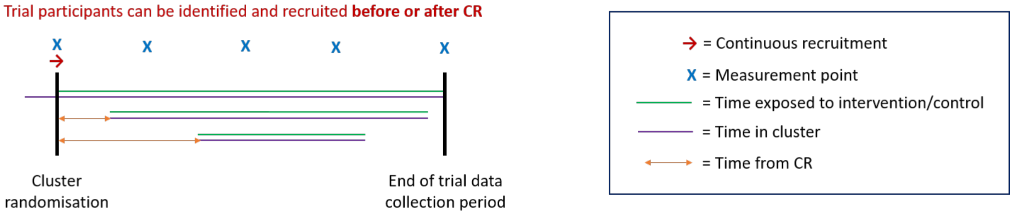

Trial participants can be identified and recruited before or after CR
X
X
X
X
X
Cluster randomisation
End of trial data collection period
= Continuous recruitment
X = Measurement point
= Time exposed to intervention/control
= Time in cluster
= Time from CR

Figure 4.12: Diagram of the open cohort design with continuous recruitment.

4.5 User engagement workshop and survey results

From the user engagement workshop and survey, 48 unique examples of trials were provided by respondents as potentially having an OC design. Two of these could not be found, bringing the total to 46. Of these, 15 (33%) were based in care homes, 8 (17%) in primary care, 8 (17%) in secondary care, 6 (13%) in education, 2 (4%) in workplaces and 2 (4%) in communities. The remaining settings included sheltered accommodation, homeless centres, emergency departments, and combinations of primary and secondary care.

Just 7 of the 46 (15%) given trials had possible open cohort designs. Of those excluded from having a possible open cohort design, the most common reasons were that it was in fact a closed cohort design (46%), a NACR design (15%) or a R-CS design (13%). Further exclusions included trials that were not parallel-group (eg. stepped-wedge, cluster-crossover), individually randomised trials, non-randomised trials and one case where the unit of randomisation was rheumatologists as opposed to an institution (Table 4.1).

Of the 7 trials with possible open cohort designs, 4 were in a care home setting and 3 in education (Table 4.2). Three of the education trials reported their design as cross-sectional, but I believe the designs to be open cohort and the analyses to be cross-sectional, because there is the ability to link measurements from the same individual over time [162, 199, 200].

Three of the care home trials were also similar to each other in design. Both PiTSTOP [201] and DCM-EPIC [202] began as closed-cohort designs but changed the design to

First author	**Year**	**Setting**	**Reason for exclusion**
Davison [167]	2021	Care homes	NACR design
Graham [168]	2020	Care homes	CC design
Halek [169]	2020	Care homes	SW-CRT
Hewitt [170]	2014	Care homes	CC design
Lichtwarck [171]	2016	Care homes	CC design
Logan [160]	2019	Care homes	CC design
Pasay [172]	2019	Care homes	R-CS design
Rokstad [173]	2013	Care homes	CC design
Sackley [71]	2015	Care homes	CC design
Sampson [174]	2020	Care homes	CC design
Sluggett [175]	2020	Care homes	CC design
Campbell [176]	2013	Primary care	NACR design
Harris [177]	2017	Primary care	CC design
Harrison [178]	2019	Primary care	Non-randomised
Heaven [179]	2020	Primary care	CC design
Moore [180]	2003	Primary care	CC design
Mullis [181]	2019	Primary care	CC design
Willis [182]	2020	Primary care	R-CS design
Duhig [183]	2019	Hospitals	SW-CRT
Harding [184]	2020	Hospitals	SW-CRT
Hartley [161]	2017	Hospitals	R-CS design
Hopkins [185]	2020	Hospitals	R-CS design
Lacherade	NA[1]	Hospitals	CRXO-CRT[2]
Pourrat [186]	2014	Hospitals	CRXO-CRT
Singh [187]	2017	Secondary care	NACR design
Young [163]	2020	Hospitals	NACR design
Clemes [188]	2020	Schools	CC design
Evans [189]	2013	Schools	CC design
Pechey [190]	2021	Universities	Non-randomised
Clemes [191]	2019	Workplaces	CC design
Vasiljevic [192]	2018	Worksites	SW-CRT
Fairhall [193]	2015	Community	Individually randomised
Neuman [194]	2018	Communities	R-CS design
Cox [195]	2021	Homeless centres	CC design
Gale[3]	NA	Secondary & primary care	NACR design
Lamb [196]	2013	Emergency departments	NACR design
Martineau [197]	2015	Sheltered accommodation	CC design
Rodríguez-Mañas [198]	2014	Unclear[4]	CC design

Table 4.1: Summary of the 39 trials provided by trialists in the user engagement workshop and survey that were excluded as having an open cohort design, organised by setting. [1]ClinicalTrials.gov Identifier given by respondent but no relevant publications found. [2]ClinicalTrials.gov entry identifies this trial as having a cross-over design. [3]No published protocol or results were available for this trial, so the protocol was obtained from authors; the Chief Investigator rather than the first author is given. [4]Setting could not be identified and no response from the author was received.

First author	Year	Trial name	Setting	Design reported as
Siddiqi [201]	2016	PiTSTOP	Care homes	Unclear
Surr [202]	2020	DCM-EPIC	Care homes	Open cohort
Tadrous [203]	2020	APDP	Care homes	Unclear
Underwood [70]	2013	OPERA	Care homes	Unclear
Bonell [204]	2019	INCLUSIVE	Schools	Cross-sectional
Kipping [199]	2016	NAP SACC	Nurseries	Cross-sectional
Shinde [162]	2018	SEHER	Schools	Cross-sectional

Table 4.2: Summary of the 7 trials provided by trialists in the user engagement workshop and survey with a possible open cohort design, organised by setting.

include a further recruitment point to overcome attrition; these designs then resemble an open cohort design, with the minimum number of 2 recruitment points. The OPERA trial [70] also had a closed cohort design with an additional cross-sectional recruitment point at final follow-up, but cohort and cross-sectional analyses were reported rather than designs. For the APDP trial, the protocol [205] implied a R-CS design (and described the data as ‘repeated cross-sectional’), but the results [203] reported using a repeated measures analysis with random effects for individuals in mixed effects models, alluding to the linkage of individuals’ outcomes over time.

Clearly there still exists confusion about parallel-group CRT designs and how to classify them. Trialists working in CRTs were only able to correctly provide 7 out of 46 trials that were actually open cohort, and of the 7 that were, only one of these reported the design as open cohort; this was the DCM-EPIC trial, the motivating example of this thesis.

4.6 Discussion

The proposed classification system is primarily intended as a tool for trialists to more clearly specify the design of parallel-group or factorial CRTs. This was achieved by disentangling the complex processes that occur in CRTs, identifying the key features of each design and, where designs are similar, describing what sets them apart. Results of the scoping review (Chapter 2) and user engagement online survey both identified a lack of consensus and awareness of the components of different parallel-group CRT designs, and this work provides the first step towards a solution to this problem. The classification system can also be used to assess possible biases that can occur in CRTs with cluster-level interventions.

The importance of timescales and their connections to components of trial design are also major results to come out of this chapter. Timescales are represented by different colour lines and arrows on the diagrams. If recruitment is continuous, usually this means measurement timescales are anchored to that of individuals rather than CR. The level of intervention delivery also affects the choice of timescales. If interventions are delivered at an individual-level only, the timescale of the cluster becomes less important, whereas for cluster-level interventions the time since CR is also of interest in addition to the timescale of the individual. The concept of these two timescales in the open cohort design with discrete recruitment is continued into Chapters 5 and 6 where a data generating model for open cohort data uses both timescales, for the first time. The timing of the trial end also depends on the design. The CC, R-CS, and OC designs include measurement timings dependent on CR, not specific to individuals, whereas the NACR and 'non-standard' CC designs depend on timings of individuals and so the trial end is defined by the last individual.

4.6.1 Comparison to existing literature

This chapter adds to the work of Copas *et al.* [32] who provided a framework for classifying SW-CRTs by assessing three components: the measurement process, the timing of the start of exposure and the duration of exposure. I have extended this by also including the identification and recruitment processes (if applicable), and consideration of the level of intervention delivery which I believe influences design choice.

The continuous recruitment design presented here is not categorised by the length of the exposure period as in Copas' work; instead, a broad NACR design is presented with different types of exposure period possible. Given that CRTs for complex interventions often include multicomponent interventions operating at different levels, it was not always simple to categorise exposure periods distinctly as fixed or variable. For example, the individual level part may be fixed and the cluster level part variable (as in PEARL), or vice versa. Another reason for avoiding naming a design 'continuous recruitment, fixed exposure' is that the definition of a fixed exposure period is not always entirely clear. In the first stage of the MINT trial, patients attending emergency departments were given an 'active management' consultation in the intervention arm; this could be classed as fixed in that it was a set procedure, but the actual duration of these consultations could have been different. Fixed exposure periods that are unambiguous may be more prevalent in CRT settings with larger cluster sizes, such as global health, where it would not be feasible to offer tailoring of an intervention to individuals, whereas variable exposure periods were common amongst the settings of the trials from the survey. Short and long exposure periods have also been grouped in this NACR terminology. In SW-CRTs the implications are more important and relate to exposure to one or both treatment conditions, but for parallel-group trials the main implication is that when exposure periods are very short and measurements are collected outside of a participant's exposure period, data collection could take place outside of the cluster, and could have to rely on other methods of data collection such as by post, for example.

Many of the biases associated with individually- and cluster-randomised trials have already been discussed in detail and frameworks exist for identifying them, for example identification and recruitment bias in the Risk of Bias tool [206]. However, there are other potential sources of bias presented in this work which, despite their prevalence, have not received much attention in the literature. This includes the time exposed to a cluster-level intervention before participant recruitment, variable exposure periods across participants, timing of intervention delivery and measurement and presence as a co-intervention. Awareness of all of the above biases is crucial as they can reduce the quality of evidence gained from CRTs, rendering them less robust than other types of trials when used in systematic reviews [158, 165].

Copas *et al.* [32] consider bias mainly in the form of carry-over effects and the choice of timing of data collection relative to the roll-out period, both of which are not applicable to parallel-group and factorial CRTs. Regarding cluster-level learning, embedding periods have also been called "implementation periods" in the SW-CRT literature [207] where the cross-over from control to intervention conditions can occur during a trial. In this context, Hooper and Copas recommend that no recruitment takes place during an implementation period so that all participants in the intervention arm receive a "full-strength" intervention. The authors do not mention bias but see the intervention participants experiencing a sub-optimal intervention dose as "contaminated". Bias caused by intervention drift is often discussed in relation to fidelity [208].

Frameworks also already exist to promote better reporting of CRTs, as previously discussed, and this work complements them. Given that Item 11 was reported poorly, this could be considered for future revisions of the CONSORT guidance [152]. Currently the closest item of the CONSORT extension to Item 11 ('timing of intervention delivery and measurement') is an acknowledgement that "clusters can be randomised all at once (or in batches) rather than one at a time", and this is provided in the elaboration on the checklist only, not the checklist itself. I feel that both elements provide much greater transparency of trial processes, and given that CRTs are often complex, more information is always beneficial. The TIDieR checklist [209], a tool to aid better reporting of interventions in health research, already encourages trialists to report in more detail approaches taken to mitigate intervention drift, such as refresher training or monitoring. These are captured in their items 4 and 8, under "enabling or support activities" and "when and how much" respectively.

This work also adds to that of Caille *et al.* who developed the Timeline cluster [3], a graphical tool which can be used by trialists to improve reporting and assess risk of different biases associated with CRTs. The Timeline cluster emphasises the order of processes, specifically identification, recruitment, randomisation and assessment, and the levels at which they occur. Whilst my diagrams also include detail on the same four elements, the levels are not differentiated on the diagrams but they are in the system itself. However, my diagrams reference specific designs and include other biases not covered in the Timeline cluster, as well as differentiating between three different timescales. Given that details on blinding are provided in Caille's tool but not here, use of both tools is recommended for full transparency.

Caille *et al.* refer to one of the designs discussed here, the repeated cross-sectional design, and how the Timeline cluster tool could be edited to represent this design with recruitment post-randomisation, by using a repeating loop of identification, recruitment and assessment processes. Figure 4.13 shows a generic example of how the Timeline cluster could be used to complement this classification system when reporting (repeated) cross-sectional and open cohort (with discrete recruitment) designs which have post-randomisation recruitment. In the particular example of Figure 4.13, cluster identification and recruitment and individual identification and recruitment as well as a baseline assessment occurs before cluster-randomisation for some participants, but at 6 and 12 months post-randomisation there are discrete recruitment/measurement windows where participants are again identified, recruited and measured. The black boxes before cluster randomisation tell us that these processes are fully blinded, and therefore pose no risk of selection bias, whereas those repeated at 6 and 12 months potentially have no blinding and are at high risk (dependent on other factors). For the OC (with continuous recruitment) and NACR designs where there is also recruitment post-randomisation but the recruitment is a continuous process, "at 6 and 12 months" could be replaced with "continuously".

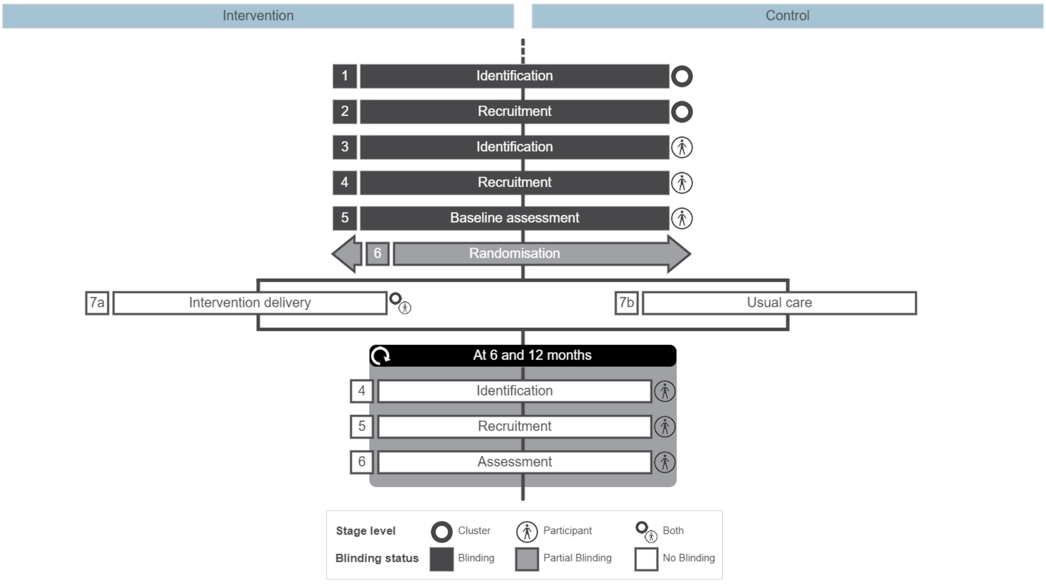


Figure 4.13: Example of how the Timeline cluster tool [3] could be used to portray R-CS and OC designs with recruitment post-randomisation. This figure was produced using a template provided by the author.

4.6.2 Strengths, limitations and directions for future research

The classification system has been reviewed in both a user engagement workshop and an online survey completed by experts in the field of CRTs, which is a strength of this work. It has also been tested on real CRTs, making it directly applicable and useful rather than purely theoretical. The terminology used in this classification system is also accessible and clear for those working in CRTs. Misleading or ambiguous statements were removed following the survey, and known language has been adopted where it already exists. For example, in Caille et al., intervention delivery is also classified as being at cluster-level or participant-level [3]. Current guidance in the CONSORT statement: extension to cluster randomised trials item 5 requires trialists to report “whether interventions pertain to cluster level, individual participant level, or both”. Despite extensive piloting, a limitation is that there may be parallel-group or factorial CRTs which are too unique to fully classify in this system, but this could lead to future development. There may exist further variations on each design, or possibly new designs entirely, but this work provides a solid foundation on which to build.

Although the possibility of an open cohort design with continuous recruitment is suggested as part of the classification system, no examples of this design were found amongst the trials provided by trialists in the workshop and survey. Chapters 5 and 6 therefore focus on the open cohort design with discrete recruitment only, with the continuous recruitment version set aside as an avenue for future research. The continuous recruitment version of the OC design will be returned to briefly in the final conclusions of Chapter 7.

The NACR design is a broad classification in this work for the reasons given in the previous section. With future research into this type of design, it could be that different subtypes warrant different statistical methods or sample size calculations; in this case, further work distinguishing between the different types of NACR could be an area for future research.

A limitation of this chapter is that it does not consider how the designs vary with respect to whether full- or sub-samples are taken. This element was omitted from the classification system for simplicity, but given that in Chapter 3 broad types of CC, R-CS and OC designs were defined, this could still be applied to full or sub-samples for each type. In Chapter 3, the CC, R-CS and OC designs were set out with how they would look for full-samples and sub-samples respectively. An area for future work would then be to investigate how the other sub-designs in this classification system could interact with full- and sub-samples, in particular the OC with continuous recruitment and NACR sub-designs. Both of these have continuous recruitment, which could mean they are more suited to taking full-samples rather than sub-samples.

The work presented here adds greater clarity to the field of CRT design, and provides a framework for previously unnamed designs that are used frequently in health research. As a result, this opens up opportunities for further research into each of the designs and their associated analyses, as well as further discussion of some of the lesser known biases identified herein.

The design diagrams are useful when paired with the classification system as trials classed as having the same design might still look different pictorially. The purple and green lines in the diagrams were originally intended as a way to distinguish between total length of stay in the cluster (purple) and time exposed to the intervention (green). However, due to the different levels of intervention delivery, this could be an over-simplification of the actual processes involved. The PEARL trial diagram gives an example of how the cluster- and individual-levels of delivery could be distinguished with a third coloured line (pink). However, complex interventions are rarely simple in nature, so whilst these diagrams attempt to illustrate the designs more clearly, in some cases the diagrams may need further extension to fully capture unique designs.

4.7 Guidance for improving reporting

The guidance provided in this chapter adds to the guidance already provided in the previous chapter.

Firstly, following this classification system, trialists are encouraged to use known terminology for trial design type where this exists such as that provided in this chapter, such as "open cohort with discrete recruitment". When reporting measurement schedules, it is recommended that instead of stating that measurements are taken 'at 6 months', that the timescale in use is clearly referenced; for example, "12 months post-randomisation" [182], "10 days post-recruitment" [166].

Trialists are also encouraged to report in detail any strategies used to counteract dilution of intervention effects, and when embedding periods have been used; adoption of a common language will help this in future research, as they may also be called piloting or implementation periods. The final component, timing of intervention delivery and measurement, was not reported in any of the examples. Timing of cluster recruitment is also poorly reported, so the former is likely to be even more so, especially with limited space in publications to describe trial procedures. More information should be provided on *all* timings to provide a clearer picture for readers; again this could be done with a timeline or table as previously discussed in Chapter 2. For a clear depiction of trial design, the diagrams presented in this chapter help to quickly communicate the many aspects of trial design, including measurement and recruitment schedules, as well as differentiating between the differing timescales for the cluster and the individual.

Chapter 5

Comparison of designs for parallel-group CRTs: a simulation study

5.1 Introduction

In Chapter 1 a distinction was made between the population, design and analysis of a parallel-group, cluster-randomised trial. A closed population loses members only to death, whereas an open population's members fluctuate over time. The two existing designs for parallel-group CRTs were also introduced; the closed cohort (CC) and repeated cross-sectional (R-CS) designs. Briefly, a CC design recruits and measures individuals at baseline, and follows only these individuals over time, with no further recruitment after cluster-randomisation (CR). The R-CS design allows recruitment post-randomisation and samples one or more cross-sections of individuals at different time points, where the individuals sampled at each time point may either be completely independent or have some overlap. Each design has its own corresponding estimand(s) and potential inference to be made. The CC design allows assessment of individual change over time, whereas the R-CS design due to its cross-sectional nature can only provide cluster-level inference at specific time points. As described in Chapter 1, the R-CS design I refer to in this section permits overlaps of individuals at consecutive time points but does not link outcomes across time.

The focus of this thesis is on the novel open cohort (OC) design, a hybrid between the existing designs, which allows for recruitment post-randomisation and follows individuals for as long as they remain in the cluster. For parallel-group CRTs conducted in open populations, the existing CC design is not ideal due to the drop-out of individuals, and the R-CS design is unable to provide inference on individual change over time; the OC design could bridge this gap as a potential solution. The CC design, and the R-CS design in some situations, are subsets of the OC design, so the OC design can be used for either OC and CC estimands or all three of OC, CC and R-CS estimands which could be appealing to trialists. In this simulation study I aim to define estimands for the OC design and compare the existing designs to the OC design with discrete recruitment and measurement. As no examples of the OC design with continuous recruitment were identified in Chapter 4, this design is not considered but will be returned to in Chapter 7.

As previously stated, in this thesis an underlying open population is assumed. I will therefore simulate data from an open population, where individuals move in and out of clusters over time. I will evaluate the bias and precision under a common analysis approach with regard to several estimands under each of the three designs, whilst also varying other study parameters and complications which may exist in practice. As the designs will be varied, the analysis method must be fixed; three analysis models will be used. A simulation study has been chosen over an analytic comparison or the use of real data, because this allows investigation of a wide range of scenarios, and the underlying population can be generated as an open population.

Key themes of timescales and estimands will be introduced subsequently, as well as an overview of the structure of the simulation study. This section then concludes with the aims and research questions for this chapter.

5.1.1 Timescales

In an individually randomised trial (IRT), say for a drug, participants usually have their own individual timescale, where time zero is the time of (individual) randomisation. In CRTs, timescales exist that pertain to both the cluster and the individual; which of these timescales is deemed as the primary timescale depends on the design as discussed in Chapter 4. For example, when cluster-level interventions are investigated, the cluster timescale may be of primary interest, but in NACR designs the individual timescale is likely to be more important.

In this thesis I propose that both of these timescales need to be taken into account for an open cohort design. The individual timescale represents the amount of time an individual has been exposed to the intervention/control condition, and thus begins either at the time of cluster-randomisation if they are members of the original cohort, or when they join the cluster if they are members of the additional cohort. The cluster timescale begins at cluster-randomisation. By including both timescales I am able to partition the total intervention effect a participant receives into the cluster-level intervention effect, as a result of changes at the cluster-level, and the individual-level intervention effect, based on their exposure time. For the CC design the two timescales are equivalent so this design only uses time from cluster-randomisation. For the R-CS and OC designs, the timescales are distinct and so both are required, except for members of the original cohort.

A third possible timescale is 'length of stay', which commonly describes the total amount of time an individual is present in a cluster, not just the time exposed to the intervention, and so includes any time an individual has been present in the cluster before CR if they are in the original cohort. Time spent in the cluster before CR could affect outcomes; this consideration was Item 11 of the classification system in Chapter 4. There is also a fourth possible timescale of calendar time if clusters were randomised at different times which could also affect timing of intervention delivery and measurement across clusters, as discussed in Item 10 of the classification system. These timescales are not considered as integral as the first two and so are omitted from the present simulation study for simplicity.

The individual and cluster timescales will be assumed in the data generating model, but it is of interest whether or not a benefit is gained by also including both timescales in the analysis model, or whether one is sufficient. The use of just the cluster timescale is the approach currently taken in practice, as two timescales do not appear to have been implemented previously in parallel-group CRTs.

5.1.2 Estimands

Detail on specific estimands to use with an open cohort model is lacking in the literature [36]. The choice of design could be affected by the estimand of interest, so several will be investigated in this study. The introduction of a second timescale also opens up possibilities for new open cohort estimands. Whilst there were lots of estimands to consider, a mix of both mortal and immortal estimands was desired, as well as at least one estimand relating to each design. Mortal estimands are included to contrast against the immortal estimands, to determine whether mortal estimands are ever equal to immortal estimands, and because mortal estimands may favour the R-CS design.

5.1.3 Base case and complications

I will examine a variety of scenarios in this simulation study, made up of study parameters and complications. Study parameters are properties of the design that can be chosen by trialists, whereas complications are issues relating to the intervention or the setting which cannot be controlled, and which could exist on their own or in combination with each other.

Whether full- or sub-samples of clusters are measured is a key distinction often overlooked. Sub-samples are often assumed when cluster sizes are large, but for small cluster sizes as in DCM-EPIC, sampling as much of the cluster as possible is vital. Furthermore, as described in Chapter 3, full- and sub-samples imply different relationships between the OC and R-CS designs, so this was important to include as a study parameter. Similarly, the number of follow-ups and number and size of clusters were investigated as conclusions cannot be assumed to apply equally across variations of these.

Feldman and McKinlay [14] have previously compared the CC and R-CS designs but did not take into account serious barriers that are encountered in reality by trialists; for example, high levels of drop-out, missing data mechanisms that are not MCAR or even MAR, and differing shapes of intervention effect. For this reason a base case scenario will be investigated to look at situations with relatively few problems, and complications scenarios with more complex circumstances. The base case fixes the complications to their least problematic version, then in the complications they are looked at in different combinations with each other, from just one complication 'switched on' up to the most complex situation. The specific case of DCM-EPIC will be given special attention amongst the variations of study parameters and complications.

5.1.4 Study parameters and complications

The following study parameters will be varied in the simulations:

1. The trial design
2. Number of follow-ups
3. The number/size of clusters

4. Full- or sub-samples of the clusters measured
5. Relative value of individual and cluster level intervention effect rates

The complications are:

1. Missing data mechanism
2. Turnover rate
3. Shape of cluster-level intervention effect

5.1.5 Aims and research questions

The overarching aim of this chapter is to explore the conditions under which an OC design with discrete recruitment is superior to the conventional CC and R-CS designs, with respect to several estimands, in terms of bias and precision, when data is from an inherently open population. A base case scenario where no complications are present will be investigated first. The robustness of each design will then be evaluated across a variety of complications.

This chapter will be structured around answering the following research questions (RQ). The first set of research questions concerns a comparison of designs:

RQ1-4: For each estimand, when is the OC design superior to the CC and R-CS designs?

The remainder of the research questions relate to study parameters:

RQ5: How many measurement/recruitment points should be used?

RQ6: When is it beneficial to use two timescales over one?

The rest of this chapter is organised as follows. The methods section firstly describes the study parameters in more detail. Data generation mechanisms, estimands, analysis models and performance measures are then addressed using a variation of the 'ADEMP' structure [4] for reporting simulation studies. The complications are described in Section 5.2.2 as each requires its own data generating mechanism. At the end of the methods section, the research questions are returned to in more detail along with my hypotheses. The results of the simulation study are then presented, structured around the research questions. The discussion begins with comments for each estimand and answers to the remaining research questions. This work is finally discussed in the context of the wider literature, concluding with strengths, limitations and avenues for future work.

5.2 Methods

5.2.1 Study parameters

The study parameters introduced in the previous section are described here in more detail.

5.2.1.1 Trial design and number of follow-ups

The trial design and number of follow-up measurements within each design were varied. Figure 5.1 shows a breakdown of the number of recruitment, baseline and follow-up points for each design. Individuals recruited before cluster randomisation are referred to as the 'original cohort', and those recruited after as the 'additional cohort'. The CC design always has one recruitment point only (at $t = 0$ before cluster-randomisation), one baseline measurement taken on the original cohort only, and follow-up measurements taken on the original cohort only, because no further participants are recruited. I assumed that recruitment and measurement take place at the same times in the R-CS and OC designs, so in these designs the total number of measurement points (baseline plus follow-ups) equals the number of recruitment points. The main difference between the R-CS and OC in comparison to CC is that follow-up measurements can be taken from a combination of both original and additional cohort participants, due to the additional recruitment after cluster-randomisation.

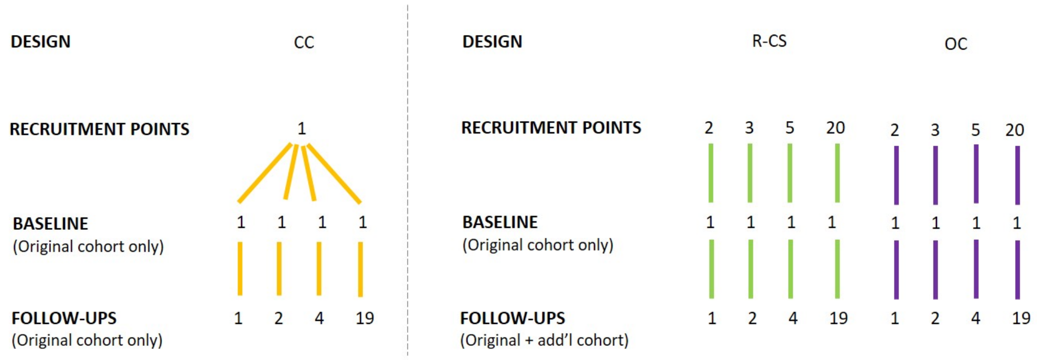


Figure 5.1: The 3 trial designs with different numbers of follow-up measurements and recruitment points.

For the discrete time points, 2, 3 and 5 follow-up points were used because smaller numbers were most common in the scoping review. Originally recruitment and measurement on a weekly basis was also included, but due to the time taken to fit models with such a high number of time points, this was reduced to 20. Daily measurement would have been even closer to ‘continuous’ but with 548 time points, computational issues were inevitable. It would also be extremely difficult in practice to recruit and measure daily in an institution such as a care home, where processes are labour and time intensive, unless outcome measures are measured routinely.

The length of the trial was fixed at 18 months, or equivalently 78 weeks, roughly three times longer than the median duration from the scoping review (26.1 weeks, IQR 20.4-50.0). As several of the trials in the scoping review purposely compromised their length of follow-up, I wanted to use a long enough trial period to see substantial turnover of participants. If the trial period were too short the benefit of an OC design may not be obvious, and one purpose of this study is to enable trialists to be able to conduct trials with longer follow-up. Having said this, the unit of months is arbitrary and is the same as if days or months were chosen instead, whereas the distribution of measurement points within the 78 weeks does matter as well as the turnover. When there were an odd number of time points, more were situated in the first half of the trial rather than the second, because in some of the scenarios there is more change in the earlier stages. Table 5.1 gives the fixed timings of the follow-up measurements.

Number of follow-ups	Spacing
1	78 weeks
2	26, 78 weeks
4	13, 26, 52, 78 weeks
19	3, 6, 9, 13, 16, 19, 23, 26, 29, 32, 35, 38, 41, 44, 52, 58, 65, 71, 78 weeks

Table 5.1: Fixed timings of follow-up measurements in the simulation study.

To ensure fair comparison, the datasets for each design needed to be subsets of the same, larger dataset, akin to samples from an overall population. A dataset with the maximum number of 79 measurement and recruitment points was created under each scenario, referred to as the population dataset, with each design being some subset of the population dataset. This aligns with what would happen in reality; an underlying inherently open cohort of individuals, with each design selecting some subset. None of the designs contained the full weekly information and instead extracted data only from those present at the included measurement time points.

5.2.1.2 The number/size of clusters and full- versus sub-samples

Table 5.2 displays the combinations of cluster sample sizes and cluster population sizes included in the simulation study. As described in Chapter 3, ‘cluster sample size’, m, refers to the number of individuals sampled at each measurement point in each cluster, whereas ‘cluster population size’, M, refers to the full *available* size of the clusters at each measurement point. If the cluster sample size is the same as the cluster population size, the cell reads ‘full’ because all eligible participants are sampled. If a sub-sample of the entire cluster population size was made, the cell reads ‘sub’. Sizes of 15, 50 and 100 were chosen because the settings that would potentially benefit from a new design in the scoping review of Chapter 2 had such sizes; very large cluster sizes for settings such as communities are not included. The first cell in Table 5.2 where the sample size and cluster size are both 15 describes the situation in DCM-EPIC, where the issue of small clusters with a fixed upper limit means that all eligible individuals were sampled rather than a sub-sample.

Sample size calculations in this simulation study were based on the number of individuals required for a CC design, using the same assumptions that were made in the DCM-EPIC trial, with a 1:1 allocation ratio instead of 3:2. Assuming a normally distributed outcome and balance in cluster size, I fixed the desired power at 90%, the two-sided type I error rate at 5% and the overall variance of the outcome (σ^2) at 1. The sample size was then calculated to detect a standardised effect size of 0.4 for the between arm difference in mean scores at final follow-up. In practice, sample size calculations may assume that LTFU will occur and as a result inflate the required number of participants; I did not do this, assuming instead that all observable participants are observed.

In the cluster sample size column, the overall sample size and overall number of clusters were found by specifying the required value of m and working out the design effect due to clustering given by $1 + (m - 1)\rho$ [210]. Table 5.2 assumes the ICC, $\rho = 0.1$. This was then multiplied by the number of participants required for an individually randomised trial, where the power, type I error, overall variance and standardised effect size are as given previously. This number was then rounded up to the next multiple of the required cluster sample size and the number of clusters needed was calculated. Each column of Table 5.2 has differing cluster sample sizes, total participants and clusters, but the power is the same. The full calculation is provided in Appendix Table B.1.

The sample size, m, for the original cohort is the same across all designs, using the calculation described above. In the OC and R-CS designs, the size of m was maintained throughout the trial, so these datasets contain the same overall number of measurements. As the CC design does not allow recruitment after CR, the value of m over the trial period is monotone decreasing, depending on the level of turnover. The specific individuals

<table>
<tr><th rowspan="2">Cluster population size, M</th><th colspan="3">Cluster sample size, m</th></tr>
<tr><th><b>15</b>
660 participants
44 clusters</th><th><b>50</b>
1600 participants
32 clusters</th><th><b>100</b>
3000 participants
30 clusters</th></tr>
<tr><td><b>15</b></td><td>Full</td><td></td><td></td></tr>
<tr><td><b>50</b></td><td>Sub</td><td>Full</td><td></td></tr>
<tr><td><b>100</b></td><td>Sub</td><td>Sub</td><td>Full</td></tr>
</table>

Table 5.2: Combinations of cluster sample sizes and cluster population sizes. Overall numbers of participants and clusters are also given, assuming the ICC = 0.1. Blank cells are impossible combinations. On the 'full' diagonal, the OC designs result in exactly the same datasets as the R-CS design.

who contribute measurements to each dataset will also differ by design. The CC dataset will only ever contain measurements from individuals in the original cohort. The R-CS and OC datasets will contain measurements from at least this many people, and could be much greater depending on the turnover rate of individuals. Although the OC and R-CS datasets have the same number of measurements, the cluster population size M dictates whether they are from the same or different individuals, as will be described shortly.

Both cluster sample size and cluster population size were varied within the simulations because whether a full-sample or a sub-sample is taken is an important part of the design, as described previously, and because this clearly distinguishes between the R-CS and OC designs. For full-samples, drop-outs were directly replaced by new people entering the same bed. Alternatively, for a sub-sample with $M > m$, the CC design uses a CC sampling scheme, the R-CS design uses a R-CS sampling scheme, and the OC design uses the OC (beds) sampling scheme, all of which were previously defined in Chapter 3.

Following the definitions in Chapter 3, it is important to note that in the 'full' diagonal in Table 5.2, the R-CS and OC datasets are exactly the same, whereas for sub-samples the R-CS and OC designs sample different individuals. When the datasets are the same, the only difference would occur at the analysis stage, when one could opt for either an analysis which links repeated measurements from individuals, or a cross-sectional analysis which does not, instead treating the samples as though they are independent samples (which they are not, because overlaps exist). As the intervention is delivered at the cluster-level, for sub-samples there are participants in each cluster who are exposed to the intervention but not recruited.

5.2.1.3 Base case and complications

The base case has 144 scenarios and consists of the full range of study parameters given in Table 5.3, with all the complications set to their lowest or simplest level.

After the base case a full factorial of the complications was investigated (Table 5.4). In order to make this manageable, when investigating the complications some of the study parameters were fixed. I fixed the number of measurement points as 3, the number of clusters as 44, the cluster sample size as 15 and the cluster population size as either 15 or 50. The design, full- versus sub-samples, and the relative value of intervention effects were still varied within the complications. This resulted in 216 scenarios.

5.2.2 Data generating models

Multilevel models were used to simulate repeated measurements from individuals within clusters. The mean model assumed for the data generating model (DGM) includes two timescales instead of the usual one. An overview of these timescales is given first, followed by an explanation of the mean model focusing on the fixed effects, followed by the random effects.

Supplementary code in R for the data generation and other relevant examples are provided at `https://github.com/LEMarsden/Thesis/`.

5.2.2.1 The mean model

Let *cr.time* be time from cluster-randomisation and *ind.time* be the time an individual has been exposed to the intervention/control condition. For the additional cohort, both timescales were included in the mean model for the DGM, as follows:

$$\mu = A\, \mathit{ind.time} + B\, (\mathit{cr.time} \times \mathit{trt}) + C\, (\mathit{ind.time} \times \mathit{trt}). \tag{5.1}$$

A is the fixed coefficient for *ind.time*. The *cr.time* term was assumed to have no effect otherwise initial outcomes for new control participants would have been different for those joining later in time. A linear effect of *ind.time* was assumed with $A = 0.01$, such that the outcome was assumed to increase/worsen by 0.01 for every week of exposure in either arm. In practice, time zero on the *ind.time* timescale is the individual's date of entry to

Factor	Levels
Fixed	
Study parameters	
Allocation ratio	1:1
Total variance of outcome (σ^2)	1
Power	90%
Type I error rate (two-sided)	5%
Total standardised effect size by end of study	0.4
ICC	0.1
Length of intervention period	78 weeks
Timing of follow-ups	See Table 5.1
Complications	
Missing data mechanism	MCAR
Turnover rate	10%
Cluster-level intervention effect rate	Constant
Varied	
Study parameters	
Design	Closed-cohort / (Repeated) cross-sectional / Open-cohort
Number of follow-ups	1 / 2 / 4 / 19
Number of clusters/cluster size	See Table 5.2
Full or sub-sample	See Table 5.2
Relative value of individual and cluster-level intervention effect rates	See Table 5.5

Table 5.3: Details on fixed and varied factors and their levels for the base case. This includes $3 \times 4 \times 6 \times 2 = 144$ scenarios in total.

Factor	Levels
Fixed	
Study parameters	
Allocation ratio	1:1
Total variance of outcome (σ^2)	1
Power	90%
Type I error rate (two-sided)	5%
Total standardised effect size by end of study	0.4
ICC	0.1
Length of intervention period	78 weeks
Timing of follow-ups	See Table 5.1
Number of clusters	44
Cluster sample size	15
Cluster population size	15 or 50
Number of follow-ups	2
Varied	
Study parameters	
Design	Closed-cohort / (Repeated) cross-sectional / Open-cohort
Full or sub-sample	See Table 5.2
Relative value of individual and cluster-level intervention effect rates	See Table 5.5
Study parameters	
Missing data mechanism	MCAR / MAR / MNAR
Turnover rate	10% / 20% / 40%
Cluster-level intervention effect rate	Constant / Non-constant

Table 5.4: Details on fixed and varied factors and their levels for the complications. This includes $3 \times 2 \times 2 \times 3 \times 3 \times 2 = 216$ scenarios in total.

the cluster as opposed to the time of recruitment; the practical implication of this will be returned to in the discussion.

B and C are fixed coefficients for the arm by time interactions, and trt is a dummy variable for treatment arm which is equal to 0 for the control arm and 1 for the intervention arm. The $(cr.time \times trt)$ term is the cluster-level intervention effect rate, and $(ind.time \times trt)$ the individual-level intervention effect rate. The 'total intervention effect rate' is the sum of the cluster-level and individual-level intervention effect rates.

For the original cohort, $cr.time = ind.time$, and substituting $ind.time$ for $cr.time$ reduces the mean model to:

$$\mu = A\, cr.time + D\, (cr.time \times trt), \tag{5.2}$$

with A as above and $D = B + C$, where D represents the total intervention effect rate. Substituting $cr.time$ for $ind.time$ would have resulted in the equivalent expression in $ind.time$.

The $cr.time \times ind.time$ interaction was not included in either model as this would have affected the slopes of the control arm participants who were assumed to follow the same trajectory throughout, nor the $cr.time \times ind.time \times trt$ interaction for simplicity.

It was the aim of this simulation study to be generalisable to other settings and not specific to one trial, therefore the coefficient values were not directly based on estimates from DCM-EPIC. Only the ICC and standardised effect size from DCM-EPIC were used in this chapter to calculate sample sizes. Instead, values were chosen to ensure that trajectories over the trial period were realistic. Aiming for a standardised effect size of 0.4 at 78 weeks, in the base case I set $D = B + C = -0.00514$, as $-0.00514 \times 78 = -0.40092$, then setting $A = 0.01$ meant that the proportion of deterioration that an original cohort participant in the intervention arm avoided was 48.6%. The value of A was fixed throughout; B and C were varied as described in Section 5.2.2.2, as well as when the cluster-level intervention effect rate was non-constant (Section 5.2.2.7).

5.2.2.2 Relative value of individual and cluster-level intervention effect rates

In initial tests, the relative value of the individual and cluster-level intervention effect rates in the DGM affected performance measures. The total intervention effect rate, D, was fixed at -0.00514 per week but the coefficient values were weighted as either 3:1 or 1:3 (Table 5.5). The total intervention effect rate is additive because I assumed no interaction terms in the DGM of model 5.1.

Cluster:individual ratio	Cluster-level intervention effect rate	Individual-level intervention effect rate
3:1	-0.003855	-0.001285
1:3	-0.001285	-0.003855

Table 5.5: Coefficient values used in the DGM for the individual and cluster-level intervention effect rates.

5.2.2.3 Random effects

The outcome in all scenarios including the base case is given by

$$Y_{tijk} = \mu + \mathbf{CT}_{tjk} + \mathbf{E}_{tijk} \tag{5.3}$$

where cluster-period random effects are assumed to follow a multivariate Normal distribution given by

$$\begin{pmatrix} \mathbf{CT}_{0jk} \\ \mathbf{CT}_{1jk} \\ \mathbf{CT}_{2jk} \\ \vdots \\ \mathbf{CT}_{Tjk} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2_{CT} \begin{pmatrix} 1 & & & & \\ \pi & 1 & & & \\ \pi & \pi & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \pi & \pi & \pi & & 1 \end{pmatrix} \right), \tag{5.4}$$

and correlated individual-level random effects are assumed to follow a multivariate Normal distribution given by

$$\begin{pmatrix} \mathbf{E}_{0ijk} \\ \mathbf{E}_{1ijk} \\ \mathbf{E}_{2ijk} \\ \vdots \\ \mathbf{E}_{Tijk} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \sigma^2_{E} \begin{pmatrix} 1 & & & & \\ \tau & 1 & & & \\ \tau & \tau & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \tau & \tau & \tau & & 1 \end{pmatrix} \right), \tag{5.5}$$

for time $t = 0, \ldots, T$, individual $i = 0, \ldots, I$, cluster $j = 0, \ldots, J$ and two treatment arms with $k = 0$ for control and $k = 1$ for intervention. For simplicity, the error term is assumed equal across arms. The mean model is μ as described previously. Both matrices are symmetric so only the lower triangle is given.

Random cluster-period effects in model (5.3) are given by $\mathbf{CT}_{tjk}$, where each cluster has a random coefficient for each period (or time point), $t = 0, \ldots, T$ [211]. Similarly, each individual has a random coefficient for each time point, $\mathbf{E}_{tijk}$, which is a combination of individual-level variability and measurement error. Cluster-period random effects are assumed to use the *cr.time* timescale because this is the timescale the cluster operates on and the cluster is defined from $cr.time = 0$. Only one measurement per individual per time point is assumed here, so there is no individual-period random effect in addition to the residual because these effects would be confounded and lead to an identifiability problem. The *ind.time* timescale is therefore not used for random effects.

Random effects at different levels in the model are assumed mutually independent. Whilst the size of the covariance matrix in (5.4) will be $T \times T$, the covariance matrix in (5.5) for each individual i will be of size $n_i \times n_i$, where n_i is the number of measurements provided by individual i.

The matrices in (5.4) and (5.5) represent block exchangeable, or compound symmetry, within-cluster and within-individual structures, and are again assumed for simplicity. This model assumes that for any two periods p and q where $p \neq q$, cell $[p, q]$ of the cluster-level covariance matrix is equal to π, and similarly cell $[p, q]$ of the individual-level covariance matrix is equal to τ. This means that measurements in the same cluster at the same time are more correlated than those in the same cluster at different times, and that measurements at different times have the same correlation regardless of the time between them. The variance at both individual and cluster level is assumed constant over time, so σ^2_{CT} and σ^2_E can be factored out [36].

The covariance structures above for within-cluster and within-individual measurements over time are equivalent to including a random intercept for both cluster and individual, a random coefficient for cluster-period and a residual:

$$Y_{tijk} = \mu + \mathbf{C}_{jk} + \mathbf{CT}_{tjk} + \mathbf{I}_{ijk} + \mathbf{E}_{tijk} \tag{5.6}$$

$$\mathbf{C}_{jk} \sim N(0, \sigma^2_C), \quad \mathbf{CT}_{tjk} \sim N(0, \sigma^2_{CT}), \quad \mathbf{I}_{ijk} \sim N(0, \sigma^2_I), \quad \mathbf{E}_{tijk} \sim N(0, \sigma^2_E). \tag{5.7}$$

The random effects $\mathbf{C}_{jk}$ and $\mathbf{I}_{ijk}$ denote random intercepts at cluster and individual level respectively, and $\mathbf{CT}_{tjk}$ and $\mathbf{E}_{tijk}$ denote the random cluster-period effects and residuals as above. The same variance-covariance structures are assumed across arms.

This model has the cross-classified random effects and constrained baseline approach of the Hooper and Kasza models presented in Section 1.3.8.4, and is specific to parallel-group designs rather than being general to SW and CRXO designs.

5.2.2.4 Variance components

Using model (5.3), the total variance of the outcome is $\sigma^2 = \sigma^2_{CT} + \sigma^2_E$, and the ICC is

$$\rho = \frac{\sigma^2_{CT}}{\sigma^2_{CT} + \sigma^2_E}, \tag{5.8}$$

or using equivalent model (5.6) the ICC is

$$\rho = \frac{\sigma^2_C + \sigma^2_{CT}}{\sigma^2_C + \sigma^2_{CT} + \sigma^2_I + \sigma^2_E}. \tag{5.9}$$

The individual autocorrelation (IAC) and cluster autocorrelation (CAC) are given by the parameters τ and π respectively and will both be set at values of 0.5 [46], as they are likely to be higher than ICCs. Higher values of the IAC may also benefit the CC/OC designs over R-CS. By specifying the ICC ρ and the overall variance of y_{tijk}, σ^2, the variance components σ^2_{CT} and σ^2_E can be found. Specification of τ and π, then allows covariance matrices to be formed and used to generate the vectors $\mathbf{CT}_{tjk}$ and $\mathbf{E}_{tijk}$.

5.2.2.5 Data generating process

The models above use a constrained baseline approach for the data generation which will also be used in the analysis models. As the models do not contain a main effect for treatment arm *trt*, they are constrained to be equal at baseline. The choice of this approach over a longitudinal ANCOVA approach is discussed in Section 1.3.4. Briefly, the two DGMs are used to generate all 79 outcomes, including baseline, as opposed to including the baseline outcome as a covariate in the model as in a longitudinal ANCOVA approach.

A full set of 79 measurements ($t = 0, \ldots, 78$) was generated for each individual in the original cohort. Missing data mechanisms (Section 5.2.2.6) were then applied to the original cohort, which provided a time of exiting the cluster for each individual. Individuals leaving a cluster were replaced the next day by individuals of the additional cohort. Outcomes were then generated for these additional cohort members, again using the chosen DGM, but instead of a full set of outcomes the first outcome is assumed to be taken at the next measurement point following their entry to the cluster. For example, an individual in the additional cohort entering a cluster at week 5 and 3 days had measurements calculated for week 6 through to final follow-up ($t = 6, \ldots, 78$). The same missing data mechanisms were applied to these individuals' measurements, and this process continued until every 'bed' in each cluster was accounted for over the 78 week trial period.

The full data generation process is summarised in Appendix, Figure B.1.

5.2.2.6 Missing data mechanisms and turnover rate

Missing data mechanisms describe *why* data is missing, as discussed previously in Section 2.3.2.1. Missing data patterns describe *which* data is missing. Simplifying notation and letting y_t be an outcome measurement at time t, data is monotone missing if y_t is missing and $y_{t+1}, \ldots, y_T$ are also missing, for all $t = 1, \ldots, T-1$ [57]. This occurs in trials with longitudinal data collection where an individual drops out, say at time t, and therefore any measurements after time t are also missing. If this does not hold, missingness can be called general or intermittent. Only monotone missingness is considered in this simulation study, so rather than specifying a percentage of missing data to delete as with intermittent missingness, instead the turnover rate of individuals will be varied. For example, with a turnover of 10%, this means that the parameters in the following models are chosen such that the mean drop-out rate over all simulations is as close to 10% as possible. Turnover rates will be varied as in Table 5.4.

As well as varying the turnover rate, the missing data mechanisms (MCAR, MAR or MNAR) will be varied. The following approach is an extension of that used by Thomadakis *et al.* [212]. Drop-out times are assumed to be exponentially distributed and are calculated using the inverse CDF theorem (Appendix, Section B.2). In the following models, the term $b_j \sim \mathrm{N}(0, \sigma_b^2)$ represents a random effect for cluster j, creating a clustering effect in drop-out times for individuals within the same cluster. The b_j term is independent of all other random effects in the mean model. Due to the inclusion of b_j in the following models, standard definitions of MCAR, MAR and MNAR will be adapted slightly to be known as *cluster-dependent* MCAR, MAR and MNAR, which is a novel definition. This is because the cluster effect b_j is unobserved and under standard definitions would be seen as MNAR; instead, I define the mechanisms based not on the cluster effect but on the observed and unobserved measurements only.

For the following three hazard functions, δ_0, δ_1 and δ_2 are used to denote constants which will be varied for different drop-out rates. Missing data is defined as (cluster-dependent) MCAR if the risk of drop-out is not dependent on any observed or unobserved measurements (outcomes or covariates). The number of measurements and visit times are allowed to differ across individuals, with measurements for individual i in cluster j given by $y_{ij,1}, \ldots, y_{ij,d_i}$ at times $t_{i,1}, \ldots, t_{i,d_i}$, where d_i is the total number of measurements for individual i. The corresponding hazard for an individual in cluster j will be calculated using

$$h_{ij}(t) = \exp(\delta_0 + b_j), \quad t_{i,t} \leq t < t_{i,t+1} \tag{5.10}$$

where only δ_0 needs to be set. This hazard for individual i at time t is defined from time t up to, but not including, time $t + 1$.

Missing data is defined as (cluster-dependent) MAR if the risk of drop-out at time t is assumed dependent on observed outcomes only. Though some describe MAR as being dependent on measured covariates, this case is not considered here. The hazard for individual i in cluster j will therefore be calculated using

$$h_{ij}(t) = \exp(\delta_0 + \delta_1 y_{ij,t} + b_j), \quad t_{i,t} \leq t < t_{i,t+1} \tag{5.11}$$

where $y_{ij,t}$ is the individual's most recent measurement. For this mechanism δ_0 and δ_1 need to be set.

Missing data is defined as (cluster-dependent) MNAR if the risk of drop-out at time t is assumed dependent on both observed and unobserved outcomes. The hazard for individual i in cluster j is then calculated using

$$h_{ij}(t) = \exp(\delta_0 + \delta_1 y_{ij,t} + \delta_2 y_{ij,t+1} + b_j), \quad t_{i,t} \leq t < t_{i,t+1} \tag{5.12}$$

where $y_{ij,t}$ and $y_{ij,t+1}$ are the individual's most recent *and* next (unobserved) measurement. For this mechanism δ_0, δ_1 and δ_2 need to be set. Values of δ_0, δ_1 and δ_2 used to generate the different missing data mechanisms and turnover rates are given in Table 5.6.

Mechanism	Turnover	δ_0	δ_1	δ_2
MCAR	10%	-7.05		
	20%	-6.226		
	40%	-5.238		
MAR	10%	-7.05	-0.583	
	20%	-7.05	-1.823	
	40%	-7.05	-2.99	
MNAR	10%	-7.05	-0.583	-0.05
	20%	-7.05	-0.583	-1.63
	40%	-7.05	-0.583	-2.945

Table 5.6: Values of δ_0, δ_1 and δ_2 used to generate the different missing data mechanisms and turnover rates.

5.2.2.7 Cluster-level intervention effect rate

As described in Equation (5.1), the total intervention effect rate consists of the cluster-level and individual-level intervention effect rates. The individual-level part, represented by C in mean model (5.1), is assumed to be constant over time in both the base case and complications, resulting in a linear effect. For the cluster-level part, I also assume a constant rate over time in the base case, represented by B in (5.1). The intervention therefore improved by the same amount each week in the base case.

For the complications, the cluster-level intervention effect will also be allowed to be non-constant to assess how this affects results, with B becoming B_t so that the intervention effect changes at different time points t. Table 5.7 shows the two scenarios and how the B_t are calculated, with negative values of the intervention effect indicating improvement.

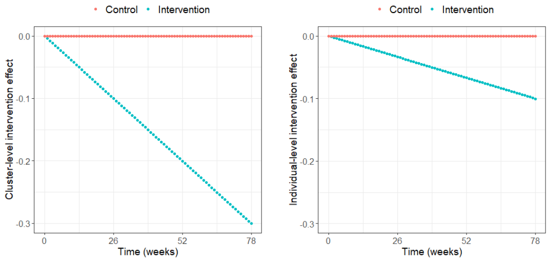
For simplicity, just one non-constant cluster-level intervention effect was chosen to compare against the constant case. For this reason, the shape of the effect had to be sufficiently different from the constant case and a realistic possibility in practice. After discussion with the supervisory team, a shape incorporating an improvement and a plateau was chosen, with an added delay of 8 weeks at the start to mimic the lag often seen in interventions. Exponential functions of time commonly used in the literature on learning effects in surgery trials [155, 156] are used to approximate the desired shapes, with values chosen using trial and error.

Figures 5.2 and 5.3 show the cases where the cluster-level and individual-level intervention effects are linear, and Figures 5.4 and 5.5 show the cases where the cluster-level intervention effect rate is non-constant, with two versions of each due to the different values. In all scenarios the intervention effect is the same at 78 weeks. Across all scenarios, the effect rate in the control arm is zero.

Scenario	**Formulae for B_t**
Constant (base case)	$B = -0.40092$ for all $t = 1, \ldots, 78$
Delay, improvement then plateau (C>I)	$B_t = \begin{cases} 0, & \text{if } 0 \leq \text{t} < 8 \\ 0.30095\exp(-0.1(t-8)) - 0.30095, & \text{if t} \geq 8 \end{cases}$
Delay, improvement then plateau (I>C)	$B_t = \begin{cases} 0, & \text{if } 0 \leq \text{t} < 8 \\ 0.10032\exp(-0.1(t-8)) - 0.10032, & \text{if t} \geq 8 \end{cases}$

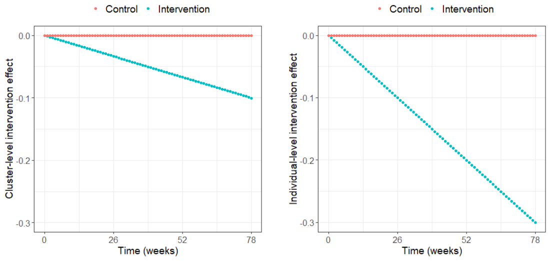
Table 5.7: Functions used to generate the B_t coefficients for the cluster-level intervention effects. Time, t, is a sequence of 78 time points coded in weeks.

This complication focuses on non-constant intervention effect rates at the cluster-level only. There could exist more complicated learning effects at the individual-level, possibly related to an individual's characteristics, but these will not be investigated here and the individual-level intervention effect rate is assumed constant throughout. Intervention effects are also assumed to be the same across all clusters in the intervention arm, equivalent to assuming no treatment by time by cluster interaction.



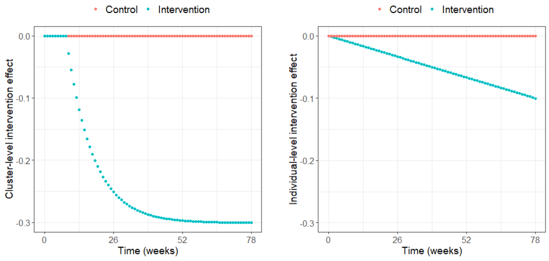
(a) (b)

Figure 5.2: Constant cluster-level intervention effect rate, larger than individual-level (3:1).



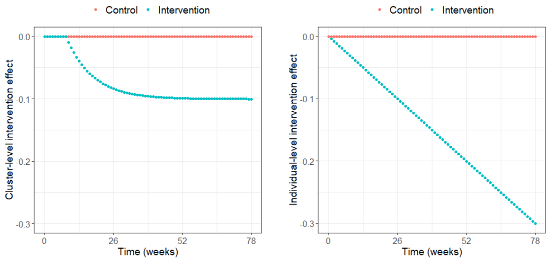
(a) (b)

Figure 5.3: Constant cluster-level intervention effect rate, smaller than individual-level (1:3).



(a) (b)

Figure 5.4: Non-constant cluster-level intervention effect rate, larger than individual-level (3:1).



(a) (b)

Figure 5.5: Non-constant cluster-level intervention effect rate, smaller than individual-level (1:3).

5.2.3 Estimands

Use of different estimands is important in this simulation study because the analysis models used may provide unbiased results for some estimands and biased results for others. Moreover, the three designs give rise to different estimands so it would be an unfair comparison to only look at one.

5.2.3.1 Immortal and mortal estimands

The estimands will be split into whether they are immortal, meaning individuals who drop out are still included, and mortal, where individuals who drop out are not included; these will be noted by -I and -D suffixes respectively. The immortal estimands assume that individuals do not leave the cluster, due to death or any other reason, and use of ‘immortal’ in this chapter does not distinguish between different reasons for leaving the cluster. Immortal estimands ultimately are hypothetical, so how attractive they would be to trialists depends on how strong the assumptions are and whether they are realistic or not for a particular situation. If an intervention causes differential drop-out between arms, immortal estimands are not affected, whereas the -D estimands are as they assess the individuals who remain, rather than making predictions about those who are present at other times throughout the trial but not necessarily at the end. Similarly, if the intervention causes different missing data mechanisms between arms, this would not be accounted for with immortal estimands but more so with -D estimands.

The ICH E9 framework refers to a ‘hypothetical’ strategy as one of the ways to overcome intercurrent events, such as death, when defining an estimand [213]. The estimands that take into account drop-out could be seen as a type of ‘whilst alive’ approach in the ICH E9 framework. Mortal cohort inference and how partly conditional or fully conditional models can be used as an analysis tool for this type of inference were previously discussed in Chapter 2, and will also be returned to in Chapter 6 where some of the alternative analysis models presented use a similar approach.

5.2.3.2 Description of estimands

The immortal closed cohort (CC-I, to avoid confusion with the ICC) estimand, θ_{CC-I}, is defined as the treatment effect at 78 weeks, where an individual has received a full 78 weeks of exposure at both individual and cluster level. This is a closed cohort estimand because only members of the closed cohort experience 78 weeks' exposure at individual and cluster level. This estimand is 'immortal' as it assumes that these individuals do not leave the cluster, due to death or any other reason, where the use of 'immortal' in this chapter does not distinguish between different reasons for leaving the cluster. The true value of this estimand is taken directly from the DGM as $\theta_{CC-I} = -0.40092$, and is the same for both C>I and I>C because the total intervention effect at 78 weeks is the same irregardless of the relative size of the individual- and cluster-level intervention effects (Table 5.8).

The closed-cohort with drop-out (CC-D) estimand, θ_{CC-D}, is defined as the treatment effect for those present at 78 weeks who were also present at baseline. The cross-sectional or open-cohort with drop-out (CS-OC-D) estimand, $\theta_{CS-OC-D}$, is the treatment effect for those present at 78 weeks regardless of whether they were present at baseline or not. These estimands do not assume immortality and instead take into account drop-out, so the true values cannot be calculated directly from the parameters of the DGM. True values are instead calculated by simulating one very large dataset of 4,998,000 individuals, the largest number of individuals that could feasibly be generated in the 48 hour limit of the High Powered Computing facilities. This was done by simulating 13,328 clusters with a cluster sample size and population size of 15 with an open cohort design, providing 199,920 individuals, running this in parallel 25 times with random number seeds that are 1 million apart, and combining these datasets together. The dataset must be large enough that the variance of the estimand is "negligible" [4]; the standard deviation of the difference between arms in this case was never more than 0.001 (3 s.f.). The true value for θ_{CC-D} is calculated using the difference in mean outcomes between arms at 78 weeks from this very large dataset, for those in the original cohort only. The true value for $\theta_{CS-OC-D}$ is calculated using the difference in means from the very large dataset, but for all individuals present at 78 weeks.

Finally, I define the open-cohort with exactly 52 weeks of individual exposure (OC-52-I) estimand, $\theta_{OC-52-I}$, and the open-cohort with exactly 26 weeks of individual exposure (OC-26-I) estimand, $\theta_{OC-26-I}$, as the treatment effects at 78 weeks on the cluster-randomisation timescale (cr.time) and either 52 or 26 weeks on the individual timescale (ind.time), respectively. These are immortal estimands that correspond to the treatment effects from being present in the cluster at 78 weeks but having only been present for either 52 or 26 weeks, rather than from cluster-randomisation. These estimands use two timescales as opposed to the others which only require one, that timescale being time from cluster-randomisation. The true values of these estimands are derived in a similar way to that of the CC-I estimand, using values from the DGM. The CC-I estimand is a special case of the OC estimand, with 78 weeks on both timescales (OC-78-I).

For the OC-52-I estimand with C>I, for example, the true value is calculated as:

$$78 \times -0.003855 + 52 \times -0.001285 = -0.36751.$$

This calculation assumes that individuals are present in the cluster at 78 weeks but joined 52 weeks from the end, and with no drop-out assumed this is again an immortal estimand. The other OC estimands with different values are given in Table 5.8. These are not as large in magnitude as the CC-I estimand of -0.40092, as might be expected, because individuals are only present in the cluster for 52 or 26 weeks (for OC-52-I and OC-26-I respectively), as opposed to the full 78 weeks with CC-I.

Given that the true values of the immortal estimands are taken from the underlying DGM, the drop-out mechanisms are not important as they are not taken into account. For the mortal estimands, the true value of the estimand will differ according to the type of missing data mechanism and the turnover rate (Table 5.9).

For example, the true value of the estimand for MCAR sub-samples with C>I has the same value for 10%, 20% and 40% turnover. If it were not an immortal estimand, there would be different true values for each of the turnover rates, as there is for the CC-D and CS-OC-D estimands.

	CC-I	OC-52-I	OC-26-I
C>I	-0.40092	-0.36751	-0.3341
I>C	-0.40092	-0.30069	-0.20046

Table 5.8: True values of the immortal estimands with different values.

An overview of which designs and how many timescales will be used for each estimand is given in Table 5.10. The analysis models used to estimate the different estimands are

Mechanism	Values	Intervention effect	Turnover	CS-OC-D	CC-D
MCAR	C>I	Constant	10	-0.393282	-0.3990512
			20	-0.3879463	-0.3986262
			40	-0.3764333	-0.3995109
		Non-constant	10	-0.393282	-0.3990512
			20	-0.3879463	-0.3986262
			40	-0.3764333	-0.3995109
	I>C	Constant	10	-0.3829711	-0.3990512
			20	-0.3665646	-0.3986262
			40	-0.3300396	-0.3995109
		Non-constant	10	-0.3829711	-0.3990512
			20	-0.3665646	-0.3986262
			40	-0.3300396	-0.3995109
MAR	C>I	Constant	10	-0.396564	-0.3962919
			20	-0.3866165	-0.3736403
			40	-0.3774256	-0.3455625
		Non-constant	10	-0.3976866	-0.3951148
			20	-0.3819215	-0.3600174
			40	-0.3626033	-0.3135752
	I>C	Constant	10	-0.3864624	-0.3962919
			20	-0.3696398	-0.3736403
			40	-0.3491395	-0.3455625
		Non-constant	10	-0.3863231	-0.395881
			20	-0.3673629	-0.3692838
			40	-0.3440042	-0.3354733
MNAR	C>I	Constant	10	-0.398452	-0.3971206
			20	-0.3843441	-0.3691626
			40	-0.3692244	-0.3408626
		Non-constant	10	-0.398999	-0.395756
			20	-0.3750389	-0.3515806
			40	-0.3509729	-0.3058638
	I>C	Constant	10	-0.387865	-0.3971206
			20	-0.3691923	-0.3691626
			40	-0.346645	-0.3408626
		Non-constant	10	-0.3879901	-0.3967288
			20	-0.3652782	-0.3635612
			40	-0.3394025	-0.3298923

Table 5.9: True values of the mortal estimands with different values, missing data mechanisms, turnovers and intervention effects.

given in Section 5.2.4.

5.2.4 Analysis models

In most simulation studies, different statistical methods are assessed over a range of scenarios. Here, the ‘methods’ to be compared are the trial designs. As the designs will be

	CC-I	CC-D	CS-OC-D	OC-52-I	OC-26-I
Single timescale discrete model	✓	✓	✓	✗	✗
Two timescale discrete model	✓	✓	✓	✗	✗
Two timescale continuous model	✗[1]	✗[1]	✗[1]	✓	✓
CC design	✓	✓	✓	✗	✗
R-CS design	✓	✓	✓	✓	✓
OC design	✓	✓	✓	✓	✓

Table 5.10: Summary of which designs and models will be used to estimate the different estimands. [1]This model could in theory have been used for these estimands but only the discrete models were used.

varied, the method of analysis will be fixed within each comparison. A model is therefore required that is flexible enough to be applied to all designs. Kasza's model [36], a special case of Feldman and McKinlay's unified model [14] and described previously in Section 1.3.8.4.3, permits analysis of R-CS, CC and OC designs. Whilst this model is unified in the sense that it encompasses all three designs in equation form, in practice to fit the R-CS model the individual random effects must be removed. Furthermore, for the R-CS design, even where there is an overlap of participants at different time points, I assume that this information is not available.

The following three analysis models are all mixed effects models. Mixed effects models assume that missing data is MAR. As a result, correctly specified models should yield unbiased estimates under MCAR and MAR, but MNAR estimates are expected to be biased. Moreover, due to the implicit imputation that occurs, mixed effects models should provide unbiased estimates for immortal estimands, but not mortal estimands.

In terms of random effects, the CC and OC designs are correctly specified as they contain the same random effects as the underlying DGM. The R-CS design has individual IDs changed; for example, in the CC/OC datasets where the first three measurements might be linked to an individual with ID number 1, the R-CS design changes these to 1.1, 1.2 and 1.3 so that linkage cannot be made. This was in an attempt to make the dataset closer to an actual cross-sectional dataset. Given this and the omission of the individual random effect, there is an attempt to make the R-CS model more correctly specified to the DGM, but they are not perfectly aligned.

In this section, the analysis models that will be used and their technical details are described. `R` code for the analysis models is provided in Appendix, Section B.6 as well as equivalent codes in `Stata` and `SAS`.

5.2.4.1 Kasza's single-timescale analysis model with discrete time

Kasza's single-timescale model [36], as previously described in Section 1.3.8.4.3, is written in a general form such that it can be applied to all types of longitudinal CRTs such as stepped-wedge and cross-over designs. The model used here uses arm by time interactions as in Hooper's model [51] (Section 1.3.8.4.2) instead of cluster by time interactions to make it specific to parallel-group CRTs, and a subscript k for arm is included which is omitted in the Kasza model.

This model contains just one timescale, time from cluster-randomisation, denoted previously as *cr.time*. The individual random effect is dropped when applying this model to the R-CS design, but retained for CC and OC designs. This model is the same as the DGM of (5.6) and mean model (5.2), but here time is treated as discrete, which is indicated with a subscript t. The subscript 1 after a comma denotes that these coefficients are from the model with one timescale only:

$$\mu = A_{t,1}\, cr.time + D_{t,1}\,(cr.time \times trt). \tag{5.13}$$

In this model, $D_{t,1}$ at $t = 78$, or equivalently $D_{78,1}$, is the estimate of interest to be extracted for the CC-I, CC-D, and CS-OC-D estimands.

This model is mis-specified for all three designs because the underlying DGM contains *cr.time* and *ind.time* timescales, and this model contains just *cr.time*.

5.2.4.2 Two-timescale analysis model with discrete time

This model extends Kasza's model by adding an individual exposure timescale, and as a result aligns with the underlying DGM (5.6) and mean model (5.1). Time is again treated as discrete, indicated by a subscript t, but this time the subscript 2 after the comma denotes that these coefficients are from the model with two timescales:

$$\mu = A_{t,2}\, cr.time + B_{t,2}\, ind.time + C_{t,2}\,(cr.time \times trt) + D_{t,2}\,(ind.time \times trt). \tag{5.14}$$

In this model, $C_{t,2}+D_{t,2}$ at $t = 78$, or equivalently $C_{78,2} + D_{78,2}$, is the estimate of interest to be extracted for the CC-I, CC-D, and CS-OC-D estimands, representing the total intervention effect and equivalent to $D_{78,1}$ in the single-timescale model. As these estimates are found separately before being combined, the standard error (SE) for the combined total intervention effect is calculated post-hoc using standard formulae [214].

For both the discrete single- and two-timescale models, one estimate is obtained for each scenario. As there are true values for each estimand, the bias for each estimand can be calculated separately, but the empirical standard error is only estimated once for all estimands.

This analysis model is a simple but novel extension of the existing Kasza model; in Chapter 6 other possible analysis methods for OC datasets are considered that explicitly incorporate the missing data mechanism.

This model is correctly specified for the OC design because both the underlying DGM and analysis model contain *cr.time* and *ind.time* timescales. The R-CS design has more mis-specification than the OC design due to the data not being truly cross-sectional, as described in Section 5.2.4. When the cluster-level intervention effect rate is non-constant, due to the use of discrete time this model is closer to being correctly specified than the continuous time version in the following section.

5.2.4.3 Two-timescale analysis model with continuous time

The final model contains both the individual exposure timescale and time from cluster randomisation timescales, but time is treated as continuous and denoted by '*cts*' subscripts rather than t subscripts as in the discrete models:

$$\mu = A_{cts}\, cr.time + B_{cts}\, ind.time + C_{cts}\,(cr.time \times trt) + D_{cts}\,(ind.time \times trt). \tag{5.15}$$

This model is used for the OC-52-I and OC-26-I estimands because of issues encountered when trying to estimate these with the discrete model (5.14) above. With discrete time, an unforeseen issue was that it was not always possible to estimate $D_{t,2}$ at the desired time points of $t = 52$ or 26. As individuals entered clusters at different times, the fixed times of measurements dictated the individual time at which measurements were taken, which were not always at the exact times of $ind.time = 52$ or 26. By treating time as continuous, any value of *ind.time* and *cr.time* can be estimated from the model using multiplication. The estimates of interest are therefore:

$$C_{cts} \times 78 + D_{cts} \times 52, \quad \text{for the OC-52-I estimand}$$

and

$$C_{cts} \times 78 + D_{cts} \times 26, \quad \text{for the OC-26-I estimand.}$$

As OC-52-I and OC-26-I are obtained using separate estimates, empirical standard error is also estimated for each of them separately.

This model is correctly specified for the OC design as above, but only when the cluster-level intervention effect rate is constant. This model is more mis-specified for both designs when it is non-constant, because continuous time assumes a linear trend. The R-CS design again has more mis-specification than the OC design due to the data not being truly cross-sectional (Section 5.2.4).

5.2.4.4 Packages

Throughout this simulation study the `lmer` function within the `lme4` package was used to fit linear mixed effects models in `R` version 4.1.0, [215, 216]. This package was chosen over `nlme` because `lme4` offers crossed random-effects capability [217]. Restricted maximum-likelihood (REML) was used for estimation.

5.2.4.5 Convergence issues

The default optimizer for `lmer` is "nloptwrap", which gave convergence errors related to a singularity (the cluster variance was near zero) in many cases. In initial tests, the "bobyqa" optimizer had fewer convergence errors and so was chosen over the default. The number of iterations was also increased to 2×10^5 as is recommended with this optimizer.

If a model fails to converge, results are unreliable, therefore the number of errors were recorded for each scenario and the non-converged results were omitted from analyses. Results should therefore be carefully considered where the number of successfully converged simulations are less than n_{sim}.

5.2.4.6 Parallel computing and random number seeds

This work was undertaken on ARC3 and ARC4, part of the High Performance Computing facilities at the University of Leeds, UK. Different scenarios were run in parallel across different cores, reducing computational time significantly.

`R`'s default Mersenne-Twister random number generator [218] is used for simulation of datasets. The same seed was set once at the beginning of all DGM simulations, and the state of the random number stored before the generation of each dataset in a repetition and after all n_{sim} repetitions to avoid accidental dependence [4]. Setting the seed guaranteed reproducibility of results, but also meant that if interruptions occurred the last seed could be located and generation could be restarted from this point.

5.2.5 Performance measures

For each estimand θ, the key performance measure of interest is bias, which assesses how far away the estimator is from the true value. I will also assess the empirical standard error, equivalent to the standard deviation of $\hat{\theta}$, which describes the spread of the estimates calculated under each scenario. As there are only finite repetitions of the simulations, there is a degree of uncertainty involved in calculating the performance measure estimates. This uncertainty is determined by the Monte Carlo standard error (MCSE) of a performance measure estimate. Table 5.11 below details the definition, estimate and MCSE of bias and empirical SE [4].

Performance measure	**Definition**	**Estimate**	**Monte Carlo SE**
Bias	$\text{E}[\hat{\theta}] - \theta$	$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\hat{\theta}_i - \theta$	$\sqrt{\frac{\text{Var}(\hat{\theta})}{n_{sim}}}$
Empirical SE	$\sqrt{\text{Var}(\hat{\theta})}$	$\sqrt{\frac{1}{n_{sim}-1}\sum_{i=1}^{n_{sim}}(\hat{\theta}_i - \bar{\theta})^2}$	$\frac{\widehat{\text{EmpSE}}}{\sqrt{2(n_{sim}-1)}}$

Table 5.11: Definition, estimate and Monte Carlo SE for bias and empirical SE. Adapted from Morris *et al.* [4]

As bias is the key performance measure, the number of required simulations (or repetitions), n_{sim}, will be calculated using the formula for Monte Carlo SE for bias and by specifying the desired level of accuracy. If the Monte Carlo SE for bias is required to be at most 0.005, the following can be solved:

$$n_{sim} = \frac{\text{Var}(\hat{\theta})}{0.005^2}.$$

Test simulations were run and an upper bound for $\text{Var}(\hat{\theta})$ was found to be 0.127^2, leading to $n_{sim} = 650$. Note that n_{sim} is the number of simulations run for *each scenario* of the base case and complications.

Model-based SEs are not investigated in this simulation study as they are likely to be inappropriate in many cases due to model mis-specification. In practice, bootstrapping methods could be used to obtain suitable model-based SEs.

5.2.6 Research questions revisited and hypotheses

The research questions (RQ) from the introduction are revisited and split by estimand now that these have been set out. Below each research question are my hypotheses (H). Bias and precision are the performance measures to be assessed as described in the previous section.

RQ1: For a CC-I estimand, when is the OC design superior to the CC and R-CS designs?
H1: For a CC-I estimand, the CC design will have the lowest bias. The OC design will have higher precision than the CC design. In full-samples, OC will have slightly higher precision than R-CS. In sub-samples, R-CS will have higher precision than OC.

RQ2: For a CC-D estimand, when is the OC design superior to the CC and R-CS designs?
H2: For a CC-D estimand, the CC design *may* have the lowest bias, but this could be affected by the implicit imputation of the mixed model as described in Section 5.2.4. The OC design will have higher precision than the CC design. In full-samples, OC will have slightly higher precision than R-CS. In sub-samples, R-CS will have higher precision than OC.

RQ3: For a CS-OC-D estimand, when is the OC design superior to the CC and R-CS designs?
H3: For a CS-OC-D estimand, the OC or R-CS designs *may* have the lowest bias, but again this could be affected by the implicit imputation of the mixed model as described in Section 5.2.4. The OC design will have higher precision than the CC design. In full-samples, OC will have slightly higher precision than R-CS. In sub-samples, R-CS will have higher precision than OC.

I expect the CC design to have the lowest bias for the CC-I and CC-D estimands and the R-CS or OC designs to have the lowest bias for the CS-OC-D estimand because this is when the design aligns with the estimand; the design and estimand are based on the same population.

I expect the CC design to have the lowest precision for the discrete estimands (RQ1-3) for both full and sub-samples because individuals lost are not replaced, so fewer people and fewer data points are available for analysis. In full-samples, as R-CS and OC have the same dataset, I would expect them to have similar bias and precision, but the OC design to have slightly better precision as it takes advantage of extra information by linking measurements where possible. In sub-samples, I expect precision and bias to be distinct for R-CS and OC because the sampling method is different; specifically I would expect

R-CS to have higher precision than OC because a greater number of distinct individuals are sampled at one or more times.

RQ4: For an OC-52-I estimand, when is the OC design superior to the R-CS design?
H4: The OC design will have the lowest bias. In full-samples, OC will have slightly higher precision than R-CS. In sub-samples, R-CS will have higher precision than OC.

RQ5: How many measurement/recruitment points should be used?
H5: I would expect the precision to increase as the number of measurements increases in the base case. For the complications, for all discrete estimands, I would expect the precision to decrease in the CC design as the turnover rate increases because individuals are not replaced and therefore there will be fewer individuals and fewer measurements overall. In contrast, for the R-CS and OC designs I would expect the precision to increase as the turnover rate increases because the number of measurements in the datasets remains the same but the number of individuals that these measurements are from increases.

RQ6: When is it beneficial to use two timescales over one?
H6: In general I would expect the two timescale discrete model to be less biased than the single timescale version because it is correctly specified as opposed to the single timescale model which is mis-specified as one of the underlying timescales is omitted. When there are more individuals in the additional cohort, or equivalently when the turnover is higher, I would expect the two timescale models to be less biased than the one timescale versions as the two timescales in the underlying DGM become more important. However, when the turnover is low and the majority of individuals are from the original cohort, the one timescale model may be a good approximation to the two timescale model with comparable bias. I would also expect the two timescale model to have lower precision as there are more parameters to estimate.

5.3 Results

5.3.1 Immortal closed cohort (CC-I) estimand

5.3.1.1 Base case

Analysis using one timescale, full samples

When is the OC design superior to the CC design?

The OC design is less biased than CC for larger cluster sizes of 100 when C>I (Figure 5.6). The CC design is least biased with small cluster sizes (15 and 50) as expected. The OC design has superior precision to the CC design for all cluster sizes and numbers of measurement points (Figure 5.7).

When is it beneficial to link measurements (when is OC superior to R-CS)?

Linking measurements in the OC design is almost always beneficial for reducing bias, but only be a small amount (Figure 5.6). In full-samples where the R-CS/OC designs include the same individuals, OC and R-CS have similar precision as expected, and linking measurements can provide small improvements in precision, but this depends on the number/size of clusters and number of measurement points (Figure 5.7).

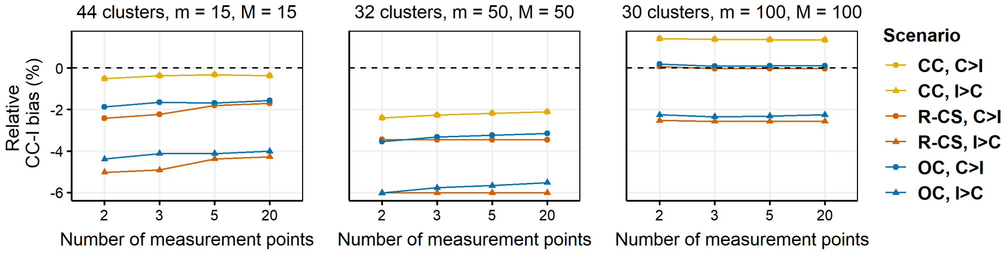


Figure 5.6: Relative CC-I bias in the base case using one timescale for full-samples.

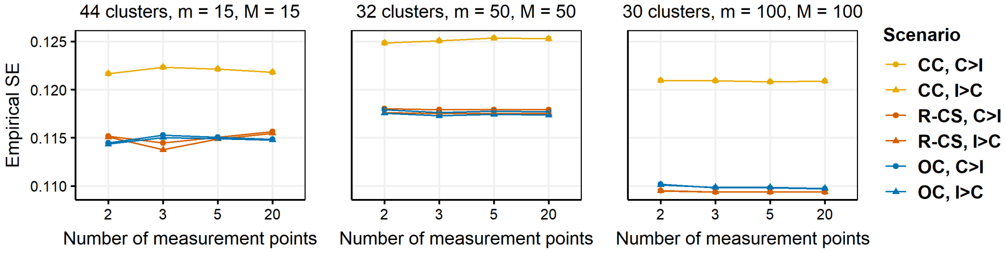


Figure 5.7: Empirical SE in the base case using one timescale for full-samples.

Analysis using one timescale, sub-samples

When is the OC design superior to the CC design?

The OC design can offer improvements in bias over CC for 15,100 and 50,100, if C>I, whereas for 15,50 the CC design is the least biased (Figure 5.8). Again OC has superior precision to CC for all cluster sizes as expected (Figure 5.9).

When is OC sampling superior to R-CS sampling for sub-samples?

OC is less biased than R-CS in most cases except for 15,100 with I>C (Figure 5.8). R-CS and OC designs are distinct for sub-samples in both bias and precision, as expected due to differences in sampling. OC has superior precision to R-CS in the case of 15,100 where the sampling proportion is smallest, but R-CS has improved precision over OC in 15,50 and 50,100 (Figure 5.9).

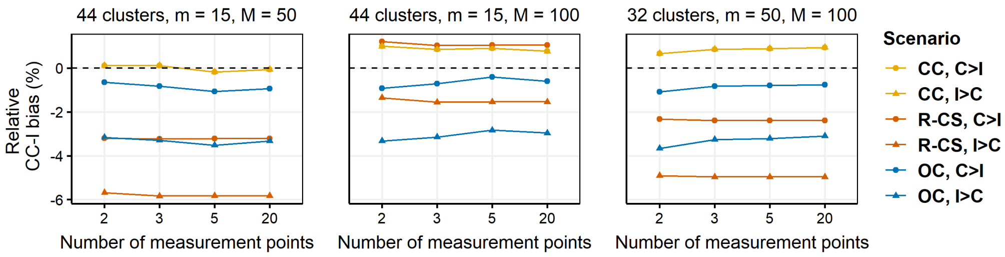


Figure 5.8: Relative CC-I bias in the base case using one timescale for sub-samples.

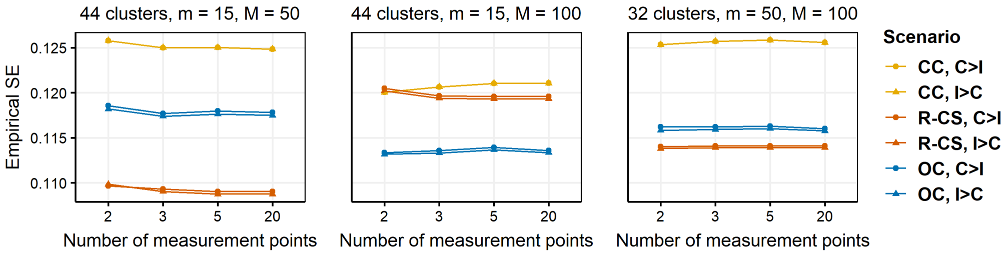


Figure 5.9: Empirical SE in the base case using one timescale for sub-samples.

5.3.1.2 Complications

Analysis using one timescale, full-samples

When is the OC design superior to the CC design?

The CC design is still the least biased design for all turnover rates under MCAR in the complications (Figure 5.10). However, under MAR and MNAR, the OC design can offer improvements in bias over the CC design in many cases when C>I, in particular where the turnover is 20% or greater (Table 5.12). As in the base case, the OC design has superior precision to the CC design for all cluster sizes and numbers of measurement points (Figure 5.11).

When is it beneficial to link measurements (when is OC superior to R-CS)?

Linking measurements in OC decreases bias compared to R-CS in most cases (Figure 5.10). The OC and R-CS designs are similar under MCAR and when the turnover is only 10%, as expected. However, for MAR and MNAR with higher turnovers the OC and R-CS designs differ more. This suggests that when turnover is higher and individuals have fewer measurements, the linkage of the open cohort design becomes more beneficial in terms of reducing bias even though the datasets are the same. Linking measurements in OC can make small improvements in precision in some cases but this depends on relative values of C and I, turnover rate, missing data mechanism and intervention effect (Figure 5.11, Table 5.13).

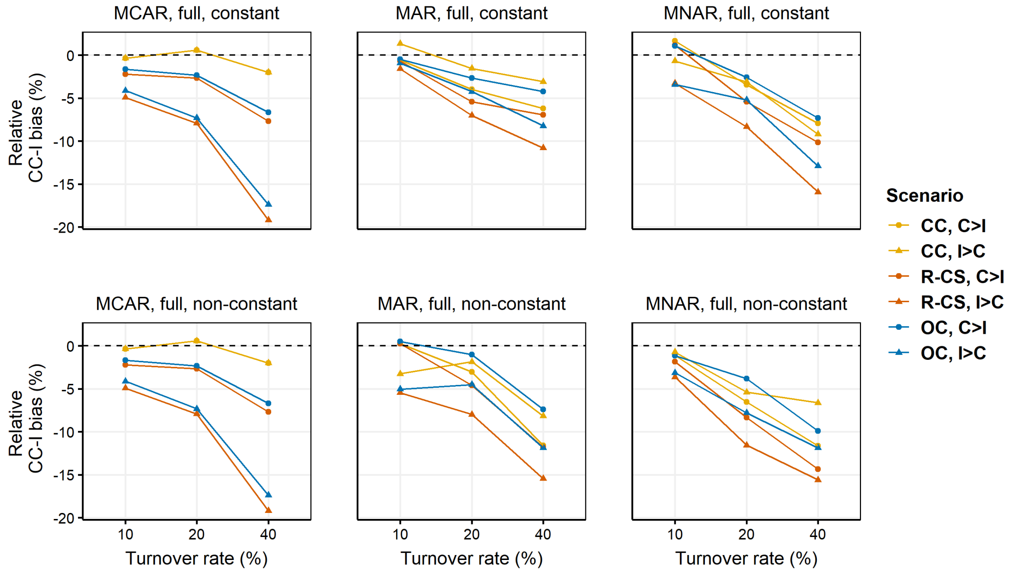


Figure 5.10: CC-I bias in the complications scenarios using one timescale for full-samples.

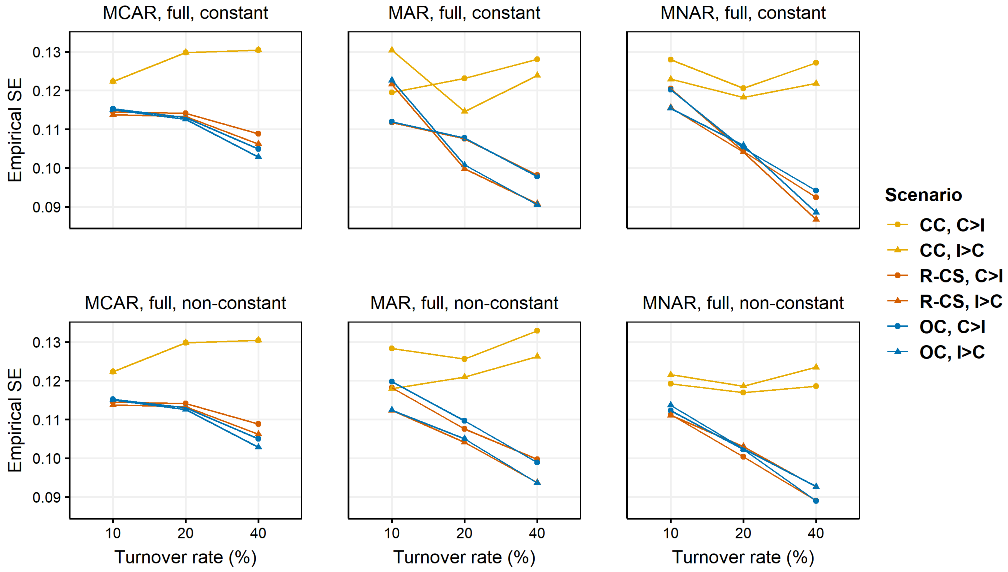


Figure 5.11: Empirical SE in the complications scenarios using one timescale for full-samples.

Superior to CC?		MCAR	MAR	MNAR
Full, bias	Constant	✗✗✗	✓C C	C C C
	Non-constant	✗✗✗	✗C C	✗C C
Full, precision	Constant	✓✓✓	✓✓✓	✓✓✓
	Non-constant	✓✓✓	✓✓✓	✓✓✓

Superior to R-CS?		MCAR	MAR	MNAR
Full, bias	Constant	✓✓✓	✓✓✓	C ✓✓
	Non-constant	✓✓✓	I ✓✓	✓✓✓
Full, precision	Constant	✗✓✓	✗✗✗	I ✗✗
	Non-constant	✗✓✓	✗✗C	✗I I

Table 5.12: A summary of whether the OC design provides an improvement over the CC and R-CS designs for the CC-I estimand in full-samples for the complications. A tick denotes where both values are better for the OC design, a cross denotes where both values are worse for the OC design, and a letter denotes that one of the two values is better for OC, with I = I>C and C = C>I.

Analysis using one timescale, sub-samples

When is the OC design superior to the CC design?

When the turnover rate is varied under MCAR in the complications, CC is still the least biased design for all turnover rates (Figure 5.12). But as in full-samples, under MAR and MNAR, the OC design can offer improvements in bias over the CC design in many cases when C>I, in particular where the turnover is 20% or greater (Table 5.13). The OC design again has superior precision to the CC design for all cluster sizes and numbers of measurement points (Figure 5.13).

When is OC sampling superior to R-CS sampling for sub-samples?

As the sampling differs in sub-samples, the R-CS and OC designs are expected to differ, but they are similar under MCAR and MAR with I>C and a constant intervention effect, which is unexpected (Figure 5.12). The OC design can provide improvements in precision over R-CS in some cases, but this is again dependent on the various complications (Figure 5.13).

Superior to CC?		MCAR	MAR	MNAR
Sub, bias	Constant	✗✗✗	✗C C	✗C C
	Non-constant	✗✗✗	C C C	✗C C
Sub, precision	Constant	✓✓✓	✓✓✓	✓✓✓
	Non-constant	✓✓✓	✓✓✓	✓✓✓

Superior to R-CS?		MCAR	MAR	MNAR
Sub, bias	Constant	✓I ✓	I ✓✓	✗✓✓
	Non-constant	I I ✓	C ✓✓	I ✓✓
Sub, precision	Constant	✗✓✗	C C ✓	✗C ✗
	Non-constant	✗✓✗	✗✓✓	✓✗I

Table 5.13: A summary of whether the OC design provides an improvement over the CC and R-CS designs for the CC-I estimand in sub-samples for the complications. A tick denotes where both values are better for the OC design, a cross denotes where both values are worse for the OC design, and a letter denotes that one of the two values is better for OC, with I = I>C and C = C>I.

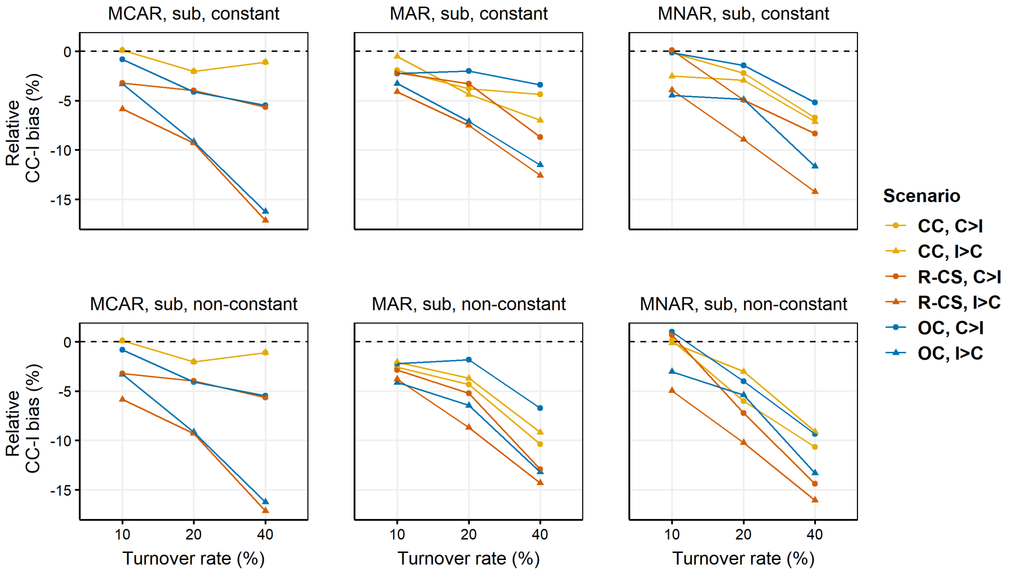


Figure 5.12: CC-I bias in the complications scenarios using one timescale for sub-samples.

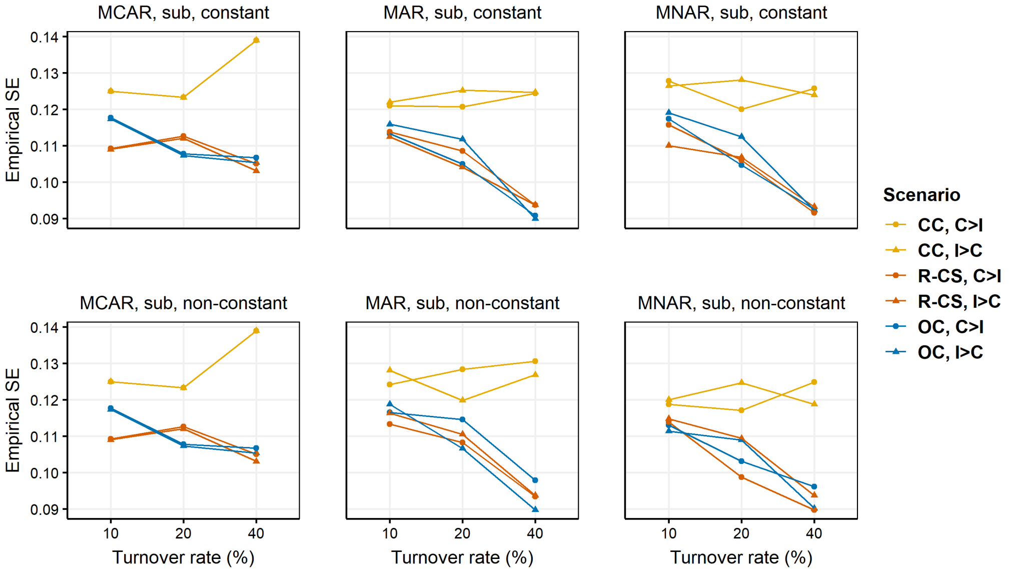


Figure 5.13: Empirical SE in the complications scenarios using one timescale for sub-samples.

5.3.2 Closed cohort with drop-out (CC-D) estimand

5.3.2.1 Base case

The true value of the CC-D estimand in the base case is -0.3990512 and for CC-I is -0.40092. This similarity means that Figures 5.6 and 5.8 are shifted slightly in the positive y-direction to give the CC-D results.

Analysis using one timescale, full-samples

With this shift, the conclusions for the CC-D estimand in the full-samples with one timescale are the same as those for the CC-I estimand in Figures 5.6 and 5.7 (figures omitted).

Analysis using one timescale, sub-samples

When is the OC design superior to the CC design?

For 15,50, CC is no longer the least biased design for all measurement points as in Figure 5.8, with the OC design with C>I having slightly lower bias for 2 and 3 measurement points, though they are similar. For 15,100 and 50,100, the OC design is less biased than CC for all measurement points when C>I.

When is OC sampling superior to R-CS sampling for sub-samples?

The OC design is less biased than R-CS in all cases except 15,100 with I>C (Figure 5.8).

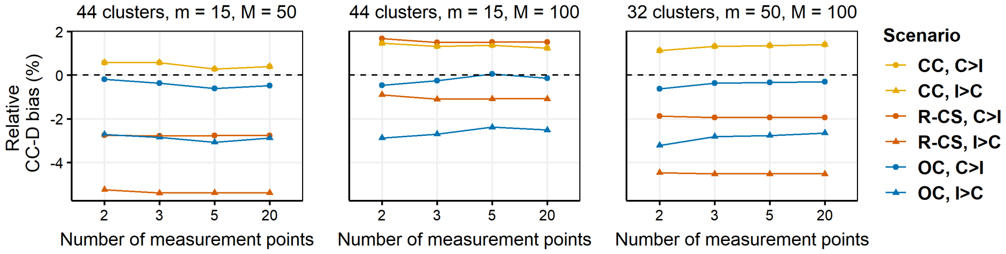


Figure 5.14: Relative CC-D bias in the base case using one timescale for sub-samples.

5.3.2.2 Complications

Analysis using one timescale, full-samples

When is the OC design superior to the CC design?

Under MCAR, the designs are distinct and CC is the least biased for all turnovers (Figure 5.15). Under MAR and MNAR with turnovers of 20% or more, the OC design is less biased than CC but in contrast to CC-I only when I>C. The OC design has superior precision to the CC design for all cluster sizes and numbers of measurement points (Figure 5.11).

When is it beneficial to link measurements (when is OC superior to R-CS)?

Under MCAR, OC is slightly less biased than R-CS by linking measurements (Figure 5.15). Surprisingly, in more challenging circumstances of MAR and MNAR with turnovers of 20% or more, the least biased design is R-CS with the OC design not able to offer as many improvements as for the CC-I estimand (Table 5.14). Linking measurements in OC can make small improvements in precision in some cases but this depends on relative values of C and I, turnover rate, missing data mechanism and intervention effect (Figure 5.11).

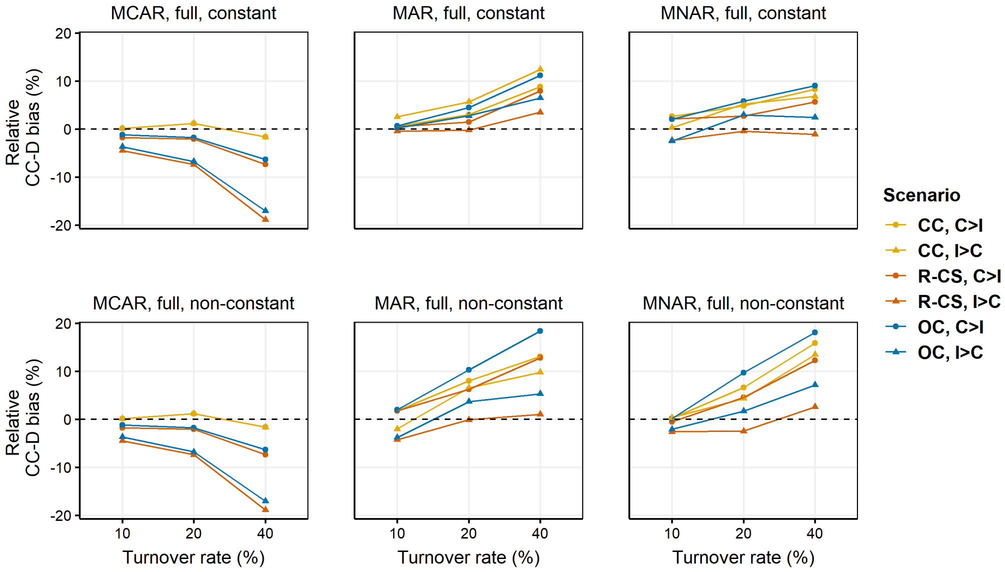


Figure 5.15: CC-D bias in the complications scenarios using one timescale for full-samples.

Superior to CC?		MCAR	MAR	MNAR
Full, bias	Constant	✗✗✗	I I I	C I I
	Non-constant	✗✗✗	✗I I	C I I
Full, precision	Constant	✓✓✓	✓✓✓	✓✓✓
	Non-constant	✓✓✓	✓✓✓	✓✓✓

Superior to R-CS?		MCAR	MAR	MNAR
Full, bias	Constant	✓✓✓	I ✗✗	C ✗✗
	Non-constant	✓✓✓	I ✗✗	✓I ✗
Full, precision	Constant	✗✓✓	✗✗✗	I ✗✗
	Non-constant	✗✓✓	✗✗C	✗I I

Table 5.14: A summary of whether the OC design provides an improvement over the CC and R-CS designs for the CC-D estimand in full-samples for the complications. A tick denotes where both values are better for the OC design, a cross denotes where both values are worse for the OC design, and a letter denotes that one of the two values is better for OC, with I = I>C and C = C>I.

Analysis using one timescale, sub-samples

Conclusions for the sub-samples are the same as those for the full-samples (Figure 5.13 and Appendix, Figure B.3).

Superior to CC?		MCAR	MAR	MNAR
Sub, bias	Constant	C ✗✗	✗I I	C I I
	Non-constant	C ✗✗	C I I	✗I I
Sub, precision	Constant	✓✓✓	✓✓✓	✓✓✓
	Non-constant	✓✓✓	✓✓✓	✓✓✓

Superior to R-CS?		MCAR	MAR	MNAR
Sub, bias	Constant	✓I ✓	I I ✗	C ✗✗
	Non-constant	✓I ✓	C ✗✗	I ✗✗
Sub, precision	Constant	✗✓✗	C C ✓	✗C ✗
	Non-constant	✗✓✗	✗✓✓	✓✗I

Table 5.15: A summary of whether the OC design provides an improvement over the CC and R-CS designs for the CC-D estimand in sub-samples for the complications. A tick denotes where both values are better for the OC design, a cross denotes where both values are worse for the OC design, and a letter denotes that one of the two values is better for OC, with I = I>C and C = C>I.

5.3.3 Cross-sectional/open cohort with drop-out (CS-OC-D) estimand

For this estimand, there is an estimand for C>I and a different one for I>C as estimated by simulation, which is why there are two lines here for CC where there would normally be one, and the different lines for the C and I values are close together.

5.3.3.1 Base case

Analysis using one timescale, full samples

When is the OC design superior to the CC design?

The OC design is less biased than CC for cluster sizes of 15 and 100 as expected, but for size 50 the CC design could be least biased if C>I (Figure 5.16). The OC design has superior precision to the CC design for all cluster sizes and numbers of measurement points (Figure 5.7).

When is it beneficial to link measurements (when is OC superior to R-CS)?

Linking measurements in OC reduces the bias sometimes depending on the number/size of clusters and number of measurement points, but only by small amounts (Figure 5.16). In full-samples where the R-CS/OC designs include the same individuals, OC and R-CS have similar precision as expected, and linking the measurements in the OC design provides small improvements in precision, but this depends on the number/size of clusters and number of measurement points (Figure 5.7).

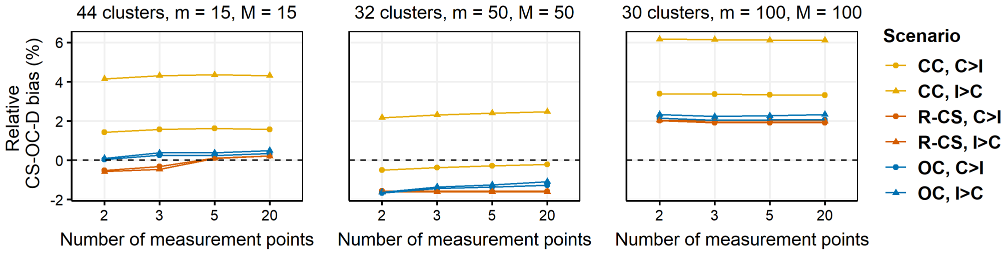


Figure 5.16: Relative CS-OC-D bias in the base case using one timescale for full-samples.

Analysis using one timescale, sub-samples

When is the OC design superior to the CC design?

The OC design is less biased than CC for all cluster sizes and measurement points (Figure 5.17). The CC and OC designs have consistent bias across cluster sizes whereas the R-CS design is more variable. Again OC has superior precision to CC for all cluster sizes as expected (Figure 5.9).

When is OC sampling superior to R-CS sampling for sub-samples?

The OC design is least biased for all measurement points in 15,50 and 15,100 where the cluster sample size is small, but R-CS is slightly superior in 50,100 (Figure 5.17). OC has superior precision to R-CS in the case of 15,100 where the sampling proportion is smallest, but R-CS has improved precision over OC in 15,50 and 50,100 (Figure 5.9).

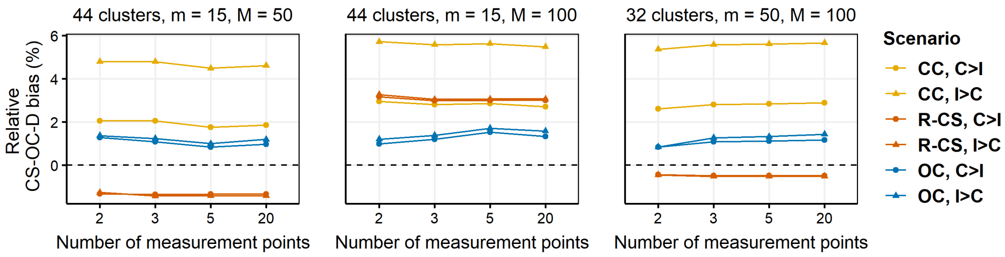


Figure 5.17: Relative CS-OC-D bias in the base case using one timescale for sub-samples.

5.3.3.2 Complications

Analysis using one timescale, full-samples

When is the OC design superior to the CC design?

Under MCAR the OC design is less biased than CC as expected for all turnover rates (Figure 5.18). Under MAR and MNAR with turnovers of 20% or more, the OC design is less biased than CC but only for I>C as with the CC-D estimand. It is surprising that in some cases with C>I, the CC design exhibits almost no bias. The OC design has superior precision to the CC design for all cluster sizes and numbers of measurement points (Figure 5.11).

When is it beneficial to link measurements (when is OC superior to R-CS)?

The OC design can provide some improvements over R-CS depending on the complications and values of C and I (Figure 5.18, Table 5.16). Linking measurements in OC can make small improvements in precision in some cases but this depends on relative values of C and I, turnover rate, missing data mechanism and intervention effect (Figure 5.11).

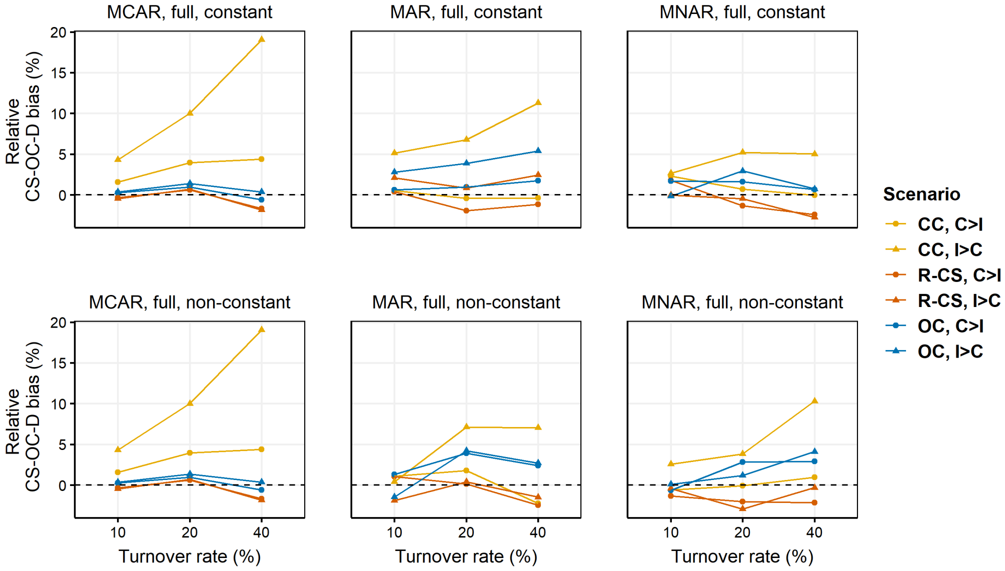


Figure 5.18: CS-OC-D bias in the complications scenarios using one timescale for full-samples.

Superior to CC?		MCAR	MAR	MNAR
Full, bias	Constant	✓✓✓	I I I	✓I I
	Non-constant	✓✓✓	✗I I	I I I
Full, precision	Constant	✓✓✓	✓✓✓	✓✓✓
	Non-constant	✓✓✓	✓✓✓	✓✓✓

Superior to R-CS?		MCAR	MAR	MNAR
Full, bias	Constant	✓✗✓	✗C ✗	C ✗✓
	Non-constant	✓✗✓	I ✗C	✓I ✗
Full, precision	Constant	✗✓✓	✗✗✗	I ✗✗
	Non-constant	✗✓✓	✗✗C	✗I I

Table 5.16: A summary of whether the OC design provides an improvement over the R-CS and CC designs for the CS-OC-D estimand in full-samples for the complications. A tick denotes where both values are better for the OC design, a cross denotes where both values are worse for the OC design, and a letter denotes that one of the two values is better for OC, with I = I>C and C = C>I.

Superior to CC?		MCAR	MAR	MNAR
Sub, bias	Constant	✓✓✓	I I I	C I I
	Non-constant	✓✓✓	✓I I	I I I
Sub, precision	Constant	✓✓✓	✓✓✓	✓✓✓
	Non-constant	✓✓✓	✓✓✓	✓✓✓

Superior to R-CS?		MCAR	MAR	MNAR
Sub, bias	Constant	✓I ✗	I ✗C	C ✗✗
	Non-constant	✓I ✗	C ✗C	I ✗✗
Sub, precision	Constant	✗✓✗	C C ✓	✗C ✗
	Non-constant	✗✓✗	✗✓✓	✓✗I

Table 5.17: A summary of whether the OC design provides an improvement over the CC and R-CS designs for the CS-OC-D estimand in sub-samples for the complications. A tick denotes where both values are better for the OC design, a cross denotes where both values are worse for the OC design, and a letter denotes that one of the two values is better for OC, with I = I>C and C = C>I.

Analysis using one timescale, sub-samples

When is the OC design superior to the CC design?

Bias conclusions for sub-samples are the same as those for full-samples (Figure 5.19). The OC design again has superior precision to the CC design for all cluster sizes and numbers of measurement points (Figure 5.13).

When is OC sampling superior to R-CS sampling for sub-samples?

Conclusions are similar to those of the full-samples, but in sub-samples the R-CS design is less biased than OC in MNAR situations with high turnover, whereas under MAR this still depends on the complications and C and I values (Figure 5.19). The OC design provides improvements in precision over R-CS in some cases, but this is again dependent on the various complications (Figure 5.13).

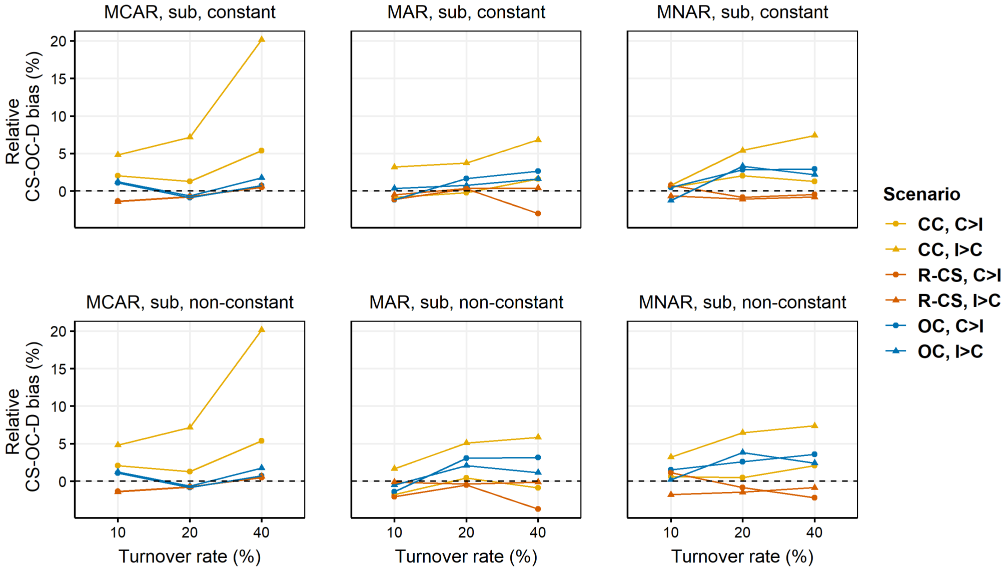


Figure 5.19: CS-OC-D bias in the complications scenarios using one timescale for sub-samples.

5.3.4 Immortal open cohort with 52 weeks of exposure (OC-52-I) estimand

5.3.4.1 Base case

Two timescales, full samples

When is it beneficial to link measurements (when is OC superior to R-CS)?

The OC design can provide less bias than R-CS by linking measurements for all measurement points for the smallest cluster size of 15, but for larger cluster sizes the bias is more similar (Figure 5.20). Linking in the OC design provides improved precision over R-CS in almost all cases except for the smallest cluster size of 15 with just 2 measurement points, but this is a small loss (Figure 5.21).

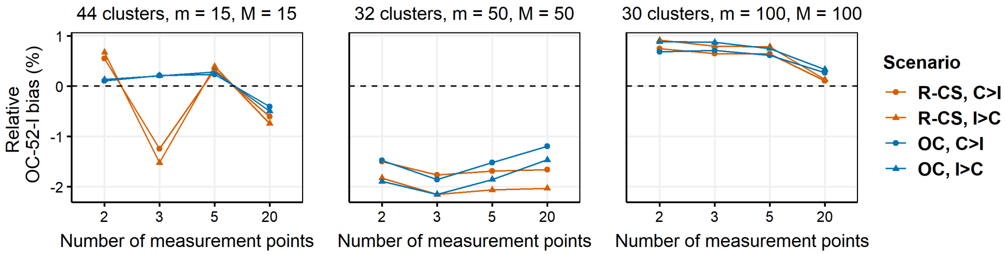


Figure 5.20: Relative OC-52-I bias in the base case using two timescales for full-samples.

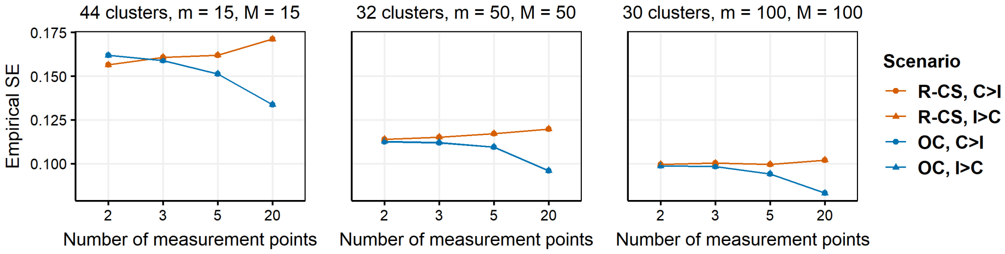


Figure 5.21: Empirical SE for OC-52-I in the base case using two timescales for full-samples.

Two timescale, sub-samples

When is OC sampling superior to R-CS sampling for sub-samples?

When the R-CS/OC designs are distinct, the OC design is less biased than R-CS for all measurement points for 15,50, for none of the measurement points for 15,100, and for 5 measurement points only in 50,100 (Figure 5.22). The OC design has improved or equal precision compared to R-CS for 2, 3 and 5 measurement points in all cases, but not 20 (Figure 5.23).

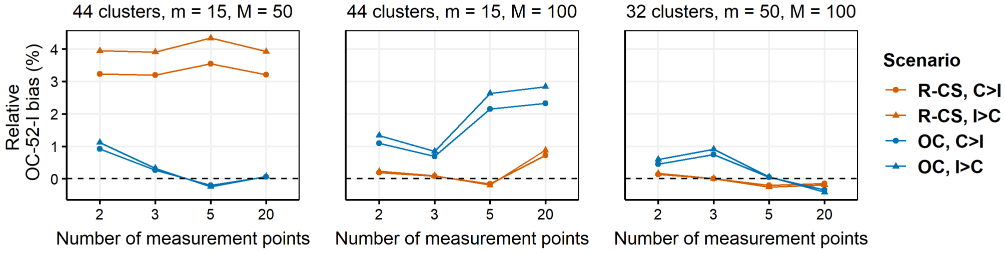


Figure 5.22: Relative OC-52-I bias in the base case using two timescales for sub-samples.

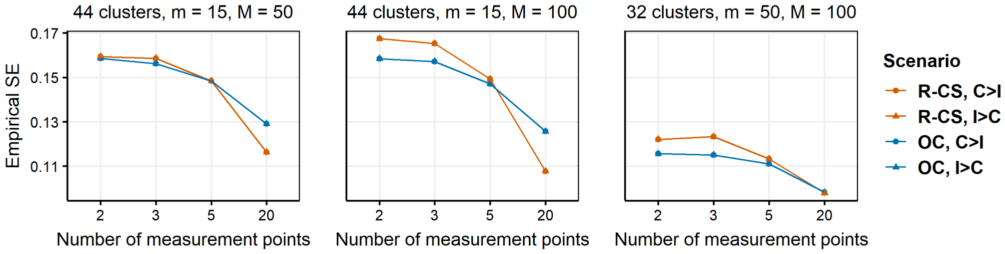


Figure 5.23: Empirical SE for OC-52-I in the base case using two timescales for sub-samples.

5.3.4.2 Complications

Two timescales, full samples

When is it beneficial to link measurements (when is OC superior to R-CS)?

In full-samples the OC design always provides an improvement in bias over R-CS when the missing data mechanism is MAR or MNAR and the turnover is greater than 20%; this also occurs in other scenarios but these depend on C and I values, turnover and the intervention effect (Figure 5.24, Table 5.18). The OC design has slightly improved precision over R-CS in all cases by linking measurements (Figure 5.25, Table 5.18).

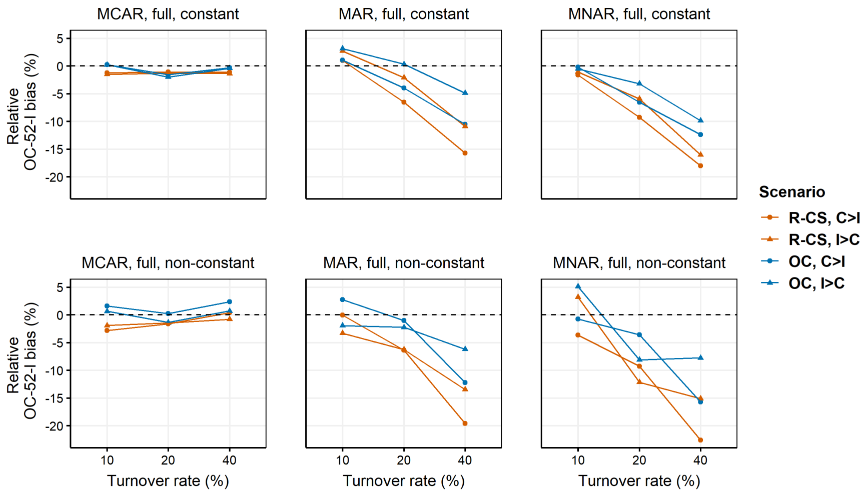


Figure 5.24: Relative OC-52-I bias in the complications scenario using two timescales for full-samples.

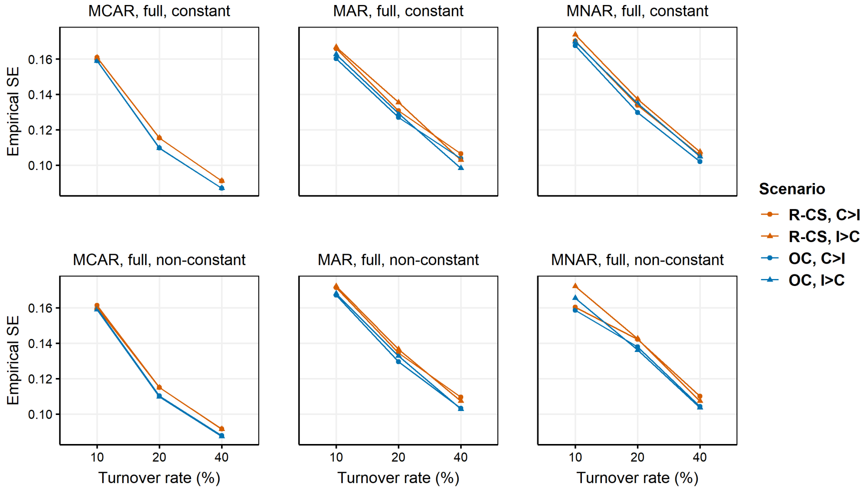


Figure 5.25: Empirical SE for OC-52-I in the complications scenario using two timescales for full-samples.

Two timescales, sub-samples

When is OC sampling superior to R-CS sampling for sub-samples?

In sub-samples, the OC design is always less biased than R-CS again for MAR and MNAR with turnover greater than 20%, as well as almost all MCAR scenarios (Figure 5.26, Table 5.19). OC and R-CS have similar precision but which one is superior depends on the missing data mechanism, turnover and intervention effect (Figure 5.27).

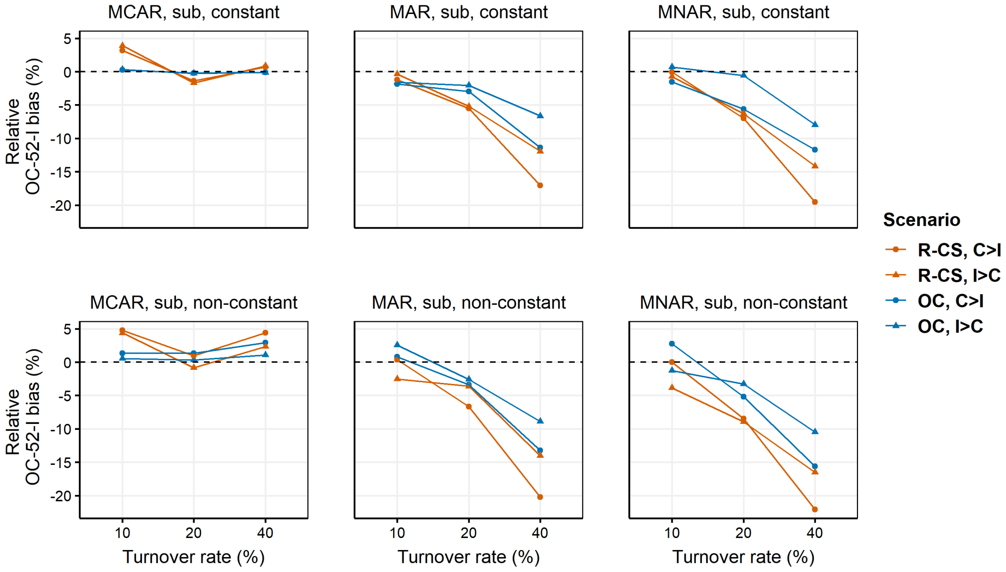


Figure 5.26: Relative OC-52-I bias in the complications scenario using two timescales for sub-samples.

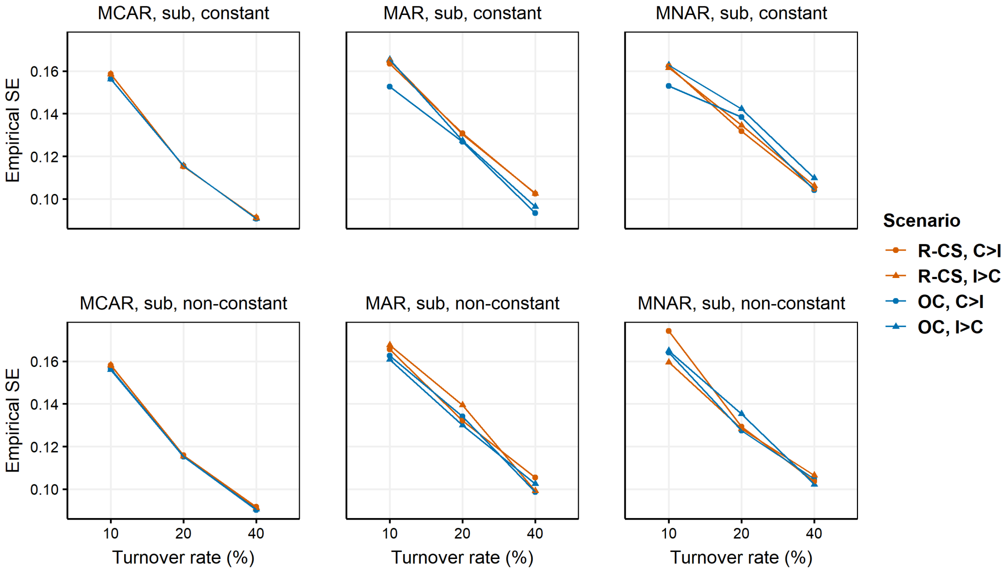


Figure 5.27: Empirical SE for OC-52-I in the complications scenario using two timescales for sub-samples.

Superior to R-CS?		MCAR	MAR	MNAR
Full, bias	Constant	✓✗✓	C ✓✓	✓✓✓
	Non-constant	✓✓I	I ✓✓	C ✓✓
Full, precision	Constant	✓✓✓	✓✓✓	✓✓✓
	Non-constant	✓✓✓	✓✓✓	✓✓✓

Table 5.18: A summary of whether the OC design provides an improvement over the R-CS design for the OC-52-I estimand in full-samples for the complications. A tick denotes where both values are better for the OC design, a cross denotes where both values are worse for the OC design, and a letter denotes that one of the two values is better for OC, with I = I>C and C = C>I.

Superior to R-CS?		MCAR	MAR	MNAR
Sub, bias	Constant	✓✓✓	✗✓✓	I ✓✓
	Non-constant	✓I ✓	✗✓✓	I ✓✓
Sub, precision	Constant	✓✗✓	C ✓✓	C ✗C
	Non-constant	✓✓✓	✓I C	C C I

Table 5.19: A summary of whether the OC design provides an improvement over the R-CS design for the OC-52-I estimand in full-samples for the complications. A tick denotes where both values are better for the OC design, a cross denotes where both values are worse for the OC design, and a letter denotes that one of the two values is better for OC, with I = I>C and C = C>I.

5.3.5 Immortal open cohort with 26 weeks of exposure (OC-26-I) estimand

Trends for the OC-52-I and OC-26-I estimands are similar but for OC-26-I the magnitudes of the empirical SEs are approximately double those of OC-52-I (Appendix, Figures B.4-B.11). For sub-samples in the base case, the OC design also has improved bias for OC-26-I compared with OC-52-I for sub-samples of 15,100 and 50,100, but for full-samples OC-26-I has higher bias for smaller sample sizes of 15 and 50.

5.3.6 Convergence

In the base case, for all designs and for both the discrete and continuous models, as the number of measurement points increase the number of convergence errors decreases and approaches zero (Tables B.6-B.10). Across different cluster sizes, the designs are all similar with convergence error rates of less than 5%, but in the case of a full-sample with the smallest cluster size of 15 and small numbers of measurement points, the R-CS has up to 15% convergence errors. In this case, linking measurements as in the OC design reduces the convergence errors to less than 5%. For small cluster sizes, estimation of cluster-level random effects, particularly cluster-period, will likely be more difficult. Whilst all of the designs will be affected by this, the R-CS dataset's independent individual ID numbers and lack of an individual random effect could also be compounding with this, resulting in more convergence errors.

For the complications the convergence differs slightly between the discrete and continuous models. Firstly, for the discrete model under the complications with one timescale, the R-CS design again exhibits worse convergence rates in full-samples which roughly decrease as turnover increases (Tables B.14-B.19). The OC and CC designs have similar convergence over full- and sub-samples. In sub-samples, all designs have increasing convergence errors as turnover increases with the OC design exhibiting the highest error rates of the designs. With two timescales the convergence is similar but R-CS has improved convergence for full-samples under MAR and MNAR as turnover increases compared to one timescale (Tables B.20-B.22).

For the continuous model under the complications, the trends are similar to those in the discrete model, but under MAR and MNAR the OC design has slightly more convergence errors, reaching between 10-12.5%, which is larger than both the one and two-timescale versions of the discrete model (Tables B.23-B.25).

5.3.7 Alternative open cohort sampling scheme

Another OC sampling scheme for sub-samples was laid out in Chapter 3, named "open cohort (wider cluster)" (Figure 3.2). When someone drops out, instead of sampling the replacement individual who enters the same bed as in the OC (beds) sampling scheme (Figure 3.1), the replacement individual could be sampled from the wider cluster population at that time. The probability of sampling an individual from the additional cohort was previously stated as $\frac{1+D_{ns}}{M-m+1}$, where D_{ns} is the number of drop-outs in those that are not sampled, equal to 4 in this example.

For open cohort estimands, this sampling scheme led to the presence of more original cohort individuals in the dataset, which subsequently meant it was more difficult for the model to distinguish between the two timescales and the worse the collinearity. This then led to larger bias, larger empirical SE's and larger corresponding MCSE's. This issue is worst when the probability fraction is minimised, which occurs when cluster population size (M) is large, the cluster sample size (m) is small, and the drop-out rate is low (represented by D_{ns}). These issues meant that whilst this sampling scheme was initially investigated, ultimately the OC (beds) sampling scheme (Figure 3.1) was chosen instead for this chapter.

5.3.8 When is it beneficial to use two timescales over one?

This question is only applicable to the estimands which can use either one or two timescales; namely, CC-I, CC-D and CS-OC-D. To answer this, figures with both the one and two timescale versions of the models are presented.

In the base case, the two timescale model for OC/R-CS in full-samples has similar bias to CC with one timescale (Figure 5.28). For sub-samples, again the OC with two timescales is similar to the CC with one timescale, but R-CS with two timescales is variable (Figure 5.29). Subsequently, in the cases where the CC design is less biased than OC, this is also when the two timescale version of OC is less biased than the one timescale version. For R-CS, sometimes the two timescale version is less biased and sometimes more, depending on cluster size.

The same conclusions apply to the precision. Given that the CC design always has the worst precision, the precision for the OC and R-CS designs with two timescales for full- and sub-samples is always worse when using two timescales compared to one (Figures 5.30 and 5.31).

In the complications, the OC design with two timescales has the same behaviour as in the base case (Figures 5.32 and 5.33). The behaviour of the R-CS design is more complicated; whilst it behaves the same as the base case under MCAR in the full-samples by having comparable bias to CC with one timescale, this is not replicated under MAR and MNAR and varies with intervention effect and turnover rate. In sub-samples it is not similar in bias to CC, even under MCAR. For precision, the OC and R-CS designs with two timescales have similar precision to CC in full-samples; again this means the designs with two timescales have worse precision than with one timescale (Figures 5.34 and 5.35).

Whilst the figures above are specific to the CC-I estimand, this behaviour is the same over the other two estimands (figures omitted). Overall, whether or not the second timescale is beneficial in terms of bias depends on the design and study parameters in some cases, as well as whether full- or sub-samples are taken. For precision, the one timescale model was always superior to the two timescale model as expected, likely due to having fewer parameters to estimate. If using just one timescale in the analysis model is better or as good as using two, the implication is that even if the underlying mechanism of cluster-level and individual-level processes is true in reality, trialists can continue to use single timescale analysis models in these cases without loss of precision or risk of further bias.

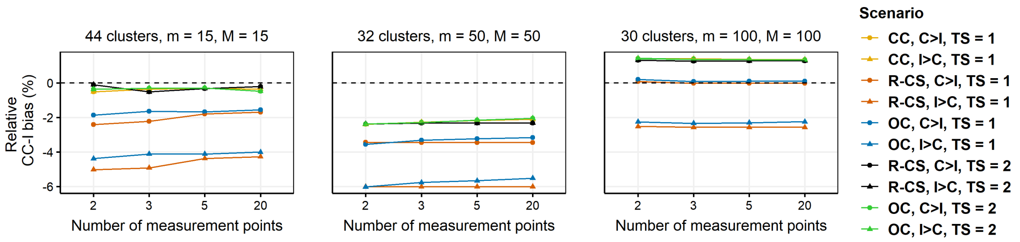


Figure 5.28: Relative CC-I bias with both timescales on the same graph, for full-samples in the base case. TS = number of timescales.

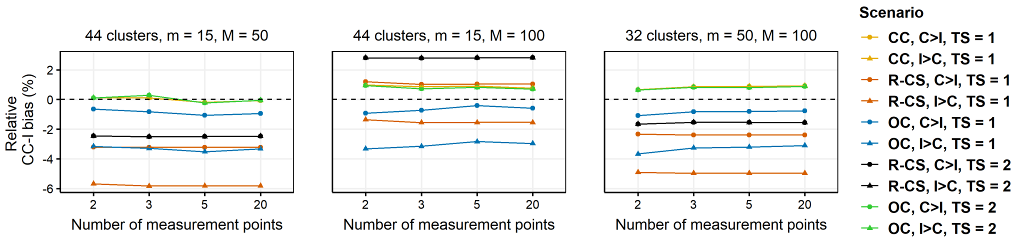


Figure 5.29: Relative CC-I bias with both timescales on the same graph, for full-samples in the base case. TS = number of timescales.

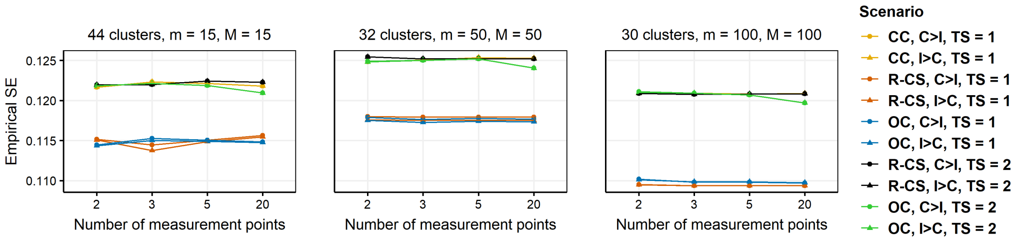


Figure 5.30: Empirical SE with both timescales on the same graph, for full-samples in the base case. TS = number of timescales.

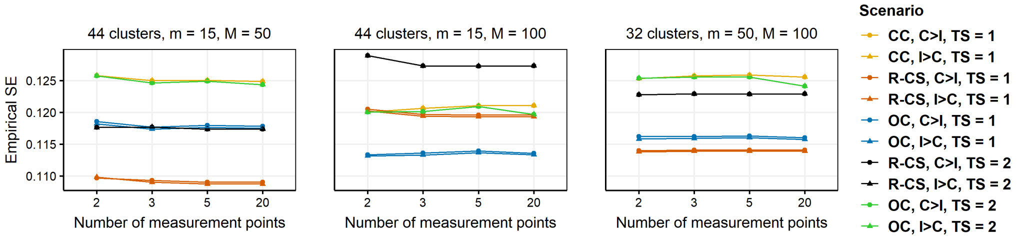


Figure 5.31: Empirical SE with both timescales on the same graph, for sub-samples in the base case. TS = number of timescales.

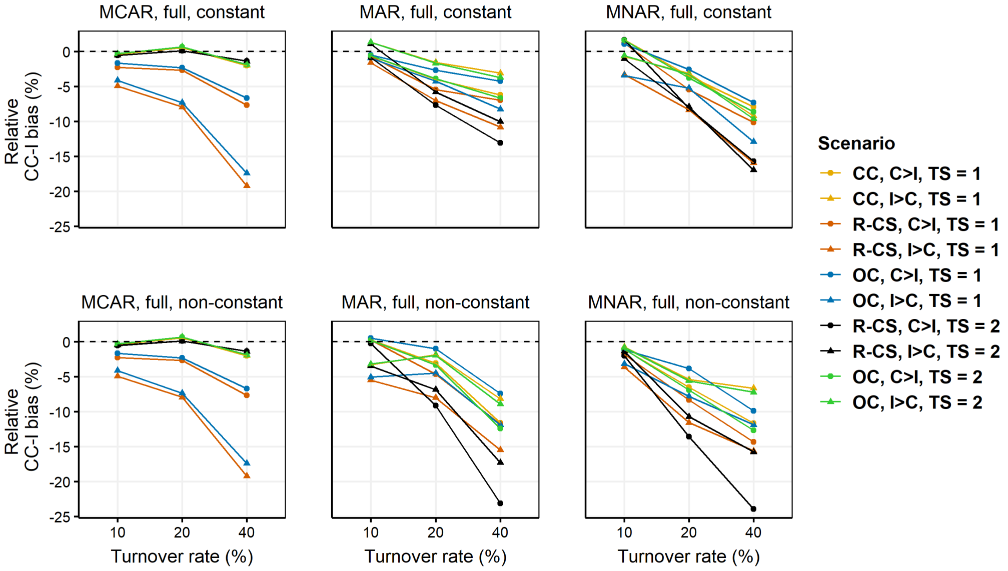


Figure 5.32: Relative CC-I bias with both timescales on the same graph, for full-samples in the complications. TS = number of timescales.

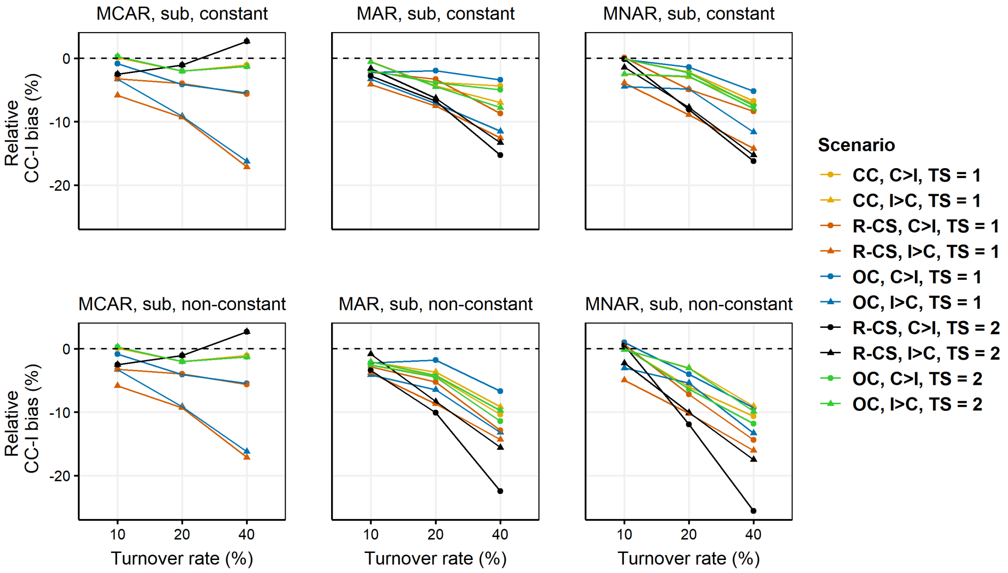


Figure 5.33: Relative CC-I bias with both timescales on the same graph, for sub-samples in the complications. TS = number of timescales.

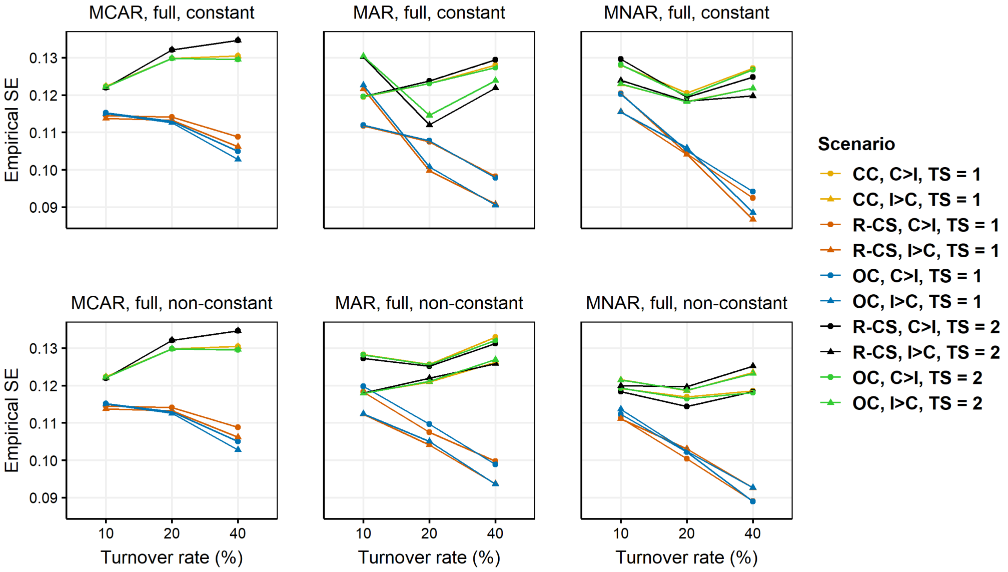


Figure 5.34: Empirical SE with both timescales on the same graph, for full-samples in the complications. TS = number of timescales.

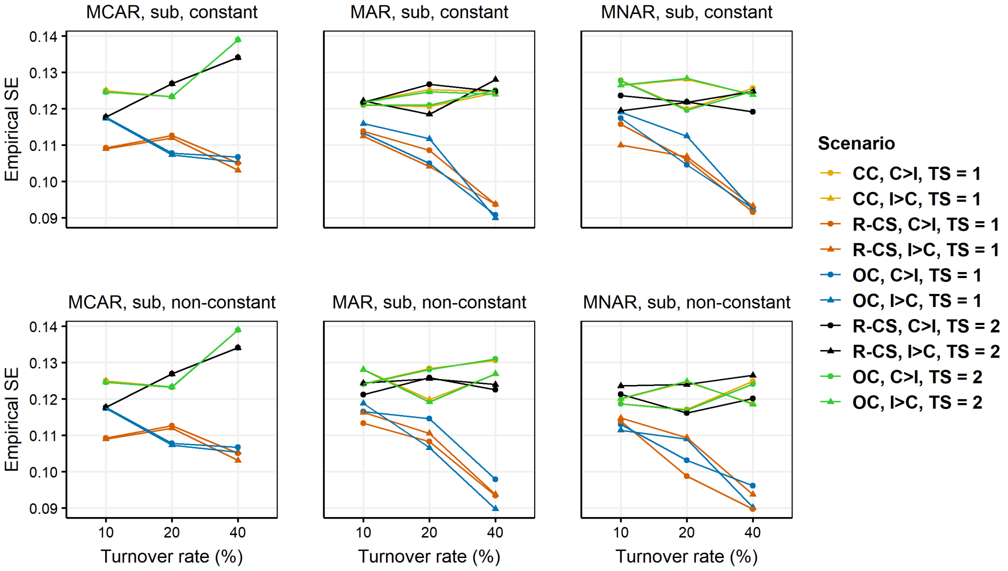


Figure 5.35: Empirical SE with both timescales on the same graph, for sub-samples in the complications. TS = number of timescales.

5.3.9 How many measurement/recruitment points should be used?

The number of measurements in the base case had no or very little effect across both precision and bias for the CC-I, CS-OC-D and CC-D estimands. This goes against the original hypothesis that more measurement points would lead to higher precision. However, all three of these estimands are based on differences between arms at a single time point, final follow-up, so this could explain why the number of measurement points preceding this does not matter. In practice, trialists considering these estimands can therefore opt for the lowest number of 2 measurement points, saving considerable time, money and resources, whilst knowing precision and bias will not be greatly affected.

In contrast, the number of measurements does have an effect on precision for the OC estimands in many cases, with more measurement points providing higher precision. The number of measurement points also affects bias for these estimands but these trends are inconsistent across scenarios. There is a possibility that the way time was treated for the different estimands could have affected how influential the number of measurement points was. For the CC-I, CS-OC-D and CC-D estimands, time was treated as discrete, but for OC estimands time was treated as continuous. If the timescales are treated as continuous variables, more measurement points provide more information for the calculation, whereas if they are treated discretely the other measurement points are calculated separately and do not improve precision. Moreover, when the CC-I, CS-OC-D and CC-D estimands were estimated using two timescales, time was treated as discrete in the model. To confirm the conclusions on number of measurement points, future work could look at repeating the base case using continuous time for the three former estimands.

5.4 Discussion

5.4.1 Summary of findings and implications

A broad conclusion of this simulation study is that the answers to the research questions posed are rarely simple, and a complex picture has been uncovered with several mechanisms operating simultaneously. Each answer depends on study parameters and complications as well as the estimand of interest, highlighting that these elements, previously not taken into account by other authors, are all crucial to consider when making comparisons between designs.

Findings and implications for each of the estimands will be summarised and discussed in turn.

5.4.1.1 CC-I estimand

These results agree somewhat with the hypothesis that CC designs would be optimal in terms of bias for CC-I estimands, with the CC design offering the lowest bias in many base case situations with low turnover and a simpler MCAR mechanism. There are however cases where the OC design could offer improvements over CC, even in the base case with a large cluster population size of 100, but more-so in the complications under MAR and MNAR mechanisms with 20-40% turnover; however, this only occurs when C>I. The OC design also had higher precision than the CC design in all cases in both the base case and complications. Therefore, if the cluster-level intervention effect is known to be stronger than the individual-level intervention effect, in situations where the turnover is high and MCAR may be too strong an assumption, the OC design could be a good choice.

With the single timescale model, the *ind.time* timescale is ignored and only the *cr.time* timescale included. I believe that the difference in bias between the C>I and I>C scenarios, for this estimand and the CC-D and CS-OC-D estimands, is directly linked to this. When C>I, the cluster-level intervention effect is stronger and so ignoring the *ind.time* timescale does not introduce as much bias as when I>C and the individual-level intervention effect makes up most of the overall intervention effect. The figures which include both timescales (Section 5.3.8) show that estimates differ in the one timescale model but in the two timescale model where the model aligns with the underlying DGM the estimates for C>I and I>C are the same.

5.4.1.2 CC-D estimand

It is not surprising that conclusions for the base case under the CC-D estimand are almost identical to those of the CC-I estimand, because the same population of the closed cohort is being targeted, and though CC-D incorporates drop-out, the drop-out is MCAR and minimal in the base case at 10%. This gives an example of how the mortal estimand can be approximately equal to its immortal version. However, turnover rate is very influential on this estimand; if low turnover is not expected and the missing data mechanism cannot be assumed to be MCAR, the R-CS design is then the design of choice. The OC design is not recommended for this estimand in terms of bias. This is an unexpected result and goes against hypotheses that the CC design would be least biased for the CC-D estimand. With a high turnover it is surprising that the R-CS design, where random sub-samples are taken of size m, has lower bias compared to the CC design where only the remaining individuals are measured. A possible explanation for this is that the linking in *both* the OC and CC designs is detrimental for the CC-D estimand at high turnovers, meaning the R-CS design comes out on top. This could also be the reason why the OC design can offer improvements over the CC design for CC-I but not for CC-D; the linkage is a key part of the designs and can potentially overpower differences in the sampling method and whether replacements are made or not. Despite its issues with bias, the OC design still had higher precision than the CC design in all cases in both the base case and complications.

Having said this, the CC-D estimand in itself could be seen as paradoxical in nature, as a closed cohort in the epidemiological sense as discussed in Chapter 1 is intended to represent a closed population from which there is no drop-out; in other words, it may not be an estimand which trialists would be interested in. However, consideration of the CC-D estimand alongside the CC-I estimand does distinguish between immortal and mortal estimands for the closed population.

5.4.1.3 Can the OC design estimate CC estimands as well as the CC design?

An interesting sub-question is whether a CC design is necessary if the aim is to estimate CC estimands, or whether an OC design can do equally well. From the previous two sections, the OC design can consistently provide improved precision under all of the scenarios explored here for both the CC-I and CC-D estimands. Under certain circumstances, the OC design can also provide improvements in bias for CC-I estimands, but not for CC-D estimands. Given that the CC design is always a subset of the OC design, opting for an OC design in practice would mean that trialists do not lose anything if interest is in the CC-I estimand, as the CC design could still be used for secondary analyses. However, as the OC design would be more expensive and require more resources to implement in practice due to recruitment post-randomisation, trialists would have to assess whether these gains are worthwhile. If interest is in the CC-D estimand, the lack of benefit in terms of bias would not warrant the extra costs and complexity associated with the OC design, despite improvements in precision.

5.4.1.4 CS-OC-D estimand

These results mostly agree with the hypothesis that the R-CS or OC designs would be least biased for the CS-OC-D estimand, but it is a surprise that in the base case for a full-sample of size 50 with one timescale the CC could be less biased, though this does depend on the intervention effect values C and I. The base case results suggest that in sub-samples with one timescale, either the OC or the R-CS design is least biased depending on study parameters, with the OC design providing improvements over OC when the cluster sample size is small. For two timescales, the OC is superior to R-CS only when the cluster sample size is small *and* the population is large.

For the complications, although the R-CS and OC designs are the least biased design in some cases, a similar unexpected result arises with the CC design with C>I offering near zero bias in some of the more complex scenarios. With one timescale the superior design is not so clear, but the OC design does not in general provide improvements for MAR/MNAR with turnovers of 20% or more. For two timescales, if relative values are known to be C>I and the missing data mechanism is not plausibly MCAR, the OC design provides improvements over R-CS, whereas if instead I>C then the R-CS design is recommended.

5.4.1.5 Precision for the discrete model

Starting with the base case, the OC design always has improved precision over the CC design in full- and sub-samples with one timescale as expected, and similar with two timescales. With one and two timescales, R-CS and OC are very similar in full-samples, again as expected. For sub-samples the OC design is only superior to R-CS with a small cluster sample size and a large population size (15/100), otherwise R-CS is superior; this applies for both timescales. Given that the gain in precision from linking measurements as seen in the full-samples is minimal, this could be due to the differing methods of sampling. In 50,100 and 15,50 the R-CS design would have more overlapping samples than in 15,100, which could be reason for its higher precision. In practice trialists should therefore be aware of the impact that cluster population and sample size have on the relative precisions of the designs when using the discrete model.

For the complications, a general improvement in precision with increasing turnover was seen with one timescale, which agrees with hypotheses. This could be driven by the presence of more individuals in the OC and R-CS datasets as turnover increases. The two timescale models consistently had poorer precision than their one timescale counterparts, and the same improvements with increasing turnover do not occur. Instead the trends differ depending on missing data mechanism, turnover rate, intervention effect shape and the study parameters. It could be that the benefit of more individuals is outweighed by the estimation of an extra timescale in the two timescale model to different degrees, depending on the situation. The loss of precision and unpredictable trends for the two timescale model are not appealing, so if the two timescale model cannot provide improvements in bias it should definitely not be used over the single timescale model.

The complications results would likely show the OC design to be more precise than R-CS if instead the situation of 15,100 had been assumed in the complications instead of 15,50, for which R-CS was more precise in the base case.

5.4.1.6 OC-52-I estimand

There is a caveat to this discussion section for the OC-52-I estimand in a base case scenario, that with a MCAR missing data mechanism and a low turnover, it is unlikely that the OC design or an OC estimand would be opted for. These results are therefore provided for completeness but are not likely to impact practice unlike the complications.

For the base case, the OC design does not lose anything for linking measurements in full-samples in terms of bias, and the only one loss in precision is small. Therefore, for full-samples linkage is recommended, which agrees with hypotheses. However, in sub-samples the OC design only has superior bias for the smallest cluster population size of 15,50, whilst having superior precision for 2, 3 and 5 measurement points. For sub-samples the OC design is therefore only recommended for this estimand if the cluster population size is small, for example sampling 15 from 50, as it is superior in terms of both bias and precision, but for larger cluster population sizes though precision is improved, the OC design can have higher bias.

In the complications, for full-samples linkage in the OC design gives equal or improved bias to R-CS, especially under MAR/MNAR with high turnover. Linking measurements is likely to have been beneficial in this case because the underlying DGM included repeated measurements from individuals, and the estimand is immortal. Precision is also better in every scenario for the OC design, albeit by a small amount. For sub-samples, though the precision is not always better for OC, the bias is always better in the open cohort scenarios and the few losses in precision are not notable. The OC design is therefore recommended over R-CS for the OC-52-I estimand in both full- and sub-samples, agreeing with hypotheses. The OC design is a good starting point for this novel estimand, particularly as it is useful in both full- and sub-samples. Having said this, it is important to note that the model applied to the OC design in this comparison is more correctly specified than the model applied to the R-CS design, so it is possible that the OC design had an advantage.

The OC design is mis-specified under MNAR and it is in this case that it exhibits the highest levels of relative bias of up to 15%. Alternative models should be explored that can provide lower bias for these estimands under such scenarios (see Chapter 6).

5.4.1.7 OC-26-I estimand

In the OC-26-I estimand results, similar trends to the OC-52-I results can be seen but the magnitudes of precision are approximately double in size. Indeed, after further investigation with a test dataset using OC-13-I and OC-65-I for comparison, it was found that estimates were less precise as the individual-time exposure decreased, with OC-65-I having the lowest empirical SEs and OC-13-I having the highest. When the two different intervention effects estimated have a greater magnitude of difference, and given that *cr.time* is multiplied by 78, multiplying *ind.time* by a larger value of 52 ‘cancels out’ this discrepancy more than a lower value of 26, resulting in less extreme estimates and improving precision.

This is an important property of the OC estimands that trialists should take into account if this estimand is of interest. However, as the value of individual-time increases the OC estimand becomes more similar to the CC-I estimand, which could be thought of a special case of the OC estimand with both individual- and cluster-time at 78 weeks, so there is a need to reduce the chance of having low precision whilst also selecting a value of individual-time that is distinct enough from the CC-I estimand.

5.4.1.8 Linkage of repeated measurements

CC-I is an immortal estimand and so assumes no drop-out. It therefore makes sense that, in full-samples, linking repeated measurements of individuals using an individual random effect in the model is beneficial in terms of bias compared to not linking, because the intra-individual correlation structure induces implicit imputation whereby predictions are made of individuals even if they drop out. In contrast, for the CC-D estimand where the same population is targeted but drop-out is taken into account, implicit imputation of the mixed model by including individual random effects is not desirable, so this makes sense again as to why, for full-samples, not linking measurements is preferable to linking measurements as in the CC-I estimand. Moreover, linking was beneficial for the OC estimands which are also immortal, and linking generally worsened bias for the CS-OC-D estimand in the complications. These results therefore suggest that linkage of repeated measurements is beneficial for immortal estimands, but can be detrimental for estimands which incorporate drop-out.

5.4.1.9 Implications for DCM-EPIC and other care home trials

DCM-EPIC had small cluster sizes of 15, and the average cluster size of the care home CRTs sampled in Chapter 2 was 11.9 (IQR 6.3-21.2). In DCM-EPIC, the limited cluster sizes meant that full samples had to be taken. Though a range of cluster sizes and both full- and sub-samples are presented in this chapter, in this section the specific case of DCM-EPIC will be focused on to assess the implications for the motivating example. DCM-EPIC is closest to the case of full-samples with a cluster sample size of 15, and given the difficulties already discussed from DCM-EPIC, the base case will be ignored and the complications with 40% turnover and a MNAR missing data mechanism focused on (for either a constant or non-constant intervention effect, and either C>I or I>C, and one or two timescales).

If a CC-I estimand is of interest, with one timescale the OC design has equal or less bias than the CC design but for C>I only, independent of intervention effect. With both one and two timescales, OC is less biased than R-CS for both C>I and I>C, independent of intervention effect. If a CC-D estimand is of interest, the R-CS design is less biased for both C>I and I>C and both intervention effects, over both timescales. If a CS-OC-D estimand is of interest, with one timescale OC is less biased than CC but for I>C weights only, independent of intervention effect. With one timescale, whether OC is less biased than R-CS is dependent on values of C and I and intervention effect, but with two timescales, OC is less biased than R-CS for C>I only. For the CC-I, CC-D and CS-OC-D estimands and one timescale, OC has better precision than CC, but R-CS and OC are very similar and which is better depends on the intervention effect and values of C and I. The same applies for two timescales.

Finally, for the OC-52-I estimand, the OC design is notably less biased than R-CS for both intervention effects and both C>I and I>C. Precision for this estimand is similar for OC and R-CS but OC is superior in all cases by a small amount.

In summary, whether or not the OC design provides benefits over the existing designs in the case of DCM-EPIC firstly depends on the estimand of interest, and can also depend on the relative value of intervention effects at individual- and cluster-levels, the shape of the intervention effect and number of timescales used. However, it is clear that the OC design is not useful in the case of the CC-D estimand, and definitely is useful in the case of the OC-52-I estimand. Noting that in full-samples the OC and R-CS designs have the same dataset but OC includes linkage means that linkage is beneficial for the OC-52-I estimand and is not for the CC-D estimand. These conclusions may also be applicable to the prison and palliative care settings which were most similar to the care homes in Chapter 2.

5.4.1.10 OC sampling scheme

The sampling of beds as opposed to sampling from the wider population (Section 5.3.7) was the chosen sampling scheme in this simulation study because this method provided less bias and improved precision for the OC estimands, which were of key interest. The choice of scheme also has practical implications as previously discussed in Chapter 3. Although the results were better for the OC estimands for this sampling scheme, for some estimands such as CC-I under some circumstances, sampling from the wider population provided less bias and improved precision. This makes sense as more of the original cohort would be sampled, which is exactly who the CC-I estimand is targeting.

5.4.2 Comparison to existing literature

This simulation study compares the novel OC design to existing R-CS and CC designs in terms of bias and precision for the first time. Feldman and McKinlay [14] compared CC and R-CS designs from an analytical point of view without simulation, with an emphasis on precision only. Whilst precision is clearly an important property of an estimator, arguably unbiasedness is of higher importance, but the authors only discuss bias in a general way, perhaps due to the lack of reference to estimands. This work adds to that of Feldman and McKinlay by suggesting possible estimands that trialists may be interested in targeting with the three compared designs, and as such recommendations are specific to estimands.

Based on bias and representativeness, from a general standpoint Feldman and McKinlay recommend the R-CS design when turnover is high because the CC design would "no longer be representative of the cluster" [14]. The results of this study, however, suggest that this depends on the estimand of interest; for the CC-D estimand the R-CS design was superior for high turnovers, but for the CC-I estimand the OC design was superior with high turnover. It is possible that, without considerations of whether estimands are mortal or immortal and the implications of the linkage that occurs in a mixed model, their conclusions are simply too broad.

There are other key differences between the work presented in this study and that of Feldman and McKinlay. Whilst treatment effect estimates in this chapter involved a constrained baseline approach, Feldman and McKinlay used a difference of differences approach, as discussed in Section 1.3.4.2. They also calculated the sample sizes of the R-CS and CC designs separately, allowing them to differ, whereas the approach taken herein created the CC and R-CS datasets from an underlying open population of individuals with a fixed sample size. Feldman and McKinlay's work also has a clear emphasis on 'large field trials', with the authors mentioning cluster examples of worksites, communities and schools. Though worksites were not included in the scoping review, communities and schools were, with communities in particular having much larger cluster sizes on average. The challenge of inherently small cluster sizes in DCM-EPIC does not appear to be a consideration in their work which, with a focus on larger cluster sizes, has less restrictions. A major drawback of Feldman and McKinlay's work is the very strong assumption of no missing data, which may not be as much of an issue in settings such as communities but is an inevitable part of trials in open cohort settings; in contrast, this work considers how recommendations would change based on different levels of turnover.

For some estimands, the R-CS design was found to be superior to the other designs when cluster sizes were small. Another of Feldman and McKinlay's general recommendations is that the R-CS design should not be used with small cluster sizes because the chance of obtaining overlapping samples would be high. This disagreement could be due again to the lack of specific estimands, but also to the difference between their version of the R-CS design (and McKinlay's [25]) and the one presented here. In the former, there is an emphasis on independent samples and a need to obtain non-overlapping samples, with a possible unwritten assumption that sub-samples are being taken from larger populations. In this work, I have firstly made a distinction between full- and sub-samples, to cater for small clusters where taking a sub-sample would not be possible. In my full-samples the whole cluster is sampled and so there is a guaranteed overlap in samples. The samples in the former's work appear to be closer to my version of sub-samples, where the R-CS sampling does actually take random samples as opposed to OC sampling. Overlaps will be minimised when the cluster sample size is small, the cluster population size is large and drop-out is high, but again with the difference in cluster sizes between this work and the former, overlap is something that is seen here as inevitable but as an inconvenience to avoid in theirs. For instance, McKinlay considers cluster sizes ranging from 148 to 771, whereas only 15, 50 and 100 are considered here.

Feldman and McKinlay add that the CC design should theoretically be more precise than R-CS for a given sample size when the correlation between an individual's responses at different time points (the IAC) is positive. On the contrary, this study found that in most scenarios the CC design was the least precise design of the three. This is likely to be due to the presence of at least 10% missing data in every scenario of this simulation study, something that Feldman and McKinlay did not consider, so the CC design would have less data than the other designs. They also go on to say that although theoretically the CC can be more precise, this can be outweighed by other parameters in the model, namely a small CAC, large cluster-level and residual variance and large cluster size. It could also be that the values they were fixed at in this simulation study, for example an IAC and CAC of 0.5, led to the CC design having the lowest precision in many cases.

Another disparity between this work and Feldman and McKinlay's is their recommendation against the CC design for large samples of the population due to the costs involved, and the recommendations presented here for the OC design with large cluster sizes for some estimands. Cost has not been considered in this work which is a limitation, but nevertheless this is an important factor for trialists to bear in mind in addition to bias and precision when choosing between the designs. With its extra waves of recruitment (at least one), the OC design is guaranteed to incur higher costs than the CC design. It could be argued however that the ability to answer more research questions with an OC design makes the extra costs worthwhile under some circumstances.

McKinlay [25] extended the work of Feldman and McKinlay by combining the unified model and a cost equation to compare the relative cost of CC and R-CS designs with equal precision. Unlike the former, they do take drop-out into account and incorporate the additional costs of tracing and measuring those lost to follow-up in their cost equation. They found that the CC design was more cost-effective for fixed precision when the IAC and CAC were high (greater than 0.75) and the trial period short. The one occasion where the R-CS design was more cost-effective than CC was when the trial period was long (5 years), the IAC and CAC were lower (less than 0.5) and drop-out was high, with the CC design incurring extra costs to follow up those lost. There is, however, an assumption made here that drop-outs are still alive, possibly as the clusters are again worksites, communities and schools. In the case of DCM-EPIC set in care homes for example, many of the drop-outs are due to death and so are unmeasurable, so there is no cost to be associated with this as it is impossible to rectify.

The recent work of Kasza [36] is most closely related to the contents of this chapter; an open cohort analysis model is presented, used herein, as well as a sample size calculation for open cohort designs. A major innovation of the work I have presented here is that my analysis models contain a second timescale, in comparison to Kasza's single-timescale open cohort analysis model. Kasza's work also presents three possible open cohort sampling schemes. As in this chapter, a steady state of individuals is assumed, and participants are allowed to contribute differing numbers of measurements, whilst keeping the number of participants sampled at each time point constant throughout the trial. Whilst the open cohort design discussed in this work and by Kasza is the same in these general terms, a key difference is that the sampling schemes appear to be motivated by different problems, and are approached with a different population or setting in mind to those discussed here, namely larger cluster sizes, which also occurred in Feldman and McKinlay's work [14]. Kasza *et al.* also appear to consider sub-samples only, which is another indication of their assumption of large cluster sizes.

The rotation sampling scheme presented by Kasza as a possible open cohort sampling scheme, described in Chapter 3, was suggested as a way of reducing measurement burden for participants. This again appears to have opposing motivations to those of this work. Whilst Kasza *et al.* may be concerned with measuring participants too much, a key concern from the perspective of this thesis is not measuring participants enough due to missing data; in other words, I anticipate missing data whereas the former do not. Having said this, measurement burden is undoubtedly important, particularly in settings such as care homes and palliative care where participants may be very vulnerable as highlighted in Chapter 2, so a balance between maximising repeated measurements and reducing measurement burden could be required.

In summary, the open cohort sampling scheme proposed here is unique and differs to all of those presented by Kasza, though it is most similar to the core group scheme; original participants are followed for as long as they remain in the trial, but newcomers outside of the core group are also retained in the sample for as long as they remain, rather than restricting them to one measurement only. Whilst Kasza's work focuses more on possible sampling schemes and analytical derivation of a sample size calculation which are not given in this thesis, their open cohort design is not compared to the established designs as it is herein. Moreover, the authors do not describe the specific estimands that would be targeted using the open cohort model, and whether or not the different sampling schemes lead to different estimands, whereas this work provides a starting point on estimands to consider in these types of design. In practice, both the work presented here and Kasza's are necessary for trialists to be able to both compare designs, decide which sampling scheme and/or design is most appropriate, which can then can be followed up with the sample size calculation.

5.4.3 Strengths, limitations and directions for future research

The use of a simulation study in this chapter is a strength as it allowed true values of underlying parameters to be set and the designs to be assessed against a known truth. It also meant that a wide range of scenarios could be studied as opposed to a single situation in any given trial. Moreover, the required number of simulations were calculated using requirements for MCSE's rather than using arbitrary numbers such as 1000 [4]. This meant that values of MCSE were controlled, as can be seen in the tables.

The sample size for each simulation was calculated loosely based on DCM-EPIC, the motivating example of this thesis, using its ICC and standardised effect size. To make the results generalisable to other settings, study parameters were also varied in the base case, to investigate larger cluster sizes as encountered in the scoping review, sub-samples in addition to full-samples, and a variety of measurement points. However, given the complexities in trials such as DCM-EPIC, it was not enough to assume simple missing data mechanisms and low turnover, so the study parameters were then compounded with complications. This allowed for assessment of a wide range of real life problems that trials in these settings can be faced with, including more complex missing data mechanisms, high turnover rates and non-linear intervention effects. The complications were varied in a factorial manner, which meant that interactions as well as main effects could be assessed; examining factors individually would have led to over-simplified conclusions, whereas in reality the findings are complex and provide opportunity for future research.

A downside of the high number of study parameters and complications is that some study parameters had to be fixed in the complications section, for instance a cluster sample size of 15, cluster population size of 15 or 50 and 3 measurement points. This means that other cluster sizes and numbers of measurement points were not investigated in combination with the complications, and the same conclusions in this study may not apply in these different situations. Similarly, some factors were only able to be varied over a small number of levels. For example, only two values of the relative cluster- and individual-level values were studied. Values that exist in different ratios or are equal to each other were not investigated and this topic alone warrants further investigation. Furthermore, only one non-linear cluster-level intervention effect shape was able to be considered alongside the linear case. Only an ICC of 0.1 is studied here because attempts with an ICC of 0.05 were much worse for convergence. The fixing of the individual and cluster autocorrelations (IAC and CAC) at values of 0.5 throughout is also a limitation that could be expanded upon. For settings where the clustering effect is not as strong, the same conclusions discussed herein may not apply; this again requires further work. Although MCAR, MAR and MNAR missing data mechanisms were investigated, these took one form only and therefore may not be applicable to some missing data mechanisms in practice.

Another strength of this work is that a variety of estimands were considered; the existing work in this area did not consider estimands at all. The estimands included were a mixture of both immortal and mortal estimands, and estimands that require either one or two timescales. Whilst this is a strength, bringing estimands into the problem could also be seen as a limitation as so many alternatives are possible, which is an area for future research. All of the estimands considered here focus on the intervention effect at a single point in time (78 weeks). Other estimands could instead potentially include only the individual timescale and make an adjustment for the cluster timescale, so an intervention effect for say six months of exposure could be estimated, without necessitating that this occurs at a particular time following cluster-randomisation. The idea of mortal OC estimands could also be explored.

The immortal estimands discussed in this chapter could be seen as addressing a hypothetical setting which may never exist and so may be less useful in the 'real world' if events such as death and drop-out are unavoidable, which is a limitation. I have also considered mortal estimands which compare outcomes of individuals who remain in the cluster at a particular time point. A disadvantage of using such a 'survivor analysis' is that, if an intervention increases the rate of death, the surviving patients in each arm will be different and thus a comparison between arms will be biased [219]. In a care home setting however, interventions are less likely to influence death or drop-out . These 'while alive' estimands may therefore suitable in a care home setting, but should be carefully considered if translating to other settings, and other strategies such as using composite endpoints could be necessary.

Several assumptions were made for the data generating mechanisms which are limitations of this work. A steady state of individuals was assumed, whereby a new individual replaces a drop-out the day after they have left a cluster; in reality this could be too strong an assumption and there may be variable gap times where the bed is empty. Compound symmetry or exchangeable correlation structures have also been assumed, such that the correlation between measurements at different time points is the same regardless of how far apart the time points are. This is a strong assumption and is likely not to be the case in reality, but this could be relaxed in future work using the methods of Kasza who illustrate how to implement structures that incorporate decay over time [36]. Equality of errors across arms was also assumed, but this could be relaxed by allowing errors to differ by arm. I also assumed equal cluster sizes and no intermittent missing data. Finally, an important drawback of designs which recruit post-randomisation is selection bias which was not investigated here. Relaxation of these strong assumptions provides potential avenues for future work.

The overall goal of this chapter was to compare designs across different estimands. This meant assuming a common analysis *method* across all designs and estimands; mixed effects models. I chose the models herein because they can be applied across all designs, and show the full capabilities of linkage using random effects; they were also the choice of Feldman and McKinlay for their comparison of designs [14]. However, a major caveat to this chapter is that mixed effects models are an analysis tool that assume immortality. The immortal estimands are therefore given an advantage in this chapter over the mortal estimands. For the mortal estimands, the results show that linkage in a mixed model is detrimental. An altogether better option for the mortal estimands may be to not use a mixed model at all and simply find the difference in means between arms at the desired time point. All of the conclusions for this chapter should therefore take into account that this is only the case with mixed effects analysis models. This will be returned to in Chapter 7.

In order to ensure fair comparison of the R-CS, CC and OC designs, I used the same underlying population and adapted it to sample the three designs. Filtering out the required number of time points gave the OC dataset; also filtering out individuals not in the original cohort gave the CC dataset. For R-CS, though there were repeated measurements from individuals in the underlying population, individual ID numbers were changed so that the R-CS model would treat the data as from completely different individuals. Because of this, the R-CS data could be at a disadvantage compared to data that is genuinely simulated from completely different individuals at different time points; if a correlation between the measurements at different time points does exist, ignoring this in the R-CS analyses could worsen precision and bias.

Interpretation of results in this work is difficult as there are so many different mechanisms at play. I believe there are at least three other important processes occurring in this simulation study which contribute to the complex results seen, in addition to the study parameters, complications and estimands. This includes the linkage of measurements, occurring in the CC and OC designs but not in the R-CS design, and whether new individuals are recruited post-randomisation or not, which occurs in the OC and R-CS designs but not in the CC design. There are also two different sampling methods used in sub-samples for the OC and R-CS designs. Attempts could be made to isolate these effects but these have their own limitations. For example, in full-samples, OC and R-CS use the same dataset, so the effect of linkage could theoretically be assessed on its own by comparing these designs. However, the effect of making the individual ID numbers independent so they can't be linked in the R-CS dataset is also included in this effect. In sub-samples, again when comparing OC and R-CS designs, there is a combined effect of linkage and sampling. If OC and R-CS full-samples are similar (indicating no effect of linkage), but a difference is seen between them in sub-samples, this effect *could* be attributed to the sampling. However, there could also possibly be an interaction effect when the linkage and sampling processes act together. In summary, it may not be possible to isolate these different effects here due to the setup of the simulation study, or it may not be possible at all for some effects due to the relatedness of the three designs. This could be an area for future research if comparison of the designs is of interest.

There are also complications of obtaining measurements in a care home setting in practice to consider. Agitation scores when a participant enters a care home are usually worse, with higher measurement error and therefore higher variance, because the individual is immersed in a new environment they are unfamiliar with which can cause distress. In practice it is therefore not ideal to take several measurements from a care home resident at the very beginning of their stay; allowing a settling in period would allow measurements to stabilise. These fluctuations at the beginning of a stay were not included in the DGM for both simplicity and because they may not have applied to other settings and outcome measures. Though complications such as this exist in real life, the main focus in the DGM is assumed to be the general trajectory over time.

Further practical considerations for trialists are necessary if two timescale models are to be used. In the simulations, if a participant left the cluster after 5 weeks and 3 days, the replacement participant would enter at exactly week 6, and their individual timescale would begin from zero at this point. There was therefore no distinction in the simulations between the time of entering the cluster and the time of recruitment. This links to Item 5 of the classification system in Chapter 4, where I said that in the OC design with discrete recruitment and the R-CS design, there could be a discrepancy between the time of entering the cluster and being recruited; this is less of an issue with continuous recruitment designs. Therefore, if the two timescales were to be used for these designs, trialists would have to collect the date of entry to the cluster in addition to the date of recruitment.

Four timescales were introduced in this chapter, but only time from CR and individual time were included. Length of stay was assumed not to affect outcomes in the simulations, but in reality an individual's presence in the cluster before CR could be a co-intervention if their length of stay before cluster-randomisation affects outcome. Individuals present for longer periods before CR in the care home setting may for example have lower agitation outcomes than those who are relatively new to the home before CR. Further work could explore the effects of length of stay, for example by using a covariate to adjust for this in the models. Likewise, the effect of calendar time was not considered in this chapter, as it is assumed that clusters are randomised all at once and that intervention delivery and measurement occurs on the same timescale across clusters. As discussed in Chapter 4, if these processes occur at different times across clusters then this could be another timescale to account for.

Another limitation in terms of time is the use of both discrete and continuous time in the analysis models. I began using discrete time in the analysis models for the CC-I, CC-D, and CS-OC-D estimands, but found afterwards that discrete time did not work for the OC estimands, and continuous time had to be used instead. It was not an intention to compare discrete versus continuous time, so this has not been done, but comparisons should not be made across estimands for this reason. Moreover, when continuous time is used in the third analysis model, a linear relationship is assumed which is not appropriate when there is a non-constant intervention effect at the cluster level. Extensions could be made to the model by treating time more flexibly; methods such as splines or interrupted time series analyses [207] could be explored.

Finally, possibly the strongest assumption that has been made in this work is that the total intervention effect in CRTs with a cluster-level intervention consists of a cluster-level intervention effect and an effect of individual time in cluster which operate on their own timescales. This assumption underpins the results of this chapter and Chapter 6 as it is a fundamental part of the DGM. Though the cluster-level intervention effect was varied between having a constant and non-constant rate, the individual-level effect was only assumed to be linear with time. The concept of the different intervention effects is believed to be novel, and this work provides a starting point for future developments.

Overall, broad generalisations are unable to be made from this work and an element of caution is advised in interpreting the results. Although different estimands were explored, they were not always hugely different in true value, and often the biases seen were fairly modest except in extreme cases. The comparison of such different designs meant that in most scenarios there was also unavoidable model mis-specification, so generalisations about the ‘best’ design are not always clear cut and should be carefully considered. To overcome this, trialists may wish to conduct simulations of their own using specific study parameters and complications known to their setting or intervention along with a chosen estimand of interest to further confirm a choice of design, using the code provided.

Chapter 6

Analysis methods for open-cohort parallel-group CRTs: a simulation study

6.1 Introduction

In Chapter 5, the novel OC design was compared to CC and R-CS designs for parallel-group CRTs in terms of bias and precision using a simulation study. The underlying DGM assumed an open population whereby individuals moved in and out of clusters over time, and two timescales: one for the time from cluster-randomisation, and another timescale for individual length of stay in a cluster. The three analysis models in Chapter 5 were all based on the model by Kasza [36], with some variations; two of these used discrete time, and included either one or both of the aforementioned timescales, and the third used continuous time. An important aspect of Chapter 5 was the introduction of five estimands, which were either mortal or immortal and varied according to the sampling design. Besides this model there are no other known models for the analysis of open cohort data in the literature.

There are three main drawbacks to using the Kasza model to analyse data from open populations. Firstly, the Kasza model is a mixed effects model, and is thus only valid when missing data are MCAR or MAR [220]. Use of the Kasza model when missing data are MNAR provides biased estimates; to obtain unbiased estimates with MNAR data, both the longitudinal and drop-out processes must be modelled jointly [221]. Secondly, because of the correlation structure used for individuals in the Kasza model, implicit imputation can occur whereby predictions are made about an individual's outcome at all times, even when they are no longer present in the cluster. For this reason, mixed effects models are a type of *unconditional* (on survival) model [60]. For open populations, where individuals move in and out of clusters over time and some are present in the cluster for short periods only, this implicit imputation would frequently occur. This 'linkage' was seen in Chapter 5, and whilst it can be advantageous for some estimands, namely immortal estimands, for mortal estimands the linkage can be detrimental and cause bias. An alternative approach, depending on the estimand, could involve making inference for individuals only when they are present in the cluster; this could be seen as being closer to mortal inference as opposed to immortal inference when using mixed models. The third potential weakness, not discussed in Chapter 5, is that the Kasza model includes a cluster random effect for all individuals with an assumed coefficient of 1, regardless of how long they have been present in the cluster. In other words, an individual present in the cluster for only 1 week and an individual present for 78 weeks would have the same cluster effect contributing to their outcome. An alternative could be to change the coefficient of the cluster random effect, to account for the differing lengths of stay for individuals.

In this chapter, I propose three types of analysis model for the analysis of parallel-group CRTs with open cohort designs which attempt to overcome these drawbacks of the Kasza model. The rationale for choosing these models and how they could overcome these drawbacks will be discussed in the subsequent sections.

6.1.1 Joint models

I will firstly look at joint longitudinal and survival models because they can theoretically overcome the first issue discussed above. Joint models can handle data that is MNAR by modelling the longitudinal and drop-out processes jointly; they are also known as *fully conditional* (on survival) models [60]. However, they do not provide a solution to the second issue discussed above as joint models still contain a mixed effects sub-model for the longitudinal data, and therefore implicit imputation can still occur when this is not desirable.

There are three main reasons why a joint model might be used [222–224]. Primary interest may be in the time to event outcome, but longitudinal outcomes, which are seen as time-dependent covariates measured with error, need to be accounted for. Alternatively, as in this case, primary interest could be in making inference on the longitudinal outcomes, but the missing data mechanism is believed to be informative/MNAR. As previously discussed, the only way to handle MNAR data and make unbiased inference is to jointly model both longitudinal and survival outcomes, acknowledging possible dependency between the two. GEE and mixed effects models assume MCAR and MAR respectively and may result in biased estimates. The third reason for using a joint model would be if interest is in the longitudinal and time to event outcomes as a joint process and it is desirable to investigate the relationship between them. As an added bonus, joint modelling can increase precision of both the longitudinal and survival estimates, particularly if there is a strong association between the two models [222].

In order to carry out a joint analysis of longitudinal data, two sub-models must be considered: a longitudinal sub-model for repeated measurements on individuals, and a time-to-event sub-model. There are three main ways to link the two models. A naive approach is to use observed longitudinal outcomes as time-varying covariates in the time-to-event sub-model alone, but this has been shown to be highly biased [222, 225]. Another method is the two-stage approach [226], where a model is fit to the longitudinal data in the first stage and the fitted values of this model are used as covariates in the time-to-event sub-model. Though the two-stage approach is superior to using raw longitudinal outcomes, it can still produce biased and inefficient results. Shared parameter models, where the two sub-models are linked by their inclusion of shared parameters, are seen as the best of the three methods [225] and are the chosen approach in this chapter.

A joint model involves the joint distribution of longitudinal and survival outcomes. There are three model families proposed in the literature for the joint distribution: shared parameter models (SPM), pattern mixture models (PMM) and selection models (SM) [227]. Using the full likelihood of the joint distribution, different factorisations lead to each of these models. Let $f()$ represent density distributions, y_i represent longitudinal outcomes from individual i, r_i represent missing data for individual i and θ represent ancillary parameters in the models (with θ_y for the distribution of y_i and θ_r for the distribution of r_i). Selection models include factors for the marginal density of the longitudinal outcomes and the density of survival conditional on longitudinal outcomes:

$$f(y_i, r_i; \theta) = f(y_i; \theta_y) f(r_i \mid y_i; \theta_r)$$

The pattern mixture model factorisation is the other way around, with factors for the marginal density of the survival outcome and the density of longitudinal outcomes conditional on survival:

$$f(y_i, r_i; \theta) = f(y_i \mid r_i; \theta_y) f(r_i; \theta_r)$$

For the SPM, given random effects b_i, the longitudinal data and missing data processes y_i and r_i are assumed independent, and so integrating over b_i results in the joint distribution:

$$f(y_i, r_i; \theta) = \int f(y_i \mid b_i; \theta_y) f(r_i \mid b_i; \theta_r) f(b_i; \theta_b) db_i.$$

If the primary interest is in the longitudinal sub-model, PMM are mainly used, whereas if primary interest is in the event sub-model, SM are usually chosen [228]. However, if the event sub-model has drop-out times which are continuous, PMM and SM cannot be used [212]. In the open population dataset that was generated in Chapter 5 and will be used again in this chapter, time to drop-out was calculated as a continuous variable in weeks, and individuals who reach the end of the study without leaving are censored. The joint model of choice for this simulation study is therefore the SPM.

6.1.2 Heteroscedastic models

In the Kasza model, a common individual-level variance is assumed across all time points, with $\mathbf{I}_{ijk} \sim N(0, \sigma_I^2)$. The second model relaxes this assumption and partitions the individual level variance to give a separate variance for each time point. I will refer to this model as the heteroscedastic (at the individual-level) model, or heteroscedastic model for short. This technique has previously been used in the literature on partially nested therapist designs [50], where individual-level variance is partitioned by treatment arm. This model could potentially overcome the issue of implicit imputation that occurs in mixed effects models, because if a person is not MAR and is actually ‘unobservable’ at a particular time point, this is accounted for in the model, and so inference is obtained only for those who are present. Two variations of this type of model will be presented in Section 6.2.5.

The heteroscedastic models are similar to what Kurland describes as *partly* conditional on survival [60]. In partly conditional models, the expected value of the response at time t is conditional on the individual being alive at time t. They can be estimated using mixed effects models, assuming independence between repeated measurements, or using GEE with an independence working structure. In comparison to a mixed effects model without individual random effects, the heteroscedastic models do contain individual random effects but they are partitioned using dummy variables for time. Kurland adds that partly conditional models describe the longitudinal trend for a dynamic cohort of individuals at time t, as opposed to change for individuals.

6.1.3 Cluster-weighted models

As introduced in Chapter 1, multiple membership models are typically used to model situations where a lower level unit belongs to 2 or more higher level units simultaneously [1]. The key part of MM models is that weights are used to implicitly handle time rather than explicitly as with random slopes or random coefficients, with some MM models not even including a fixed effect of time [229]. These weights are usually assumed to sum to 1 for each lower level unit [1]. For example, in the case of an individual belonging to cluster 1 for a quarter of their time, and cluster 2 for the remainder, the typical approach would set the weights as 0.25 and 0.75 for this particular individual. Here there is an assumption that the weights correspond to the proportion of time spent in a particular cluster, and this is the most natural and common approach. However, other approaches could be taken in calculating the weights depending on context. For example, particular clusters could be weighted more heavily, or more weight could be given to earlier/later clusters occupied over time [230].

To explore further the third issue of adapting the coefficients of the cluster random effect, I consider a variation of a multiple membership non-hierarchical model. Whilst the heteroscedastic model focuses on variances at the individual level, the weighted model introduces weighting at the cluster level, and will be referred to as the cluster-weighted model throughout.

6.1.4 Software and convergence

As this chapter focuses on analysis models, appropriate software and convergence properties are important considerations for each of the models. If any of the models are recommended for use following the results of this chapter, in order to become easily adopted in practice the packages must be available in standard statistical software and be able to be implemented in a relatively straightforward manner. Methods requiring new programming languages would require more time to learn and could prove to be less popular with statisticians. Furthermore, models involving large numbers of assumptions or complicated 'black box' methods are likely to be less well received by funders. Similarly, methods with good convergence properties in a simulation study are attractive to users as this increases the chances of a successful implementation on real datasets.

6.1.5 Estimands

As in Chapter 5, it is not enough to pose research questions in general without reference to specific estimands. As the heteroscedastic models in this chapter could potentially reduce bias for mortal estimands, it was important to include at least one mortal and one immortal estimand in this chapter. Consideration of an estimand that could use one as well as two timescales was also of interest, to assess how the proposed models fare when this changes. In addition to these requirements, as the design will be fixed as open cohort in this chapter, the estimands chosen should also have shown promise in Chapter 5 regarding bias and precision under an open cohort design. The OC-52-I estimand is an obvious choice for an immortal estimand due to the recommended use of an open cohort design for this estimand, following Chapter 5. For the mortal estimand, either CC-D or CS-OC-D could have been chosen. Whilst neither of these were as beneficial with an open cohort design as the immortal estimands, the latter was chosen as this could be seen as of more interest to trialists than the CC-D.

6.1.6 Aims and research questions

Whilst Chapter 5 focused on a comparison of designs and fixed the analysis model, in this chapter the design is fixed as the open cohort design, and alternative analysis models to the Kasza model will be explored. I aim to determine whether three alternative types of analysis model can provide an improvement upon the Kasza model, with respect to two estimands, in terms of convergence, bias and precision, when data is again from an inherently open population. The results of Chapter 5 demonstrated that the open cohort design was more beneficial in complex situations, so in this chapter the base case scenarios of Chapter 5 are not explored and I assess the analysis models over the range of complications only.

This chapter will be structured around answering the following research questions (RQ).

RQ1: For the OC-52-I estimand, are any of the four alternative analysis models superior to the Kasza model?

RQ2: For the CS-OC-D estimand, are any of the four alternative analysis models superior to the Kasza model when using a single timescale?

RQ3: For the CS-OC-D estimand, are any of the four alternative analysis models superior to the Kasza model when using two timescales?

The remainder of this chapter is organised as follows. The methods section provides a brief reminder of the data generating mechanisms used to generate the data and the estimands of interest, both of which were provided in detail in Chapter 5. The analysis models are then described in more detail, as well as the performance measures for this chapter and details on how improvements over the Kasza model will be calculated. Results of the simulation study follow, firstly summarising convergence, then structured around the research questions in terms of bias and precision. The results section concludes with a description of variations of the analysis models that were attempted but not presented in the results section for extra information. Finally, the work of this chapter is discussed in the context of the wider literature, and concludes with strengths, limitations and directions for future research.

6.2 Methods

6.2.1 Data-generating mechanisms

The same open population datasets that were generated for the complications scenarios in Chapter 5 are also used in this chapter. In the complications I vary: the missing data mechanism across MCAR, MAR and MNAR; the turnover rate of individuals between 10, 20 and 40%; and the intervention effect rate between constant and non-constant over time. As described in Section 5.2.1.3, the complications scenarios fixed several of the study parameters to make the number of variables manageable. The ICC is fixed at 0.1, as in DCM-EPIC, and because attempts at using an ICC of 0.05 in Chapter 5 lead to poor convergence rates. The cluster *sample* size is fixed at 15, as in DCM-EPIC, which led to using 44 clusters as per the sample size calculation given in Chapter 5. The cluster *population* size is varied between either 15 (for full-samples) or 50 (for sub-samples). All datasets in the complications considered three measurement points, at 0, 26 and 78 weeks. This was deemed suitable for this chapter because increasing to 5 or 20 measurement points led to computational issues with the joint model. As the design is fixed as the open cohort design in this chapter, and there were 216 scenarios when all three designs were used in Chapter 5, this results in $216/3 = 72$ scenarios in this chapter.

6.2.2 Estimands

In this chapter I will focus on two of the five estimands from Chapter 5. OC-52-I uses 2 timescales only, and targets the difference between arms for a cr.time of 78 weeks and an ind.time of 52 weeks. OC-52-I is an immortal estimand as its true value is taken directly from the DGM, assuming no drop-out. The CS-OC-D estimand targets the difference between arms at 78 weeks, and this can be estimated using either 1 or 2 timescales. This estimand is mortal as it takes into account drop-out.

6.2.3 Analysis models

In this section, details for each of the models are given in turn.

In Chapter 5, continuous time was used in addition to discrete time for fixed effects in the analysis models because use of discrete time for the open cohort estimands was unfeasible (Section 5.2.4). This did not pose a problem for answering the majority of the research questions where designs were compared within the same analysis model and within each estimand. One limitation however was that the number of measurement points was found to have no effect for the models and estimands where time was measured discretely, but did have an effect for the OC estimands in the model where time was measured continuously; as both estimand and the way time was treated were varied the effects could not be disentangled. Learning from this, all of the models used in this chapter use continuous time for the fixed effects, to ensure that any conclusions made about time are not due to whether time is treated as fixed or continuous.

6.2.4 Joint model

In the majority of joint modelling literature published to date, the individual is assumed to be the only clustering factor, with measurements (level 1) clustered within individuals (level 2) [231]. Most of the proposed methodology applies therefore to individually randomised trials only. I begin this section with an overview of standard joint models where this assumption is made, and then extend to the situation where clustering can also occur above the individual (level 3), as in CRTs. Following this I describe how the use of a second timescale in the mean model is implemented in the joint model, as this is also potentially novel within this field.

6.2.4.1 Standard model setup

The standard SPM consists of an event sub-model and a longitudinal sub-model as follows.

Event sub-model

The hazard of an event at time t for individual i is

$$h_i(t \mid M_i(t), w_i) = h_0(t) \exp\{w_i^T \gamma + \alpha m_i(t)\} \tag{6.1}$$

given $M_i(t) = \{m_i(s), 0 \leq s < t\}$, the history of the true unobserved longitudinal process for individual i up to time t, and a vector w_i consisting of covariates [227]. This is a proportional hazards model, meaning that the covariates w_i are assumed to have a multiplicative effect on the hazard. The baseline hazard function is given by $h_0(t)$ and corresponds to the hazard for an individual whose covariates w_i are zero. The vector of regression coefficients corresponding to w_i is given by γ. Let $m_i(t)$ represent the true, unobserved value of the longitudinal outcome for participant i at time t with $m_i(t) = y_i(t) - \epsilon_i(t)$. The difference between $m_i(t)$ and $y_i(t)$ is that the latter is contaminated with measurement error whereas the former is not.

Longitudinal sub-model

Longitudinal measurements $y_i(t)$ for individual i at time t are given by

$$y_i(t) = \mu(t) + \epsilon_i(t), \quad \epsilon_i(t) \sim N(0, \sigma^2), \tag{6.2}$$

where $\mu(t)$ is the mean function and the error terms $\epsilon_i(t)$ are mutually independent and independent of the random effects.

The association parameter, α, captures the association between the underlying longitudinal outcome and the time to event outcome. If $\alpha = 0$ this indicates a lack of association between the two, whereas if $\alpha < 0$, as the longitudinal outcome increases the hazard decreases, meaning that higher values of the longitudinal outcome lead to increases in the time to event [232]. A value of $\alpha = 0$ means that the extra information supplied by the longitudinal model does not improve on the estimate found when using just the survival outcomes.

6.2.4.2 CRT model setup

In CRTs, an additional level of clustering exists above the individual, with clusters (for example, care homes) at level 3. Extending model (6.1), the hazard of an event at time t for individual i *in cluster* j is

$$h_{ij}(t \mid M_{ij}(t), w_i, B_j) = h_0(t)\exp\{w_i^T\gamma + B_j + \alpha m_{ij}(t)\} \tag{6.3}$$

where $M_{ij}(t) = \{m_{ij}(s), 0 \leq s < t\}$ is now the history of the true unobserved longitudinal process for individual i *in cluster* j up to time t. Let $m_{ij}(t)$ represent the true, unobserved value of the longitudinal outcome for participant i at time t with $m_{ij}(t) = y_{ij}(t) - \epsilon_{ij}(t)$. A shared frailty term could have been included in model (6.3) to induce correlation in the event times of participants within a level 3 cluster, but this model did not converge (see Section 6.3.4). As an alternative, B_j, a fixed effect for cluster j, is included in the event sub-model [231].

Longitudinal sub-model

The effects of clustering at level 3 are also contained within the longitudinal sub-model; this is sufficient because these cluster effects are assumed not to be associated with the hazard of the event.

Longitudinal measurements $y_{ij}(t)$ for individual i *in cluster* j at time t are given by

$$y_{ij}(t) = \mu(t) + u_j + \epsilon_{ij}(t), \quad \epsilon_{ij}(t) \sim N(0, \sigma^2), \tag{6.4}$$

where u_j is the cluster random effect which is time-independent. The error terms $\epsilon_{ij}(t)$ are mutually independent and independent of the random effects. The final model (6.4) does not contain a random effect for individuals for convergence reasons (see Section 6.3.4).

6.2.4.3 Sharing structures

A sharing structure refers to the way the longitudinal and event sub-models are linked. Model (6.3) uses a current value (or expected value) sharing structure due to its inclusion of $m_{ij}(t)$ in the event sub-model. There are a variety of sharing structures that can be chosen to link the two sub-models. In a review by Sudell, sharing or association structures were divided into two main groups, either containing both fixed and random effects, or random effects only [233]. The current value parameterisation is an example of a sharing structure with both fixed and random effects, because the value of the longitudinal outcome, including both fixed and random effects, is assumed to affect risk of drop-out. When using random effects only in the sharing structure, the amount by which an individual deviates from the population mean longitudinal trajectory affects their risk of drop-out. In other words, an individual's random effects directly affects risk of drop-out, and is independent of time.

Choice of sharing structure should be based primarily on the data, in particular the clinical context [233]. When choosing a sharing structure, thought should be given to what is the possible cause of informative drop-out. In DCM-EPIC, risk of drop-out was likely to be higher for individuals who were more agitated, so in this case the agitation outcome is linked to drop-out, and a current value or random effects sharing structure could be chosen. In contrast, sudden changes in agitation might be thought to precede drop-out, and an alternative such as the slope of the longitudinal trajectory (not to be confused with the random slope) could be used. Though many more exist, none of the commonly used sharing structures available as built-in options in joint modelling software match the underlying MNAR DGM from Chapter 5 exactly. The current value sharing structure was chosen here because it appears to be the closest match, but I will highlight in the following section the differences between them.

6.2.4.4 Comparison to the underlying MNAR DGM

Given that the joint model models the missingness process to be able to provide valid inference for MNAR data, a reminder of the underlying MNAR DGM from Chapter 5 (Section 5.2.2.6) and a comparison to the form of the hazard function used in the joint model is necessary.

The underlying hazard function for the MNAR DGM in Chapter 5 was given previously in (5.12) by

$$h_{ij}(t) = \exp(\delta_0 + \delta_1 y_{ij,t} + \delta_2 y_{ij,t+1} + b_j), \quad t_{i,t} \leq t < t_{i,t+1}.$$

The hazard for a particular individual is therefore influenced by the realised values (that is, including measurement error) of their most recent (observed) and next (unobserved) measurement. Contrasting this with (6.3) there are three main differences between the underlying MNAR DGM and the association structure used for the joint model.

Firstly, the standard current value association structure in (6.3) uses $m_{ij}(t)$ whereas the underlying DGM (5.12) uses $y_{ij,t}$. Given that $m_{ij}(t) = y_{ij}(t) - \epsilon_{ij}(t)$ there is a difference here of measurement error. Secondly, the current value association structure does not account for the previous value also present in the DGM. The third difference is that the random effect for cluster in the underlying DGM, b_j, is not replicated exactly in (6.3), but a fixed effect for cluster, B_j, is used instead.

6.2.4.5 Mean function

The mean function in both the standard and CRT version of the joint model above refers to the fixed effects portion of the longitudinal model, and is a *function* because hazard models are time specific and the mean model contains time-specific terms. As in Chapter 5, the mean function will vary depending on whether a one- or two-timescale version of the model is used.

From Chapter 5, equations (5.1) and (5.2), if a 2 timescale model is used, we have

$$\mu = A\, ind.time + B\,(cr.time \times trt) + C\,(ind.time \times trt) \tag{6.5}$$

where *cr.time* is time from cluster-randomisation and *ind.time* is the time an individual has been exposed to the intervention/control condition. A is the fixed coefficient for *ind.time*, and B and C are fixed coefficients for the arm by time interactions, where *trt* represents treatment arm k. The cluster-level intervention effect rate is given by $(cr.time \times trt)$, and $(ind.time \times trt)$ is the individual-level intervention effect rate. The 'total intervention effect rate' is the sum of the cluster-level and individual-level intervention effect rates.

For 1 timescale only, $cr.time = ind.time$, which reduces the mean model to:

$$\mu = A\, cr.time + D\,(cr.time \times trt), \tag{6.6}$$

with A as above and $D = B{+}C$, where D represents the total intervention effect rate.

Although the mean function is used in the same form within the joint models as in the Kasza model in Chapter 5, the extra timescale could affect the performance of the joint model (see Section 6.2.4.7).

6.2.4.6 Choice of baseline hazard function

To avoid making assumptions about its true form, the baseline hazard is often left unspecified. However, non-parametric methods are not able to provide absolute measures of risk as the parametric methods are, and non-parametric methods can underestimate standard errors of the parameter estimates [222]. For the parametric approach, standard survival distributions (eg. exponential, weibull, gamma) can be used to model the baseline hazard. I will assume an exponential baseline hazard because the DGM in Chapter 5 [212] used to generate the drop-out times also assumed an exponential baseline hazard and exponentially distributed survival times.

6.2.4.7 Software

There is a variety of software available for joint modelling with a frequentist approach - that is, using maximum likelihood - but many of the popular options could not be used in this project for different reasons. The `JM` package in `R` [234] is commonly used but cannot handle nested random effects. Similarly, the `joineR` package in `R` [235] does not mention clustering in its documentation. The `stan_jm` package in `R` [236] is one of the few packages that has functionality for more complicated nesting structures in the longitudinal sub-model, but unfortunately this package is no longer being updated by the author and so is not recommended for use. Both `stjm` [237] and `merlin` [238] in `Stata` can handle nested random effects within joint models, but `merlin` is arguably much more flexible and has less data preparation than `stjm`, so despite it being more recently developed, `merlin` was the choice for this project.

One limitation of `merlin` is that it does not support non-hierarchical random effects structures in the longitudinal sub-model, (that is, cross-classified or multiple membership structures). The closest nested model to the desired one- and two-timescale versions of the Kasza model for the longitudinal sub-model would have included times (repeated measures) nested in participants who are in turn nested in clusters; in other words, random effects for individual and cluster would be included. Random slopes for cluster could also be considered. However, `merlin` could not cope with a three-level nested model or with a random slope for cluster and resulted in non-convergence, so the final model includes only a random intercept for cluster.

When the sharing structure depends on time, integration for the survival sub-model needs to be carried out with respect to time, so the timescale being used needs to be specified in merlin. Currently, merlin does not have the option for users to specify more than one time variable. Whilst we can be confident that merlin will work as expected with the single timescale model, this creates an issue for the two timescale model. As such, the scenarios with the two timescale model in this chapter should be taken as an illustrative example only because both time variables are not able to be specified, leading to issues with the integration and the model potentially not estimating what it should be, even if it runs without errors.

The joint model will from here on be referred to in shorthand as JM. More technical details on model fitting are provided along with code in Appendix, Section C.2.

6.2.5 Heteroscedastic model

In this model the variance at the individual level, in previous models assumed to be common across time points with $\mathbf{I}_{ijk} \sim N(0, \sigma_I^2)$, is partitioned so that an individual variance exists for each time point.

6.2.5.1 Model setup

The response Y_{tijk} is given by

$$\begin{aligned} Y_{tijk} &= \mu + \mathbf{C}_{jk} + \mathbf{CT}_{tjk} + \sum_{t=1}^{T} \mathbf{I}_{tijk} D_{ti} + \mathbf{E}_{tijk} && (6.7) \\ &= \mu + \mathbf{C}_{jk} + \mathbf{CT}_{tjk} + \mathbf{I}_{0ijk} D_{1i} + \mathbf{I}_{1ijk} D_{2i} + \ldots + \mathbf{I}_{Tijk} D_{Ti} + \mathbf{E}_{tijk} && (6.8) \end{aligned}$$

where $\mathbf{C}_{jk} \sim N(0, \sigma_C^2)$ is the random effect for cluster j and $\mathbf{E}_{tijk} \sim N(0, \sigma_E^2)$ is the residual error. Random effects for individual i at time t are denoted by $\mathbf{I}_{tijk}$, and are constrained to be independent such that the covariance between them is zero:

$$\begin{pmatrix} \mathbf{I}_{1ijk} \\ \mathbf{I}_{2ijk} \\ \vdots \\ \mathbf{I}_{Tijk} \end{pmatrix} \sim N \left(0, \begin{pmatrix} \sigma_{I1}^2 & & & \\ 0 & \sigma_{I2}^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \ldots & \sigma_{IT}^2 \end{pmatrix} \right).$$

This zero covariance constraint improves estimation (see Section 6.3.4) but also means that $\mathbf{I}_{tijk}$ is not estimated for individuals for times t when they are not present in the cluster, which could be seen as a way of overriding implicit imputation. It could however be seen as unnatural to assume no correlation between measurements from the same individual, which is assumed in the other models (except the JM). As in Section 6.2.4.1, let μ represent the fixed effects which differ depending on whether the model has two timescales or one, given by (6.5) and (6.6) respectively. The term D_{ti} is an indicator which is equal to 1 if person i is present in the cluster at time t, and 0 if not.

As the partitioning of variance occurs at the individual level and there is only a simple random effect for cluster being used, there is the possibility to include the cluster-period random effect $\mathbf{CT}_{tjk} \sim N(0, \sigma^2_{CT})$, that was used in the Kasza model of Chapter 5. Two versions of the heteroscedastic model will be used; one with the $\mathbf{CT}_{tjk}$ term and one without, and will be referred to as HETW and HETWO respectively throughout.

6.2.5.2 Software

This model requires simple dummy variables to be created in a dataset prior to fitting the model which can be done in any standard statistical software package. The `mixed` function in `Stata` 16.1 [239] is then used to fit the model; this should also be straightforward to code in other programs. Code is provided in Appendix, Section C.3.

6.2.6 Cluster-weighted model

In this chapter, the general MM approach will be used, but with some key differences to the other assumptions. For the open population datasets of Chapter 5 I previously assumed that whilst individuals can move in and out of a cluster over time, they cannot re-enter once they have left, and they can only ever belong to one cluster. The weight is assumed to be the proportion of the entire trial period that an individual is present. This is calculated by subtracting their entry time from their leave time and dividing by 78 (in weeks). This was chosen because it takes into account each individual's length of stay and weights the cluster random effect accordingly.

6.2.6.1 Calculation of weights

Using the matrix notation of Cafri *et al.* [240], let Y_i be the vector of responses for individual i:

$$Y_i = X_i\theta + A_ib_i + Z_ic + \epsilon_i. \qquad (6.9)$$

The design matrix for fixed effects and corresponding vector of coefficients are given by X_i and θ, the random effects design matrix for individual effects and corresponding vector of random effects by A_i and b_i, and the random effects design matrix for cluster membership and corresponding vector of random effects as Z_i and c, with ϵ_i denoting the residual vector. Z_i is illustrated for three particular individuals $i = 1, 2, 3$, in the open cohort dataset with 3 measurement points at 0, 26 and 78 weeks. As the dataset is in long format, each individual has as many rows as they do measurements, with a maximum of 3. The corresponding vector of random effects for clusters, c, is given alongside, with u_j representing the cluster random effect for clusters $j = 1, \ldots, J$. As individuals do not move between clusters, there is only one entry in each row for an individual at a particular time point.

Individual 1 is a member of the original cohort and remains in the cluster until the end of the trial period. They therefore have a full set of 3 measurements with a '1' as the weight each time, and as they are a member of cluster 1 only these appear in the first column:

$$Z_1 = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 0 \end{pmatrix}, \quad c = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_J \end{pmatrix}.$$

Individual 2 is also a member of cluster 1 but doesn't enter the cluster until just after 10 weeks. Their calculated weight is therefore 0.866, and they miss the first measurement point at $t = 0$, so their two rows of the design matrix look like:

$$Z_2 = \begin{pmatrix} 0.866 & 0 & 0 & \ldots & 0 \\ 0.866 & 0 & 0 & \ldots & 0 \end{pmatrix}, \quad c = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_J \end{pmatrix}.$$

Individual 3 is a member of cluster 2 and so their entries appear in column 2 of the design matrix. They enter the cluster at just after 29 weeks, and so miss the measurements at $t = 0$ and $t = 26$, and only have a measurement at $t = 78$, with a calculated weight as

0.619:

$$Z_3 = \begin{pmatrix} 0 & 0.619 & 0 & \ldots & 0 \end{pmatrix}, \quad c = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_J \end{pmatrix}.$$

As individuals are only ever a member of one cluster, this formulation is equivalent to a nested model as given by Cafri [240], except the weights do not have to be 1.

6.2.6.2 Model setup

The response at time t for individual i is given by

$$Y_{tijk} = \mu + \mathbf{I}_{ijk} + \mathbf{C}_{jk} W_{ijk} + \mathbf{E}_{tijk} \tag{6.10}$$

where again μ represents the fixed effects in either the two timescales or one timescale models in (6.5) and (6.6) respectively. The individual random effect is given by $\mathbf{I}_{ijk} \sim N(0, \sigma_I^2)$, the cluster random effect by $\mathbf{C}_{jk} \sim N(0, \sigma_C^2)$, and the residual error by $\mathbf{E}_{tijk} \sim N(0, \sigma_E^2)$. Cluster and individual effects and residual errors are assumed independent of one another. In standard MM models the weights may be represented by W_{tij} with a t subscript for time to denote that weights may be different at each time point. In this case, the t subscript is dropped because it is assumed that individuals' weights are the same across all time points, so W_{ijk} represents the weight for individual i in cluster j and arm k. Standard MM models also have a summation over the weights in the model, but again as there is only one weight per individual, this is omitted. The cluster-period random effects of the Kasza model do not appear in this model, with time in cluster appearing instead as a coefficient for the cluster random effect which does not vary over time.

6.2.6.3 Software

This cluster-weighted model, which will be referred to as CW, requires design matrices Z_i for all individuals to be created in a dataset and the generation of weights prior to fitting the model. The `xtmixed` function in `Stata` is then used to fit the model; this should also be straightforward to code in other programs.

Code is provided in Appendix, Section C.4.

6.2.7 Recap of Kasza model

The Kasza model from Chapter 5 was previously given in (5.6) as

$$\begin{aligned}&Y_{tijk} = \mu + \mathbf{C}_{jk} + \mathbf{CT}_{tjk} + \mathbf{I}_{ijk} + \mathbf{E}_{tijk}\\&\mathbf{C}_{jk} \sim N(0, \sigma_C^2), \quad \mathbf{CT}_{tjk} \sim N(0, \sigma_{CT}^2), \quad \mathbf{I}_{ijk} \sim N(0, \sigma_I^2), \quad \mathbf{E}_{tijk} \sim N(0, \sigma_E^2),\end{aligned}$$

where the mean function, μ, is the same as across the other models. Continuous time was used for the OC-52-I estimand and discrete time for all other estimands.

6.2.8 Estimates extracted from each model

All of the models considered in this chapter include the same mean model of fixed effects, μ. With the exception of the Kasza model, all models also treat time as continuous in the fixed effects.

For the two timescale model, both the individual exposure timescale and time from cluster randomisation timescales are included as in (5.15):

$$\mu = A_{cts,2}\, cr.time + B_{cts,2}\, ind.time + C_{cts,2}\,(cr.time \times trt) + D_{cts,2}\,(ind.time \times trt). \quad (6.11)$$

The 'cts' subscripts again denote that these are the coefficients obtained when continuous time is used rather than discrete, and the '2' for two timescales. This model is used for both the OC-52-I and CS-OC-D estimands. The estimate of interest is then:

$$C_{cts,2} \times 78 + D_{cts,2} \times 52, \quad \text{for the OC-52-I estimand}$$

and

$$C_{cts,2} \times 78 + D_{cts,2} \times 78, \quad \text{for the CS-OC-D estimand.}$$

For the single timescale model, only one timescale is included:

$$\mu = A_{cts,1}\, cr.time + B_{cts,1}\,(cr.time \times trt), \quad (6.12)$$

where this time the '1' subscript denotes coefficients for the one-timescale model. The estimate extracted is $B_{cts,1} \times 78$ and only the CS-OC-D estimand can use this single-timescale model.

6.2.9 Hypotheses and model mis-specification

In this section I return to the research questions (RQ) and provide some hypotheses (H) as to what I think will be the case.

The sharing structure of the JM hazard function does not align exactly with the one in the underlying MNAR DGM as described in Section 6.2.4.4. Moreover, the longitudinal sub-model itself does not align exactly with the underlying DGM due to issues with convergence (see Section 6.3.4). Due to the additional complexities in specifying a joint model, the JM is therefore the most mis-specified of the proposed models. Across the missing data mechanisms, the JM is closer to being correctly specified when the underlying data mechanism is MNAR.

The CW and HETW/HETWO models do not have the complexities of the JM, using a single longitudinal model only without the event sub-model. For the CW model, the cluster random effect of the DGM is included but with a different coefficient, the individual random effect is included but the cluster-period random effect is missing.

In the HET models, both include the cluster random effect, and both include the individual random effect but this is split by time point with no covariance between time points. HETW also includes the cluster-period random effect whereas HETWO does not. HETWO is therefore more mis-specified than the HETW model compared to the underlying DGM.

The Kasza model aligns most closely to the underlying DGM, with all of the same random effects in common.

As explained in Chapter 5, mixed effects models assume that missing data is MAR, so correctly specified models should yield unbiased estimates under MCAR and MAR, but biased estimates under MNAR. The JM is the only model of those proposed that can theoretically provide unbiased estimates under MNAR.

In terms of fixed effects, as in Chapter 5 the two timescale model is correctly specified with regards to including both timescales, and the single timescale model mis-specified. When the intervention effect rate is constant, all of the models using continuous time are correctly specified with regards to this term, but when the rate is non-constant the models with continuous time assume a linear trend and are therefore mis-specified.

Mixed effects models, that is, the Kasza model and the CW models, should also provide unbiased estimates for immortal estimands, but not mortal estimands due to the implicit imputation that occurs. Typically, joint models, due to the longitudinal sub-model containing individual random effects, would also have this implicit imputation occur, but as the individual random effects could not be fitted in merlin the JM may actually predict more towards a mortal rather than an immortal cohort. Whilst the HET models are also mixed models, I hypothesise that due to the partitioning of the variance by time points, they avoid prediction to an immortal cohort so may provide more unbiased estimates for the mortal CS-OC-D estimand.

Unlike Chapter 5 where different designs were used, in this chapter the same open cohort design is used across all scenarios so the same number of individuals and measurements exist in the dataset. The precision is therefore entirely dependent on the analysis model.

RQ1: For the OC-52-I estimand, are any of the four alternative analysis models superior to the Kasza model?

H1: For the above reasons, the JM could offer improvements in bias over the Kasza model in the case of MNAR, but the mis-specification of the JM could outweigh this benefit. The omission of the random subject effect in the JM and therefore lack of implicit imputation may also disadvantage the JM for an immortal estimand. The CW model could also offer improvements in bias for this immortal estimand. In terms of convergence, the most complex model is the JM and with two timescales may exhibit more convergence errors. The JM may provide improvements in precision when there is a strong association between the longitudinal and event sub-models, or equivalently, if the missing data mechanism is MNAR.

RQ2: For the CS-OC-D estimand, are any of the four alternative analysis models superior to the Kasza model when using a single timescale?

H2: The HET models could potentially offer improvements in bias over the Kasza model for mortal estimands, for the reasons given above. The omission of the random subject effect in the JM could mean that implicit imputation does not occur which would benefit inference for a mortal estimand. As in H1, the JM may provide improvements in precision if the missing data mechanism is MNAR.

RQ3: For the CS-OC-D estimand, are any of the four alternative analysis models superior to the Kasza model when using two timescales?

H3: Same as H2. In terms of convergence, the most complex model is the JM and with two timescales may exhibit more convergence errors.

6.2.10 Performance measures

As in Chapter 5, the performance measures of interest are bias and empirical standard error. In this chapter, convergence is also considered a key performance measure as this is an important consideration when comparing different analysis models.

6.2.10.1 Meaningful improvement in performance measures

Figures for convergence will be provided which include results from all models including the Kasza model. For the other performance measures of bias and empirical SE, the Kasza model is not given in the figure but instead the *improvement* in relative bias and *improvement* in empirical SE are given on the y-axis by subtracting the equivalent Kasza model estimate from the other models using the formulae below. Figures containing the Kasza model and actual empirical SE's and biases are provided in the Appendix (Figures C.4 - C.9).

Improvements in relative bias are found by subtracting the absolute value of the relative bias under the new proposed model from the absolute value of the relative bias under the Kasza model:

$$\text{Improvement in relative bias (\%)} = |\text{relbias}_{Kasza}| - |\text{relbias}_{new}|,$$

where relative bias is given by relbias and calculated by substracting the true estimand, $\hat{\theta}$, from the value of the estimate, θ, and dividing through by the true estimand

$$\text{relbias} = \frac{\hat{\theta} - \theta}{\theta} \times 100.$$

Similarly, improvement in empirical SE is calculated by subtracting the empirical SE of the new proposed model from the empirical SE of the Kasza model and dividing through by the empirical SE of the Kasza model, where empSE represents the empirical SE:

$$\text{Improvement in empirical SE (\%)} = \frac{\text{empSE}_{Kasza} - \text{empSE}_{new}}{\text{empSE}_{Kasza}} \times 100.$$

In order to clearly separate models where improvements are small and not meaningful in practice from those which are larger and more impactful, the figures in this section have reference lines at 5% improvement, 0% and 5% worse. If a model can provide greater than 5% improvement over the Kasza model, this is deemed to be a meaningful improvement.

6.3 Results

Convergence results are presented first, followed by separate sections for each estimand, with bias and empirical SE discussed in each.

6.3.1 Convergence

Convergence for the 5 models taken forward with 1 timescale are shown in Figure 6.1. The CW model is the clear winner on convergence for both 1 and 2 timescales with no errors in almost every scenario. The heteroscedastic models also have consistently low convergence error rates across all scenarios. None of the complications, even in the most complex scenario where all complications are 'on' and at their worst level, affect the CW and heteroscedastic models.

The JM and Kasza models are clearly affected by turnover rate and missing data mechanism. The Kasza model has more convergence issues as turnover increases. For MCAR with 40% turnover and MAR/MNAR with 20% turnover or more, the JM has better convergence rates than the Kasza model.

The main convergence error for the Kasza models was related to singularities, which is where one or more of the random effects variances are estimated at zero, or close to zero. The cluster random effect was the effect to be estimated with zero variance. It could be that the Kasza model, with its cluster-period and individual in cluster cross-classified random effects in addition to a random effect for cluster mean that the model is overfitted, or too complex. The heteroscedastic models differ from the Kasza model by the partitioning of the variance of the individual random effect, and in these models there is no covariance between measurements from individual. When the turnover is higher, the Kasza model has more work to do in trying to estimate a covariance between measurements from the same individual, which could be difficult with few or only one measurement per individual, whereas the heteroscedastic models assume this is zero. The CW model is also simpler than the Kasza model, as it does not include the cluster-period random effect, which could again explain why its convergence is better.

Conclusions for the sub-samples are the same as those for the full-samples, but the JM also has better convergence than the Kasza model for MCAR with 20% turnover.

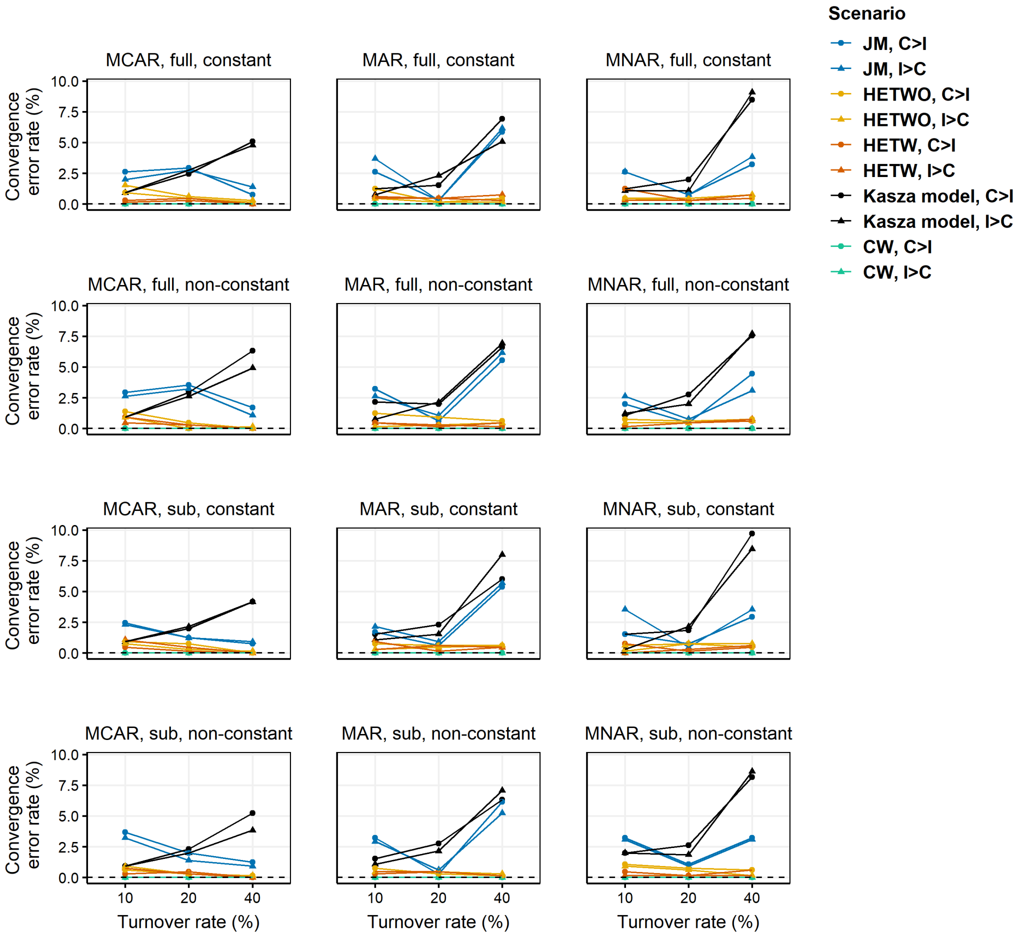


Figure 6.1: Convergence error rates for all models using 1 timescale. 'Full' denotes samples of 15 taken from cluster populations of 15; 'sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

While the CW, HETW, HETWO and Kasza models are fairly similar when changing from 1 to 2 timescales, the JM worsens massively for MAR/MNAR with 10 and 20% turnover, though it is still better than the Kasza model for MAR/MNAR with 40% turnover (Figure 6.2). This is expected due to the issues with `merlin` not currently being able to deal with more than one timescale, discussed in Section 6.2.4.7.

Conclusions for the sub-samples are the same as those for the full-samples.

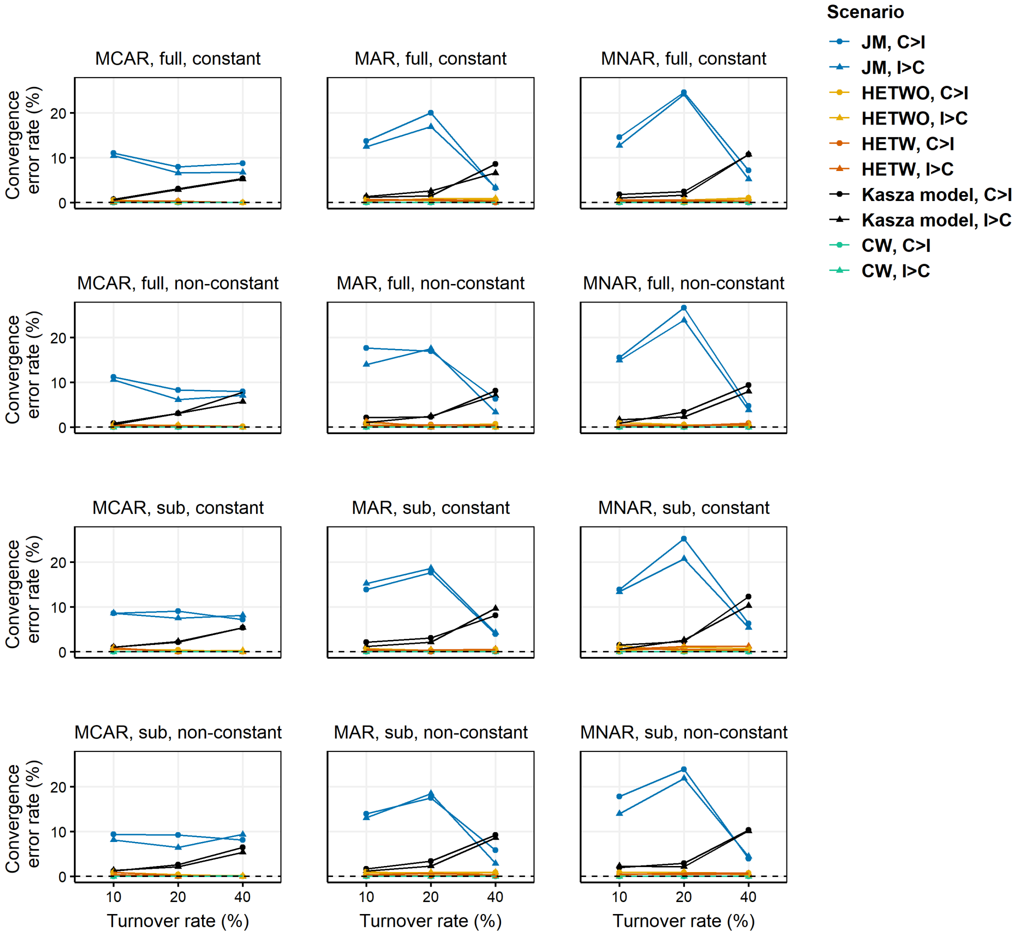


Figure 6.2: Convergence error rates for all models using 2 timescales for full-samples. ‘Full’ denotes samples of 15 taken from cluster populations of 15; ‘sub’ denotes samples of 15 taken from cluster populations of 50; ‘constant’ (‘non-constant’) denote a constant (non-constant) intervention effect rate.

6.3.2 OC-52-I estimand

6.3.2.1 OC-52-I bias

JM under every missing data mechanism with 40% turnover has very high bias (Figure 6.3). JM with C>I for 10 and 20% turnover has the highest bias in 10 cases out of 12. JM with I>C is generally less biased across 10 and 20% turnovers.

The only model that provides a meaningful improvement of just over 5% is the CW model with I>C in the case of MNAR with a constant intervention effect rate and 40% turnover. The CW model is the best model for MAR/MNAR with a constant intervention effect rate and turnover of 40%, but for a non-constant intervention effect HETW looks superior. Comparing constant to non-constant and excluding the JM, the HET and CW models are much closer to zero improvement for constant, whereas they are more spread out for non-constant. For constant, HETWO has a slight advantage over HETW for MAR/MNAR with 40% turnover but the reverse is true for non-constant.

Conclusions for the sub-samples are the same as those for the sub-samples, but none of the models provide a meaningful improvement of over 5% (Appendix, Figure C.1).

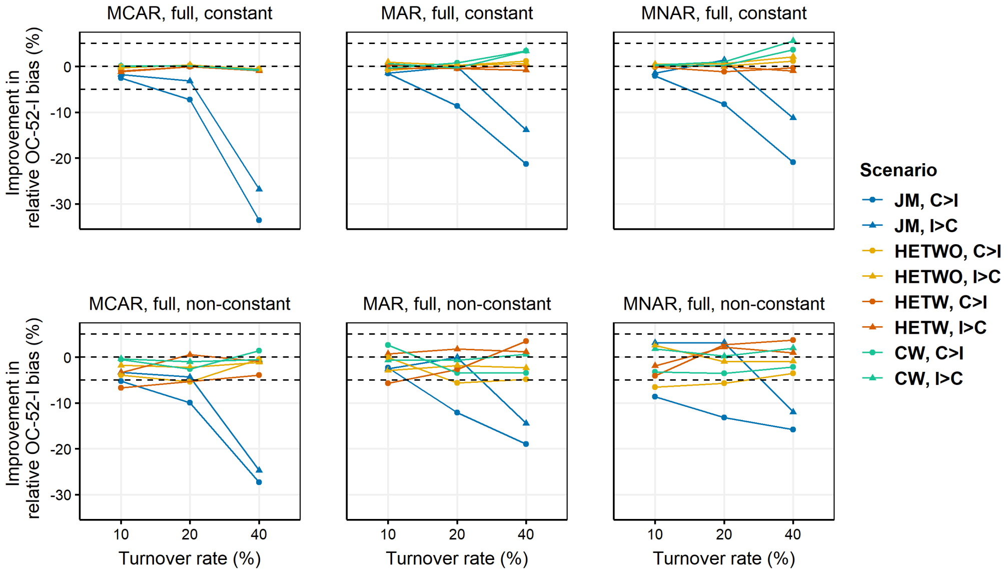


Figure 6.3: Improvement in relative bias (%) for all models for the OC-52-I estimand in full-samples in comparison to the Kasza model. ‘Full’ denotes samples of 15 taken from cluster populations of 15; ‘constant’ (‘non-constant’) denote a constant (non-constant) intervention effect rate.

6.3.2.2 OC-52-I empirical SE

The models are similar for MCAR, but under MAR and MNAR the JM has clear gains in precision, and gains are larger as the turnover increases (Figures 6.4 and 6.5). This gain in precision is outweighed by the high bias for the JM in Figure 6.3.

The CW model is the best under MCAR but for MAR/MNAR, and as the turnover increases, the CW model loses precision and aside from the JM, the HETW model is the best of the remaining three. The HETW model appears consistently better than the HETWO model in all of these scenarios.

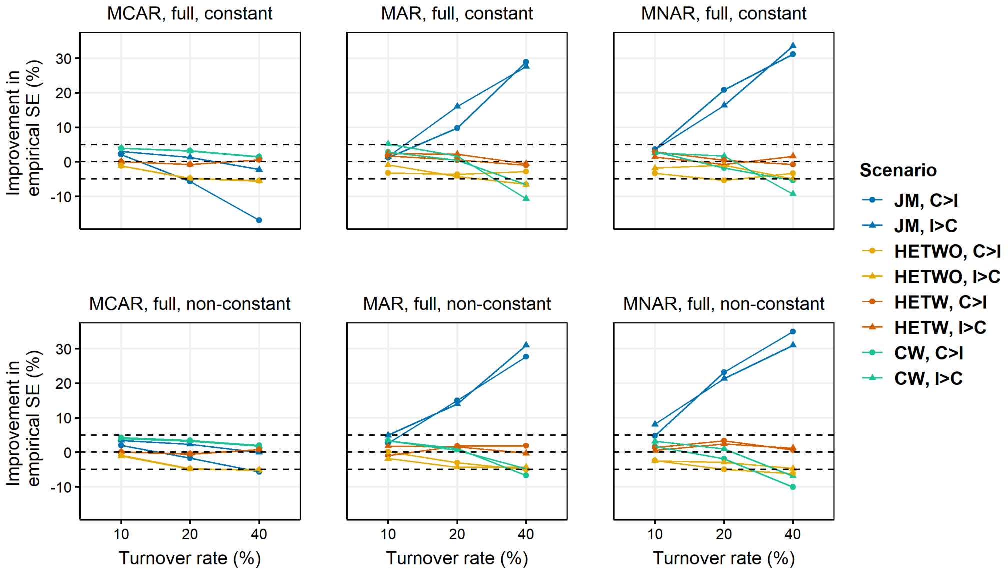


Figure 6.4: Improvement in empirical SE (%) for all models for the OC-52-I estimand in full-samples in comparison to the Kasza model. ‘Full’ denotes samples of 15 taken from cluster populations of 15; ‘constant’ (‘non-constant’) denote a constant (non-constant) intervention effect rate.

Whilst for the full-samples the CW model is most precise for MCAR, for sub-samples this is different with the HETW and JM also providing improvements. The CW model again loses precision under MAR/MNAR as the turnover increases.

The HETW model again appears consistently better than the HETWO model in all of these scenarios, and the JM shows large gains in precision for MAR/MNAR with turnover of 20% or greater.

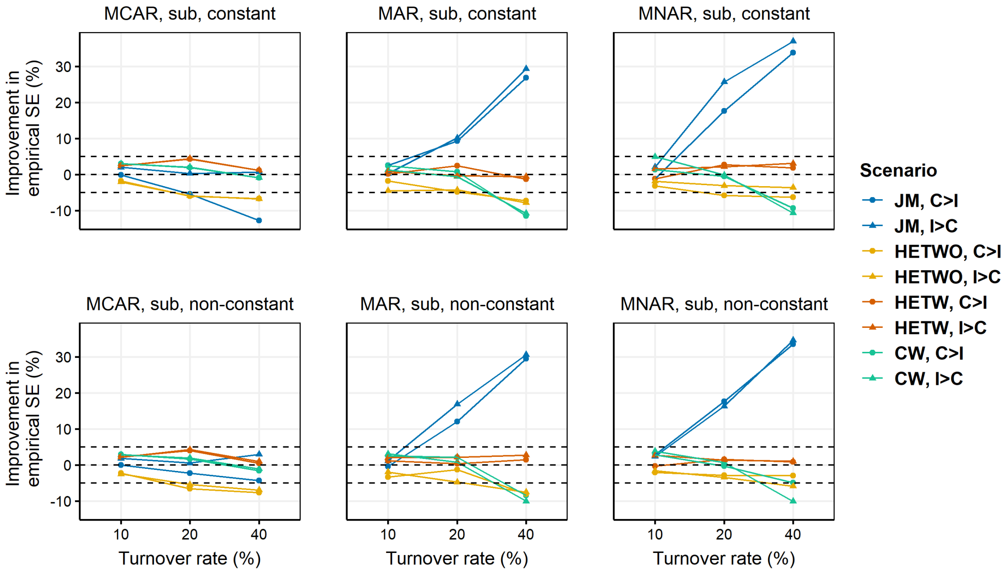


Figure 6.5: Improvement in empirical SE (%) for all models for the OC-52-I estimand in sub-samples in comparison to the Kasza model. 'Sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

6.3.3 CS-OC-D estimand

6.3.3.1 CS-OC-D bias with 1 timescale

With only 1 timescale, the JM is not as biased as in the previous section for the OC-52-I estimand with 2 timescales, but for C>I and constant intervention effect with 40% turnover it is 5% more biased than the Kasza model (Figure 6.6). In most of the scenarios, models are within the -5 to 5% improvement zone. The improvement exceeds 5% only for the CW model for MAR with I>C, with a constant intervention effect rate and 40% turnover. For a non-constant intervention effect the CW model under MAR/MNAR with 40% turnover the CW has 5% or more worse bias than the Kasza model in 3 cases out of 4. The HETW/HETWO models are closer to zero improvement for a constant intervention effect, but are much more spread out for non-constant, with HETWO looking superior for the majority of the non-constants.

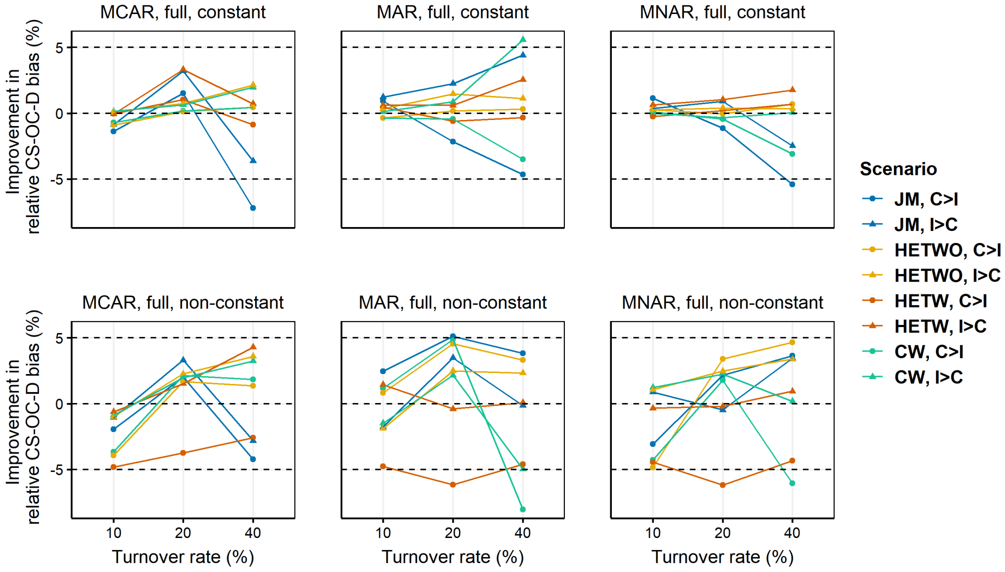


Figure 6.6: Improvement in relative bias (%) for all models for the CS-OC-D estimand in full-samples with 1 timescale in comparison to the Kasza model. ‘Full’ denotes samples of 15 taken from cluster populations of 15; ‘constant’ (‘non-constant’) denote a constant (non-constant) intervention effect rate.

Conclusions for the sub-samples are similar to those for the full-samples. The HETW/HETWO model improvements are again much closer to zero for constant intervention effect than they are for non-constant, and these trends are similar over full- and sub-samples. In most scenarios, models are within the -5 to 5% improvement zone, with none over 5%.

However, for the sub-samples the constant cases have improvements much closer to zero, particularly for MAR (Figure 6.7). The JM and CW models behave differently under sub-samples but are still within the -5 to 5% improvement zone. The CW model does not provide an improvement over 5% as it did for the full-samples, and performs generally worse. For sub-samples with non-constant intervention effect under MAR/MNAR with 40% turnover, the CW has worse bias by 5% or more than the Kasza model in 4 cases out of 4. The JM is much more stable and closer to zero for sub-samples compared to full-samples.

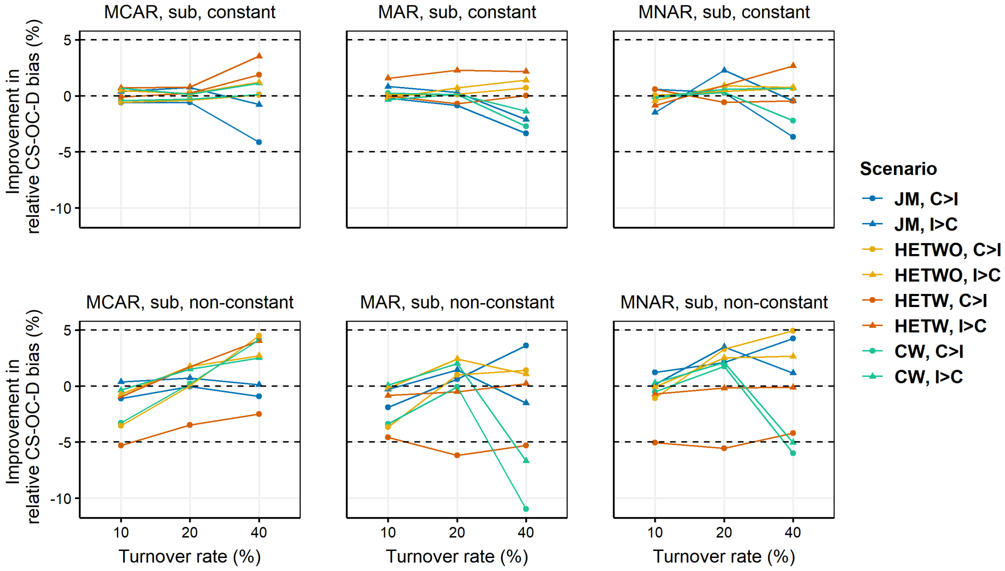


Figure 6.7: Improvement in relative bias (%) for all models for the CS-OC-D estimand in sub-samples with 1 timescale in comparison to the Kasza model. ‘Sub’ denotes samples of 15 taken from cluster populations of 50; ‘constant’ (‘non-constant’) denote a constant (non-constant) intervention effect rate.

6.3.3.2 CS-OC-D empirical SE with 1 timescale

For MCAR, there are no notable improvements in precision of the alternative models over Kasza's model, and the JM at 40% turnover is worse by more than 5% (Figure 6.8). For MAR/MNAR there is an obvious fanning effect as the turnover increases, with JM as the best, followed by HETW, HETWO and the CW worst.

For MAR/MNAR with 40% turnover, the JM has higher precision than Kasza, and this also occurs in some MAR/MNAR scenarios with lower turnover for the JM; only in MNAR with 40% turnover is this improvement by 5% or more. The other models do not provide any notable improvements in precision. The HETW/HETWO models generally only provide reductions in precision except for HETW which has small improvements under MAR/MNAR with 40% turnover. The CW has much lower precision than the other models in the cases of MAR/MNAR with 20 and 40% turnover.

There is more divergence in the models as the missing data mechanism becomes more complex, so the choice of model under a MNAR mechanism is more important than under MCAR.

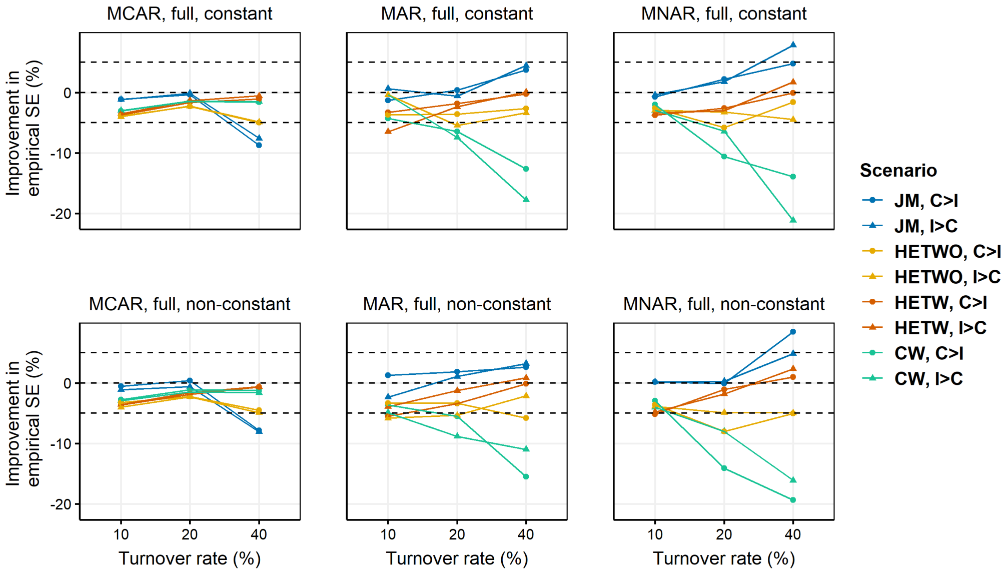


Figure 6.8: Improvement in empirical SE (%) for all models for the CS-OC-D estimand in full-samples with 1 timescale in comparison to the Kasza model. 'Full' denotes samples of 15 taken from cluster populations of 15; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

For sub-samples the conclusions are the same as for the full-samples, but the JM also provides some more improvements of over 5% in the case of MAR with 40% turnover (Figure 6.9). The improvements by the HETW model with I>C under MAR/MNAR with 40% turnover and non-constant intervention effect are also greater for the sub-samples.

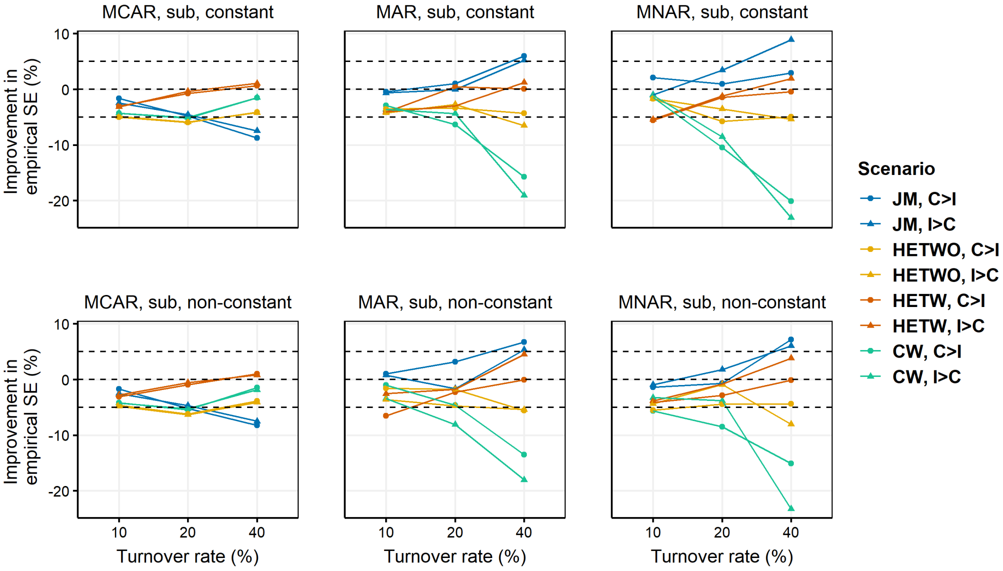


Figure 6.9: Improvement in empirical SE (%) for all models for the CS-OC-D estimand in sub-samples with 1 timescale in comparison to the Kasza model. 'Sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

6.3.3.3 CS-OC-D bias with 2 timescales

The JM performs poorly in terms of bias for this estimand with 2 timescales with C>I, yet makes a large improvement over the Kasza model in the case of 40% turnover under MCAR with I>C (Figure 6.10). Most other models fall between the -5 to 5% improvement zone, except the CW model with I>C for MAR/MNAR with constant intervention effect and 40% turnover and the HETWO model with C>I with non-constant and 40% turnover, which are both more than 5% worse than Kasza. As with this estimand using 1 timescale, the HETW/HETWO models are much closer to zero improvement in the constant intervention effect case compared to non-constant. Compared to 1 timescale, the difference between the HETW/HETWO models is not so obvious here and depends on the relative values of C and I, but the HETW model is better for both C>I and I>C under MAR/MNAR with a non-constant intervention effect.

Conclusions for the sub-samples are very similar to those of the full-samples (Appendix, Figure C.2).

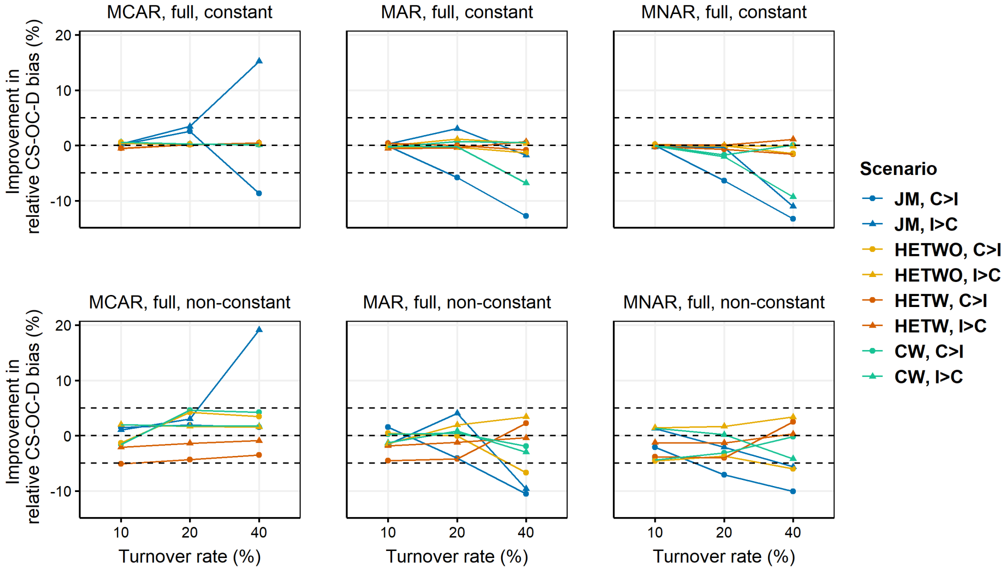


Figure 6.10: Improvement in relative bias (%) for all models for the CS-OC-D estimand in full-samples with 2 timescales in comparison to the Kasza model. 'Full' denotes samples of 15 taken from cluster populations of 15; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

6.3.3.4 CS-OC-D empirical SE with 2 timescales

The JM is slightly more precise than Kasza in some cases for MCAR but this is not meaningful, though it is also more than 5% biased in 3 cases out of 4 for 40% turnover (Figure 6.11). As with 1 timescale, JM is always more precise for MAR/MNAR with 40% turnover, and it is only at some of these where an improvement of 5% or greater is seen. HETW provides a small improvement (<5%) under MAR with I>C, 40% turnover and a non-constant intervention effect rate. The CW has much lower precision than the other models in the cases of MAR/MNAR with 20 and 40% turnover.

The fanning effect in MAR/MNAR as turnover increases also seen for the empirical SE with 1 timescale is clear here, with the same trend of JM being the best, followed by the HETW then HETWO and the CW model the worst.

For sub-samples, the conclusions are generally the same as those for the full-samples, but the JM in MCAR is not more than 5% worse (Appendix, Figure C.3).

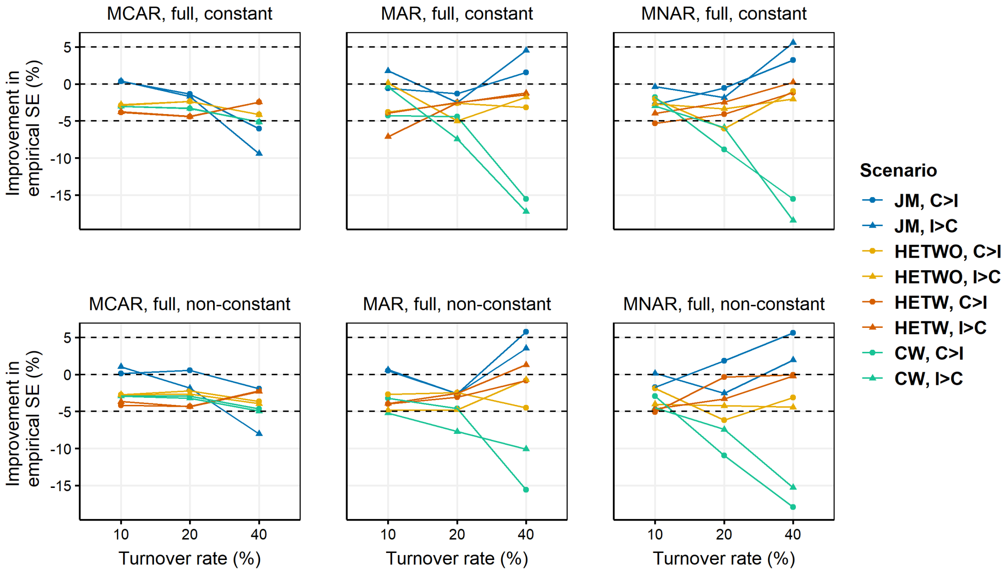


Figure 6.11: Improvement in empirical SE (%) for all models for the CS-OC-D estimand in full-samples with 2 timescales in comparison to the Kasza model. ‘Full’ denotes samples of 15 taken from cluster populations of 15; ‘constant’ (‘non-constant’) denote a constant (non-constant) intervention effect rate.

6.3.4 Variations of the models

6.3.4.1 Joint models

A restriction of the high powered computing services is that single jobs must run for no longer than 48 hours. For each scenario, this meant that each of the 650 simulations must have finished in less than 4.5 minutes for both the single and two timescale models. In practice, trialists may be willing to accept longer run times but it was not feasible to do this in a large simulation study. Several joint models were attempted, with different random effects in the longitudinal model, different association structures and different terms in the event sub-model.

For the longitudinal sub-model, models were attempted with random intercepts and random slopes for individual and cluster in different combinations. The only version that converged in a reasonable time was with a random intercept for cluster. For the event sub-model, a random (frailty) effect for cluster did not converge, but a fixed effect for cluster did. For association structures, the current value and random effect (for cluster) structures converged in a reasonable time. The cumulative exposure structure was also attempted, which is the integral of the current value with respect to time, but this did not converge in a reasonable time on its own or in combination with other structures. A combination of both the current value and random effect (for cluster) structures was also attempted but again this did not converge in time. In summary, the options for a joint model that converged in an acceptable time included a fixed effect for cluster in the event sub-model or not, a current value or random effect (for cluster) association structure, and a random intercept for cluster in the longitudinal sub-model.

Prior to running the joint models, it was believed that the current value association structure would most closely fit the underlying DGM in comparison to the random effect (for cluster) structure. Results were obtained for both association structures, and the presented results are for those of the current value structure. The random effects structure in general provided much less bias but worse precision than the current value structure. However, these results are not presented because in reality a structure would not be able to be selected based on its performance as the truth is unknown.

The version of the JM presented in the methods and results has a fixed effect for cluster in the event sub-model. A version without this cluster effect was also investigated, but was dropped from the main results due to the JM with cluster performing better in terms of convergence, bias and precision. With 1 timescale, the JM with cluster effect had less or the same number of convergence errors than JM without cluster effect in 75% of scenarios. With 2 timescales, this happened in 58% of the scenarios. Biases and empirical SE's were very similar for the two models but in the majority of cases the JM with cluster had a slight advantage.

6.3.4.2 CW models

The CW model presented in the methods and results has a constant term at the level of the cluster weightings. For typical multiple membership models, this constant term is omitted because it interferes with the constraint that the weightings sum to 1 and leads to issues with collinearity [241]. As this is not a constraint in my version of the model, CW models were fitted with and without the constant term to assess whether one was superior. Comparing the two CW models, biases across the scenarios were exactly the same, but the model with a constant offered slightly better precision for some of the empirical SE's. The CW model without a constant was therefore dropped.

6.3.4.3 Heteroscedastic models

Heteroscedastic models with and without the cross-classification of time in cluster were also compared, but both were retained as the models performed differently depending on the intervention effect. A version of the heteroscedastic models was also attempted with an unstructured variance-covariance matrix for the individual by time random effects instead of an independent structure:

$$\begin{pmatrix} \mathbf{I}_{1ijk} \\ \mathbf{I}_{2ijk} \\ \vdots \\ \mathbf{I}_{Tijk} \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma^2_{I1} & & & \\ \sigma_{I1,I2} & \sigma^2_{I2} & & \\ \vdots & \vdots & \ddots & \\ \sigma_{I1,IT} & \sigma_{I2,IT} & \cdots & \sigma^2_{IT} \end{pmatrix}\right).$$

These models had very poor convergence with up to 60% of models not converging in different scenarios. This could be due to the covariances between time points in the unstructured variance-covariance matrix; for an individual who is present at only one time point, the covariances mean that the effects at the other time points are also estimated, even though they were not present at these times. The individual random effects at times they were not present would therefore be estimated poorly and lead to negative or close to zero variances, causing convergence issues.

6.4 Discussion

6.4.1 Summary of findings

This simulation study aimed to explore possible analysis methods for CRTs with open cohort designs and underlying open populations as an alternative to the Kasza mixed model with one and two timescales used in Chapter 5, over a range of missing data mechanisms, turnover rates, and intervention effects. None of the models studied were superior to the Kasza model over all performance measures, but the heteroscedastic models (HETW and HETWO), particularly when combined with a random effect for cluster-period (HETW), were good 'all rounders'.

In terms of convergence, the CW and heteroscedastic models were superior to the Kasza model. The CW models had almost no issues with convergence, followed by the heteroscedastic models with consistently low non-convergence. The JM improved on the Kasza model in cases with high turnover, but the trends for JM were less predictable and worsened significantly when changing from 1 to 2 timescales.

6.4.1.1 OC-52-I estimand

For the OC-52-I estimand, the JM exhibits high bias but does offer greatly improved precision in the case of MAR/MNAR. Despite the improvements in precision, bias is not a desirable property of an estimator and so the JM is not recommended for this estimand. The CW model makes small improvements in bias over the Kasza model for constant intervention effects, but not for non-constant, and loses precision at high turnovers for MAR/MNAR. If the intervention effect was known to be constant, the CW model could be used if lower bias was deemed more important than increased precision. The HETW model has very similar precision to the Kasza model, with some small improvements, and for bias is also similar to the Kasza model for a constant intervention effect. When the intervention effect is non-constant, under MAR/MNAR with 40% turnover (ie. in the setting of DCM-EPIC or similar settings), HETW is less biased than the Kasza model. Crucially, the HETW model does not lose anything in precision for what is gained in bias in the non-constant cases, and so is the overall best alternative to the Kasza model for the OC-52-I estimand.

6.4.1.2 CS-OC-D estimand with 1 timescale

For the CS-OC-D estimand with 1 timescale under MAR/MNAR with high turnover, again as in DCM-EPIC or other OC settings, the HETW or HETWO models could be used as an alternative to Kasza's model. If an intervention effect is known to be constant, HETW could provide some improvements in bias, though only for I>C, and similar or slightly better precision than the Kasza model. If there is a possibility that the intervention effect is non-constant, however, although the HETWO model does lose precision compared to the Kasza model, in terms of bias it can provide improvements when the intervention effect is non-constant unlike HETW which does not cope as well; again this is a trade-off between reducing bias and losing precision. CW and JM are more unstable for bias over the different scenarios, though for JM the bias is not entirely compromised as in the OC-52-I and CS-OC-D (2 timescales) estimands, particularly for sub-samples. CW has consistently poor precision for MAR/MNAR with high turnovers, but JM is the most precise of the four methods and can provide some notable improvements over Kasza, however, this does not outweigh the issues with bias.

6.4.1.3 CS-OC-D estimand with 2 timescales

For the CS-OC-D estimand with 2 timescales, for MAR/MNAR with high turnover, as in DCM-EPIC or other OC settings, I would again recommend the HETW or HETWO models. With a constant intervention effect they are comparable in terms of bias to the Kasza model and are consistent across missing data mechanisms. Under a non-constant intervention effect, although under MCAR the HETW model is clearly more biased than HETWO, under MAR/MNAR which are arguably of more interest, the HETW model at a turnover of 40% is the same as or better than the Kasza model for both C>I and I>C, whereas for the HETWO model this depends on the relative values of C and I, with I>C doing better and C>I doing worse. The HETW model is also superior to HETWO in terms of precision as in the 1 timescale case. The JM for the CS-OC-D estimand with 2 timescales is also worse by 5% or more in terms of bias for MAR/MNAR with higher turnovers, and as for the OC-52-I estimand this is counteracted by an increase in precision in these cases, though these are not notable and around 5%.

6.4.2 Implications for DCM-EPIC

As in Chapter 5, for the implications for DCM-EPIC and related trials, the results for full-samples with 40% turnover and a MNAR missing data mechanism will be focused on. In general, there is not one model that is superior for both bias and precision, and the values and shape of the intervention effects are important.

For the OC-52-I estimand with constant cluster-level intervention effect rates, the CW and HETWO can provide improvements in bias but both of these lead to a reduction in precision compared to the Kasza model. For non-constant, HETW can provide small improvements in bias without losing precision. Given that the HETW model also has much better convergence rates than the Kasza model, this model is worth considering as an alternative.

For the CS-OC-D estimand with one timescale, for constant cluster-level intervention effect rates the HETW model can provide comparable or small improvements in bias and precision and large improvements in convergence to the Kasza model; for non-constant, the JM can provide more notable improvements in both bias and precision and smaller improvements in convergence. For the CS-OC-D estimand with two timescales, only the HET models can provide small improvements in bias without losing too much precision, but this again depends on the shape and value of the intervention effects at play.

6.4.3 Joint model

The poor performance of the JM in terms of bias for the OC-52-I and CS-OC-D estimands with 2 timescales could be expected due to the issues with `merlin` not currently being able to deal with more than one timescale, discussed in Section 6.2.4.7. However, the JM bias for the CS-OC-D estimand is still poor in some cases with 1 timescale, so this aspect of the software cannot be entirely to blame.

A joint model with an individual random effect in the longitudinal sub-model would be expected to provide immortal cohort inference, but given the lack of the individual random effect here it was hypothesised that the joint model could be less biased for mortal inference. The JM did in fact exhibit lower bias, and even improvements over the Kasza model in some cases, for the mortal CS-OC-D estimand with one timescale. Unfortunately, issues with the software with the two timescale models mean it is not possible to say whether the same is true when two timescales are used.

The JM's poor performance could also be due to the other elements of the longitudinal sub-model not being correctly specified to match the underlying Kasza model; specifically, as well as the individual random effect, the cluster-period cross-classification was also missing. In an ideal world with no implementation issues and better alignment of the underlying DGM and JM specification, the JM may have provided improved bias under MNAR scenarios as hypothesised, but mis-specification and other issues could have obscured this.

For the OC-52-I estimand, the bias for JM also worsens as turnover increases. As the turnover increases, there are more individuals in the dataset who have entered after CR and so the distinction between the two timescales is more important. If the software is unable to deal correctly with the second timescale then it would make sense that the bias increases as turnover increases.

Ultimately, however, even if the JM did offer improvements in both bias and precision, a major drawback is its convergence error rate when two timescales are used and the difficulty of implementation. At the present time and until appropriate frequentist software is developed that can handle the complex non-hierarchical structures of CRTs, the JM should therefore be considered for estimands with 1 timescale only. Even then, users should be aware of the increased level of complexity involved in specifying the model due to more assumptions having to be made.

6.4.4 Heteroscedastic models

The heteroscedastic models differ by HETW's inclusion of a cluster-period random effect. The results suggest that this cluster-period random effect interacts with the constant or non-constant intervention effect rate and affects bias. In a very general way this could be because both relate to time. However, the HETW model is less biased for non-constant intervention effects under the OC-52-I estimand and the CS-OC-D estimand with 2 timescales, but conversely under the CS-OC-D estimand with 1 timescale the HETWO model is less biased in these situations. This could be related to the fact that individual and time are crossed, and the individual variance is partitioned in the HET models, though it does also appear to depend on timescales too. As with the results of Chapter 5, there are several mechanisms at play here, therefore further work focusing on the workings of the heteroscedastic models is required to disentangle them.

The HETW model always has better precision than HETWO for all estimands under

MAR and MNAR with 20% turnover or more. As the datasets are the same this is not due to sample size, nor is it due to one model estimating fewer parameters; in fact, the HETW model estimates one more variance parameter than HETWO and provides better precision. This fanning effect in precision for the CS-OC-D estimands is clearly driven by turnover, but for the OC-52-I estimand the precision of the HETW and HETWO models is approximately constant as turnover increases. It is not clear why this is the case, therefore further investigation is warranted.

Improvements made by the HET models over the Kasza model in terms of bias and precision do not exceed 5%, so it could be argued that the Kasza model should not be discounted. Comparisons should be based on the specific characteristics of a trial, and in some cases the Kasza model could be preferable. However, users should also bear in mind that the HET models have consistently good convergence, even in the most complex circumstances with high turnover and all complications turned ‘on’, unlike the Kasza model. They are also simple to implement using established and well-developed functions in statistical programs.

6.4.5 Cluster-weighted models

The CW model has very poor empirical SE’s for the CS-OC-D estimand with 1 and 2 timescales under MAR/MNAR with 20 and 40% turnovers; this also occurs for the OC-52-I estimand but not to the same magnitude. Further exploration involving variations of the CW model could provide more information on why this is the case.

This is the first time weighting of cluster effects has been used in this way. The CW model is simple to implement using the code provided in the appendix, which as with the HET models can be done using established functions in existing statistical programs. Weights need to be calculated first, and a CW model can be implemented using standard functions in `Stata`. The most appealing property of this model is its almost zero non-convergence rate across all scenarios, even in the most difficult circumstances with high turnover and complex missing data mechanism, unlike the Kasza model. However, whilst this model can offer some improvements in special cases it still has issues with precision that would potentially be unappealing to trialists, so more development is needed.

6.4.6 Comparison to existing literature

Joint models in theory could be ideal for analysing open cohort datasets due to their ability to deal with informative drop-out, but there are serious drawbacks relating to their implementation. Firstly, as demonstrated in this work, frequentist software able to deal with clustering and cross-classification or other non-hierarchical structures in the longitudinal sub-model is not readily available. Moreover, software cannot currently support more than one timescale and requires further adaptations to support estimands where two timescales are important.

If suitable software does exist, there are then several strong assumptions to be made, including the form of the baseline hazard, the association structure and the form of the longitudinal trajectory. Crowther *et al.* [242] studied the effects of misspecifying the longitudinal trajectory in joint models under two different association structures, finding that the current value association structure was more robust to mis-specification than the rate of change association structure, with the latter giving highly biased parameter estimates. In both cases, the true underlying trajectory included only random intercepts and slopes at the individual level. Given that in this work the trajectory is more complex with the inclusion of higher-level clustering terms, the same results may not apply here. Moreover, the DGMs used by Crowther *et al.* were very similar to the analysis models applied, thus unbiased results could be expected. More recently, others have also found joint models to be sensitive to the longitudinal equation form, with SPMs exhibiting much larger bias compared to mixed models even when the underlying DGM is a SPM [243]. Arisido *et al.* also found that joint models were not robust against mis-specification of the baseline hazard and shape of the longitudinal trajectory, giving highly biased results [244]. The biased estimates of the JM in this work could have been caused by the disparity between standard association structures catered for in the software and the actual longitudinal trajectory in the underlying DGM.

Miller *et al.* also conducted a simulation study to compare linear mixed models to SPMs with either random effects or current value association structures [245]. When the underlying missing data DGM was MNAR based on shared random effects and a strong association value, the mixed model produced biased results compared to the SPM (current value) and SPM (random effects) which were unbiased. However, when the underlying DGM was MNAR using a deterministic approach which depended on observed values of the outcome, both SPM (current value) and SPM (random effects) approaches were as biased as the linear mixed model for many parameters. This latter underlying DGM is similar to the one used in this simulation study in that it uses observed values of the outcome rather than the underlying true value (ie. without measurement error), which could point to another reason for the JM's poor performance.

Whilst the joint model presented in the main results exhibited high bias, it also provided large improvements in precision over the Kasza model. This agrees with previous findings [222].

6.4.7 Strengths, limitations and directions for future research

The strengths of Chapter 5 relating to the generation of data also apply to this chapter. As only the complications scenarios were investigated, this meant that many of the study parameters were fixed, which could be extended in future work. This included the number of time points in the datasets, an ICC of 0.1, fixed cluster sizes of 15 for full-samples and 50 for sub-samples. However, a strength is that a variety of relevant real-life complications were considered.

Another strength of this simulation study is that several different versions of the models presented in the main results were also attempted in Section 6.3.4 for completeness.

The joint model was by far the most complex of the models explored in this chapter and required more assumptions to be made, specifically for the baseline hazard function and the sharing structure. An exponential baseline hazard function was chosen as this aligned with the DGM, as opposed to other parametric choices such as weibull or gamma. As the baseline hazard function was not varied this could be seen as a limitation. The current value sharing structure was the only structure used for the results presented in this chapter, which again could be seen as a limitation, but in fact many other structures were attempted, on their own and in different combinations with each other. However, only the current value and random effects structures converged within an acceptable amount of time. As described in Section 6.3.4, a decision was made to present the current value results as in reality a choice of structure would have to be made based on clinical context rather than bias and precision as this would be unknown, and the current value structure was believed to be closer to the underlying MNAR DGM *a priori*. A further complication is that the sharing structures that were possible in `merlin` did not align with the original DGM used to generate the data in Chapter 5.

There were further limitations relating to the joint modelling software used in this study. The longitudinal sub-model of the joint model had to be simplified to a simple nested model because software does not currently exist to deal with multiple levels of clustering and cross-classification. Only frequentist approaches were investigated, though Bayesian joint modelling is also possible. Another limiting factor of joint modelling software is that it does not currently include an offset for two timescales instead of one in the longitudinal sub-model. Furthermore, the open cohort dataset was not generated using a function specifically developed for joint models, such as `simjoint` in the `joineR` package or `simulate` in the `JM` package (both in R), and instead used a cross-classified mixed model aligning with the Kasza model. This could have given the other models, particularly the Kasza model, an advantage, as these models were closer to the underlying DGM, and the JM may have performed better if the underlying DGM was also generated using a joint model. Given these challenges, the joint model could be at more of a disadvantage in this study in comparison to the other models. In terms of the bigger picture, whilst the JM exhibited poor performance under the specific circumstances of this simulation study, the various mis-specifications should be taken into account and it should not be discounted in general.

To overcome issues faced with the joint models, Bayesian joint modelling techniques could be attempted in future work. Such methods were avoided in this simulation study as they can be more computationally complex and can require the ability to code in programs such as `winBUGS`. However, the flexibility of a Bayesian approach could be a way of overcoming the restrictions seen in frequentist software relating to clustering above the individual level and non-hierarchical structures in the longitudinal model, and could also allow a more flexible version of the sharing structure than frequentist software currently permits. With the exception of the joint model, all other analysis models were able to be implemented simply using widely used functions in standard software programs.

The CW model used here is an ‘acute’ model as opposed to ‘cumulative’ [240], meaning that the weights for a person are the same at every time point. Although this method takes into account length of stay for each individual, the weighted cluster random effect is the same at all of their measurement points, and the ‘future’ known value of the length of stay influences a current value, which could be seen as unnatural. An alternative and perhaps more natural approach could be to set the weight at a particular time point as the time spent in the cluster up until that time. This could be described as a cumulative model as the weights for the same person across each row are not equal, but instead increase in size as time goes on, and more weight would be given to the random cluster effect at later time points. In the model presented here, the weights are used simply for cluster random effects. This could also be extended so the weightings are instead for cluster-period random effects. Future developments of the CW model could focus on these two adaptations.

Although all of the new models presented in this chapter used continuous time for fixed effects, they are being compared to the Kasza model from Chapter 5 which used discrete time for the CS-OC-D estimand. The effect of using the Kasza model and the effect of using discrete time rather than continuous time are therefore entangled for comparisons across models for the CS-OC-D estimand. As the OC-52-I estimand used continuous time in Chapter 5, this is not an issue under this estimand.

Finally, in practice there are different reasons for drop-out from a cluster; in DCM-EPIC, these reasons were moving to another care home and death, for example, and can be called competing risks. All reasons for drop-out were treated in the same way in this simulation study and that of Chapter 5, but future research could instead differentiate between competing risks of drop-out in the data generation and analysis.

Chapter 7

Discussion

7.1 Introduction

The overarching aims of this thesis were to develop the design and analysis of novel open cohort parallel-group CRTs where the underlying population is open. Whilst development of design and analysis has largely been achieved in Chapters 5 and 6 respectively, this work also makes significant contributions to how reporting of CRT designs can be improved and the classification of CRT designs in Chapters 2, 3 and 4. The latter two outputs were not part of the original aims but transpired from obstacles encountered in the scoping review of Chapter 2.

The scoping review of Chapter 2, due to poor reporting, was split into searches for institutional settings and explicit mentions of open and dynamic cohorts. The institutional settings search suggested that the need for an open cohort design could be greater in some settings more than others, namely care homes, palliative care and prisons. The second search for explicit use of open and dynamic cohort terminology was carried out both for epidemiological studies and CRTs. No specific open cohort methods were found from the epidemiological literature, where the open cohort design was already more established. The CRTs search found two cases for analysis of open cohort CRTs with continuous outcomes; a mixed effects model with random effect for individual was used in both cases. In both CRTs and epidemiology, the open/dynamic cohort terminology was used to describe both populations and designs but never specifically open cohort analyses. Moreover, there was a distinct lack of agreement as to what an OC design consists of with many differences in definition and explanation. Not being able to determine a design from its individual components inspired the following chapter.

Following the results of the scoping review, I still found there was a lack of clarity as to what a design, for a CRT specifically, actually consists of. In Chapter 3, I laid out what I believe to be the essential components of a CRT design, with an entire section dedicated to sampling schemes, commenting on their pros and cons and their ease of implementation in practice. This chapter also presented the novel idea of full- versus sub-samples when sampling from clusters, and how the distinction between these can lead to differences in design.

Given the lack of named designs in the scoping review, I presented a classification system in Chapter 4 for parallel-group (and by extension, factorial) CRTs, when institutions are the unit of randomisation. This chapter includes user engagement work carried out as part of the larger research project, consisting of a workshop and an online survey involving trialists working in CRTs. An initial version of the classification system was presented to workshop attendees, and updates were made based on feedback. An improved version was then included in an online survey whereby trialists were asked to fill it out for their own trials. The aims of this were twofold: trialists suggested possible OC trials, and to test the classification system. A final round of updates were made to the system following the feedback and I classified the designs provided by trialists. Despite the respondents' expertise in CRTs, only 7 of the 46 (15%) provided designs were possibly OC; many were in fact CC, R-CS or other designs, demonstrating again the lack of clarity on design components within the field. The provided examples were used to illustrate six identified CRT designs in Chapter 4 (Figure 7.1). No examples of the OC design with continuous recruitment were found.

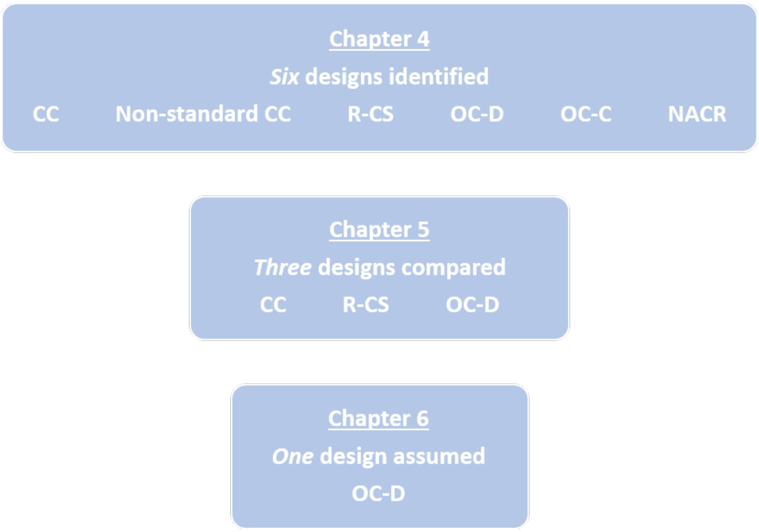


Figure 7.1: Scope of designs considered throughout the thesis.

In Chapter 5 I compared the conventional CC and R-CS designs to the novel OC design with discrete recruitment. Variations of the mixed effects model seen in the scoping review and in the previous work of others were assumed in this chapter. A major element of this chapter was the assumption that for CRTs where intervention effects can occur at cluster-level and individual-level that there are (at least) two timescales in operation, one of which is usually ignored. Conclusions were highly dependent on the estimand of interest, and in some cases on study parameters. Timescales, estimands and full- versus sub-sampling are discussed in further detail in the following section.

Finally, in Chapter 6, the downfalls of mixed effects models used in others' work and in Chapter 5 were highlighted and alternative analysis models explored. An OC design with discrete recruitment was assumed. Given the lack of exploration of other methods seen in the scoping review, for both CRTs and epidemiology, this is totally novel.

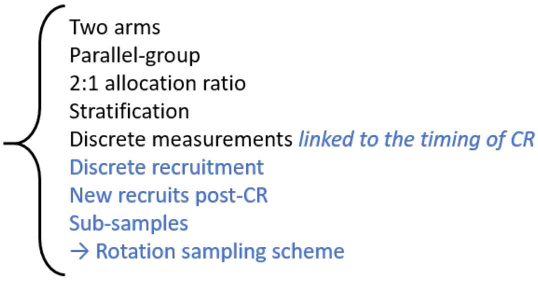
In the remainder of this chapter I provide an overview of the key themes, implications and general limitations for this work, before concluding with suggestions for future work and final conclusions.

7.2 Key themes and their implications

The introductory chapter began by splitting up the concepts for this thesis into population, design and analysis sections. An underlying open population was then assumed for the remainder of the thesis, with Chapter 5 focusing on design and Chapter 6 focusing on analysis separately. I believe that the partitioning of these three elements is key to obtaining a suitable design and analysis that align with the underlying population. Appropriate estimands cannot be obtained without consideration of the underlying population. Following determination of a suitable estimand, designs then facilitate trialists to be able to collect the necessary data at the necessary times from the necessary individuals, and the analysis model is a tool to be used at the end of that process to obtain statistical inference. Seeing these three concepts as linking processes over time means that the inherent nature of the population is considered early on and accounted for, rather than trying to fix issues later at the analysis stage. It is possible that previously, open populations have been seen as a nuisance and methods for closed populations used blindly, when in fact specialist methods are required in order to make correct inference for open populations.

One of the biggest issues encountered in the scoping review that has potentially been overlooked in the field of CRTs to date, is the ambiguity over what a *design* consists of.

Examples were found where trialists conflated the sampling scheme and the design. For example, random sampling does not necessarily mean the entire design is R-CS; there are other elements to consider, which are laid out in Chapter 3. Instead, I argue that the sampling scheme falls under the 'umbrella' of design, but that they are not one and the same thing. Similarly, in the search for 'open cohort designs', trialists conflated the ability for new individuals to be recruited post-CR with the design itself. Whilst this is an important element that I also propose falls under the umbrella of design, this characteristic does not define the design entirely, and a lot of the smaller details are unreported and unclear. It is therefore imperative that full disclosure is given as to what components are involved in a design, and if a shorthand name is given, to provide extra elaboration. The introduction of full- versus sub-samples as part of a design is a simple concept that could also bring about further clarity in description. It is common in the literature for sub-samples to be assumed, for example for CRTs in countries or large communities, but this is not always the case as was highlighted by DCM-EPIC. Distinguishing full- and sub-samples, and the requirement of a sampling scheme if sub-samples are used, allows there to essentially be two different versions of a general trial design. The R-CS design defined in this thesis for example differed in full- versus sub-samples. Given the importance of the design of CRTs, I believe the concept of the design umbrella term should be recognised by commonly used reporting guidelines such as the CONSORT statement [64]. Figure 7.2 gives an example of one overall design; seeing a design in this way illustrates the many combinations possible and how it can be fine tuned to meet the needs of a specific CRT.



Two arms
Parallel-group
2:1 allocation ratio
Stratification
Discrete measurements *linked to the timing of CR*
Discrete recruitment
New recruits post-CR
Sub-samples
→ Rotation sampling scheme

Figure 7.2: An example of one overall design broken into its components.

Estimands have been discussed in this thesis with consideration of whether they are immortal or not; immortal estimands assume that nobody dies and make predictions into the future, even for those that die. Immortal estimands link to unconditional models, as survival is not taken into account, whereas mortal estimands link to conditional or partly conditional models [60]. Ultimately, which estimand is of interest comes down to the aims of the research.

In analysis models for parallel-group CRTs, just one timescale of time from cluster-randomisation is typically assumed. In this thesis, I proposed that an equally important timescale, the timescale of individual length of stay post-randomisation, should also be considered in the underlying DGM and analysis models. The other two timescales introduced but not incorporated into the underlying DGM of Chapters 5 and 6 were calendar time and total stay in cluster, to take into account time present in the cluster before cluster-randomisation. For each timescale that is to be included in the modelling, trialists would have to plan to collect this information in advance. Routinely collected data may be able to assist with dates that are already part of existing processes, such as dates of admission or discharge from hospital. As more timescales are added, the more complex models become and the more assumptions are required. This could lead to issues with convergence or, as was seen with the joint model in Chapter 5, some functions may not be able to cope with multiple timescales. It is possible however that simpler ways of adjusting for multiple timescales exist, and these are avenues for future work.

As discussed in Chapter 5, it is important to reiterate that, given mixed effects models were the analysis model of choice in Chapter 5, the results concerning the comparison of designs are not equally applicable across other methods of analysis. As mixed effects models naturally provide inference for an immortal cohort, estimands which are instead mortal would have been at a disadvantage in this chapter. If a particular combination of design and analysis is not appropriate for an estimand of interest, this does not necessarily mean that the design on its own is not a good choice for that estimand; rather, the combination of design and analysis must be taken into account. For example, use of a R-CS design and a mixed effects model was biased for the CS-OC-D estimand, but the R-CS design may have been unbiased using a different analysis model such as a simple difference between arms.

There was a clear need for improvement in reporting of trial designs within CRTs. Guidance was provided for this in Chapters 2, 3 and 4 which is summarised in Box 6. Ultimately, such improvements in reporting aid the interpretation of CRT results papers, as it can often be difficult to understand the full workings of complex trials. This would also impact future research as relevant articles would appear in electronic searches, and greater transparency means researchers conducting reviews and meta-analyses could more easily locate and extract information. Suggested changes to the existing CONSORT statement with extension to CRTs [152] are also summarised in Box 7.

A separate but related issue to that of reporting is the development of the classifica-

Box 6: Guidance for trialists reporting CRTs

Wording

- Use ‘continuous’ or ‘discrete’ when describing measurement and recruitment processes.
- If using open or dynamic cohort terminology, be specific as to whether this relates to the population, design or analysis (or combinations of these).
- Report measurements clearly using a timescale of reference (eg. 10 days post-recruitment, 12 months post-randomisation).
- Use known terminology for trial design type where this exists (eg. closed cohort design - see Chapter 4).

Placement

- Report trial design, or open or dynamic cohort terminology, in the abstract where possible.

Illustrative tables and diagrams

- Use diagrams to illustrate trial design such as those used in Chapter 4.
- Use diagrams to illustrate the ordering of trial processes such as the Timeline Cluster diagram [3].
- Clearly depict the specific timings of trial processes using a tool such as the SPIRIT schedule [151]. Horizontal lines could be used for continuous processes and X’s for discrete.
- Use CONSORT diagrams to report in-migration and out-migration at each time point, clearly stating how many of the original cohort from before cluster-randomisation remain at each time point.

Content

- Clearly state whether repeated measurements from the same individual are linked over time, and if not, provide a reason as to why.
- If data collection and/or recruitment occur through the use of EHR or other external sources, this should be stated clearly, and whether this is a retrospective or prospective process.
- Report any use of embedding periods of an intervention and any strategies used to counteract dilution of intervention effects.
- Report whether intervention delivery and measurement schedules took place at the same time across clusters, in batches or varied on an individual cluster basis.

tion system to help trialists better identify and name designs for parallel-group CRTs. Separating different trial designs will allow more focused research to be carried out and provides a starting point for new methodologies to be developed unique to each design. Further research could then determine whether some designs are better suited to particular settings or interventions and the reasons for this based on individual components. Ultimately, the classification system could lead to a deeper understanding of why trials

Box 7: Suggested changes to existing CONSORT statement

Item 3a

Current: "description of trial design (such as parallel, factorial) including allocation ratio" and "definition of cluster and description of how design features apply to clusters"

Suggested change: state whether the CRT type is parallel-group, factorial, SW or CRXO. State whether new participants are able to be recruited after cluster-randomisation or not. Provide a named sub-design if this exists, for example, "open cohort with discrete recruitment".

New Item 3c:

State whether measurements are collected discretely or in a continuous fashion.

New Item 3d:

State whether recruitment occurs discretely or continuously.

Item 10b

Current: "mechanism by which individual participants were included in clusters for the purposes of the trial (such as complete enumeration, random sampling)"

Suggested change: whether full- or sub-samples of participants are taken from clusters, and if sub-samples are taken, the sampling scheme used. Make the reader aware that many variations of sampling scheme exist besides simple random sampling.

are designed in certain ways and demystify the process of which trial design to use in particular situations.

7.3 Implications for DCM-EPIC and care home trials

The DCM-EPIC CRT where clusters were care homes was the motivational example for this work, and as such is intertwined throughout. DCM-EPIC influenced many of the study parameters and complications scenarios that were varied in the simulation studies, and the trial is returned to in the discussions of Chapters 5 and 6 to consider the design and analysis implications. Without going into specific details, the OC design was beneficial for the OC-52-I estimand and not useful for the CC-D estimand when a mixed model analysis was assumed. For other estimands there were scenarios where the OC design with a mixed effects analysis model was superior, but this depended on the estimand of interest and the complications. In Chapter 6, depending again on the scenario, some of the alternative models proposed provided improvements over the Kasza single-timescale model. In summary, as the results cover a vast range of scenarios, trialists would have to have a specific estimand in mind and prior knowledge about their interventions, for example the shape of the intervention effect over time, in order to navigate to an appropriate section similar to their work. The comparison of designs and analysis is a complex picture with many elements to untangle.

7.4 General limitations and future work

As limitations specific to each chapter of the thesis have already been discussed, I conclude here with a number of general limitations of the thesis overall and subsequently how this could be overcome with further research.

This thesis focused on parallel-group and factorial CRTs. The classification system of Chapter 4 and results of the design and analysis in Chapters 5 and 6 are therefore not applicable to other CRT designs such as SW or CRXO, in particular as these designs are more complicated due to the randomising of clusters to sequences involving time. The findings are also not applicable to individually-randomised trials as they generally do not involve many of the complications of CRTs, though some elements of the classification system could be considered in the context of IRTs, such as whether calendar time should be adjusted for. In Chapter 4 I assumed that the classification system only applied to clusters that are institutions, such as schools or care homes, not to CRTs where the unit of randomisation is the individual delivering treatment. Some elements of the classification system may apply to this case, but testing and development for this is outside the scope of this thesis and an area for future work. In Chapters 5 and 6 however, as the underlying DGM is assumed to have cluster-level and individual-level intervention effects, the results of these chapters are applicable to CRTs with cluster-level interventions, not CRTs with individual-level interventions only. In this simpler case, it may be sufficient to consider a single timescale only.

Only continuous outcomes have been considered in general in this thesis, with the exception of the classification system of Chapter 4 which was amended to be independent of outcome type. The scoping review of Chapter 2 did examine outcome type in the explicit CRT search, but ultimately only methods for continuous outcomes were sought. Conclusions from Chapters 5 and 6, particularly relating to analysis models, do not apply to other outcome types such as binary or time-to-event, and this is an avenue for future development.

A further limitation of this work is the omission of selection bias in the simulation studies of Chapters 5 and 6, despite the risk of selection (and recruitment) bias being identified in the classification system of Chapter 4. Open cohort designs, or any design where individuals are recruited post-randomisation, are at risk of imbalance between arms. Given the level of the complexity in the simulation studies and the underlying DGM, selection bias could not have been thoroughly investigated in Chapter 5 as one of the complications, and would have required extra methods in the analysis to be dealt with in Chapter 5. The findings of this thesis should therefore be considered in conjunction with the possibility of selection bias, and readers are directed to research which focuses entirely on this subject, such as the work of Leyrat *et al.* [246]. Methods for trialists to reduce the risk of selection bias have already been discussed in Chapter 4.

For the first time in the field of parallel-group CRTs, six designs have been outlined in terms of their components. Whilst the closed cohort and (repeated) cross-sectional designs were previously well known, four new designs have been identified. Following Chapter 4, the remainder of the thesis focused only on CC, R-CS and the OC design with discrete recruitment, which leaves a huge opportunity for more research into the OC design with continuous recruitment, the NACR design and the non-standard CC design. This is not to say that these designs have not been previously used; the NACR design for example, as shown from the user engagement work of Chapter 4, has been used frequently to this day, but has previously not been recognised as a 'type'. The only design that an example could not be found for, which may only be a theoretical design for this reason, is the OC design with continuous recruitment. This work paves the way for a review, for example, of settings where the NACR design is most commonly carried out, or the particular characteristics of the CRTs that tend to use this design. There may also be specific methodologies to develop in relation to each design type.

7.5 Conclusion

CRTs, where clusters are randomised to different arms of a clinical trial rather than individuals, have become more increasingly more common since their initial use in the 1980s [247]. Notably, research activity surged in this area in the early 2000s when the first CONSORT extension to CRTs was published [152], with significant developments made since. However, due to their increased complexity in comparison to individually-randomised trials, there is still work to be done in this field. In this thesis I have focused mainly on the design of CRTs, delving into what constitutes a CRT design, as well as identifying subtypes of CRT design and breaking down their components to aid the selection of design for trialists in practice. More research is required to fully understand the capabilities of each design and the circumstances in which each is most appropriate.

Amongst the CRT designs discussed in this thesis, I have paid particular attention to the open cohort design, a novel CRT design in the early stages of development that has been used fairly sparsely to date. Through the use of simulation, I have shown when this new design is superior to existing designs, and highlighted the potential of analysis models for such designs, which have not previously been used for this application. In terms of both design and the corresponding analysis that follows, the open cohort design opens up a multitude of opportunities for future research.

Appendix A

Appendix for Chapter 2

A.1 Search strategies

1. Randomized Controlled Trial/
2. CRCT.tw.
3. (cluster* adj3 RCT).tw.
4. (cluster* adj3 randomi*).tw.
5. (cluster* adj3 trial*).tw.
6. (crt and cluster*).tw.
7. “group randomi*”.tw.
8. cluster-unit*.tw.
9. “randomi* unit*”.tw.
10. “unit of randomi?ation”.tw.
11. 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10
12. 1 and 11

Table A.1: Search strategy for CRTs. These 12 lines were included in all of the setting searches.

13. care?home*.tw.
14. “assisted living”.tw.
15. “homes for aged”.tw.
16. “care home*”.tw.
17. “long term care”.tw.
18. “group care”.tw.
19. “residential facilit*”.tw.
20. “residential home*”.tw.
21. “residential aged care”.tw.
22. “nursing home*”.tw.
23. “nursing facilit*”.tw.
24. “retirement home*”.tw.
25. “retirement village*”.tw.
26. “retirement communit*”.tw.
27. “old people* home*”.tw.
28. “home* for the elderly”.tw.
29. “congregate living facilit*”.tw.
30. “group home*”.tw.
31. “residential care”.tw.
32. Nursing Homes/
33. Homes for the Aged/
34. 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23
    or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34
35. 12 and 35
36. limit 36 to yr=“2011 -Current”

Table A.2: Search strategy for the care home CRTs.

13. Prisons/
14. Criminals/
15. Prisoners/
16. prison*.tw.
17. jail*.tw.
18. offend*.tw.
19. reoffend*.tw.
20. convict*.tw.
21. inmate*.tw.
22. detainee*.tw.
23. cellmate*.tw.
24. incarcerat*.tw.
25. felon*.tw.
26. remand*.tw.
27. penitentiary.tw.
28. 13 or 14 or 15 or 16 or 17 or 18 or 19 or
    20 or 21 or 22 or 23 or 24 or 25 or 26 or 27
29. 12 and 28

Table A.3: Search strategy for the prison CRTs.

13. Hospitals/
14. “hospital*”.tw.
15. 13 or 14
16. 12 and 15

Table A.4: Search strategy for the hospital CRTs.

13. Schools/
14. school*.tw.
15. 13 or 14
16. 12 and 15

Table A.5: Search strategy for the school CRTs.

13. Primary Health Care/
14. General Practice/
15. “general practice*”.tw.
16. GP.tw.
17. “primary care”.tw.
18. 13 or 14 or 15 or 16 or 17
19. 12 and 18

Table A.6: Search strategy for the primary care CRTs.

13. Hospices/
14. “Hospice and Palliative Care Nursing”/
15. Palliative Care/
16. palliative.tw.
17. hospice*.tw.
18. 13 or 14 or 15 or 16 or 17
19. 12 and 18

Table A.7: Search strategy for the palliative care CRTs.

13. rural population/ or suburban population/ or urban population/
14. Cities/
15. communit*.tw.
16. village*.tw.
17. town*.tw.
18. 13 or 14 or 15 or 16 or 17
19. 12 and 18

Table A.8: Search strategy for the community CRTs.

1. Randomized Controlled Trial/
2. CRCT.tw.
3. (cluster* adj3 RCT).tw.
4. (cluster* adj3 randomi*).tw.
5. (cluster* adj3 trial*).tw.
6. (crt and cluster*).tw.
7. “group randomi*”.tw.
8. cluster-unit*.tw.
9. “randomi* unit*”.tw.
10. “unit of randomi?ation”.tw.
11. 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10
12. 1 and 11
13. “open cohort*”.tw.
14. “open?cohort*”.tw.
15. “dynamic?cohort*”.tw.
16. “dynamic cohort*”.tw.
17. 13 or 14 or 15 or 16
18. 12 and 17

Table A.9: Search strategy for the CRTs (explicit usage).

1. Epidemiology/
2. Non-Randomized Controlled Trials as Topic/
3. Observational Study/
4. Epidemiologic Studies/
5. 1 or 2 or 3 or 4
6. “open cohort*”.tw.
7. “open?cohort*”.tw.
8. “dynamic cohort*”.tw.
9. “dynamic?cohort*”.tw.
10. “dynamic population*”.tw.
11. 6 or 7 or 8 or 9 or 10
12. 5 and 11

Table A.10: Search strategy for epidemiological studies (explicit usage).

A.2 PRISMA flow diagrams

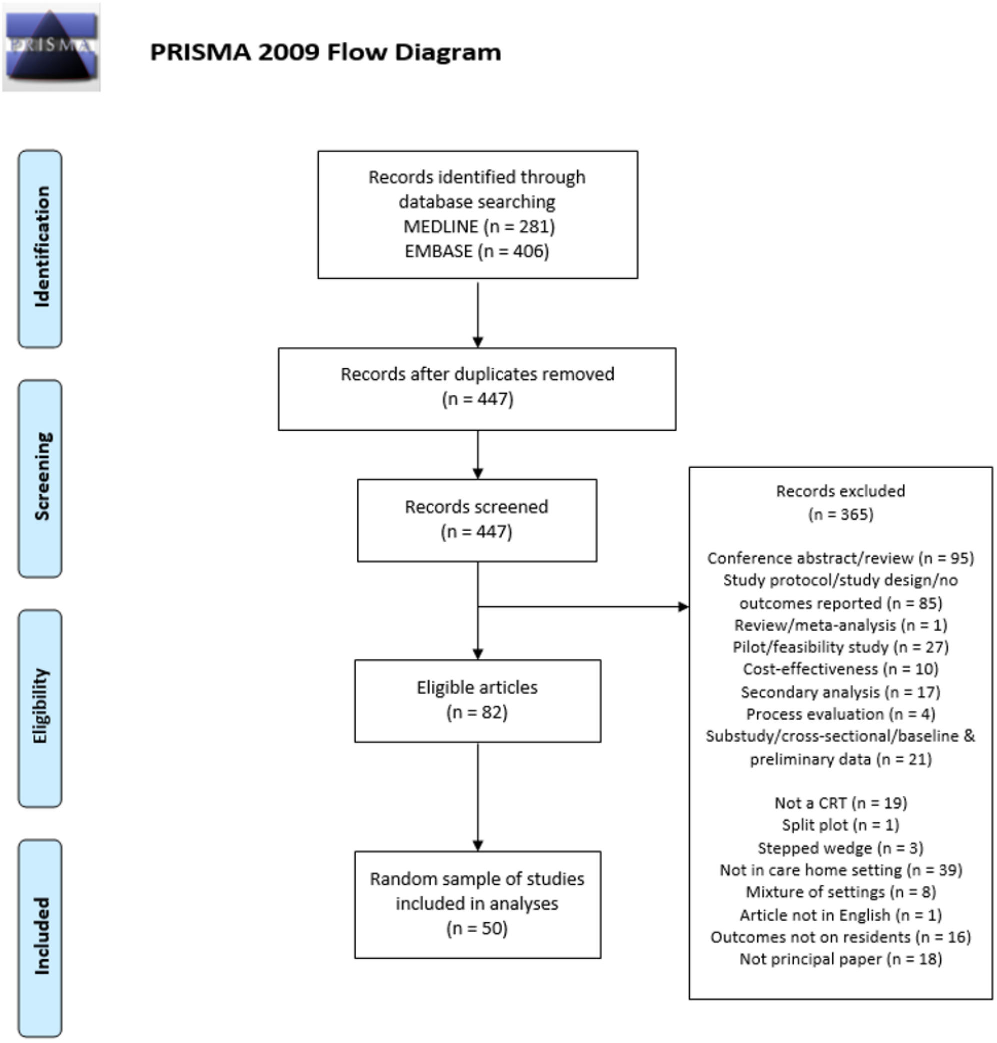

PRISMA
PRISMA 2009 Flow Diagram
Identification
Screening
Eligibility
Included
Records identified through database searching MEDLINE (n = 281) EMBASE (n = 406)
Records after duplicates removed (n = 447)
Records screened (n = 447)
Records excluded (n = 365)
Conference abstract/review (n = 95)
Study protocol/study design/no outcomes reported (n = 85)
Review/meta-analysis (n = 1)
Pilot/feasibility study (n = 27)
Cost-effectiveness (n = 10)
Secondary analysis (n = 17)
Process evaluation (n = 4)
Substudy/cross-sectional/baseline & preliminary data (n = 21)
Not a CRT (n = 19)
Split plot (n = 1)
Stepped wedge (n = 3)
Not in care home setting (n = 39)
Mixture of settings (n = 8)
Article not in English (n = 1)
Outcomes not on residents (n = 16)
Not principal paper (n = 18)
Eligible articles (n = 82)
Random sample of studies included in analyses (n = 50)

Figure A.1: PRISMA flowchart for the care homes search.

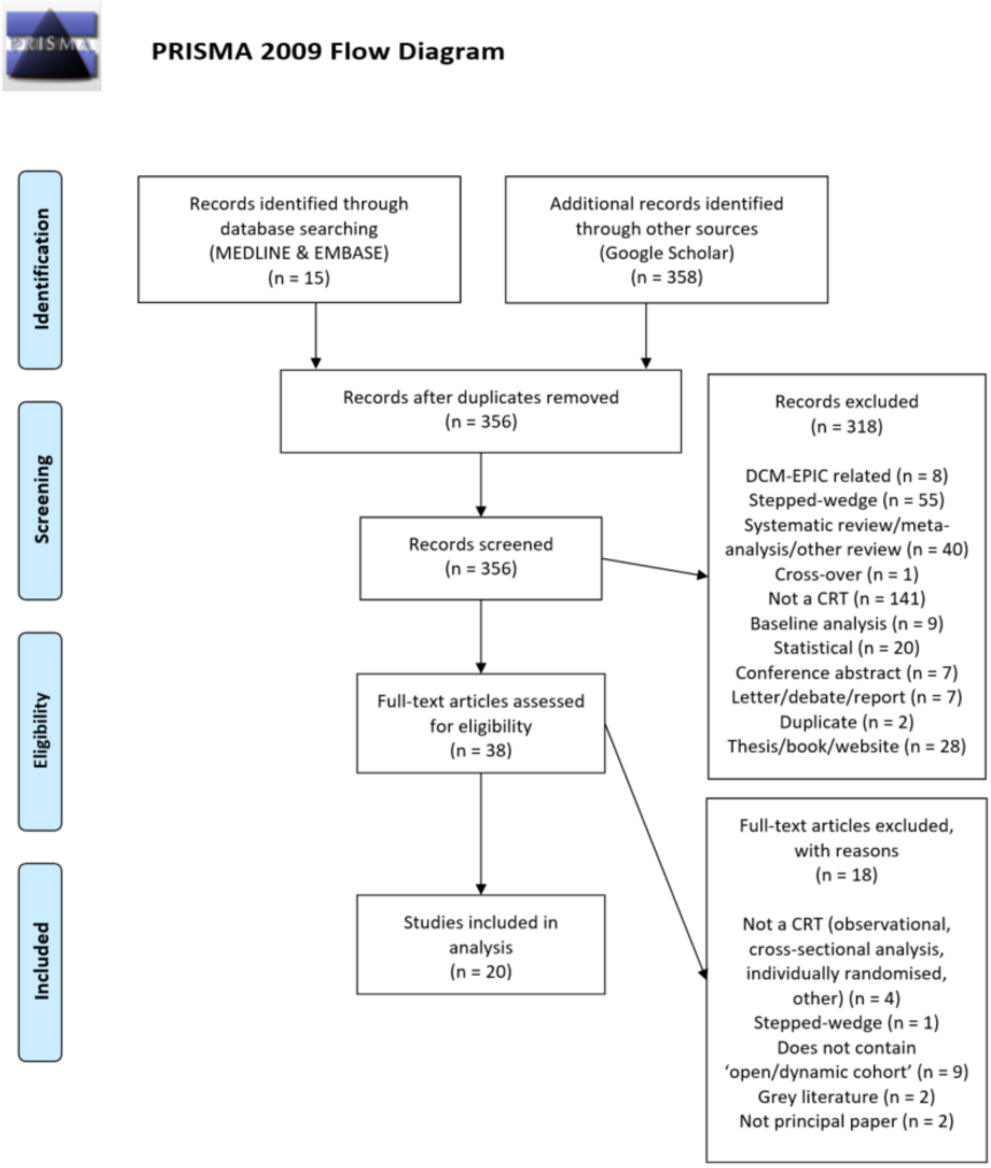

PRISMA
PRISMA 2009 Flow Diagram
Identification
Screening
Eligibility
Included
Records identified through database searching (MEDLINE & EMBASE) (n = 15)
Additional records identified through other sources (Google Scholar) (n = 358)
Records after duplicates removed (n = 356)
Records screened (n = 356)
Records excluded (n = 318)
DCM-EPIC related (n = 8)
Stepped-wedge (n = 55)
Systematic review/meta-analysis/other review (n = 40)
Cross-over (n = 1)
Not a CRT (n = 141)
Baseline analysis (n = 9)
Statistical (n = 20)
Conference abstract (n = 7)
Letter/debate/report (n = 7)
Duplicate (n = 2)
Thesis/book/website (n = 28)
Full-text articles assessed for eligibility (n = 38)
Full-text articles excluded, with reasons (n = 18)
Not a CRT (observational, cross-sectional analysis, individually randomised, other) (n = 4)
Stepped-wedge (n = 1)
Does not contain 'open/dynamic cohort' (n = 9)
Grey literature (n = 2)
Not principal paper (n = 2)
Studies included in analysis (n = 20)

Figure A.2: PRISMA flowchart for the CRTs search (explicit usage).

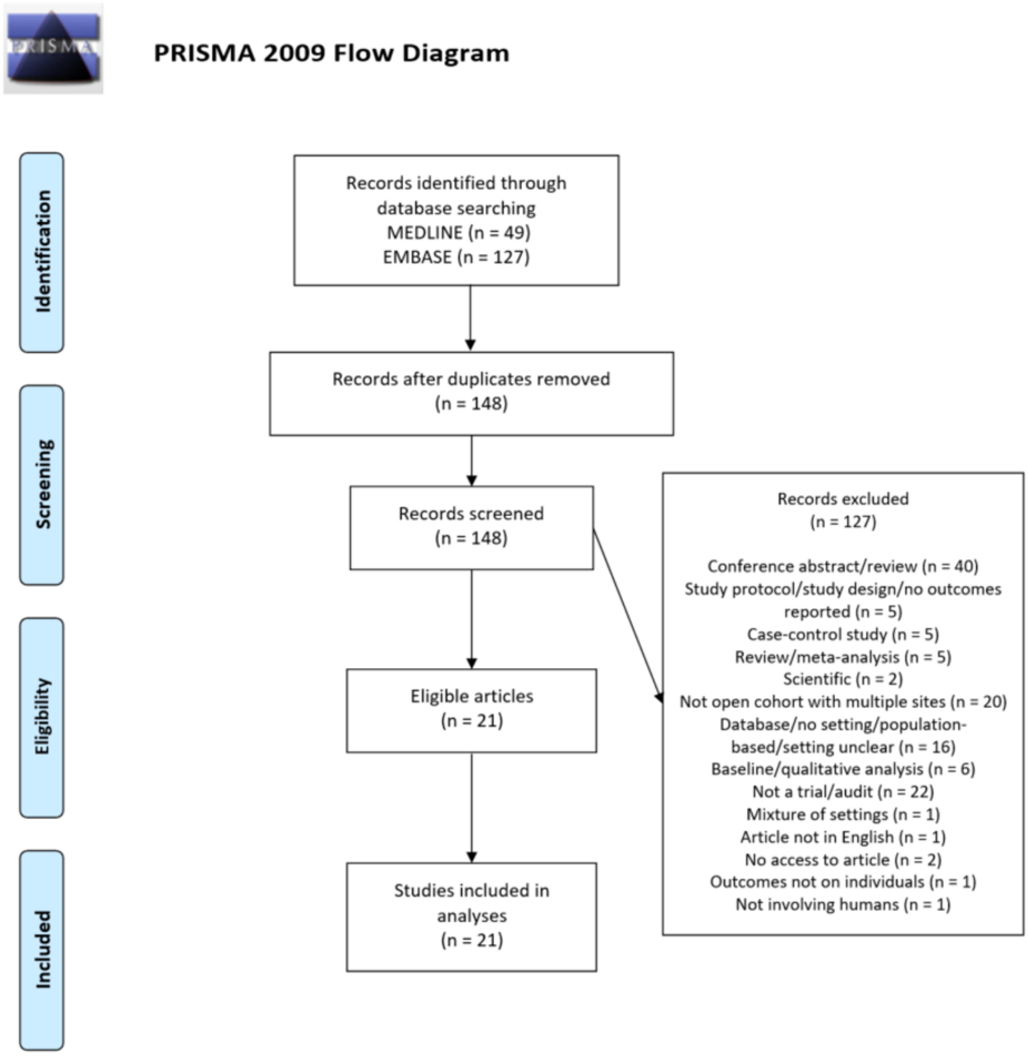

PRISMA
PRISMA 2009 Flow Diagram
Identification
Screening
Eligibility
Included
Records identified through database searching MEDLINE (n = 49) EMBASE (n = 127)
Records after duplicates removed (n = 148)
Records screened (n = 148)
Records excluded (n = 127)
Conference abstract/review (n = 40)
Study protocol/study design/no outcomes reported (n = 5)
Case-control study (n = 5)
Review/meta-analysis (n = 5)
Scientific (n = 2)
Not open cohort with multiple sites (n = 20)
Database/no setting/population-based/setting unclear (n = 16)
Baseline/qualitative analysis (n = 6)
Not a trial/audit (n = 22)
Mixture of settings (n = 1)
Article not in English (n = 1)
No access to article (n = 2)
Outcomes not on individuals (n = 1)
Not involving humans (n = 1)
Eligible articles (n = 21)
Studies included in analyses (n = 21)

Figure A.3: PRISMA flowchart for the epidemiological studies search (explicit usage).

A.3 Data extraction forms

1. First author
2. Year of publication
3. Description of setting
4. Parallel/factorial
5. Number of arms
6. Is the design explicitly named?
7. Total length of follow-up (from randomisation)
8. Overall LTFU rate of *residents* (primary outcome only)
9. Total number of clusters
10. Average cluster size

Table A.11: Data extraction form for the care homes CRT search.

1. First author
2. Year of publication
3. Description of setting
*4. Parallel/factorial
*5. Number of arms
*6. Protocol/results
*7. Primary outcome type
8. What is described as "open/dynamic cohort"?
*9. Is there linkage of repeated measurements (open cohort analysis)?
*10. Data collection
*11. If discrete, number of measurement points
12. Recruitment
13. If discrete, number of recruitment points
14. Analysis: missing data
15. Analysis: length of stay
16. Analysis: time
17. Analysis: steady state
18. Anything related to an open cohort estimand or research question
19. Any rationale for an open cohort design

Table A.12: Data extraction form for the CRTs and epidemiological studies searches (explicit usage). Elements with an asterisk were not extracted from the epidemiological studies.

A.4 Included studies

First author	**Year**	**Setting**	**Parallel/ factorial**	**No. arms**	**Design named?**	**Total LOFU (from CR)**	**Overall LTFU rate of residents[1] (%)**	**Total no. clusters**	**Avg. cluster size**
Arendts [77]	2018	RAC facilities	Parallel	2	No	32 months (max)	NA	6	NA
Ballard [72]	2018	Nursing homes	Parallel	2	No	9 months	34.7	69	12.3
Boorsma [80]	2011	Residential care facilities	Parallel	2	No	6 months	52.2	10	46.2
Booy [248]	2012	Aged Care Facilities	Parallel	2	No	30 months	NA	16	NA
Chami [249]	2012	Nursing homes	Parallel	2	No	5 months	NA	50	NA
Chen [250]	2016	Dementia special-care units	Parallel	2	No	3 months	12.8	6	32.5
Cheng [251]	2012	Nursing homes	Parallel	3	No	6 months	Unclear	3	12
Chenoweth [68]	2014	RAC homes	Factorial	2 x 2	Yes	14 months	50.7	38	15.8
Colon-Emeric [252]	2017	Nursing homes	Parallel	2	No	12 months	Unclear	24	74.8
Connolly [253]	2015	RAC facilities	Parallel	2	No	14 months	NA	36	NA
Daly [74]	2014	Retirement villages	Parallel	2	No	4 months	9	15	6.7
Davison [254]	2013	Aged care facilities	Parallel	3	No	12 months	Unclear	7	30.9
De Visschere [255]	2012	Nursing homes	Parallel	2	No	6 months	20.4	12	31.1
Ersek [256]	2016	Nursing homes	Parallel	2	No	34 weeks	20.6	27	18
Galik [257]	2015	Assisted living facilities	Parallel	2	No	6 months	15.6	4	24
Gravenstein [258]	2017	Nursing homes	Factorial	2 x 2	No	7 months	NA	823	NA
Hewitt [259]	2018	Long-term RAC facilities	Parallel	2	No	12 months	14	16	13.8
Hodl [260]	2019	Nursing homes	Parallel	2	No	12 weeks	9.4	12	31.8
Husebo [78]	2019	Nursing homes	Parallel	2	No	4 months	27.2	67	8.1
Jancey [75]	2017	Retirement villages	Parallel	2	No	6 months	27.3	38	9.6
Joranson [261]	2015	Nursing homes	Parallel	2	No	6 months	16.7	10	6
Kerr [262]	2018	Retirement communities	Parallel	2	No	12 months	21.5	11	27.9
Koczy [81]	2011	Nursing homes	Parallel	2	No	93 days	22.6	45	9.6
Kon [73]	2017	Wards in a long-term care facility	Parallel	2	No	2 weeks	0	6	5.5
Konner [263]	2015	Nursing homes	Parallel	2	No	6 months	25.9	12	19.9
Kopke [264]	2012	Nursing homes	Parallel	2	No	6 months	NA	36	NA
Kuck [265]	2014	Long-term care facilities	Parallel	2	No	8 weeks	20.6	32	3.3
Lapane [266]	2011	Nursing homes	Parallel	2	No	12 months	NA	25	NA
Leslie [267]	2013	Residential care homes	Parallel	2	No	12 weeks	24.4	18	2.3
Low [268]	2013	Nursing homes	Parallel	2	No	26 weeks	13.8	36	11.1
Mamhidir [269]	2017	Nursing homes	Parallel	2	No	7 months	22.1	10	21.3
Meeks [270]	2015	Nursing homes	Parallel	2	No	34 weeks	30.5	24	3.4
Mitchell [271]	2018	Nursing homes	Parallel	2	No	6 months	1.5	64	6.3
Moyle [272]	2017	Long-term care facilities	Parallel	3	No	10 weeks	15.9	28	16.4
Olsen [273]	2016	Nursing homes	Parallel	2	No	24 weeks	17.2	10	5.8
O'Shea [274]	2014	Long-stay care	Parallel	2	No	20 weeks (avg.)	17.1	18	16.9
Pasay [172]	2019	Nursing homes	Parallel	2	No	12 months	NA	42	NA
Rapp [275]	2013	Nursing homes	Parallel	2	No	10 months	15.1	18	16.9
Roets-Merken [276]	2018	Long-term care homes	Parallel	2	No	9 months	34.8	30	3
Roos [82]	2016	Residential facilities for older people	Parallel	2	No	8 months	25.2	18	5.9
Sackley [71]	2015	Care homes	Parallel	2	No	3 months	38.6	228	4.6
Sambrook [79]	2012	RAC facilities	Parallel	3	No	12 months	13.1	51	11.8
Sinclair [277]	2012	Residential care homes	Parallel	2	No	6 months	18.6	51	2
Testad [278]	2016	Care homes	Parallel	2	No	7 months	26.6	24	11.4
Underwood [70]	2013	Care homes	Parallel	2	No	12 months	44.7	78	11.4
van der Putten [279]	2013	Care homes	Parallel	2	No	6 months	32.4	12	28.6
Weintraub [69]	2018	Nursing homes	Parallel	2	Yes	24 months	Unclear	14	54.4
Williams [280]	2017	Nursing homes	Parallel	2	No	3 months	67.5	13	6.4
Wouters [76]	2017	Nursing home wards	Parallel	2	No	4 months	14.3	59	7.2
Yokoi [281]	2015	Residential care facilities	Parallel	2	No	12 months	21	5	21

Table A.13: Included studies for the care homes search. RAC = residential aged care. LOFU = length of follow-up. [1]NA refers to the trials where the definition of LTFU was unclear. LTFU rate is the individuals missing the primary outcome of all those randomised overall, over the entire trial period.

First author	Year	Setting	Parallel/ factorial	No. arms	Design named?	Total LOFU (from CR)	Overall LTFU rate of residents[1] (%)	Total no. clusters	Avg. cluster size
Diallo [109]	2019	Villages	Parallel	2	No	9 months	0	20	388.3
Lee [112]	2019	Villages	Parallel	2	No	NA	2.4	24	344.5
Loha [110]	2019	Villages	Factorial	2 x 2	No	121 weeks	8.6	176	196.3
Rahman [113]	2019	Neighbourhood of ∼150 households	Parallel	2	No	3 months	5.6	34	18
von Seidlein [111]	2019	Remote village populations	Parallel	2	Yes	1 year	NA	16	527.8
Beeler [104]	2019	Tertiary care academic medical centre	Parallel	2	No	NA	NA	29	NA
Bernitz [107]	2019	Obstetric units in hospitals	Parallel	2	No	NA	0	14	448.4
Chia-Hui Chen [105]	2019	Rooms in a GI ward of a medical center	Parallel	2	No	NA	9.8	38	9.9
Cuypers [108]	2019	Hospitals	Parallel	2	No	52	19.4	18	21.3
Wang [106]	2019	Hospitals	Parallel	2	No	52	1.8	301	36.5
Abernethy [91]	2013	Community-based palliative care service	Factorial	2 x 2 x 2	No	8 weeks	54	105	4.4
Beernaert [88]	2017	Acute geriatric wards in hospitals	Parallel	2	No	NA	NA	10	28.2
McCorkle [89]	2015	Disease-specific clinics in a cancer hospital	Parallel	2	No	3 months	37	4	36.5
Vermandere [92]	2016	Regional offices of a home nursing organization	Parallel	2	No	3-4 weeks	14	18	3.2
Zimmermann [90]	2014	Medical oncology clinics	Parallel	2	No	3 months	38	24	19.2
Beratarrechea [101]	2019	Primary care clinics	Parallel	2	No	6 months	13.9	8	94.4
Ferreira [102]	2019	Healthcare centres	Parallel	2	No	12 months	NA	20	35.8
Flanagan [98]	2019	General practices	Parallel	5	No	NA	27.3	63	197.5
Kristiansen [99]	2019	General practices	Parallel	2	No	17 months	0.9	340	39
Siebenhofer [100]	2019	General practices	Parallel	2	No	24 months	Unclear	52	14.2
Adane [84]	2019	Prisons	Parallel	2	No	1 year	Unclear	16	NA
Donenberg [87]	2018	Probation	Parallel	2	No	6 months	14.8	Unclear	Unclear
Sleed [85]	2013	Mother and baby units in prisons	Parallel	2	No	13 weeks	83	7	23.3
Umbach [86]	2018	Dormitories of a correctional facility	Parallel	2	No	21 weeks (avg.)	26.5	Unclear	Unclear
Karimli [96]	2019	Primary schools	Parallel	3	No	48 months	Unclear	48	29.4
Low [94]	2019	Elementary schools	Parallel	2	No	2 years	NA	61	150
Martinsen [95]	2019	Schools	Parallel	2	No	10 weeks	12.7	36	22.1
Sanchez-Lopez [93]	2019	Schools	Parallel	2	No	10 months	10.8	21	76.6
Sundhot-Borgen [97]	2019	High schools	Parallel	2	No	12 months	55.8	30	81.5

Table A.14: Included studies for the other institutional settings search.

First author	**Year**	**Setting**	**Parallel/ factorial**	**No. arms**	**Protocol/ results**	**Primary outcome type**^{1,2}	**Open or dynamic**	**What is described as "open" or "dynamic cohort"?**	**Linkage of repeated msmts (open cohort analysis)?**	**Data collection**	**If discrete, number of msmt points**
Agarwal [136]	2019	Communities	Parallel	2	Protocol	Continuous, single	Open	Design	NA	Discrete	2
Agarwal [131]	2019	Community-based social housing	Parallel	2	Protocol	Count, repeated	Open	Study	No	Discrete	24
Azad [123]	2010	Communities	Factorial	2 x 2	Results	Single/terminal event	Open	Study population	NA	Continuous	NA
Baiocchi [130]	2017	Schools	Parallel	2	Results	Single/terminal event	Open	Design and study	NA	Discrete	2
Bavarian [133]	2013	Schools	Parallel	2	Results	Continuous, repeated	Dynamic	Study population	Yes	Discrete	8
Bell [115]	2019	Food service operations in school districts	Parallel	2	Results	Event, repeated	Dynamic	Study population	Yes	External sources	NA
Chaboyer [129]	2016	Tertiary referral hospitals	Parallel	2	Results	Time to event	Open	Study	NA	Discrete	28
Clasen [124]	2012	Villages	Parallel	2	Protocol	Event, repeated	Open	Design	No	Discrete	9
Finkelstein [127]	2008	Communities	Parallel	2	Results	Count, repeated	Dynamic	Study population	No	External sources	NA
Greiver [128]	2019	Community-based primary care practices	Parallel	2	Protocol	Count, single time period	Open	Design	NA	External sources	NA
Houweling [125]	2011	Unions	Parallel	2	Protocol	Single/terminal event	Open	Study population	NA	Continuous	NA
Ivers [118]	2017	Care homes	Factorial	2 x 2	Protocol	Continuous^{3}, repeated	Open	Design	Yes	Discrete	12
Lippman [119]	2017	Villages	Parallel	2	Protocol	Event, repeated	Open	Study population	Yes	Continuous	NA
Overgaard [138]	2012	Primary schools	Factorial	2 x 2	Protocol	Count, repeated	Open	Study population	No	Discrete	105
Pape [121]	2011	Primary care clinics	Parallel	2	Results	Binary, single	Open	Unclear	NA	External sources	NA
Pickering [114]	2019	Shared water points in communities	Parallel	2	Results	Event, repeated	Open	Design and trial	No	Discrete	5 to 7
Piotrowski [120]	2020	Care homes	Parallel	2	Protocol	Count, recurrent events in single time period	Open	Design	NA	External sources	NA
Staedke [132]	2016	Public health centres	Parallel	2	Results	Binary, single	Dynamic	Study population	NA	Discrete	2
Tobe [137]	2018	Unions	Parallel	3	Protocol	Single/terminal event	Open	Study population	NA	Discrete	2
Tripathy [126]	2010	Communities	Parallel	2	Results	Single/terminal event	Open	Study population	NA	Discrete	37

Table A.15: Included studies for the CRTs (explicit usage) search. ^{1}If there are multiple primary outcomes the first has been taken. ^{2}If baseline outcomes are treated as outcomes in their own right then they are included as a repeated measure; if baselines are adjusted for only, they have not been treated as a repeated measure. ^{3}Count outcome treated as continuous

First author	**Year**	**Setting**	**Open or dynamic**	**What is described as “open” or “dynamic cohort”?**
Atzeni [143]	2018	Secondary and tertiary care units	Open	Population
Black-Tiong [282]	2021	General practices	Open	Study
Cairney [139]	2017	Public schools	Open	Population, design and study
Cannata-Andia [149]	2013	Dialysis centers	Open	Study and design
d’Arminio Monforte [122]	2019	Infectious disease clinics	Open	Population
de Burgos-Lunar [283]	2013	Health centers	Dynamic	Study
Feakins [146]	2019	General practices	Open	Study
Fernandez-Martin [284]	2015	Dialysis centres	Open	Study and design
Fontela [144]	2012	Intensive care units	Dynamic	Population and study
Hechter [148]	2015	Medical centers	Open	Population
Hermans [116]	2017	Refugee sites	Dynamic	Study
Hippisley-Cox [140]	2008	General practices	Open	Population and study
Islam [285]	2019	Cancer center sites	Open	Study
Klein [117]	2019	Football clubs	Open	Study
Ntouva [286]	2019	GP practices	Open	Study
Oguz [287]	2012	Neonatal intensive care units	Dynamic	Population
Robson [145]	2015	General practices	Open	Population and study
Sawicki [135]	2021	GP practices	Open	Population
Thompson [134]	2017	General practices	Open	Study and design
Toulis [142]	2017	General practices	Open	Study
Xie [147]	2020	Hospitals	Open	Population and study

Table A.16: Included studies for the epidemiological studies (explicit usage) search.

Appendix B

Appendix for Chapter 5

B.1 Sample size

Assumptions	**M = 15**	**M = 50**	**M = 100**
Standardised effect size	0.4	0.4	0.4
Significance level (two-sided)	5%	5%	5%
Power	90%	90%	90%
Sample size without clustering	266	266	266
ICC	0.1	0.1	0.1
Design effect	1+(15-1)*0.1 = 2.4	1+(50-1)*0.1=5.9	1+(100-1)*0.1=10.9
Allocation ratio	1:1	1:1	1:1
Sample size	266x2.4=638.4 660 residents 44 clusters	266x5.9= 1569.4 1600 residents 32 clusters	266x10.9=2899.4 3000 residents 30 clusters

Table B.1: Sample size calculations for different cluster population sizes.

B.2 Generating exponential drop-out times

The general form of the Cox proportional hazards model is

$$h(t \mid x) = h_0(t)\exp(\beta^T x),$$

where $h_0(t)$ denotes the baseline hazard function, β the vector of regression parameters and x the vector of covariates that are associated with the hazard of drop-out. To generate drop-out times for individuals, I assume here that survival times are exponentially distributed, meaning that the baseline hazard function is also exponentially distributed, so that $h_0(t) = \lambda$, where $\lambda > 0$ is the scale parameter.

The survival function corresponding to (B.2) is then

$$S(t \mid x) = \exp[-H_0(t)\exp(\beta^T x)],$$

where the cumulative baseline hazard function is

$$H_0(t) = \int_0^t h_0(r)\mathrm{d}r,$$

and the cumulative distribution function (CDF) of the Cox model is

$$F(t \mid x) = 1 - \exp[-H_0(t)\exp(\beta^T x)].$$

As described in [288], drop-out times are generated using the 'Inverse CDF' theorem as follows. First, let Y be a random variable with CDF F. I can then set $U = F(Y)$, where $U \sim \mathrm{U}(0,1)$ is a uniformly distributed variable, due to the properties of a CDF. Furthermore, if $U \sim \mathrm{U}(0,1)$, then it follows that $(1-U) \sim \mathrm{U}(0,1)$. If T is the survival time according to model (B.2), the CDF (B.2) leads to

$$U = \exp[-H_0(t)\exp(\beta^T x)] \sim \mathrm{U}(0,1).$$

The cumulative baseline hazard H_0 can be inverted as long as the baseline hazard $h_0(t) > 0$ for all t, and as a result

$$T = H_0^{-1}[-\log(U)\exp(-\beta^T x)].$$

For the exponential distribution, the inverse baseline hazard function is simply $H_0^{-1}(t) = \lambda^{-1}t$. It follows from (B.2) that the survival time T is then

$$T = \lambda^{-1}[-\log(U)\exp(-\beta^T x)].$$

The exact form of $\beta^T x$ in the hazard models for the different missing data mechanisms are given previously in Section 5.2.2.6. The vector β is either δ_0, (δ_0, δ_1) or $(\delta_0, \delta_1, \delta_2)$ depending on whether the missing data mechanism is MCAR, MAR or MNAR respectively. Similarly, in these three scenarios x is either a column vector of ones, a matrix of ones and previous y values, or a matrix of ones and previous and future y values. I have also assumed that $\lambda = 1$.

B.3 Simulation study flowchart

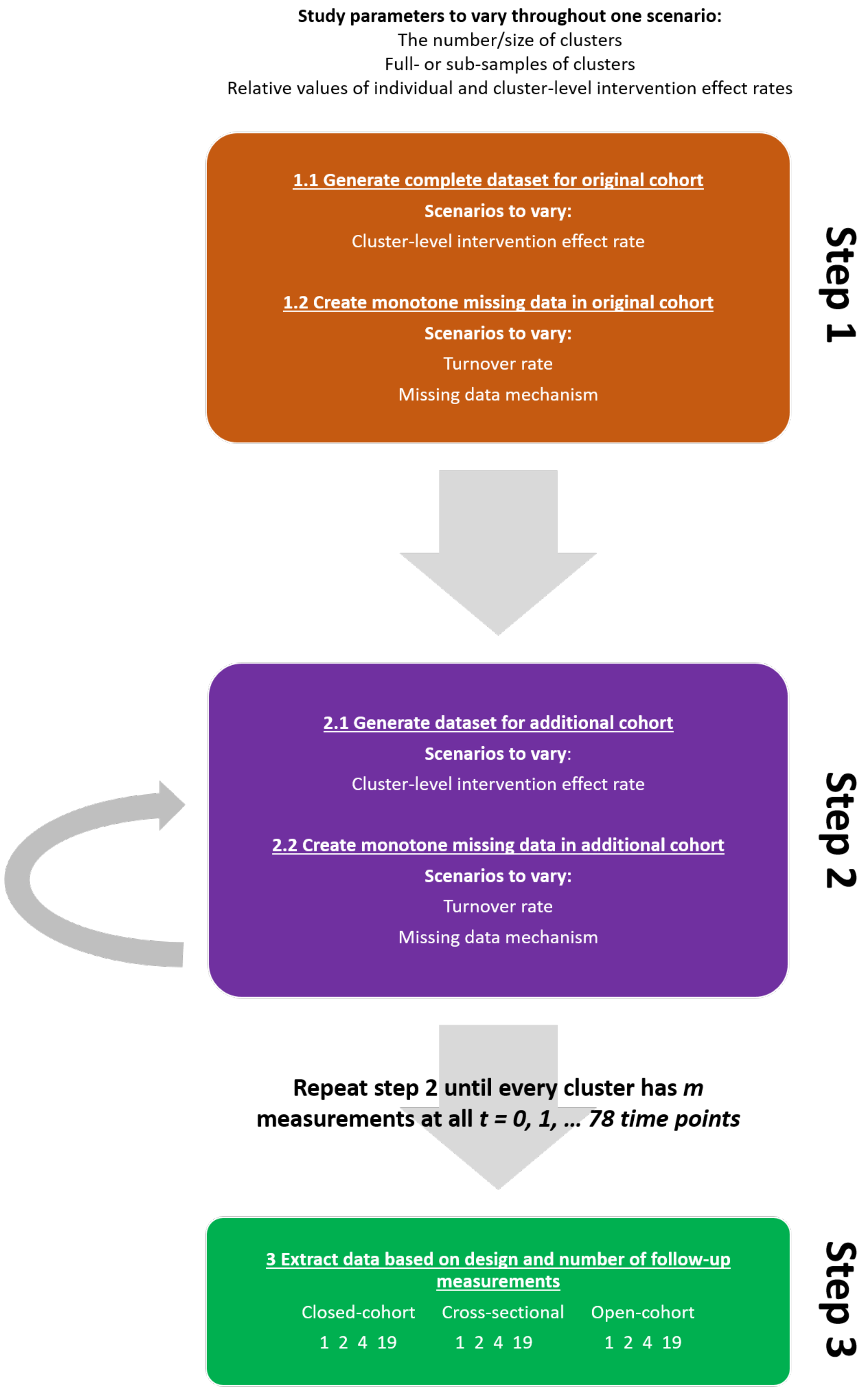

Study parameters to vary throughout one scenario:
The number/size of clusters
Full- or sub-samples of clusters
Relative values of individual and cluster-level intervention effect rates
1.1 Generate complete dataset for original cohort
Scenarios to vary:
Cluster-level intervention effect rate
1.2 Create monotone missing data in original cohort
Scenarios to vary:
Turnover rate
Missing data mechanism
Step 1
2.1 Generate dataset for additional cohort
Scenarios to vary:
Cluster-level intervention effect rate
2.2 Create monotone missing data in additional cohort
Scenarios to vary:
Turnover rate
Missing data mechanism
Step 2
Repeat step 2 until every cluster has *m* measurements at all *t = 0, 1, … 78 time points*
3 Extract data based on design and number of follow-up measurements
Closed-cohort
1 2 4 19
Cross-sectional
1 2 4 19
Open-cohort
1 2 4 19
Step 3

Figure B.1: The sequence of data generation in the simulation study including the study parameters and complications that can be varied at each stage.

B.4 Figures

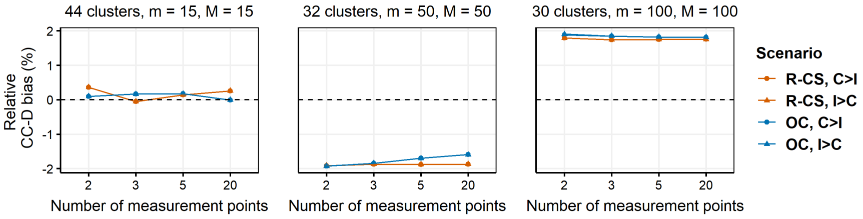


Figure B.2: Relative CC-D bias in the base case using 2 timescales for full-samples.

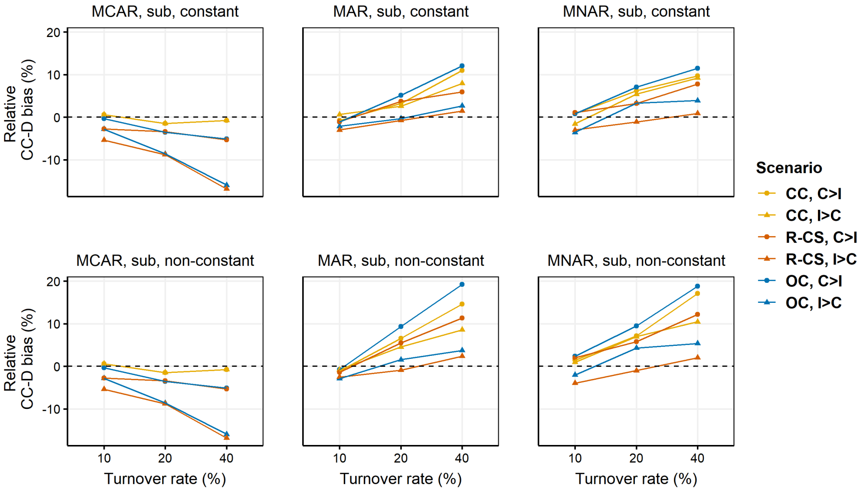


Figure B.3: CC-D bias in the complications scenarios using 1 timescale for sub-samples.

B.5 OC-26-I estimand

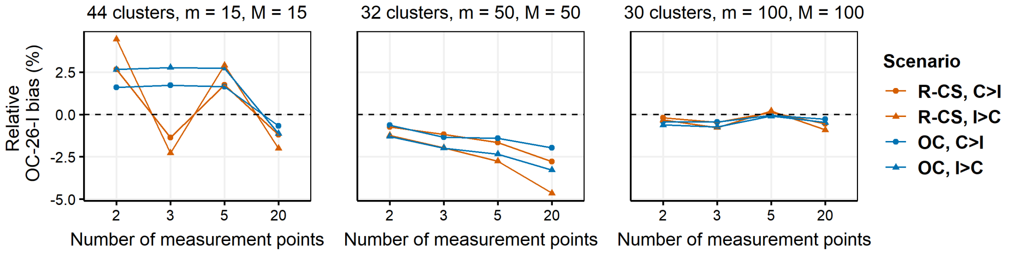


Figure B.4: Relative OC-26-I bias in the base case using 2 timescales for full-samples.

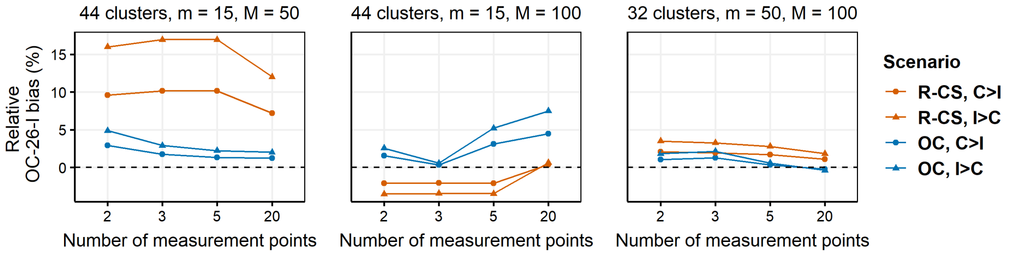


Figure B.5: Relative OC-26-I bias in the base case using 2 timescales for sub-samples.

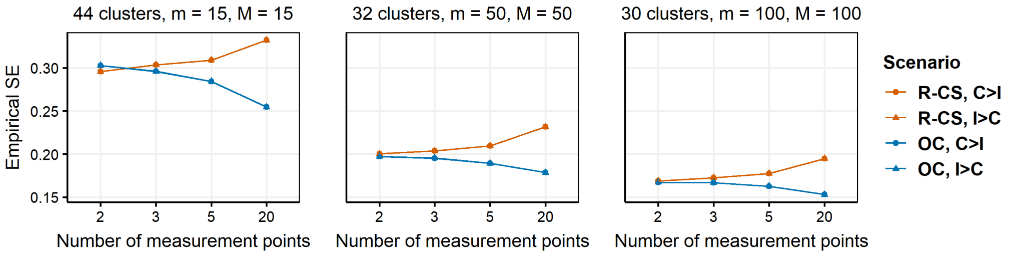


Figure B.6: Empirical SE for OC-26-I in the base case using 2 timescales for full-samples.

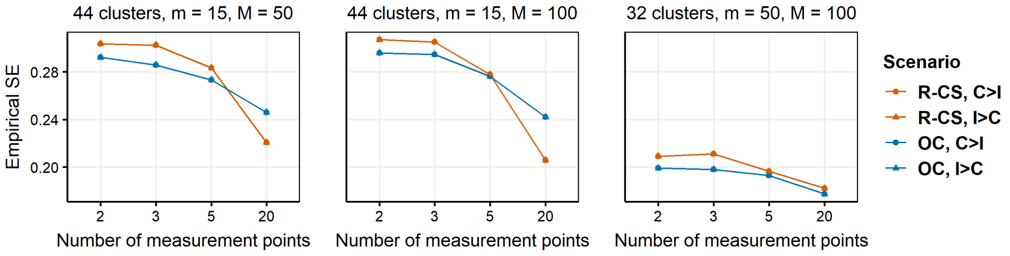


Figure B.7: Empirical SE for OC-26-I in the base case using 2 timescales for sub-samples.

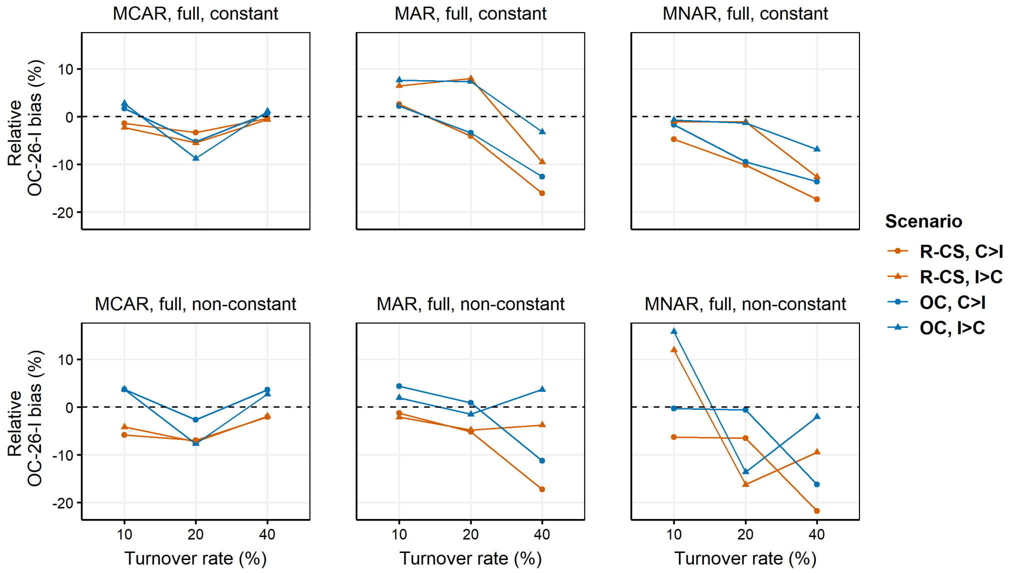


Figure B.8: Relative OC-26-I bias in the complications scenario using 2 timescales for full-samples.

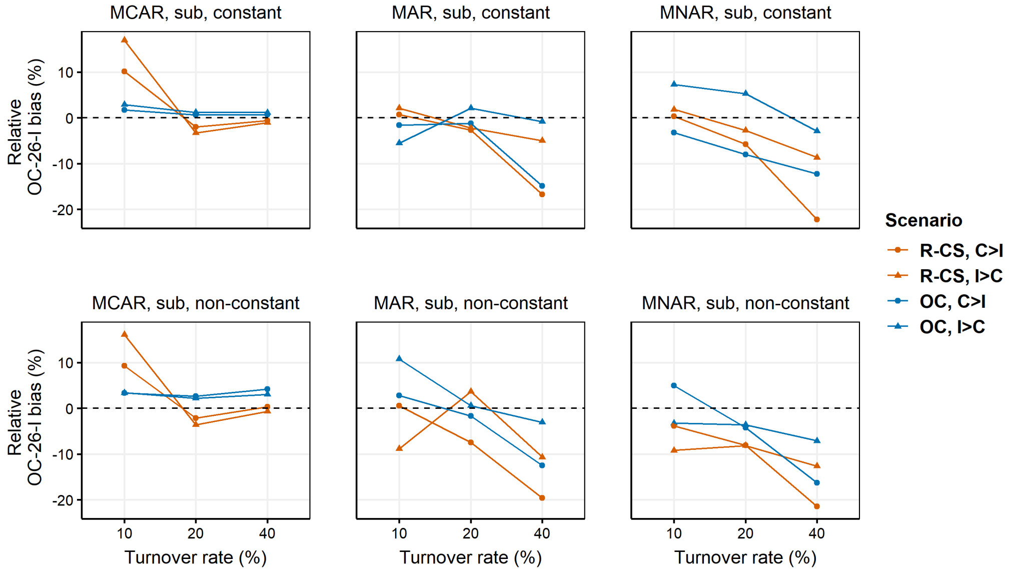


Figure B.9: Relative OC-26-I bias in the complications scenario using 2 timescales for sub-samples.

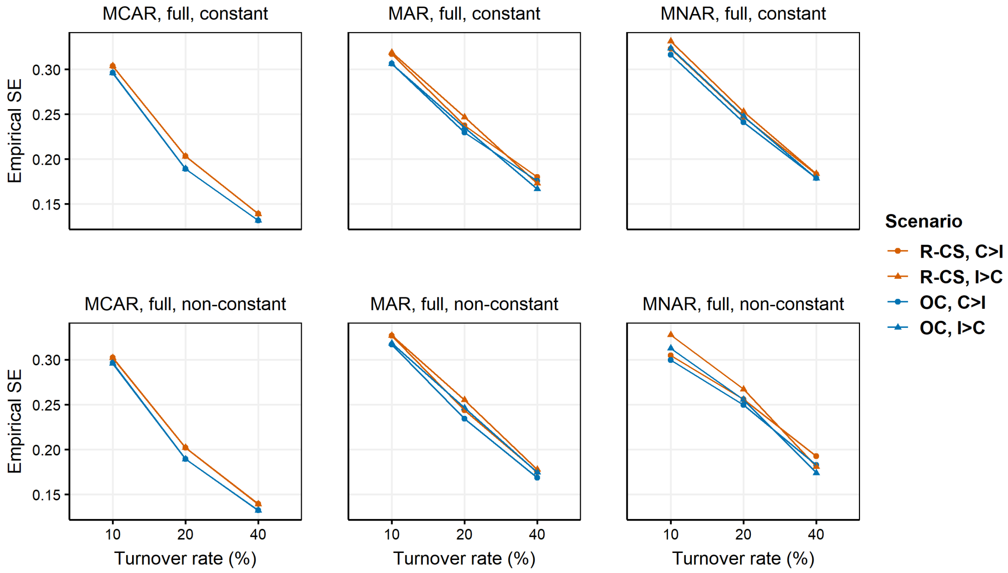


Figure B.10: Empirical SE for OC-26-I in the complications scenario using 2 timescales for full-samples.

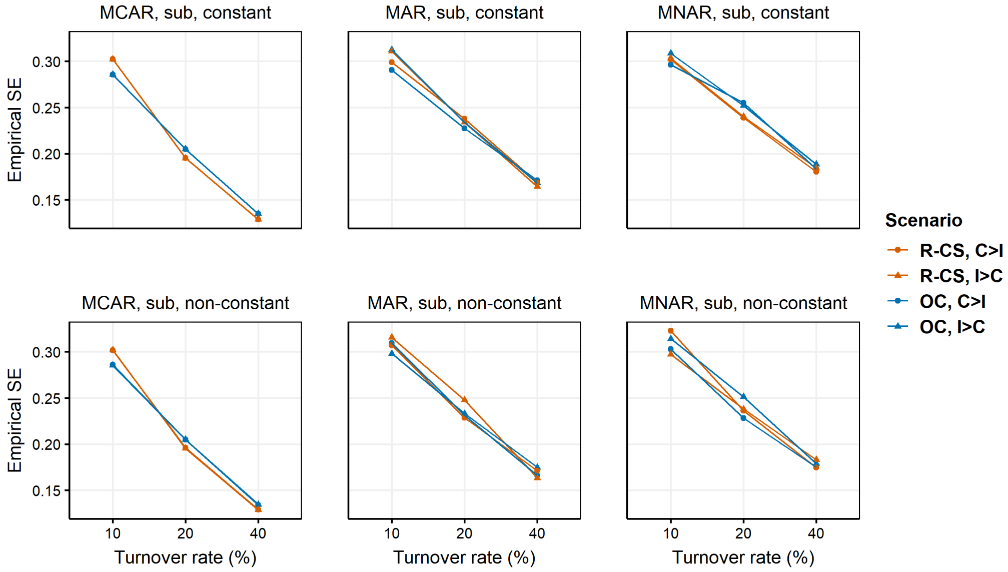


Figure B.11: Empirical SE in OC-26-I in the complications scenario using 2 timescales for sub-samples.

B.6 R, Stata and SAS code for Kasza's single-timescale model

R code

For the closed and open cohort designs:

```
lmer(y ~ timept + trt:timept + (1|cluster) + (1|cluster:timept) + (1|cluster:subject), dataframe,
            control=lmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)))
```

For the cross-sectional design, drop the subject random effect:

```
lmer(y ~ timept + trt:timept + (1|cluster) + (1|cluster:timept), dataframe,
            control=lmerControl(optimizer="bobyqa", optCtrl=list(maxfun=2e5)))
```

Stata code

For the closed and open cohort designs:

```
mixed y i.timept trt#timept || cluster: || cluster: R.timept || subject:, reml
```

For the cross-sectional design, drop the subject random effect:

```
mixed y i.timept trt#timept || cluster: || cluster: R.timept, reml
```

SAS code

For the closed and open cohort designs, method 1:

```
proc mixed;
class trt cluster timept;
model y = timept trt*timept /s;
random int timept/subject=cluster(trt);
random int/subject=subject(cluster*trt);
run;
```

For the closed and open cohort designs, method 2:

```
proc mixed;
class trt cluster timept;
model y = timept trt*timept /s;
random int timept/subject=cluster(trt);
repeated timept/subject=subject(cluster*trt) type=cs;
run;
```

For the cross-sectional design:

```
proc mixed;
class trt(ref="0") cluster timept(ref="0");
```

```
model y = timept trt*timept /s;
random int timept/subject=cluster(trt);
run;
```

B.7 Tables

No. clusters	m	M	Design	Time points	Intvn. effects	Avg. total estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	Total empSE	Total MCSE (bias)	Total MCSE (empSE)	Non-convergence (%)
44	15	15	R-CS	2	C>I	-0.4	-0.107	1.834	0.361	0.122	0.005	0.004	15.077
44	15	15	OC	2	C>I	-0.399	-0.371	1.564	0.096	0.122	0.005	0.003	2.308
44	15	15	R-CS	2	I>C	-0.4	-0.107	4.575	0.361	0.122	0.005	0.004	15.077
44	15	15	OC	2	I>C	-0.399	-0.371	4.299	0.096	0.122	0.005	0.003	2.308
44	15	50	OC	2	C>I	-0.401	0.105	2.049	0.574	0.126	0.005	0.004	2.923
44	15	50	R-CS	2	C>I	-0.391	-2.463	-0.569	-2.006	0.118	0.005	0.003	3.538
44	15	50	OC	2	I>C	-0.401	0.105	4.797	0.574	0.126	0.005	0.004	2.923
44	15	50	R-CS	2	I>C	-0.391	-2.463	2.108	-2.006	0.118	0.005	0.003	3.538
44	15	100	OC	2	C>I	-0.405	0.946	2.906	1.419	0.12	0.005	0.003	3.385
44	15	100	R-CS	2	C>I	-0.412	2.807	4.803	3.288	0.129	0.005	0.004	4.462
44	15	100	OC	2	I>C	-0.405	0.946	5.677	1.419	0.12	0.005	0.003	3.385
44	15	100	R-CS	2	I>C	-0.412	2.807	7.625	3.288	0.129	0.005	0.004	4.462
32	50	50	R-CS	2	C>I	-0.391	-2.375	-0.479	-1.917	0.125	0.005	0.003	0.462
32	50	50	OC	2	C>I	-0.391	-2.387	-0.491	-1.93	0.125	0.005	0.003	1.385
32	50	50	R-CS	2	I>C	-0.391	-2.375	2.201	-1.917	0.125	0.005	0.003	0.462
32	50	50	OC	2	I>C	-0.391	-2.387	2.188	-1.929	0.125	0.005	0.003	1.231
32	50	100	OC	2	C>I	-0.403	0.639	2.594	1.111	0.125	0.005	0.003	1.077
32	50	100	R-CS	2	C>I	-0.394	-1.652	0.258	-1.191	0.123	0.005	0.003	0.308
32	50	100	OC	2	I>C	-0.403	0.639	5.356	1.111	0.125	0.005	0.003	1.077
32	50	100	R-CS	2	I>C	-0.394	-1.652	2.958	-1.191	0.123	0.005	0.003	0.308
30	100	100	R-CS	2	C>I	-0.406	1.315	3.283	1.79	0.121	0.005	0.003	0.462
30	100	100	OC	2	C>I	-0.407	1.406	3.375	1.881	0.121	0.005	0.003	0.923
30	100	100	R-CS	2	I>C	-0.406	1.315	6.064	1.79	0.121	0.005	0.003	0.462
30	100	100	OC	2	I>C	-0.407	1.431	6.184	1.906	0.121	0.005	0.003	1.077

Table B.2: Results for the base case discrete model with two timescales for two time points only.

No. clusters	**m**	**M**	**Design**	**Time points**	**Intvn. effects**	**Avg. total estimate**	**CC-I bias (%)**	**CS-OC-D bias (%)**	**CC-D bias (%)**	**Total empSE**	**Total MCSE (bias)**	**Total MCSE (empSE)**	**Non-convergence (%)**
44	15	15	R-CS	3	C>I	-0.399	-0.52	1.412	-0.054	0.122	0.005	0.003	6.308
44	15	15	OC	3	C>I	-0.4	-0.302	1.634	0.165	0.122	0.005	0.003	0.923
44	15	15	R-CS	3	I>C	-0.399	-0.52	4.143	-0.054	0.122	0.005	0.003	6.308
44	15	15	OC	3	I>C	-0.4	-0.302	4.37	0.165	0.122	0.005	0.003	0.923
44	15	50	OC	3	C>I	-0.402	0.277	2.224	0.746	0.125	0.005	0.003	1.077
44	15	50	R-CS	3	C>I	-0.391	-2.515	-0.622	-2.059	0.118	0.005	0.003	0.462
44	15	50	OC	3	I>C	-0.402	0.277	4.976	0.746	0.125	0.005	0.003	1.077
44	15	50	R-CS	3	I>C	-0.391	-2.515	2.053	-2.059	0.118	0.005	0.003	0.462
44	15	100	OC	3	C>I	-0.404	0.73	2.686	1.201	0.12	0.005	0.003	1.385
44	15	100	R-CS	3	C>I	-0.412	2.79	4.787	3.272	0.127	0.005	0.004	0.154
44	15	100	OC	3	I>C	-0.404	0.73	5.451	1.201	0.12	0.005	0.003	1.385
44	15	100	R-CS	3	I>C	-0.412	2.79	7.608	3.272	0.127	0.005	0.004	0.154
32	50	50	R-CS	3	C>I	-0.392	-2.33	-0.433	-1.873	0.125	0.005	0.003	0
32	50	50	OC	3	C>I	-0.392	-2.301	-0.403	-1.843	0.125	0.005	0.003	0.462
32	50	50	R-CS	3	I>C	-0.392	-2.33	2.247	-1.873	0.125	0.005	0.003	0
32	50	50	OC	3	I>C	-0.392	-2.301	2.278	-1.843	0.125	0.005	0.003	0.462
32	50	100	OC	3	C>I	-0.404	0.825	2.783	1.297	0.126	0.005	0.003	0.308
32	50	100	R-CS	3	C>I	-0.395	-1.54	0.372	-1.079	0.123	0.005	0.003	0
32	50	100	OC	3	I>C	-0.404	0.825	5.55	1.297	0.126	0.005	0.003	0.308
32	50	100	R-CS	3	I>C	-0.395	-1.54	3.075	-1.079	0.123	0.005	0.003	0
30	100	100	R-CS	3	C>I	-0.406	1.267	3.234	1.741	0.121	0.005	0.003	0
30	100	100	OC	3	C>I	-0.406	1.371	3.339	1.845	0.121	0.005	0.003	0
30	100	100	R-CS	3	I>C	-0.406	1.267	6.013	1.741	0.121	0.005	0.003	0
30	100	100	OC	3	I>C	-0.406	1.371	6.122	1.845	0.121	0.005	0.003	0

Table B.3: Results for the base case discrete model with two timescales for three time points only.

No. clusters	m	M	Design	Time points	Intvn. effects	Avg. total estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	Total empSE	Total MCSE (bias)	Total MCSE (empSE)	Non-convergence (%)
44	15	15	R-CS	5	C>I	-0.4	-0.325	1.61	0.141	0.122	0.005	0.003	1.077
44	15	15	OC	5	C>I	-0.4	-0.293	1.643	0.174	0.122	0.005	0.003	0.154
44	15	15	R-CS	5	I>C	-0.4	-0.325	4.346	0.141	0.122	0.005	0.003	1.077
44	15	15	OC	5	I>C	-0.4	-0.293	4.38	0.174	0.122	0.005	0.003	0.154
44	15	50	OC	5	C>I	-0.4	-0.227	1.711	0.24	0.125	0.005	0.003	0.154
44	15	50	R-CS	5	C>I	-0.391	-2.502	-0.608	-2.045	0.117	0.005	0.003	0
44	15	50	OC	5	I>C	-0.4	-0.227	4.449	0.24	0.125	0.005	0.003	0.154
44	15	50	R-CS	5	I>C	-0.391	-2.502	2.067	-2.045	0.117	0.005	0.003	0
44	15	100	OC	5	C>I	-0.404	0.813	2.771	1.286	0.121	0.005	0.003	0.462
44	15	100	R-CS	5	C>I	-0.412	2.804	4.801	3.286	0.127	0.005	0.004	0
44	15	100	OC	5	I>C	-0.404	0.813	5.538	1.286	0.121	0.005	0.003	0.462
44	15	100	R-CS	5	I>C	-0.412	2.804	7.622	3.286	0.127	0.005	0.004	0
32	50	50	R-CS	5	C>I	-0.392	-2.336	-0.439	-1.878	0.125	0.005	0.003	0
32	50	50	OC	5	C>I	-0.392	-2.159	-0.259	-1.701	0.125	0.005	0.003	0
32	50	50	R-CS	5	I>C	-0.392	-2.336	2.242	-1.878	0.125	0.005	0.003	0
32	50	50	OC	5	I>C	-0.392	-2.159	2.427	-1.701	0.125	0.005	0.003	0
32	50	100	OC	5	C>I	-0.404	0.798	2.756	1.27	0.126	0.005	0.003	0
32	50	100	R-CS	5	C>I	-0.395	-1.545	0.367	-1.084	0.123	0.005	0.003	0
32	50	100	OC	5	I>C	-0.404	0.798	5.522	1.27	0.126	0.005	0.003	0
32	50	100	R-CS	5	I>C	-0.395	-1.545	3.069	-1.084	0.123	0.005	0.003	0
30	100	100	R-CS	5	C>I	-0.406	1.276	3.243	1.75	0.121	0.005	0.003	0
30	100	100	OC	5	C>I	-0.406	1.347	3.315	1.821	0.121	0.005	0.003	0
30	100	100	R-CS	5	I>C	-0.406	1.276	6.023	1.75	0.121	0.005	0.003	0
30	100	100	OC	5	I>C	-0.406	1.347	6.097	1.821	0.121	0.005	0.003	0

Table B.4: Results for the base case discrete model with two timescales for five time points only.

No. clusters	m	M	Design	Time points	Intvn. effects	Avg. total estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	Total empSE	Total MCSE (bias)	Total MCSE (empSE)	Non-convergence (%)
44	15	15	R-CS	20	C>I	-0.4	-0.212	1.726	0.255	0.122	0.005	0.003	0
44	15	15	OC	20	C>I	-0.399	-0.475	1.458	-0.009	0.121	0.005	0.003	0
44	15	15	R-CS	20	I>C	-0.4	-0.212	4.465	0.255	0.122	0.005	0.003	0
44	15	15	OC	20	I>C	-0.399	-0.475	4.189	-0.009	0.121	0.005	0.003	0
44	15	50	OC	20	C>I	-0.401	-0.048	1.894	0.421	0.124	0.005	0.003	0.154
44	15	50	R-CS	20	C>I	-0.391	-2.468	-0.573	-2.011	0.117	0.005	0.003	0
44	15	50	OC	20	I>C	-0.401	-0.048	4.637	0.421	0.124	0.005	0.003	0.154
44	15	50	R-CS	20	I>C	-0.391	-2.468	2.103	-2.011	0.117	0.005	0.003	0
44	15	100	OC	20	C>I	-0.404	0.705	2.66	1.176	0.12	0.005	0.003	0.308
44	15	100	R-CS	20	C>I	-0.412	2.821	4.818	3.303	0.127	0.005	0.004	0
44	15	100	OC	20	I>C	-0.404	0.705	5.424	1.176	0.12	0.005	0.003	0.308
44	15	100	R-CS	20	I>C	-0.412	2.821	7.64	3.303	0.127	0.005	0.004	0
32	50	50	R-CS	20	C>I	-0.392	-2.326	-0.429	-1.869	0.125	0.005	0.003	0
32	50	50	OC	20	C>I	-0.393	-2.053	-0.15	-1.594	0.124	0.005	0.003	0
32	50	50	R-CS	20	I>C	-0.392	-2.326	2.251	-1.869	0.125	0.005	0.003	0
32	50	50	OC	20	I>C	-0.393	-2.053	2.538	-1.594	0.124	0.005	0.003	0
32	50	100	OC	20	C>I	-0.404	0.883	2.842	1.355	0.124	0.005	0.003	0
32	50	100	R-CS	20	C>I	-0.395	-1.553	0.359	-1.092	0.123	0.005	0.003	0
32	50	100	OC	20	I>C	-0.404	0.883	5.611	1.355	0.124	0.005	0.003	0
32	50	100	R-CS	20	I>C	-0.395	-1.553	3.061	-1.092	0.123	0.005	0.003	0
30	100	100	R-CS	20	C>I	-0.406	1.279	3.246	1.754	0.121	0.005	0.003	0
30	100	100	OC	20	C>I	-0.406	1.339	3.307	1.814	0.12	0.005	0.003	0
30	100	100	R-CS	20	I>C	-0.406	1.279	6.026	1.754	0.121	0.005	0.003	0
30	100	100	OC	20	I>C	-0.406	1.339	6.089	1.814	0.12	0.005	0.003	0

Table B.5: Results for the base case discrete model with two timescales for twenty time points only.

No. clusters	m	M	Design	Time points	Intvn. effects	Avg. estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	EmpSE	MCSE (bias)	MCSE (empSE)	Non-convergence (%)
44	15	15	CC	2	C>I	-0.399	-0.514	1.419	-0.048	0.122	0.005	0.003	2.615
44	15	15	R-CS	2	C>I	-0.391	-2.412	-0.517	-1.955	0.115	0.005	0.003	14.308
44	15	15	OC	2	C>I	-0.393	-1.867	0.039	-1.408	0.114	0.005	0.003	3.231
44	15	15	CC	2	I>C	-0.399	-0.514	4.149	-0.048	0.122	0.005	0.003	2.615
44	15	15	R-CS	2	I>C	-0.381	-5.032	-0.581	-4.587	0.115	0.005	0.003	14.769
44	15	15	OC	2	I>C	-0.383	-4.387	0.094	-3.939	0.114	0.005	0.003	3.385
44	15	50	CC	2	C>I	-0.401	0.103	2.047	0.572	0.126	0.005	0.004	2.923
44	15	50	OC	2	C>I	-0.398	-0.651	1.278	-0.186	0.119	0.005	0.003	4.615
44	15	50	R-CS	2	C>I	-0.388	-3.203	-1.324	-2.75	0.11	0.004	0.003	4.923
44	15	50	CC	2	I>C	-0.401	0.103	4.795	0.572	0.126	0.005	0.004	2.923
44	15	50	OC	2	I>C	-0.388	-3.165	1.373	-2.712	0.118	0.005	0.003	4.615
44	15	50	R-CS	2	I>C	-0.378	-5.681	-1.26	-5.239	0.11	0.004	0.003	4.923
44	15	100	CC	2	C>I	-0.405	0.988	2.95	1.461	0.12	0.005	0.003	3.231
44	15	100	OC	2	C>I	-0.397	-0.935	0.989	-0.471	0.113	0.005	0.003	4.154
44	15	100	R-CS	2	C>I	-0.406	1.199	3.165	1.673	0.12	0.005	0.003	4.769
44	15	100	CC	2	I>C	-0.405	0.988	5.722	1.461	0.12	0.005	0.003	3.231
44	15	100	OC	2	I>C	-0.388	-3.325	1.206	-2.872	0.113	0.005	0.003	4
44	15	100	R-CS	2	I>C	-0.395	-1.358	3.266	-0.896	0.12	0.005	0.003	4.615
32	50	50	CC	2	C>I	-0.391	-2.404	-0.508	-1.947	0.125	0.005	0.003	1.231
32	50	50	R-CS	2	C>I	-0.387	-3.445	-1.57	-2.993	0.118	0.005	0.003	0.462
32	50	50	OC	2	C>I	-0.387	-3.554	-1.681	-3.102	0.118	0.005	0.003	1.846
32	50	50	CC	2	I>C	-0.391	-2.404	2.17	-1.947	0.125	0.005	0.003	1.231
32	50	50	R-CS	2	I>C	-0.377	-6.013	-1.608	-5.572	0.118	0.005	0.003	0.462
32	50	50	OC	2	I>C	-0.377	-6.032	-1.628	-5.592	0.118	0.005	0.003	1.846
32	50	100	CC	2	C>I	-0.404	0.652	2.607	1.124	0.125	0.005	0.003	0.615
32	50	100	OC	2	C>I	-0.397	-1.085	0.836	-0.622	0.116	0.005	0.003	1.692
32	50	100	R-CS	2	C>I	-0.392	-2.333	-0.436	-1.876	0.114	0.004	0.003	0.462
32	50	100	CC	2	I>C	-0.404	0.652	5.369	1.124	0.125	0.005	0.003	0.615
32	50	100	OC	2	I>C	-0.386	-3.671	0.844	-3.219	0.116	0.005	0.003	1.692
32	50	100	R-CS	2	I>C	-0.381	-4.91	-0.453	-4.465	0.114	0.004	0.003	0.462
30	100	100	CC	2	C>I	-0.407	1.422	3.392	1.897	0.121	0.005	0.003	0.462
30	100	100	R-CS	2	C>I	-0.401	0.078	2.022	0.547	0.11	0.004	0.003	0.615
30	100	100	OC	2	C>I	-0.402	0.2	2.146	0.669	0.11	0.004	0.003	1.538
30	100	100	CC	2	I>C	-0.407	1.422	6.175	1.897	0.121	0.005	0.003	0.462
30	100	100	R-CS	2	I>C	-0.391	-2.525	2.043	-2.069	0.109	0.004	0.003	0.769
30	100	100	OC	2	I>C	-0.392	-2.253	2.328	-1.795	0.11	0.004	0.003	1.385

Table B.6: Results for the base case discrete model with one timescale for two time points only.

No. clusters	m	M	Design	Time points	Intvn. effects	Avg. estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	EmpSE	MCSE (bias)	MCSE (empSE)	Non-convergence (%)
44	15	15	CC	3	C>I	-0.399	-0.361	1.574	0.105	0.122	0.005	0.003	0.769
44	15	15	R-CS	3	C>I	-0.392	-2.228	-0.329	-1.77	0.114	0.005	0.003	6.308
44	15	15	OC	3	C>I	-0.394	-1.647	0.263	-1.186	0.115	0.005	0.003	1.077
44	15	15	CC	3	I>C	-0.399	-0.361	4.309	0.105	0.122	0.005	0.003	0.769
44	15	15	R-CS	3	I>C	-0.381	-4.921	-0.465	-4.476	0.114	0.005	0.003	6.462
44	15	15	OC	3	I>C	-0.384	-4.12	0.374	-3.671	0.115	0.005	0.003	0.923
44	15	50	CC	3	C>I	-0.401	0.104	2.048	0.573	0.125	0.005	0.003	0.769
44	15	50	OC	3	C>I	-0.398	-0.833	1.092	-0.369	0.118	0.005	0.003	0.923
44	15	50	R-CS	3	C>I	-0.388	-3.229	-1.349	-2.776	0.109	0.004	0.003	0.769
44	15	50	CC	3	I>C	-0.401	0.104	4.796	0.573	0.125	0.005	0.003	0.769
44	15	50	OC	3	I>C	-0.388	-3.296	1.237	-2.843	0.117	0.005	0.003	0.923
44	15	50	R-CS	3	I>C	-0.378	-5.839	-1.425	-5.398	0.109	0.004	0.003	0.769
44	15	100	CC	3	C>I	-0.404	0.851	2.81	1.324	0.121	0.005	0.003	0.923
44	15	100	OC	3	C>I	-0.398	-0.719	1.21	-0.254	0.114	0.004	0.003	1.846
44	15	100	R-CS	3	C>I	-0.405	1.021	2.983	1.495	0.12	0.005	0.003	0.154
44	15	100	CC	3	I>C	-0.404	0.851	5.578	1.324	0.121	0.005	0.003	0.923
44	15	100	OC	3	I>C	-0.388	-3.154	1.385	-2.701	0.113	0.004	0.003	1.846
44	15	100	R-CS	3	I>C	-0.395	-1.562	3.052	-1.101	0.119	0.005	0.003	0.154
32	50	50	CC	3	C>I	-0.392	-2.27	-0.372	-1.812	0.125	0.005	0.003	0.308
32	50	50	R-CS	3	C>I	-0.387	-3.457	-1.582	-3.005	0.118	0.005	0.003	0
32	50	50	OC	3	C>I	-0.388	-3.322	-1.444	-2.869	0.118	0.005	0.003	0.462
32	50	50	CC	3	I>C	-0.392	-2.27	2.311	-1.812	0.125	0.005	0.003	0.308
32	50	50	R-CS	3	I>C	-0.377	-6.02	-1.615	-5.579	0.118	0.005	0.003	0
32	50	50	OC	3	I>C	-0.378	-5.774	-1.358	-5.333	0.117	0.005	0.003	0.462
32	50	100	CC	3	C>I	-0.404	0.854	2.813	1.326	0.126	0.005	0.003	0
32	50	100	OC	3	C>I	-0.398	-0.837	1.089	-0.372	0.116	0.005	0.003	0.462
32	50	100	R-CS	3	C>I	-0.391	-2.388	-0.492	-1.931	0.114	0.004	0.003	0
32	50	100	CC	3	I>C	-0.404	0.854	5.581	1.326	0.126	0.005	0.003	0
32	50	100	OC	3	I>C	-0.388	-3.263	1.271	-2.81	0.116	0.005	0.003	0.462
32	50	100	R-CS	3	I>C	-0.381	-4.966	-0.512	-4.521	0.114	0.004	0.003	0
30	100	100	CC	3	C>I	-0.407	1.394	3.363	1.869	0.121	0.005	0.003	0
30	100	100	R-CS	3	C>I	-0.401	-0.023	1.919	0.445	0.109	0.004	0.003	0
30	100	100	OC	3	C>I	-0.401	0.091	2.035	0.56	0.11	0.004	0.003	0
30	100	100	CC	3	I>C	-0.407	1.394	6.146	1.869	0.121	0.005	0.003	0
30	100	100	R-CS	3	I>C	-0.391	-2.574	1.992	-2.118	0.109	0.004	0.003	0
30	100	100	OC	3	I>C	-0.392	-2.346	2.231	-1.889	0.11	0.004	0.003	0

Table B.7: Results for the base case discrete model with one timescale for three time points only.

No. clusters	m	M	Design	Time points	Intvn. effects	Avg. estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	EmpSE	MCSE (bias)	MCSE (empSE)	Non-convergence (%)
44	15	15	CC	5	C>I	-0.4	-0.315	1.621	0.151	0.122	0.005	0.003	0.154
44	15	15	R-CS	5	C>I	-0.394	-1.803	0.104	-1.343	0.115	0.005	0.003	1.385
44	15	15	OC	5	C>I	-0.394	-1.678	0.232	-1.217	0.115	0.005	0.003	0.154
44	15	15	CC	5	I>C	-0.4	-0.315	4.357	0.151	0.122	0.005	0.003	0.154
44	15	15	R-CS	5	I>C	-0.383	-4.381	0.101	-3.933	0.115	0.005	0.003	1.385
44	15	15	OC	5	I>C	-0.384	-4.122	0.372	-3.673	0.115	0.005	0.003	0.154
44	15	50	CC	5	C>I	-0.4	-0.185	1.753	0.282	0.125	0.005	0.003	0
44	15	50	OC	5	C>I	-0.397	-1.074	0.847	-0.611	0.118	0.005	0.003	0.154
44	15	50	R-CS	5	C>I	-0.388	-3.22	-1.341	-2.767	0.109	0.004	0.003	0
44	15	50	CC	5	I>C	-0.4	-0.185	4.493	0.282	0.125	0.005	0.003	0
44	15	50	OC	5	I>C	-0.387	-3.518	1.004	-3.066	0.118	0.005	0.003	0.154
44	15	50	R-CS	5	I>C	-0.378	-5.83	-1.417	-5.389	0.109	0.004	0.003	0
44	15	100	CC	5	C>I	-0.405	0.902	2.862	1.375	0.121	0.005	0.003	0.308
44	15	100	OC	5	C>I	-0.399	-0.411	1.524	0.056	0.114	0.004	0.003	0.308
44	15	100	R-CS	5	C>I	-0.405	1.036	2.998	1.509	0.12	0.005	0.003	0
44	15	100	CC	5	I>C	-0.405	0.902	5.631	1.375	0.121	0.005	0.003	0.308
44	15	100	OC	5	I>C	-0.39	-2.837	1.716	-2.382	0.114	0.004	0.003	0.308
44	15	100	R-CS	5	I>C	-0.395	-1.548	3.067	-1.087	0.119	0.005	0.003	0
32	50	50	CC	5	C>I	-0.392	-2.182	-0.282	-1.723	0.125	0.005	0.003	0
32	50	50	R-CS	5	C>I	-0.387	-3.457	-1.582	-3.005	0.118	0.005	0.003	0
32	50	50	OC	5	C>I	-0.388	-3.241	-1.362	-2.788	0.118	0.005	0.003	0
32	50	50	CC	5	I>C	-0.392	-2.182	2.403	-1.723	0.125	0.005	0.003	0
32	50	50	R-CS	5	I>C	-0.377	-6.02	-1.615	-5.579	0.118	0.005	0.003	0
32	50	50	OC	5	I>C	-0.378	-5.67	-1.249	-5.228	0.117	0.005	0.003	0
32	50	100	CC	5	C>I	-0.404	0.88	2.839	1.352	0.126	0.005	0.003	0
32	50	100	OC	5	C>I	-0.398	-0.803	1.124	-0.338	0.116	0.005	0.003	0.154
32	50	100	R-CS	5	C>I	-0.391	-2.388	-0.492	-1.931	0.114	0.004	0.003	0
32	50	100	CC	5	I>C	-0.404	0.88	5.608	1.352	0.126	0.005	0.003	0
32	50	100	OC	5	I>C	-0.388	-3.21	1.327	-2.756	0.116	0.005	0.003	0.154
32	50	100	R-CS	5	I>C	-0.381	-4.966	-0.512	-4.521	0.114	0.004	0.003	0
30	100	100	CC	5	C>I	-0.406	1.366	3.335	1.841	0.121	0.005	0.003	0
30	100	100	R-CS	5	C>I	-0.401	-0.023	1.919	0.445	0.109	0.004	0.003	0
30	100	100	OC	5	C>I	-0.401	0.106	2.05	0.575	0.11	0.004	0.003	0
30	100	100	CC	5	I>C	-0.406	1.366	6.117	1.841	0.121	0.005	0.003	0
30	100	100	R-CS	5	I>C	-0.391	-2.574	1.992	-2.118	0.109	0.004	0.003	0
30	100	100	OC	5	I>C	-0.392	-2.311	2.268	-1.853	0.11	0.004	0.003	0

Table B.8: Results for the base case discrete model with one timescale for five time points only.

No. clusters	m	M	Design	Time points	Intvn. effects	Avg. estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	EmpSE	MCSE (bias)	MCSE (empSE)	Non-convergence (%)
44	15	15	CC	20	C>I	-0.399	-0.362	1.573	0.104	0.122	0.005	0.003	0
44	15	15	R-CS	20	C>I	-0.394	-1.695	0.214	-1.235	0.116	0.005	0.003	0
44	15	15	OC	20	C>I	-0.395	-1.565	0.346	-1.104	0.115	0.005	0.003	0.154
44	15	15	CC	20	I>C	-0.399	-0.362	4.307	0.104	0.122	0.005	0.003	0
44	15	15	R-CS	20	I>C	-0.384	-4.275	0.212	-3.826	0.115	0.005	0.003	0
44	15	15	OC	20	I>C	-0.385	-4.011	0.488	-3.562	0.115	0.005	0.003	0
44	15	50	CC	20	C>I	-0.401	-0.074	1.866	0.393	0.125	0.005	0.003	0.154
44	15	50	OC	20	C>I	-0.397	-0.948	0.976	-0.484	0.118	0.005	0.003	0
44	15	50	R-CS	20	C>I	-0.388	-3.22	-1.341	-2.767	0.109	0.004	0.003	0
44	15	50	CC	20	I>C	-0.401	-0.074	4.609	0.393	0.125	0.005	0.003	0.154
44	15	50	OC	20	I>C	-0.388	-3.331	1.2	-2.878	0.117	0.005	0.003	0
44	15	50	R-CS	20	I>C	-0.378	-5.83	-1.417	-5.389	0.109	0.004	0.003	0
44	15	100	CC	20	C>I	-0.404	0.76	2.717	1.232	0.121	0.005	0.003	0.154
44	15	100	OC	20	C>I	-0.399	-0.599	1.332	-0.133	0.114	0.004	0.003	0.308
44	15	100	R-CS	20	C>I	-0.405	1.036	2.998	1.509	0.12	0.005	0.003	0
44	15	100	CC	20	I>C	-0.404	0.76	5.482	1.232	0.121	0.005	0.003	0.154
44	15	100	OC	20	I>C	-0.389	-2.966	1.582	-2.511	0.113	0.004	0.003	0.308
44	15	100	R-CS	20	I>C	-0.395	-1.548	3.067	-1.087	0.119	0.005	0.003	0
32	50	50	CC	20	C>I	-0.392	-2.11	-0.209	-1.651	0.125	0.005	0.003	0
32	50	50	R-CS	20	C>I	-0.387	-3.457	-1.582	-3.005	0.118	0.005	0.003	0
32	50	50	OC	20	C>I	-0.388	-3.16	-1.279	-2.706	0.118	0.005	0.003	0
32	50	50	CC	20	I>C	-0.392	-2.11	2.478	-1.651	0.125	0.005	0.003	0
32	50	50	R-CS	20	I>C	-0.377	-6.02	-1.615	-5.579	0.118	0.005	0.003	0
32	50	50	OC	20	I>C	-0.379	-5.526	-1.098	-5.083	0.117	0.005	0.003	0
32	50	100	CC	20	C>I	-0.405	0.929	2.89	1.402	0.126	0.005	0.003	0
32	50	100	OC	20	C>I	-0.398	-0.76	1.167	-0.295	0.116	0.005	0.003	0
32	50	100	R-CS	20	C>I	-0.391	-2.388	-0.492	-1.931	0.114	0.004	0.003	0
32	50	100	CC	20	I>C	-0.405	0.929	5.66	1.402	0.126	0.005	0.003	0
32	50	100	OC	20	I>C	-0.388	-3.102	1.439	-2.649	0.116	0.005	0.003	0
32	50	100	R-CS	20	I>C	-0.381	-4.966	-0.512	-4.521	0.114	0.004	0.003	0
30	100	100	CC	20	C>I	-0.406	1.355	3.323	1.829	0.121	0.005	0.003	0
30	100	100	R-CS	20	C>I	-0.401	-0.023	1.919	0.445	0.109	0.004	0.003	0
30	100	100	OC	20	C>I	-0.401	0.103	2.047	0.572	0.11	0.004	0.003	0
30	100	100	CC	20	I>C	-0.406	1.355	6.105	1.829	0.121	0.005	0.003	0
30	100	100	R-CS	20	I>C	-0.391	-2.574	1.992	-2.118	0.109	0.004	0.003	0
30	100	100	OC	20	I>C	-0.392	-2.252	2.329	-1.794	0.11	0.004	0.003	0

Table B.9: Results for the base case discrete model with one timescale for twenty time points only.

No. clusters	**m**	**M**	**Design**	**Time points**	**Intvn. effects**	**OC-52-I bias (%)**	**OC-26-I bias (%)**	**OC-52-I empSE**	**OC-26-I empSE**	**OC-52-I MCSE (bias)**	**OC-26-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**OC-26-I MCSE (empSE)**	**Non-convergence (%)**
44	15	15	R-CS	2	C>I	0.549	2.671	0.156	0.296	0.007	0.013	0.005	0.009	15.846
44	15	15	OC	2	C>I	0.108	1.592	0.162	0.303	0.006	0.012	0.005	0.009	3.385
44	15	15	R-CS	2	I>C	0.671	4.451	0.156	0.296	0.007	0.013	0.005	0.009	15.846
44	15	15	OC	2	I>C	0.132	2.653	0.162	0.303	0.006	0.012	0.005	0.009	3.385
44	15	50	OC	2	C>I	0.914	2.919	0.159	0.292	0.006	0.012	0.005	0.008	4.462
44	15	50	R-CS	2	C>I	3.227	9.586	0.159	0.304	0.006	0.012	0.005	0.009	4.769
44	15	50	OC	2	I>C	1.117	4.866	0.159	0.292	0.006	0.012	0.005	0.008	4.462
44	15	50	R-CS	2	I>C	3.944	15.977	0.159	0.304	0.006	0.012	0.005	0.009	4.769
44	15	100	OC	2	C>I	1.094	1.534	0.158	0.296	0.006	0.012	0.004	0.008	4.308
44	15	100	R-CS	2	C>I	0.187	-2.122	0.167	0.307	0.007	0.012	0.005	0.009	4.308
44	15	100	OC	2	I>C	1.337	2.556	0.158	0.296	0.006	0.012	0.004	0.008	4.308
44	15	100	R-CS	2	I>C	0.228	-3.537	0.167	0.307	0.007	0.012	0.005	0.009	4.308
32	50	50	R-CS	2	C>I	-1.499	-0.742	0.114	0.2	0.004	0.008	0.003	0.006	0.462
32	50	50	OC	2	C>I	-1.475	-0.628	0.113	0.197	0.004	0.008	0.003	0.006	1.846
32	50	50	R-CS	2	I>C	-1.832	-1.237	0.114	0.2	0.004	0.008	0.003	0.006	0.462
32	50	50	OC	2	I>C	-1.896	-1.3	0.113	0.197	0.004	0.008	0.003	0.006	1.692
32	50	100	OC	2	C>I	0.452	1.058	0.116	0.199	0.005	0.008	0.003	0.006	1.846
32	50	100	R-CS	2	C>I	0.135	2.087	0.122	0.209	0.005	0.008	0.003	0.006	0.462
32	50	100	OC	2	I>C	0.597	1.817	0.116	0.199	0.005	0.008	0.003	0.006	1.692
32	50	100	R-CS	2	I>C	0.165	3.478	0.122	0.209	0.005	0.008	0.003	0.006	0.462
30	100	100	R-CS	2	C>I	0.749	-0.196	0.1	0.169	0.004	0.007	0.003	0.005	0.462
30	100	100	OC	2	C>I	0.686	-0.417	0.099	0.167	0.004	0.007	0.003	0.005	1.231
30	100	100	R-CS	2	I>C	0.915	-0.326	0.1	0.169	0.004	0.007	0.003	0.005	0.462
30	100	100	OC	2	I>C	0.886	-0.622	0.099	0.167	0.004	0.007	0.003	0.005	1.077

Table B.10: Results for the base case continuous model with two timescales for two time points only.

No. clusters	m	M	Design	Time points	Intvn. effects	OC-52-I bias (%)	OC-26-I bias (%)	OC-52-I empSE	OC-26-I empSE	OC-52-I MCSE (bias)	OC-26-I MCSE (bias)	OC-52-I MCSE (empSE)	OC-26-I MCSE (empSE)	Non-convergence (%)
44	15	15	R-CS	3	C>I	-1.246	-1.37	0.161	0.304	0.007	0.012	0.005	0.009	7.077
44	15	15	OC	3	C>I	0.215	1.726	0.159	0.296	0.006	0.012	0.004	0.008	0.769
44	15	15	R-CS	3	I>C	-1.523	-2.283	0.161	0.304	0.007	0.012	0.005	0.009	7.077
44	15	15	OC	3	I>C	0.205	2.772	0.159	0.296	0.006	0.012	0.004	0.008	0.615
44	15	50	OC	3	C>I	0.266	1.738	0.156	0.286	0.006	0.011	0.004	0.008	1.077
44	15	50	R-CS	3	C>I	3.197	10.166	0.159	0.302	0.006	0.012	0.004	0.008	0.769
44	15	50	OC	3	I>C	0.325	2.896	0.156	0.286	0.006	0.011	0.004	0.008	1.077
44	15	50	R-CS	3	I>C	3.907	16.943	0.159	0.302	0.006	0.012	0.004	0.008	0.769
44	15	100	OC	3	C>I	0.693	0.33	0.157	0.295	0.006	0.012	0.004	0.008	1.692
44	15	100	R-CS	3	C>I	0.078	-2.083	0.165	0.305	0.006	0.012	0.005	0.008	0.154
44	15	100	OC	3	I>C	0.847	0.551	0.157	0.295	0.006	0.012	0.004	0.008	1.692
44	15	100	R-CS	3	I>C	0.095	-3.471	0.165	0.305	0.006	0.012	0.005	0.008	0.154
32	50	50	R-CS	3	C>I	-1.766	-1.178	0.115	0.204	0.005	0.008	0.003	0.006	0
32	50	50	OC	3	C>I	-1.857	-1.338	0.112	0.195	0.004	0.008	0.003	0.005	0.615
32	50	50	R-CS	3	I>C	-2.158	-1.964	0.115	0.204	0.005	0.008	0.003	0.006	0
32	50	50	OC	3	I>C	-2.152	-1.994	0.112	0.195	0.004	0.008	0.003	0.005	0.462
32	50	100	OC	3	C>I	0.742	1.281	0.115	0.198	0.005	0.008	0.003	0.006	0.154
32	50	100	R-CS	3	C>I	0.004	1.949	0.123	0.211	0.005	0.008	0.003	0.006	0
32	50	100	OC	3	I>C	0.907	2.134	0.115	0.198	0.005	0.008	0.003	0.006	0.154
32	50	100	R-CS	3	I>C	0.005	3.248	0.123	0.211	0.005	0.008	0.003	0.006	0
30	100	100	R-CS	3	C>I	0.651	-0.471	0.1	0.172	0.004	0.007	0.003	0.005	0
30	100	100	OC	3	C>I	0.713	-0.442	0.098	0.167	0.004	0.007	0.003	0.005	0
30	100	100	R-CS	3	I>C	0.796	-0.784	0.1	0.172	0.004	0.007	0.003	0.005	0
30	100	100	OC	3	I>C	0.871	-0.737	0.098	0.167	0.004	0.007	0.003	0.005	0

Table B.11: Results for the base case continuous model with two timescales for three time points only.

No. clusters	**m**	**M**	**Design**	**Time points**	**Intvn. effects**	**OC-52-I bias (%)**	**OC-26-I bias (%)**	**OC-52-I empSE**	**OC-26-I empSE**	**OC-52-I MCSE (bias)**	**OC-26-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**OC-26-I MCSE (empSE)**	**Non-convergence (%)**
44	15	15	R-CS	5	C>I	0.322	1.742	0.162	0.309	0.006	0.012	0.005	0.009	1.077
44	15	15	OC	5	C>I	0.232	1.633	0.151	0.284	0.006	0.011	0.004	0.008	0.154
44	15	15	R-CS	5	I>C	0.394	2.904	0.162	0.309	0.006	0.012	0.005	0.009	1.077
44	15	15	OC	5	I>C	0.283	2.722	0.151	0.284	0.006	0.011	0.004	0.008	0.154
44	15	50	OC	5	C>I	-0.198	1.324	0.148	0.273	0.006	0.011	0.004	0.008	0.308
44	15	50	R-CS	5	C>I	3.549	10.16	0.148	0.283	0.006	0.011	0.004	0.008	0
44	15	50	OC	5	I>C	-0.241	2.206	0.148	0.273	0.006	0.011	0.004	0.008	0.308
44	15	50	R-CS	5	I>C	4.337	16.934	0.148	0.283	0.006	0.011	0.004	0.008	0
44	15	100	OC	5	C>I	2.155	3.121	0.147	0.276	0.006	0.011	0.004	0.008	0.462
44	15	100	R-CS	5	C>I	-0.16	-2.09	0.149	0.278	0.006	0.011	0.004	0.008	0
44	15	100	OC	5	I>C	2.634	5.202	0.147	0.276	0.006	0.011	0.004	0.008	0.462
44	15	100	R-CS	5	I>C	-0.195	-3.484	0.149	0.278	0.006	0.011	0.004	0.008	0
32	50	50	R-CS	5	C>I	-1.691	-1.652	0.117	0.21	0.005	0.008	0.003	0.006	0
32	50	50	OC	5	C>I	-1.52	-1.408	0.11	0.189	0.004	0.007	0.003	0.005	0
32	50	50	R-CS	5	I>C	-2.067	-2.753	0.117	0.21	0.005	0.008	0.003	0.006	0
32	50	50	OC	5	I>C	-1.858	-2.346	0.11	0.189	0.004	0.007	0.003	0.005	0
32	50	100	OC	5	C>I	0.046	0.344	0.111	0.193	0.004	0.008	0.003	0.005	0
32	50	100	R-CS	5	C>I	-0.207	1.679	0.113	0.196	0.004	0.008	0.003	0.005	0
32	50	100	OC	5	I>C	0.056	0.574	0.111	0.193	0.004	0.008	0.003	0.005	0
32	50	100	R-CS	5	I>C	-0.253	2.798	0.113	0.196	0.004	0.008	0.003	0.005	0
30	100	100	R-CS	5	C>I	0.642	0.123	0.1	0.177	0.004	0.007	0.003	0.005	0
30	100	100	OC	5	C>I	0.612	-0.061	0.094	0.163	0.004	0.006	0.003	0.005	0
30	100	100	R-CS	5	I>C	0.784	0.205	0.1	0.177	0.004	0.007	0.003	0.005	0
30	100	100	OC	5	I>C	0.748	-0.102	0.094	0.163	0.004	0.006	0.003	0.005	0

Table B.12: Results for the base case continuous model with two timescales for five time points only.

No. clusters	**m**	**M**	**Design**	**Time points**	**Intvn. effects**	**OC-52-I bias (%)**	**OC-26-I bias (%)**	**OC-52-I empSE**	**OC-26-I empSE**	**OC-52-I MCSE (bias)**	**OC-26-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**OC-26-I MCSE (empSE)**	**Non-convergence (%)**
44	15	15	R-CS	20	C>I	-0.607	-1.205	0.171	0.332	0.007	0.013	0.005	0.009	0
44	15	15	OC	20	C>I	-0.408	-0.675	0.134	0.254	0.005	0.01	0.004	0.007	0
44	15	15	R-CS	20	I>C	-0.742	-2.009	0.171	0.332	0.007	0.013	0.005	0.009	0
44	15	15	OC	20	I>C	-0.498	-1.126	0.134	0.254	0.005	0.01	0.004	0.007	0
44	15	50	OC	20	C>I	0.061	1.207	0.129	0.246	0.005	0.01	0.004	0.007	0
44	15	50	R-CS	20	C>I	3.208	7.206	0.116	0.221	0.005	0.009	0.003	0.006	0
44	15	50	OC	20	I>C	0.075	2.011	0.129	0.246	0.005	0.01	0.004	0.007	0
44	15	50	R-CS	20	I>C	3.921	12.011	0.116	0.221	0.005	0.009	0.003	0.006	0
44	15	100	OC	20	C>I	2.32	4.479	0.126	0.242	0.005	0.01	0.003	0.007	0.308
44	15	100	R-CS	20	C>I	0.723	0.373	0.108	0.206	0.004	0.008	0.003	0.006	0
44	15	100	OC	20	I>C	2.835	7.466	0.126	0.242	0.005	0.01	0.003	0.007	0.308
44	15	100	R-CS	20	I>C	0.883	0.621	0.108	0.206	0.004	0.008	0.003	0.006	0
32	50	50	R-CS	20	C>I	-1.663	-2.79	0.12	0.231	0.005	0.009	0.003	0.006	0
32	50	50	OC	20	C>I	-1.2	-1.965	0.096	0.178	0.004	0.007	0.003	0.005	0
32	50	50	R-CS	20	I>C	-2.033	-4.651	0.12	0.231	0.005	0.009	0.003	0.006	0
32	50	50	OC	20	I>C	-1.466	-3.275	0.096	0.178	0.004	0.007	0.003	0.005	0
32	50	100	OC	20	C>I	-0.339	-0.233	0.098	0.177	0.004	0.007	0.003	0.005	0
32	50	100	R-CS	20	C>I	-0.153	1.088	0.098	0.182	0.004	0.007	0.003	0.005	0
32	50	100	OC	20	I>C	-0.415	-0.388	0.098	0.177	0.004	0.007	0.003	0.005	0
32	50	100	R-CS	20	I>C	-0.187	1.814	0.098	0.182	0.004	0.007	0.003	0.005	0
30	100	100	R-CS	20	C>I	0.109	-0.548	0.102	0.195	0.004	0.008	0.003	0.005	0
30	100	100	OC	20	C>I	0.271	-0.287	0.083	0.153	0.003	0.006	0.002	0.004	0
30	100	100	R-CS	20	I>C	0.133	-0.914	0.102	0.195	0.004	0.008	0.003	0.005	0
30	100	100	OC	20	I>C	0.331	-0.478	0.083	0.153	0.003	0.006	0.002	0.004	0

Table B.13: Results for the base case continuous model with two timescales for twenty time points only.

No. clusters	**m**	**M**	**Design**	**MD**	**Turnover (%)**	**Shape**	**Intvn. effects**	**Avg. estimate**	**CC-I bias (%)**	**CS-OC-D bias (%)**	**CC-D bias (%)**	**EmpSE**	**MCSE (bias)**	**MCSE (empSE)**	**Non-convergence (%)**
44	15	15	CC	MCAR	10	Constant	C>I	-0.399	-0.361	1.574	0.105	0.122	0.005	0.003	0.769
44	15	15	R-CS	MCAR	10	Constant	C>I	-0.392	-2.228	-0.329	-1.77	0.114	0.005	0.003	6.308
44	15	15	OC	MCAR	10	Constant	C>I	-0.394	-1.647	0.263	-1.186	0.115	0.005	0.003	1.077
44	15	15	CC	MCAR	20	Constant	C>I	-0.403	0.576	3.939	1.155	0.13	0.005	0.004	1.538
44	15	15	R-CS	MCAR	20	Constant	C>I	-0.39	-2.656	0.599	-2.096	0.114	0.005	0.003	6.615
44	15	15	OC	MCAR	20	Constant	C>I	-0.392	-2.314	0.953	-1.752	0.113	0.004	0.003	2.462
44	15	15	CC	MCAR	40	Constant	C>I	-0.393	-2.008	4.367	-1.662	0.13	0.005	0.004	1.846
44	15	15	R-CS	MCAR	40	Constant	C>I	-0.37	-7.664	-1.658	-7.339	0.109	0.004	0.003	2.769
44	15	15	OC	MCAR	40	Constant	C>I	-0.374	-6.663	-0.591	-6.334	0.105	0.004	0.003	4.923
44	15	15	CC	MCAR	10	Constant	I>C	-0.399	-0.361	4.309	0.105	0.122	0.005	0.003	0.769
44	15	15	R-CS	MCAR	10	Constant	I>C	-0.381	-4.921	-0.465	-4.476	0.114	0.005	0.003	6.462
44	15	15	OC	MCAR	10	Constant	I>C	-0.384	-4.12	0.374	-3.671	0.115	0.005	0.003	0.923
44	15	15	CC	MCAR	20	Constant	I>C	-0.403	0.576	10.002	1.155	0.13	0.005	0.004	1.538
44	15	15	R-CS	MCAR	20	Constant	I>C	-0.369	-7.928	0.701	-7.399	0.113	0.005	0.003	6.923
44	15	15	OC	MCAR	20	Constant	I>C	-0.372	-7.29	1.399	-6.757	0.113	0.004	0.003	2.615
44	15	15	CC	MCAR	40	Constant	I>C	-0.393	-2.008	19.037	-1.662	0.13	0.005	0.004	1.846
44	15	15	R-CS	MCAR	40	Constant	I>C	-0.324	-19.201	-1.849	-18.916	0.106	0.004	0.003	3.231
44	15	15	OC	MCAR	40	Constant	I>C	-0.331	-17.379	0.364	-17.088	0.103	0.004	0.003	4.615
44	15	15	CC	MCAR	10	Non-constant	C>I	-0.399	-0.361	1.574	0.105	0.122	0.005	0.003	0.769
44	15	15	R-CS	MCAR	10	Non-constant	C>I	-0.392	-2.228	-0.329	-1.77	0.114	0.005	0.003	6.308
44	15	15	OC	MCAR	10	Non-constant	C>I	-0.394	-1.66	0.25	-1.199	0.115	0.005	0.003	0.923
44	15	15	CC	MCAR	20	Non-constant	C>I	-0.403	0.576	3.939	1.155	0.13	0.005	0.004	1.538
44	15	15	R-CS	MCAR	20	Non-constant	C>I	-0.39	-2.656	0.599	-2.096	0.114	0.005	0.003	6.615
44	15	15	OC	MCAR	20	Non-constant	C>I	-0.392	-2.314	0.953	-1.752	0.113	0.004	0.003	2.462
44	15	15	CC	MCAR	40	Non-constant	C>I	-0.393	-2.008	4.367	-1.662	0.13	0.005	0.004	1.846
44	15	15	R-CS	MCAR	40	Non-constant	C>I	-0.37	-7.664	-1.658	-7.339	0.109	0.004	0.003	2.769
44	15	15	OC	MCAR	40	Non-constant	C>I	-0.374	-6.676	-0.606	-6.347	0.105	0.004	0.003	4.615
44	15	15	CC	MCAR	10	Non-constant	I>C	-0.399	-0.361	4.309	0.105	0.122	0.005	0.003	0.769
44	15	15	R-CS	MCAR	10	Non-constant	I>C	-0.381	-4.921	-0.465	-4.476	0.114	0.005	0.003	6.462
44	15	15	OC	MCAR	10	Non-constant	I>C	-0.384	-4.12	0.374	-3.671	0.115	0.005	0.003	0.923
44	15	15	CC	MCAR	20	Non-constant	I>C	-0.403	0.576	10.002	1.155	0.13	0.005	0.004	1.538
44	15	15	R-CS	MCAR	20	Non-constant	I>C	-0.369	-7.928	0.701	-7.399	0.113	0.005	0.003	6.923
44	15	15	OC	MCAR	20	Non-constant	I>C	-0.372	-7.329	1.357	-6.796	0.113	0.004	0.003	2.462
44	15	15	CC	MCAR	40	Non-constant	I>C	-0.393	-2.008	19.037	-1.662	0.13	0.005	0.004	1.846
44	15	15	R-CS	MCAR	40	Non-constant	I>C	-0.324	-19.201	-1.849	-18.916	0.106	0.004	0.003	3.231
44	15	15	OC	MCAR	40	Non-constant	I>C	-0.331	-17.379	0.364	-17.088	0.103	0.004	0.003	4.615

Table B.14: Results for the complications discrete model with one timescale with full-samples and MCAR missing data mechanism.

No. clusters	m	M	Design	MD	Turnover (%)	Shape	Intvn. effects	Avg. estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	EmpSE	MCSE (bias)	MCSE (empSE)	Non-convergence (%)
44	15	50	CC	MCAR	10	Constant	C>I	-0.401	0.104	2.048	0.573	0.125	0.005	0.003	0.769
44	15	50	OC	MCAR	10	Constant	C>I	-0.398	-0.833	1.092	-0.369	0.118	0.005	0.003	0.923
44	15	50	CC	MCAR	20	Constant	C>I	-0.393	-2.024	1.253	-1.46	0.123	0.005	0.003	0.615
44	15	50	OC	MCAR	20	Constant	C>I	-0.384	-4.117	-0.91	-3.565	0.108	0.004	0.003	2
44	15	50	CC	MCAR	40	Constant	C>I	-0.397	-1.09	5.344	-0.742	0.139	0.005	0.004	0.615
44	15	50	OC	MCAR	40	Constant	C>I	-0.379	-5.461	0.689	-5.128	0.107	0.004	0.003	4.615
44	15	50	CC	MCAR	10	Constant	I>C	-0.401	0.104	4.796	0.573	0.125	0.005	0.003	0.769
44	15	50	OC	MCAR	10	Constant	I>C	-0.388	-3.296	1.237	-2.843	0.117	0.005	0.003	0.923
44	15	50	CC	MCAR	20	Constant	I>C	-0.393	-2.024	7.159	-1.46	0.123	0.005	0.003	0.615
44	15	50	OC	MCAR	20	Constant	I>C	-0.364	-9.151	-0.636	-8.628	0.107	0.004	0.003	2.308
44	15	50	CC	MCAR	40	Constant	I>C	-0.397	-1.09	20.152	-0.742	0.139	0.005	0.004	0.615
44	15	50	OC	MCAR	40	Constant	I>C	-0.336	-16.242	1.746	-15.946	0.105	0.004	0.003	4.769
44	15	50	CC	MCAR	10	Non-constant	C>I	-0.401	0.104	2.048	0.573	0.125	0.005	0.003	0.769
44	15	50	OC	MCAR	10	Non-constant	C>I	-0.398	-0.833	1.092	-0.369	0.118	0.005	0.003	0.923
44	15	50	CC	MCAR	20	Non-constant	C>I	-0.393	-2.024	1.253	-1.46	0.123	0.005	0.003	0.615
44	15	50	OC	MCAR	20	Non-constant	C>I	-0.385	-4.072	-0.864	-3.52	0.108	0.004	0.003	2.154
44	15	50	CC	MCAR	40	Non-constant	C>I	-0.397	-1.09	5.344	-0.742	0.139	0.005	0.004	0.615
44	15	50	OC	MCAR	40	Non-constant	C>I	-0.379	-5.461	0.689	-5.128	0.107	0.004	0.003	4.615
44	15	50	CC	MCAR	10	Non-constant	I>C	-0.401	0.104	4.796	0.573	0.125	0.005	0.003	0.769
44	15	50	OC	MCAR	10	Non-constant	I>C	-0.388	-3.296	1.237	-2.843	0.117	0.005	0.003	0.923
44	15	50	CC	MCAR	20	Non-constant	I>C	-0.393	-2.024	7.159	-1.46	0.123	0.005	0.003	0.615
44	15	50	OC	MCAR	20	Non-constant	I>C	-0.364	-9.151	-0.636	-8.628	0.107	0.004	0.003	2.308
44	15	50	CC	MCAR	40	Non-constant	I>C	-0.397	-1.09	20.152	-0.742	0.139	0.005	0.004	0.615
44	15	50	OC	MCAR	40	Non-constant	I>C	-0.336	-16.242	1.746	-15.946	0.105	0.004	0.003	4.769
44	15	50	R-CS	MCAR	10	Constant	C>I	-0.388	-3.229	-1.349	-2.776	0.109	0.004	0.003	0.769
44	15	50	R-CS	MCAR	20	Constant	C>I	-0.385	-3.961	-0.749	-3.408	0.113	0.004	0.003	0.154
44	15	50	R-CS	MCAR	40	Constant	C>I	-0.378	-5.658	0.479	-5.325	0.105	0.004	0.003	0.308
44	15	50	R-CS	MCAR	10	Constant	I>C	-0.378	-5.839	-1.425	-5.398	0.109	0.004	0.003	0.769
44	15	50	R-CS	MCAR	20	Constant	I>C	-0.364	-9.3	-0.799	-8.778	0.112	0.004	0.003	0.154
44	15	50	R-CS	MCAR	40	Constant	I>C	-0.332	-17.145	0.649	-16.853	0.103	0.004	0.003	0.462
44	15	50	R-CS	MCAR	10	Non-constant	C>I	-0.388	-3.229	-1.349	-2.776	0.109	0.004	0.003	0.769
44	15	50	R-CS	MCAR	20	Non-constant	C>I	-0.385	-3.961	-0.749	-3.408	0.113	0.004	0.003	0.154
44	15	50	R-CS	MCAR	40	Non-constant	C>I	-0.378	-5.658	0.479	-5.325	0.105	0.004	0.003	0.308
44	15	50	R-CS	MCAR	10	Non-constant	I>C	-0.378	-5.839	-1.425	-5.398	0.109	0.004	0.003	0.769
44	15	50	R-CS	MCAR	20	Non-constant	I>C	-0.364	-9.3	-0.799	-8.778	0.112	0.004	0.003	0.154
44	15	50	R-CS	MCAR	40	Non-constant	I>C	-0.332	-17.145	0.649	-16.853	0.103	0.004	0.003	0.462

Table B.15: Results for the complications discrete model with one timescale with sub-samples and MCAR missing data mechanism.

No. clusters	m	M	Design	MD	Turnover (%)	Shape	Intvn. effects	Avg. estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	EmpSE	MCSE (bias)	MCSE (empSE)	Non-convergence (%)
44	15	15	CC	MAR	10	Constant	C>I	-0.399	-0.568	0.525	0.594	0.119	0.005	0.003	0.769
44	15	15	R-CS	MAR	10	Constant	C>I	-0.398	-0.694	0.397	0.466	0.112	0.005	0.003	6
44	15	15	OC	MAR	10	Constant	C>I	-0.399	-0.493	0.6	0.669	0.112	0.004	0.003	0.923
44	15	15	CC	MAR	20	Constant	C>I	-0.385	-3.989	-0.437	3.021	0.123	0.005	0.003	0.615
44	15	15	R-CS	MAR	20	Constant	C>I	-0.379	-5.424	-1.925	1.481	0.108	0.004	0.003	4.462
44	15	15	OC	MAR	20	Constant	C>I	-0.39	-2.644	0.958	4.464	0.108	0.004	0.003	1.846
44	15	15	CC	MAR	40	Constant	C>I	-0.376	-6.209	-0.37	8.816	0.128	0.005	0.004	2.154
44	15	15	R-CS	MAR	40	Constant	C>I	-0.373	-6.944	-1.151	7.964	0.098	0.004	0.003	2.308
44	15	15	OC	MAR	40	Constant	C>I	-0.384	-4.206	1.757	11.139	0.098	0.004	0.003	5.077
44	15	15	CC	MAR	10	Constant	I>C	-0.406	1.345	5.136	2.528	0.13	0.005	0.004	0.769
44	15	15	R-CS	MAR	10	Constant	I>C	-0.395	-1.572	2.11	-0.423	0.122	0.005	0.003	5.385
44	15	15	OC	MAR	10	Constant	I>C	-0.397	-0.917	2.79	0.24	0.123	0.005	0.003	0.769
44	15	15	CC	MAR	20	Constant	I>C	-0.395	-1.542	6.79	5.646	0.115	0.005	0.003	0.923
44	15	15	R-CS	MAR	20	Constant	I>C	-0.373	-7.016	0.852	-0.228	0.1	0.004	0.003	4.615
44	15	15	OC	MAR	20	Constant	I>C	-0.384	-4.234	3.87	2.758	0.101	0.004	0.003	2.308
44	15	15	CC	MAR	40	Constant	I>C	-0.389	-3.086	11.288	12.44	0.124	0.005	0.003	1.846
44	15	15	R-CS	MAR	40	Constant	I>C	-0.358	-10.797	2.433	3.493	0.091	0.004	0.003	4
44	15	15	OC	MAR	40	Constant	I>C	-0.368	-8.228	5.383	6.474	0.091	0.004	0.003	4.154
44	15	15	CC	MAR	10	Non-constant	C>I	-0.402	0.263	1.078	1.736	0.128	0.005	0.004	1.538
44	15	15	R-CS	MAR	10	Non-constant	C>I	-0.402	0.253	1.068	1.726	0.118	0.005	0.003	5.692
44	15	15	OC	MAR	10	Non-constant	C>I	-0.403	0.513	1.33	1.989	0.12	0.005	0.003	2
44	15	15	CC	MAR	20	Non-constant	C>I	-0.389	-3.038	1.785	7.978	0.126	0.005	0.003	0.462
44	15	15	R-CS	MAR	20	Non-constant	C>I	-0.382	-4.606	0.139	6.232	0.108	0.004	0.003	4.308
44	15	15	OC	MAR	20	Non-constant	C>I	-0.397	-1	3.924	10.247	0.11	0.004	0.003	1.385
44	15	15	CC	MAR	40	Non-constant	C>I	-0.355	-11.576	-2.232	13.055	0.133	0.005	0.004	2
44	15	15	R-CS	MAR	40	Non-constant	C>I	-0.354	-11.8	-2.479	12.768	0.1	0.004	0.003	4.308
44	15	15	OC	MAR	40	Non-constant	C>I	-0.371	-7.398	2.387	18.395	0.099	0.004	0.003	5.692
44	15	15	CC	MAR	10	Non-constant	I>C	-0.388	-3.25	0.405	-2.019	0.118	0.005	0.003	1.077
44	15	15	R-CS	MAR	10	Non-constant	I>C	-0.379	-5.465	-1.893	-4.262	0.112	0.005	0.003	4.769
44	15	15	OC	MAR	10	Non-constant	I>C	-0.381	-5.048	-1.46	-3.84	0.112	0.004	0.003	0.769
44	15	15	CC	MAR	20	Non-constant	I>C	-0.394	-1.847	7.119	6.562	0.121	0.005	0.003	0.769
44	15	15	R-CS	MAR	20	Non-constant	I>C	-0.369	-7.986	0.42	-0.103	0.104	0.004	0.003	4.923
44	15	15	OC	MAR	20	Non-constant	I>C	-0.383	-4.495	4.229	3.686	0.105	0.004	0.003	2.154
44	15	15	CC	MAR	40	Non-constant	I>C	-0.368	-8.157	7.039	9.761	0.126	0.005	0.004	2.923
44	15	15	R-CS	MAR	40	Non-constant	I>C	-0.339	-15.466	-1.48	1.025	0.094	0.004	0.003	4
44	15	15	OC	MAR	40	Non-constant	I>C	-0.353	-11.885	2.694	5.305	0.094	0.004	0.003	6

Table B.16: Results for the complications discrete model with one timescale with full-samples and MAR missing data mechanism.

No. clusters	**m**	**M**	**Design**	**MD**	**Turnover (%)**	**Shape**	**Intvn. effects**	**Avg. estimate**	**CC-I bias (%)**	**CS-OC-D bias (%)**	**CC-D bias (%)**	**EmpSE**	**MCSE (bias)**	**MCSE (empSE)**	**Non-convergence (%)**
44	15	50	CC	MAR	10	Constant	C>I	-0.393	-1.938	-0.861	-0.793	0.121	0.005	0.003	1.077
44	15	50	OC	MAR	10	Constant	C>I	-0.392	-2.263	-1.19	-1.122	0.113	0.004	0.003	1.385
44	15	50	CC	MAR	20	Constant	C>I	-0.386	-3.792	-0.232	3.232	0.121	0.005	0.003	0.923
44	15	50	OC	MAR	20	Constant	C>I	-0.393	-1.985	1.641	5.171	0.105	0.004	0.003	2.462
44	15	50	CC	MAR	40	Constant	C>I	-0.383	-4.351	1.603	10.972	0.124	0.005	0.003	2.462
44	15	50	OC	MAR	40	Constant	C>I	-0.387	-3.397	2.617	12.079	0.091	0.004	0.003	5.231
44	15	50	CC	MAR	10	Constant	I>C	-0.399	-0.533	3.188	0.629	0.122	0.005	0.003	0.615
44	15	50	OC	MAR	10	Constant	I>C	-0.388	-3.271	0.347	-2.142	0.116	0.005	0.003	1.231
44	15	50	CC	MAR	20	Constant	I>C	-0.383	-4.373	3.719	2.608	0.125	0.005	0.003	1.231
44	15	50	OC	MAR	20	Constant	I>C	-0.372	-7.126	0.733	-0.346	0.112	0.004	0.003	2
44	15	50	CC	MAR	40	Constant	I>C	-0.373	-6.99	6.805	7.91	0.125	0.005	0.004	2.462
44	15	50	OC	MAR	40	Constant	I>C	-0.355	-11.515	1.609	2.66	0.09	0.004	0.003	6.308
44	15	50	CC	MAR	10	Non-constant	C>I	-0.391	-2.573	-1.781	-1.141	0.124	0.005	0.003	0.615
44	15	50	OC	MAR	10	Non-constant	C>I	-0.392	-2.212	-1.417	-0.776	0.117	0.005	0.003	1.231
44	15	50	CC	MAR	20	Non-constant	C>I	-0.384	-4.313	0.447	6.558	0.128	0.005	0.004	1.385
44	15	50	OC	MAR	20	Non-constant	C>I	-0.394	-1.804	3.08	9.352	0.115	0.005	0.003	2.462
44	15	50	CC	MAR	40	Non-constant	C>I	-0.359	-10.374	-0.903	14.591	0.131	0.005	0.004	2
44	15	50	OC	MAR	40	Non-constant	C>I	-0.374	-6.726	3.13	19.255	0.098	0.004	0.003	6.154
44	15	50	CC	MAR	10	Non-constant	I>C	-0.393	-2.054	1.646	-0.808	0.128	0.005	0.004	0.154
44	15	50	OC	MAR	10	Non-constant	I>C	-0.384	-4.125	-0.502	-2.905	0.119	0.005	0.003	1.077
44	15	50	CC	MAR	20	Non-constant	I>C	-0.386	-3.696	5.101	4.554	0.12	0.005	0.003	1.538
44	15	50	OC	MAR	20	Non-constant	I>C	-0.375	-6.458	2.087	1.556	0.107	0.004	0.003	2.462
44	15	50	CC	MAR	40	Non-constant	I>C	-0.364	-9.169	5.859	8.551	0.127	0.005	0.004	3.077
44	15	50	OC	MAR	40	Non-constant	I>C	-0.348	-13.208	1.152	3.724	0.09	0.004	0.003	6
44	15	50	R-CS	MAR	10	Constant	C>I	-0.392	-2.197	-1.123	-1.055	0.114	0.004	0.003	0.308
44	15	50	R-CS	MAR	20	Constant	C>I	-0.388	-3.307	0.27	3.753	0.109	0.004	0.003	0.154
44	15	50	R-CS	MAR	40	Constant	C>I	-0.366	-8.689	-3.005	5.938	0.094	0.004	0.003	1.846
44	15	50	R-CS	MAR	10	Constant	I>C	-0.384	-4.108	-0.521	-2.988	0.112	0.004	0.003	0.462
44	15	50	R-CS	MAR	20	Constant	I>C	-0.371	-7.511	0.316	-0.758	0.104	0.004	0.003	0.308
44	15	50	R-CS	MAR	40	Constant	I>C	-0.351	-12.572	0.394	1.433	0.094	0.004	0.003	1.846
44	15	50	R-CS	MAR	10	Non-constant	C>I	-0.389	-2.861	-2.071	-1.434	0.113	0.004	0.003	0.154
44	15	50	R-CS	MAR	20	Non-constant	C>I	-0.38	-5.235	-0.521	5.532	0.108	0.004	0.003	0.462
44	15	50	R-CS	MAR	40	Non-constant	C>I	-0.349	-12.915	-3.713	11.342	0.093	0.004	0.003	1.692
44	15	50	R-CS	MAR	10	Non-constant	I>C	-0.386	-3.776	-0.14	-2.551	0.116	0.005	0.003	0.154
44	15	50	R-CS	MAR	20	Non-constant	I>C	-0.366	-8.697	-0.357	-0.876	0.111	0.004	0.003	0.615
44	15	50	R-CS	MAR	40	Non-constant	I>C	-0.344	-14.321	-0.146	2.394	0.094	0.004	0.003	1.538

Table B.17: Results for the complications discrete model with one timescale with sub-samples and MAR missing data mechanism.

No. clusters	**m**	**M**	**Design**	**MD**	**Turnover (%)**	**Shape**	**Intvn. effects**	**Avg. estimate**	**CC-I bias (%)**	**CS-OC-D bias (%)**	**CC-D bias (%)**	**EmpSE**	**MCSE (bias)**	**MCSE (empSE)**	**Non-convergence (%)**
44	15	15	CC	MNAR	10	Constant	C>I	-0.408	1.671	2.301	2.644	0.128	0.005	0.004	0.769
44	15	15	R-CS	MNAR	10	Constant	C>I	-0.406	1.177	1.803	2.145	0.12	0.005	0.003	8.154
44	15	15	OC	MNAR	10	Constant	C>I	-0.405	1.085	1.711	2.052	0.12	0.005	0.003	1.538
44	15	15	CC	MNAR	20	Constant	C>I	-0.387	-3.457	0.707	4.849	0.121	0.005	0.003	1.231
44	15	15	R-CS	MNAR	20	Constant	C>I	-0.379	-5.414	-1.335	2.722	0.104	0.004	0.003	4.769
44	15	15	OC	MNAR	20	Constant	C>I	-0.391	-2.578	1.624	5.803	0.105	0.004	0.003	1.846
44	15	15	CC	MNAR	40	Constant	C>I	-0.369	-7.918	-0.014	8.306	0.127	0.005	0.004	2.615
44	15	15	R-CS	MNAR	40	Constant	C>I	-0.36	-10.146	-2.432	5.686	0.092	0.004	0.003	3.846
44	15	15	OC	MNAR	40	Constant	C>I	-0.372	-7.292	0.667	9.043	0.094	0.004	0.003	6.154
44	15	15	CC	MNAR	10	Constant	I>C	-0.398	-0.675	2.668	0.275	0.123	0.005	0.003	0.615
44	15	15	R-CS	MNAR	10	Constant	I>C	-0.388	-3.276	-0.021	-2.351	0.116	0.005	0.003	7.385
44	15	15	OC	MNAR	10	Constant	I>C	-0.387	-3.418	-0.167	-2.494	0.115	0.005	0.003	1.231
44	15	15	CC	MNAR	20	Constant	I>C	-0.388	-3.1	5.228	5.236	0.118	0.005	0.003	0.923
44	15	15	R-CS	MNAR	20	Constant	I>C	-0.368	-8.333	-0.455	-0.447	0.104	0.004	0.003	3.692
44	15	15	OC	MNAR	20	Constant	I>C	-0.38	-5.21	2.936	2.944	0.106	0.004	0.003	1.231
44	15	15	CC	MNAR	40	Constant	I>C	-0.364	-9.202	5.015	6.796	0.122	0.005	0.003	2.769
44	15	15	R-CS	MNAR	40	Constant	I>C	-0.337	-15.92	-2.755	-1.105	0.087	0.003	0.002	3.846
44	15	15	OC	MNAR	40	Constant	I>C	-0.349	-12.908	0.728	2.437	0.089	0.004	0.003	6.923
44	15	15	CC	MNAR	10	Non-constant	C>I	-0.397	-1.087	-0.61	0.204	0.119	0.005	0.003	0.769
44	15	15	R-CS	MNAR	10	Non-constant	C>I	-0.394	-1.818	-1.345	-0.537	0.111	0.005	0.003	6.154
44	15	15	OC	MNAR	10	Non-constant	C>I	-0.396	-1.131	-0.655	0.159	0.112	0.004	0.003	1.231
44	15	15	CC	MNAR	20	Non-constant	C>I	-0.375	-6.513	-0.062	6.607	0.117	0.005	0.003	1.692
44	15	15	R-CS	MNAR	20	Non-constant	C>I	-0.367	-8.338	-2.013	4.525	0.1	0.004	0.003	4
44	15	15	OC	MNAR	20	Non-constant	C>I	-0.386	-3.829	2.808	9.667	0.102	0.004	0.003	2.615
44	15	15	CC	MNAR	40	Non-constant	C>I	-0.354	-11.624	0.953	15.841	0.119	0.005	0.003	2.615
44	15	15	R-CS	MNAR	40	Non-constant	C>I	-0.343	-14.34	-2.15	12.281	0.089	0.004	0.003	3.077
44	15	15	OC	MNAR	40	Non-constant	C>I	-0.361	-9.902	2.919	18.098	0.089	0.004	0.003	6.615
44	15	15	CC	MNAR	10	Non-constant	I>C	-0.398	-0.73	2.578	0.319	0.122	0.005	0.003	1.231
44	15	15	R-CS	MNAR	10	Non-constant	I>C	-0.386	-3.621	-0.409	-2.603	0.111	0.005	0.003	7.538
44	15	15	OC	MNAR	10	Non-constant	I>C	-0.388	-3.125	0.103	-2.102	0.114	0.004	0.003	1.231
44	15	15	CC	MNAR	20	Non-constant	I>C	-0.379	-5.397	3.834	4.325	0.119	0.005	0.003	0.769
44	15	15	R-CS	MNAR	20	Non-constant	I>C	-0.355	-11.561	-2.932	-2.473	0.103	0.004	0.003	4.154
44	15	15	OC	MNAR	20	Non-constant	I>C	-0.37	-7.799	1.197	1.675	0.102	0.004	0.003	2.154
44	15	15	CC	MNAR	40	Non-constant	I>C	-0.374	-6.622	10.303	13.482	0.123	0.005	0.003	2
44	15	15	R-CS	MNAR	40	Non-constant	I>C	-0.338	-15.599	-0.302	2.573	0.093	0.004	0.003	2.154
44	15	15	OC	MNAR	40	Non-constant	I>C	-0.353	-11.861	4.115	7.116	0.093	0.004	0.003	6.462

Table B.18: Results for the complications discrete model with one timescale with full-samples and MNAR missing data mechanism.

No. clusters	m	M	Design	MD	Turnover (%)	Shape	Intvn. effects	Avg. estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	EmpSE	MCSE (bias)	MCSE (empSE)	Non-convergence (%)
44	15	50	CC	MNAR	10	Constant	C>I	-0.4	-0.13	0.488	0.825	0.128	0.005	0.004	0.923
44	15	50	OC	MNAR	10	Constant	C>I	-0.4	-0.155	0.463	0.8	0.117	0.005	0.003	1.692
44	15	50	CC	MNAR	20	Constant	C>I	-0.392	-2.201	2.016	6.212	0.12	0.005	0.003	1.385
44	15	50	OC	MNAR	20	Constant	C>I	-0.395	-1.413	2.838	7.067	0.105	0.004	0.003	2
44	15	50	CC	MNAR	40	Constant	C>I	-0.374	-6.732	1.274	9.701	0.126	0.005	0.004	3.385
44	15	50	OC	MNAR	40	Constant	C>I	-0.38	-5.203	2.935	11.5	0.093	0.004	0.003	6.615
44	15	50	CC	MNAR	10	Constant	I>C	-0.391	-2.514	0.767	-1.582	0.126	0.005	0.004	0.154
44	15	50	OC	MNAR	10	Constant	I>C	-0.383	-4.477	-1.261	-3.563	0.119	0.005	0.003	0.462
44	15	50	CC	MNAR	20	Constant	I>C	-0.389	-2.945	5.395	5.404	0.128	0.005	0.004	1.231
44	15	50	OC	MNAR	20	Constant	I>C	-0.381	-4.881	3.293	3.301	0.112	0.004	0.003	1.692
44	15	50	CC	MNAR	40	Constant	I>C	-0.372	-7.132	7.409	9.231	0.124	0.005	0.003	3.231
44	15	50	OC	MNAR	40	Constant	I>C	-0.354	-11.66	2.171	3.904	0.092	0.004	0.003	6.462
44	15	50	CC	MNAR	10	Non-constant	C>I	-0.402	0.149	0.631	1.456	0.119	0.005	0.003	1.077
44	15	50	OC	MNAR	10	Non-constant	C>I	-0.405	1.017	1.503	2.335	0.113	0.004	0.003	1.231
44	15	50	CC	MNAR	20	Non-constant	C>I	-0.377	-6.027	0.458	7.161	0.117	0.005	0.003	1.385
44	15	50	OC	MNAR	20	Non-constant	C>I	-0.385	-4.016	2.608	9.454	0.103	0.004	0.003	2.308
44	15	50	CC	MNAR	40	Non-constant	C>I	-0.358	-10.662	2.052	17.102	0.125	0.005	0.004	2
44	15	50	OC	MNAR	40	Non-constant	C>I	-0.363	-9.343	3.558	18.831	0.096	0.004	0.003	6.615
44	15	50	CC	MNAR	10	Non-constant	I>C	-0.4	-0.111	3.218	0.944	0.12	0.005	0.003	0.769
44	15	50	OC	MNAR	10	Non-constant	I>C	-0.389	-3.041	0.19	-2.017	0.111	0.004	0.003	2
44	15	50	CC	MNAR	20	Non-constant	I>C	-0.389	-3.011	6.453	6.956	0.125	0.005	0.003	0.615
44	15	50	OC	MNAR	20	Non-constant	I>C	-0.379	-5.406	3.824	4.315	0.109	0.004	0.003	1.692
44	15	50	CC	MNAR	40	Non-constant	I>C	-0.364	-9.105	7.37	10.465	0.119	0.005	0.003	3.846
44	15	50	OC	MNAR	40	Non-constant	I>C	-0.348	-13.308	2.405	5.357	0.09	0.004	0.003	7.077
44	15	50	R-CS	MNAR	10	Constant	C>I	-0.401	0.124	0.744	1.082	0.116	0.005	0.003	0.769
44	15	50	R-CS	MNAR	20	Constant	C>I	-0.381	-4.936	-0.836	3.242	0.106	0.004	0.003	0.154
44	15	50	R-CS	MNAR	40	Constant	C>I	-0.367	-8.35	-0.483	7.798	0.092	0.004	0.003	0.923
44	15	50	R-CS	MNAR	10	Constant	I>C	-0.385	-3.909	-0.675	-2.99	0.11	0.004	0.003	0.154
44	15	50	R-CS	MNAR	20	Constant	I>C	-0.365	-8.93	-1.103	-1.095	0.107	0.004	0.003	0.462
44	15	50	R-CS	MNAR	40	Constant	I>C	-0.344	-14.24	-0.812	0.871	0.093	0.004	0.003	1.692
44	15	50	R-CS	MNAR	10	Non-constant	C>I	-0.404	0.668	1.152	1.981	0.114	0.004	0.003	0.154
44	15	50	R-CS	MNAR	20	Non-constant	C>I	-0.372	-7.226	-0.824	5.794	0.099	0.004	0.003	0.308
44	15	50	R-CS	MNAR	40	Non-constant	C>I	-0.343	-14.398	-2.216	12.205	0.09	0.004	0.003	1.538
44	15	50	R-CS	MNAR	10	Non-constant	I>C	-0.381	-4.976	-1.81	-3.973	0.115	0.005	0.003	0.308
44	15	50	R-CS	MNAR	20	Non-constant	I>C	-0.36	-10.226	-1.466	-1.001	0.109	0.004	0.003	0.462
44	15	50	R-CS	MNAR	40	Non-constant	I>C	-0.336	-16.073	-0.861	1.997	0.094	0.004	0.003	1.846

Table B.19: Results for the complications discrete model with one timescale with sub-samples and MNAR missing data mechanism.

No. clusters	m	M	Design	MD	Turnover (%)	Shape	Intvn. effects	Avg. total estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	Total empSE	Total MCSE (bias)	Total MCSE (empSE)	Non-convergence (%)
44	15	15	R-CS	MCAR	10	Constant	C>I	-0.399	-0.52	1.412	-0.054	0.122	0.005	0.003	6.308
44	15	15	OC	MCAR	10	Constant	C>I	-0.4	-0.302	1.634	0.165	0.122	0.005	0.003	0.923
44	15	15	R-CS	MCAR	20	Constant	C>I	-0.401	0.096	3.443	0.672	0.132	0.005	0.004	7.538
44	15	15	OC	MCAR	20	Constant	C>I	-0.404	0.675	4.042	1.254	0.13	0.005	0.004	1.692
44	15	15	R-CS	MCAR	40	Constant	C>I	-0.395	-1.363	5.053	-1.015	0.135	0.005	0.004	7.077
44	15	15	OC	MCAR	40	Constant	C>I	-0.394	-1.838	4.547	-1.492	0.13	0.005	0.004	4.154
44	15	15	R-CS	MCAR	10	Constant	I>C	-0.399	-0.52	4.143	-0.054	0.122	0.005	0.003	6.308
44	15	15	OC	MCAR	10	Constant	I>C	-0.4	-0.302	4.37	0.165	0.122	0.005	0.003	0.923
44	15	15	R-CS	MCAR	20	Constant	I>C	-0.401	0.096	9.477	0.672	0.132	0.005	0.004	7.538
44	15	15	OC	MCAR	20	Constant	I>C	-0.404	0.675	10.11	1.254	0.13	0.005	0.004	1.692
44	15	15	R-CS	MCAR	40	Constant	I>C	-0.395	-1.363	19.821	-1.015	0.135	0.005	0.004	7.077
44	15	15	OC	MCAR	40	Constant	I>C	-0.394	-1.838	19.243	-1.492	0.13	0.005	0.004	4.154
44	15	15	R-CS	MCAR	10	Non-constant	C>I	-0.399	-0.52	1.412	-0.054	0.122	0.005	0.003	6.308
44	15	15	OC	MCAR	10	Non-constant	C>I	-0.4	-0.302	1.634	0.165	0.122	0.005	0.003	0.923
44	15	15	R-CS	MCAR	20	Non-constant	C>I	-0.401	0.096	3.443	0.672	0.132	0.005	0.004	7.538
44	15	15	OC	MCAR	20	Non-constant	C>I	-0.404	0.675	4.042	1.254	0.13	0.005	0.004	1.692
44	15	15	R-CS	MCAR	40	Non-constant	C>I	-0.395	-1.363	5.053	-1.015	0.135	0.005	0.004	7.077
44	15	15	OC	MCAR	40	Non-constant	C>I	-0.394	-1.838	4.547	-1.492	0.13	0.005	0.004	4.154
44	15	15	R-CS	MCAR	10	Non-constant	I>C	-0.399	-0.484	4.18	-0.018	0.122	0.005	0.003	6.154
44	15	15	OC	MCAR	10	Non-constant	I>C	-0.4	-0.302	4.37	0.165	0.122	0.005	0.003	0.923
44	15	15	R-CS	MCAR	20	Non-constant	I>C	-0.401	0.096	9.477	0.672	0.132	0.005	0.004	7.538
44	15	15	OC	MCAR	20	Non-constant	I>C	-0.404	0.675	10.11	1.254	0.13	0.005	0.004	1.692
44	15	15	R-CS	MCAR	40	Non-constant	I>C	-0.395	-1.363	19.821	-1.015	0.135	0.005	0.004	7.077
44	15	15	OC	MCAR	40	Non-constant	I>C	-0.393	-1.858	19.22	-1.511	0.13	0.005	0.004	4.308
44	15	50	OC	MCAR	10	Constant	C>I	-0.402	0.277	2.224	0.746	0.125	0.005	0.003	1.077
44	15	50	OC	MCAR	20	Constant	C>I	-0.393	-2.03	1.246	-1.466	0.123	0.005	0.003	1.538
44	15	50	OC	MCAR	40	Constant	C>I	-0.396	-1.309	5.111	-0.961	0.139	0.006	0.004	4.308
44	15	50	OC	MCAR	10	Constant	I>C	-0.402	0.277	4.976	0.746	0.125	0.005	0.003	1.077
44	15	50	OC	MCAR	20	Constant	I>C	-0.393	-2.03	7.152	-1.466	0.123	0.005	0.003	1.538
44	15	50	OC	MCAR	40	Constant	I>C	-0.396	-1.309	19.886	-0.961	0.139	0.006	0.004	4.308
44	15	50	OC	MCAR	10	Non-constant	C>I	-0.402	0.277	2.224	0.746	0.125	0.005	0.003	1.077
44	15	50	OC	MCAR	20	Non-constant	C>I	-0.393	-2.025	1.251	-1.461	0.123	0.005	0.003	1.385
44	15	50	OC	MCAR	40	Non-constant	C>I	-0.396	-1.309	5.111	-0.961	0.139	0.006	0.004	4.308
44	15	50	OC	MCAR	10	Non-constant	I>C	-0.402	0.277	4.976	0.746	0.125	0.005	0.003	1.077
44	15	50	OC	MCAR	20	Non-constant	I>C	-0.393	-2.03	7.152	-1.466	0.123	0.005	0.003	1.538
44	15	50	OC	MCAR	40	Non-constant	I>C	-0.396	-1.309	19.886	-0.961	0.139	0.006	0.004	4.308
44	15	50	R-CS	MCAR	10	Constant	C>I	-0.391	-2.515	-0.622	-2.059	0.118	0.005	0.003	0.462
44	15	50	R-CS	MCAR	20	Constant	C>I	-0.397	-1.076	2.233	-0.506	0.127	0.005	0.004	0
44	15	50	R-CS	MCAR	40	Constant	C>I	-0.412	2.683	9.362	3.045	0.134	0.005	0.004	0.615
44	15	50	R-CS	MCAR	10	Constant	I>C	-0.391	-2.515	2.053	-2.059	0.118	0.005	0.003	0.462
44	15	50	R-CS	MCAR	20	Constant	I>C	-0.397	-1.076	8.196	-0.506	0.127	0.005	0.004	0
44	15	50	R-CS	MCAR	40	Constant	I>C	-0.412	2.683	24.736	3.045	0.134	0.005	0.004	0.615
44	15	50	R-CS	MCAR	10	Non-constant	C>I	-0.391	-2.515	-0.622	-2.059	0.118	0.005	0.003	0.462
44	15	50	R-CS	MCAR	20	Non-constant	C>I	-0.397	-1.076	2.233	-0.506	0.127	0.005	0.004	0
44	15	50	R-CS	MCAR	40	Non-constant	C>I	-0.412	2.683	9.362	3.045	0.134	0.005	0.004	0.615
44	15	50	R-CS	MCAR	10	Non-constant	I>C	-0.391	-2.515	2.053	-2.059	0.118	0.005	0.003	0.462
44	15	50	R-CS	MCAR	20	Non-constant	I>C	-0.397	-1.076	8.196	-0.506	0.127	0.005	0.004	0
44	15	50	R-CS	MCAR	40	Non-constant	I>C	-0.412	2.683	24.736	3.045	0.134	0.005	0.004	0.615

Table B.20: Results for the complications discrete model with two timescales and MCAR missing data mechanism.

No. clusters	m	M	Design	MD	Turnover (%)	Shape	Intvn. effects	Avg. total estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	Total empSE	Total MCSE (bias)	Total MCSE (empSE)	Non-convergence (%)
44	15	15	R-CS	MAR	10	Constant	C>I	-0.397	-0.858	0.231	0.3	0.12	0.005	0.003	5.385
44	15	15	OC	MAR	10	Constant	C>I	-0.399	-0.581	0.511	0.58	0.12	0.005	0.003	1.077
44	15	15	R-CS	MAR	20	Constant	C>I	-0.37	-7.666	-4.25	-0.924	0.124	0.005	0.003	2.615
44	15	15	OC	MAR	20	Constant	C>I	-0.385	-3.889	-0.333	3.128	0.123	0.005	0.003	1.077
44	15	15	R-CS	MAR	40	Constant	C>I	-0.348	-13.081	-7.671	0.843	0.129	0.005	0.004	0.308
44	15	15	OC	MAR	40	Constant	C>I	-0.374	-6.648	-0.837	8.307	0.127	0.005	0.004	2.615
44	15	15	R-CS	MAR	10	Constant	I>C	-0.405	1.081	4.863	2.262	0.13	0.005	0.004	3.846
44	15	15	OC	MAR	10	Constant	I>C	-0.406	1.352	5.143	2.535	0.13	0.005	0.004	0.923
44	15	15	R-CS	MAR	20	Constant	I>C	-0.378	-5.77	2.204	1.11	0.112	0.004	0.003	3.538
44	15	15	OC	MAR	20	Constant	I>C	-0.394	-1.666	6.655	5.513	0.115	0.005	0.003	1.385
44	15	15	R-CS	MAR	40	Constant	I>C	-0.361	-10.035	3.307	4.376	0.122	0.005	0.003	1.231
44	15	15	OC	MAR	40	Constant	I>C	-0.386	-3.696	10.587	11.732	0.124	0.005	0.004	4.615
44	15	15	R-CS	MAR	10	Non-constant	C>I	-0.4	-0.255	0.556	1.211	0.127	0.005	0.004	5.077
44	15	15	OC	MAR	10	Non-constant	C>I	-0.402	0.226	1.041	1.699	0.128	0.005	0.004	1.538
44	15	15	R-CS	MAR	20	Non-constant	C>I	-0.364	-9.106	-4.585	1.221	0.125	0.005	0.004	2.308
44	15	15	OC	MAR	20	Non-constant	C>I	-0.388	-3.332	1.477	7.651	0.125	0.005	0.004	1.231
44	15	15	R-CS	MAR	40	Non-constant	C>I	-0.308	-23.099	-14.973	-1.678	0.131	0.005	0.004	0.308
44	15	15	OC	MAR	40	Non-constant	C>I	-0.351	-12.39	-3.132	12.013	0.132	0.005	0.004	4.154
44	15	15	R-CS	MAR	10	Non-constant	I>C	-0.387	-3.469	0.178	-2.24	0.118	0.005	0.003	4.923
44	15	15	OC	MAR	10	Non-constant	I>C	-0.388	-3.174	0.484	-1.942	0.118	0.005	0.003	0.923
44	15	15	R-CS	MAR	20	Non-constant	I>C	-0.374	-6.829	1.682	1.153	0.122	0.005	0.003	3.385
44	15	15	OC	MAR	20	Non-constant	I>C	-0.393	-1.954	7.002	6.445	0.121	0.005	0.003	1.692
44	15	15	R-CS	MAR	40	Non-constant	I>C	-0.332	-17.309	-3.628	-1.177	0.126	0.005	0.004	1.077
44	15	15	OC	MAR	40	Non-constant	I>C	-0.365	-8.911	6.16	8.859	0.127	0.005	0.004	3.692
44	15	50	OC	MAR	10	Constant	C>I	-0.393	-1.956	-0.879	-0.811	0.121	0.005	0.003	1.077
44	15	50	OC	MAR	20	Constant	C>I	-0.385	-3.881	-0.325	3.137	0.121	0.005	0.003	1.692
44	15	50	OC	MAR	40	Constant	C>I	-0.381	-4.977	0.938	10.245	0.125	0.005	0.004	4
44	15	50	OC	MAR	10	Constant	I>C	-0.399	-0.55	3.171	0.612	0.122	0.005	0.003	0.462
44	15	50	OC	MAR	20	Constant	I>C	-0.383	-4.488	3.594	2.485	0.125	0.005	0.003	1.692
44	15	50	OC	MAR	40	Constant	I>C	-0.37	-7.769	5.91	7.006	0.124	0.005	0.004	4.154
44	15	50	OC	MAR	10	Non-constant	C>I	-0.391	-2.546	-1.753	-1.114	0.124	0.005	0.003	0.769
44	15	50	OC	MAR	20	Non-constant	C>I	-0.383	-4.483	0.269	6.369	0.128	0.005	0.004	1.692
44	15	50	OC	MAR	40	Non-constant	C>I	-0.355	-11.436	-2.077	13.234	0.131	0.005	0.004	4.769
44	15	50	OC	MAR	10	Non-constant	I>C	-0.393	-2.098	1.602	-0.851	0.128	0.005	0.004	0.462
44	15	50	OC	MAR	20	Non-constant	I>C	-0.384	-4.185	4.568	4.024	0.119	0.005	0.003	2.615
44	15	50	OC	MAR	40	Non-constant	I>C	-0.362	-9.774	5.154	7.828	0.127	0.005	0.004	4.769
44	15	50	R-CS	MAR	10	Constant	C>I	-0.39	-2.808	-1.74	-1.673	0.122	0.005	0.003	0.462
44	15	50	R-CS	MAR	20	Constant	C>I	-0.374	-6.774	-3.325	0.033	0.127	0.005	0.004	0
44	15	50	R-CS	MAR	40	Constant	C>I	-0.34	-15.299	-10.026	-1.73	0.125	0.005	0.003	1.385
44	15	50	R-CS	MAR	10	Constant	I>C	-0.395	-1.597	2.085	-0.447	0.122	0.005	0.003	0.462
44	15	50	R-CS	MAR	20	Constant	I>C	-0.376	-6.287	1.644	0.555	0.118	0.005	0.003	0.462
44	15	50	R-CS	MAR	40	Constant	I>C	-0.348	-13.32	-0.465	0.565	0.128	0.005	0.004	2
44	15	50	R-CS	MAR	10	Non-constant	C>I	-0.387	-3.415	-2.63	-1.996	0.121	0.005	0.003	0
44	15	50	R-CS	MAR	20	Non-constant	C>I	-0.361	-10.073	-5.599	0.144	0.126	0.005	0.004	0.462
44	15	50	R-CS	MAR	40	Non-constant	C>I	-0.311	-22.501	-14.312	-0.914	0.123	0.005	0.003	1.077
44	15	50	R-CS	MAR	10	Non-constant	I>C	-0.398	-0.82	2.927	0.442	0.124	0.005	0.003	0.154
44	15	50	R-CS	MAR	20	Non-constant	I>C	-0.367	-8.343	0.03	-0.49	0.126	0.005	0.003	0.154
44	15	50	R-CS	MAR	40	Non-constant	I>C	-0.338	-15.575	-1.607	0.895	0.124	0.005	0.003	1.385

Table B.21: Results for the complications discrete model with two timescales and MAR missing data mechanism.

No. clusters	m	M	Design	MD	Turnover (%)	Shape	Intvn. effects	Avg. total estimate	CC-I bias (%)	CS-OC-D bias (%)	CC-D bias (%)	Total empSE	Total MCSE (bias)	Total MCSE (empSE)	Non-convergence (%)
44	15	15	R-CS	MNAR	10	Constant	C>I	-0.408	1.7	2.329	2.673	0.13	0.005	0.004	6.615
44	15	15	OC	MNAR	10	Constant	C>I	-0.407	1.613	2.243	2.585	0.128	0.005	0.004	0.769
44	15	15	R-CS	MNAR	20	Constant	C>I	-0.368	-8.109	-4.146	-0.204	0.119	0.005	0.003	2.769
44	15	15	OC	MNAR	20	Constant	C>I	-0.386	-3.773	0.377	4.505	0.12	0.005	0.003	1.846
44	15	15	R-CS	MNAR	40	Constant	C>I	-0.338	-15.702	-8.466	-0.849	0.125	0.005	0.003	0.769
44	15	15	OC	MNAR	40	Constant	C>I	-0.366	-8.624	-0.78	7.476	0.127	0.005	0.004	4.154
44	15	15	R-CS	MNAR	10	Constant	I>C	-0.397	-1.032	2.3	-0.085	0.124	0.005	0.004	5.846
44	15	15	OC	MNAR	10	Constant	I>C	-0.399	-0.587	2.759	0.364	0.123	0.005	0.003	0.923
44	15	15	R-CS	MNAR	20	Constant	I>C	-0.369	-7.885	0.031	0.039	0.118	0.005	0.003	2.154
44	15	15	OC	MNAR	20	Constant	I>C	-0.388	-3.274	5.039	5.047	0.118	0.005	0.003	0.923
44	15	15	R-CS	MNAR	40	Constant	I>C	-0.333	-16.93	-3.924	-2.294	0.12	0.005	0.003	1.231
44	15	15	OC	MNAR	40	Constant	I>C	-0.362	-9.634	4.515	6.288	0.122	0.005	0.003	5.077
44	15	15	R-CS	MNAR	10	Non-constant	C>I	-0.393	-2.012	-1.54	-0.733	0.118	0.005	0.003	5.846
44	15	15	OC	MNAR	10	Non-constant	C>I	-0.397	-1.034	-0.558	0.257	0.119	0.005	0.003	0.615
44	15	15	R-CS	MNAR	20	Non-constant	C>I	-0.347	-13.551	-7.585	-1.419	0.114	0.005	0.003	2.769
44	15	15	OC	MNAR	20	Non-constant	C>I	-0.373	-6.902	-0.477	6.163	0.117	0.005	0.003	2.462
44	15	15	R-CS	MNAR	40	Non-constant	C>I	-0.305	-23.916	-13.088	-0.271	0.118	0.005	0.003	1.692
44	15	15	OC	MNAR	40	Non-constant	C>I	-0.35	-12.673	-0.245	14.467	0.118	0.005	0.003	4.462
44	15	15	R-CS	MNAR	10	Non-constant	I>C	-0.396	-1.339	1.949	-0.296	0.12	0.005	0.003	6.923
44	15	15	OC	MNAR	10	Non-constant	I>C	-0.398	-0.806	2.5	0.242	0.121	0.005	0.003	1.538
44	15	15	R-CS	MNAR	20	Non-constant	I>C	-0.358	-10.711	-1.998	-1.536	0.12	0.005	0.003	2.308
44	15	15	OC	MNAR	20	Non-constant	I>C	-0.379	-5.584	3.629	4.118	0.119	0.005	0.003	1.692
44	15	15	R-CS	MNAR	40	Non-constant	I>C	-0.338	-15.761	-0.493	2.376	0.125	0.005	0.003	0.154
44	15	15	OC	MNAR	40	Non-constant	I>C	-0.372	-7.215	9.603	12.762	0.123	0.005	0.003	3.846
44	15	50	OC	MNAR	10	Constant	C>I	-0.4	-0.105	0.514	0.851	0.128	0.005	0.004	1.077
44	15	50	OC	MNAR	20	Constant	C>I	-0.392	-2.319	1.894	6.084	0.12	0.005	0.003	1.385
44	15	50	OC	MNAR	40	Constant	C>I	-0.371	-7.479	0.464	8.823	0.125	0.005	0.004	5.385
44	15	50	OC	MNAR	10	Constant	I>C	-0.391	-2.48	0.802	-1.547	0.127	0.005	0.004	0.154
44	15	50	OC	MNAR	20	Constant	I>C	-0.39	-2.825	5.526	5.534	0.128	0.005	0.004	2
44	15	50	OC	MNAR	40	Constant	I>C	-0.369	-7.946	6.467	8.273	0.124	0.005	0.004	4.923
44	15	50	OC	MNAR	10	Non-constant	C>I	-0.402	0.154	0.636	1.461	0.119	0.005	0.003	0.923
44	15	50	OC	MNAR	20	Non-constant	C>I	-0.375	-6.344	0.119	6.799	0.117	0.005	0.003	1.846
44	15	50	OC	MNAR	40	Non-constant	C>I	-0.354	-11.814	0.735	15.592	0.124	0.005	0.004	4.308
44	15	50	OC	MNAR	10	Non-constant	I>C	-0.4	-0.136	3.192	0.919	0.12	0.005	0.003	1.385
44	15	50	OC	MNAR	20	Non-constant	I>C	-0.389	-3.052	6.408	6.91	0.125	0.005	0.003	1.077
44	15	50	OC	MNAR	40	Non-constant	I>C	-0.361	-9.897	6.434	9.503	0.119	0.005	0.003	5.538
44	15	50	R-CS	MNAR	10	Constant	C>I	-0.4	-0.141	0.477	0.814	0.124	0.005	0.003	0.462
44	15	50	R-CS	MNAR	20	Constant	C>I	-0.368	-8.089	-4.125	-0.183	0.122	0.005	0.003	0.462
44	15	50	R-CS	MNAR	40	Constant	C>I	-0.336	-16.253	-9.064	-1.498	0.119	0.005	0.003	1.077
44	15	50	R-CS	MNAR	10	Constant	I>C	-0.395	-1.419	1.899	-0.476	0.119	0.005	0.003	0.154
44	15	50	R-CS	MNAR	20	Constant	I>C	-0.37	-7.765	0.161	0.169	0.122	0.005	0.003	0.615
44	15	50	R-CS	MNAR	40	Constant	I>C	-0.34	-15.293	-2.03	-0.368	0.125	0.005	0.003	2
44	15	50	R-CS	MNAR	10	Non-constant	C>I	-0.403	0.569	1.053	1.881	0.121	0.005	0.003	0
44	15	50	R-CS	MNAR	20	Non-constant	C>I	-0.353	-11.922	-5.844	0.438	0.116	0.005	0.003	0.308
44	15	50	R-CS	MNAR	40	Non-constant	C>I	-0.298	-25.596	-15.007	-2.472	0.12	0.005	0.003	1.692
44	15	50	R-CS	MNAR	10	Non-constant	I>C	-0.392	-2.266	0.991	-1.233	0.124	0.005	0.003	0.308
44	15	50	R-CS	MNAR	20	Non-constant	I>C	-0.361	-10.069	-1.294	-0.828	0.124	0.005	0.003	0.308
44	15	50	R-CS	MNAR	40	Non-constant	I>C	-0.331	-17.522	-2.572	0.236	0.127	0.005	0.004	0.769

Table B.22: Results for the complications discrete model with two timescales and MNAR missing data mechanism.

No. clusters	m	M	Design	MD	Turnover (%)	Shape	Intvn. effects	OC-52-I bias (%)	OC-26-I bias (%)	OC-52-I empSE	OC-26-I empSE	OC-52-I MCSE (bias)	OC-26-I MCSE (bias)	OC-52-I MCSE (empSE)	OC-26-I MCSE (empSE)	Non-convergence (%)
44	15	15	R-CS	MCAR	10	Constant	C>I	-1.246	-1.37	0.161	0.304	0.007	0.012	0.005	0.009	7.077
44	15	15	OC	MCAR	10	Constant	C>I	0.215	1.726	0.159	0.296	0.006	0.012	0.004	0.008	0.769
44	15	15	R-CS	MCAR	20	Constant	C>I	-1.073	-3.3	0.115	0.203	0.005	0.008	0.003	0.006	6.923
44	15	15	OC	MCAR	20	Constant	C>I	-1.598	-5.25	0.11	0.189	0.004	0.008	0.003	0.005	3.077
44	15	15	R-CS	MCAR	40	Constant	C>I	-1.12	-0.372	0.091	0.139	0.004	0.006	0.003	0.004	4.615
44	15	15	OC	MCAR	40	Constant	C>I	-0.331	0.668	0.087	0.132	0.004	0.005	0.002	0.004	5.385
44	15	15	R-CS	MCAR	10	Constant	I>C	-1.523	-2.283	0.161	0.304	0.007	0.012	0.005	0.009	7.077
44	15	15	OC	MCAR	10	Constant	I>C	0.205	2.772	0.159	0.296	0.006	0.012	0.004	0.008	0.615
44	15	15	R-CS	MCAR	20	Constant	I>C	-1.312	-5.5	0.115	0.203	0.005	0.008	0.003	0.006	6.923
44	15	15	OC	MCAR	20	Constant	I>C	-1.987	-8.759	0.11	0.189	0.004	0.008	0.003	0.005	2.923
44	15	15	R-CS	MCAR	40	Constant	I>C	-1.369	-0.62	0.091	0.139	0.004	0.006	0.003	0.004	4.615
44	15	15	OC	MCAR	40	Constant	I>C	-0.427	1.121	0.087	0.132	0.004	0.005	0.002	0.004	5.231
44	15	15	R-CS	MCAR	10	Non-constant	C>I	-2.816	-5.83	0.161	0.302	0.006	0.012	0.005	0.009	4
44	15	15	OC	MCAR	10	Non-constant	C>I	1.598	3.708	0.159	0.297	0.006	0.012	0.004	0.008	0.923
44	15	15	R-CS	MCAR	20	Non-constant	C>I	-1.641	-6.946	0.115	0.202	0.005	0.008	0.003	0.006	3.846
44	15	15	OC	MCAR	20	Non-constant	C>I	0.251	-2.628	0.11	0.19	0.004	0.008	0.003	0.005	3.077
44	15	15	R-CS	MCAR	40	Non-constant	C>I	0.409	-2.037	0.092	0.139	0.004	0.006	0.003	0.004	2.923
44	15	15	OC	MCAR	40	Non-constant	C>I	2.389	3.64	0.088	0.133	0.004	0.005	0.003	0.004	7.846
44	15	15	R-CS	MCAR	10	Non-constant	I>C	-1.918	-4.175	0.16	0.302	0.006	0.012	0.005	0.009	6.615
44	15	15	OC	MCAR	10	Non-constant	I>C	0.642	3.722	0.159	0.296	0.006	0.012	0.004	0.008	0.615
44	15	15	R-CS	MCAR	20	Non-constant	I>C	-1.458	-7.269	0.115	0.202	0.005	0.008	0.003	0.006	6.462
44	15	15	OC	MCAR	20	Non-constant	I>C	-1.379	-7.675	0.11	0.189	0.004	0.008	0.003	0.005	3.077
44	15	15	R-CS	MCAR	40	Non-constant	I>C	-0.816	-1.9	0.091	0.139	0.004	0.006	0.003	0.004	4
44	15	15	OC	MCAR	40	Non-constant	I>C	0.702	2.714	0.087	0.132	0.004	0.005	0.002	0.004	5.692
44	15	50	OC	MCAR	10	Constant	C>I	0.266	1.738	0.156	0.286	0.006	0.011	0.004	0.008	1.077
44	15	50	OC	MCAR	20	Constant	C>I	-0.214	0.672	0.115	0.205	0.005	0.008	0.003	0.006	2.154
44	15	50	OC	MCAR	40	Constant	C>I	-0.099	0.686	0.091	0.135	0.004	0.005	0.003	0.004	5.385
44	15	50	OC	MCAR	10	Constant	I>C	0.325	2.896	0.156	0.286	0.006	0.011	0.004	0.008	1.077
44	15	50	OC	MCAR	20	Constant	I>C	-0.254	1.178	0.116	0.205	0.005	0.008	0.003	0.006	2.308
44	15	50	OC	MCAR	40	Constant	I>C	-0.121	1.143	0.091	0.135	0.004	0.005	0.003	0.004	5.385
44	15	50	OC	MCAR	10	Non-constant	C>I	1.346	3.347	0.157	0.286	0.006	0.011	0.004	0.008	1.231
44	15	50	OC	MCAR	20	Non-constant	C>I	1.34	2.602	0.115	0.205	0.005	0.008	0.003	0.006	2.615
44	15	50	OC	MCAR	40	Non-constant	C>I	2.914	4.169	0.09	0.134	0.004	0.005	0.003	0.004	6.462
44	15	50	OC	MCAR	10	Non-constant	I>C	0.546	3.398	0.156	0.285	0.006	0.011	0.004	0.008	1.385
44	15	50	OC	MCAR	20	Non-constant	I>C	0.327	2.209	0.115	0.205	0.005	0.008	0.003	0.006	2.154
44	15	50	OC	MCAR	40	Non-constant	I>C	1.067	3.04	0.091	0.135	0.004	0.005	0.003	0.004	5.385
44	15	50	R-CS	MCAR	10	Constant	C>I	3.197	10.166	0.159	0.302	0.006	0.012	0.004	0.008	0.769
44	15	50	R-CS	MCAR	20	Constant	C>I	-1.362	-1.981	0.115	0.195	0.005	0.008	0.003	0.005	0
44	15	50	R-CS	MCAR	40	Constant	C>I	0.705	-0.627	0.091	0.129	0.004	0.005	0.003	0.004	0.769
44	15	50	R-CS	MCAR	10	Constant	I>C	3.907	16.943	0.159	0.302	0.006	0.012	0.004	0.008	0.769
44	15	50	R-CS	MCAR	20	Constant	I>C	-1.664	-3.301	0.115	0.195	0.005	0.008	0.003	0.005	0
44	15	50	R-CS	MCAR	40	Constant	I>C	0.861	-1.044	0.091	0.129	0.004	0.005	0.003	0.004	0.769
44	15	50	R-CS	MCAR	10	Non-constant	C>I	4.789	9.31	0.158	0.302	0.006	0.012	0.004	0.008	0.462
44	15	50	R-CS	MCAR	20	Non-constant	C>I	0.951	-2.156	0.116	0.196	0.005	0.008	0.003	0.005	0
44	15	50	R-CS	MCAR	40	Non-constant	C>I	4.398	0.289	0.092	0.129	0.004	0.005	0.003	0.004	0.308
44	15	50	R-CS	MCAR	10	Non-constant	I>C	4.395	16.078	0.158	0.302	0.006	0.012	0.004	0.008	0.462
44	15	50	R-CS	MCAR	20	Non-constant	I>C	-0.84	-3.612	0.115	0.195	0.005	0.008	0.003	0.005	0
44	15	50	R-CS	MCAR	40	Non-constant	I>C	2.337	-0.663	0.091	0.129	0.004	0.005	0.003	0.004	0.769

Table B.23: Results for the complications continuous model with two timescales and MCAR missing data mechanism.

No. clusters	m	M	Design	MD	Turnover (%)	Shape	Intvn. effects	OC-52-I bias (%)	OC-26-I bias (%)	OC-52-I empSE	OC-26-I empSE	OC-52-I MCSE (bias)	OC-26-I MCSE (bias)	OC-52-I MCSE (empSE)	OC-26-I MCSE (empSE)	Non-convergence (%)
44	15	15	R-CS	MAR	10	Constant	C>I	0.972	2.57	0.166	0.317	0.007	0.013	0.005	0.009	4.308
44	15	15	OC	MAR	10	Constant	C>I	1.032	2.202	0.16	0.306	0.006	0.012	0.004	0.009	1.231
44	15	15	R-CS	MAR	20	Constant	C>I	-6.531	-4.073	0.131	0.237	0.005	0.009	0.004	0.007	2.462
44	15	15	OC	MAR	20	Constant	C>I	-3.989	-3.423	0.127	0.229	0.005	0.009	0.004	0.006	1.538
44	15	15	R-CS	MAR	40	Constant	C>I	-15.751	-16.013	0.106	0.18	0.004	0.007	0.003	0.005	0.769
44	15	15	OC	MAR	40	Constant	C>I	-10.562	-12.534	0.104	0.176	0.004	0.007	0.003	0.005	8.615
44	15	15	R-CS	MAR	10	Constant	I>C	2.688	6.432	0.167	0.319	0.007	0.013	0.005	0.009	5.231
44	15	15	OC	MAR	10	Constant	I>C	3.142	7.62	0.163	0.306	0.006	0.012	0.005	0.009	1.385
44	15	15	R-CS	MAR	20	Constant	I>C	-2.101	7.973	0.135	0.247	0.005	0.01	0.004	0.007	2.615
44	15	15	OC	MAR	20	Constant	I>C	0.324	7.328	0.129	0.235	0.005	0.009	0.004	0.007	2.615
44	15	15	R-CS	MAR	40	Constant	I>C	-10.92	-9.503	0.103	0.173	0.004	0.007	0.003	0.005	0.615
44	15	15	OC	MAR	40	Constant	I>C	-4.912	-3.195	0.098	0.167	0.004	0.007	0.003	0.005	6.615
44	15	15	R-CS	MAR	10	Non-constant	C>I	-0.015	-1.284	0.171	0.327	0.007	0.013	0.005	0.009	3.385
44	15	15	OC	MAR	10	Non-constant	C>I	2.73	4.366	0.167	0.317	0.007	0.013	0.005	0.009	2.154
44	15	15	R-CS	MAR	20	Non-constant	C>I	-6.376	-5.143	0.134	0.244	0.005	0.01	0.004	0.007	1.077
44	15	15	OC	MAR	20	Non-constant	C>I	-1.032	0.917	0.13	0.234	0.005	0.009	0.004	0.007	2.308
44	15	15	R-CS	MAR	40	Non-constant	C>I	-19.588	-17.226	0.11	0.175	0.004	0.007	0.003	0.005	0.769
44	15	15	OC	MAR	40	Non-constant	C>I	-12.226	-11.234	0.103	0.169	0.004	0.007	0.003	0.005	8.154
44	15	15	R-CS	MAR	10	Non-constant	I>C	-3.313	-2.092	0.172	0.327	0.007	0.013	0.005	0.009	3.846
44	15	15	OC	MAR	10	Non-constant	I>C	-1.948	1.95	0.168	0.318	0.007	0.013	0.005	0.009	1.077
44	15	15	R-CS	MAR	20	Non-constant	I>C	-6.266	-4.8	0.137	0.255	0.005	0.01	0.004	0.007	2
44	15	15	OC	MAR	20	Non-constant	I>C	-2.243	-1.49	0.133	0.246	0.005	0.01	0.004	0.007	2.462
44	15	15	R-CS	MAR	40	Non-constant	I>C	-13.499	-3.77	0.107	0.178	0.004	0.007	0.003	0.005	1.077
44	15	15	OC	MAR	40	Non-constant	I>C	-6.227	3.668	0.103	0.175	0.004	0.007	0.003	0.005	7.077
44	15	50	OC	MAR	10	Constant	C>I	-1.862	-1.572	0.153	0.291	0.006	0.012	0.004	0.008	2.154
44	15	50	OC	MAR	20	Constant	C>I	-2.966	-1.189	0.127	0.228	0.005	0.009	0.004	0.006	3.077
44	15	50	OC	MAR	40	Constant	C>I	-11.382	-14.857	0.093	0.171	0.004	0.007	0.003	0.005	8.154
44	15	50	OC	MAR	10	Constant	I>C	-1.573	-5.544	0.165	0.312	0.007	0.012	0.005	0.009	1.231
44	15	50	OC	MAR	20	Constant	I>C	-2.055	2.129	0.127	0.234	0.005	0.009	0.004	0.007	2.154
44	15	50	OC	MAR	40	Constant	I>C	-6.629	-0.806	0.096	0.168	0.004	0.007	0.003	0.005	9.692
44	15	50	OC	MAR	10	Non-constant	C>I	0.806	2.815	0.163	0.309	0.006	0.012	0.005	0.009	1.692
44	15	50	OC	MAR	20	Non-constant	C>I	-3.397	-1.719	0.134	0.231	0.005	0.009	0.004	0.007	3.385
44	15	50	OC	MAR	40	Non-constant	C>I	-13.262	-12.46	0.099	0.166	0.004	0.007	0.003	0.005	9.231
44	15	50	OC	MAR	10	Non-constant	I>C	2.569	10.752	0.161	0.298	0.006	0.012	0.004	0.008	1.231
44	15	50	OC	MAR	20	Non-constant	I>C	-2.599	0.569	0.13	0.233	0.005	0.009	0.004	0.007	2.308
44	15	50	OC	MAR	40	Non-constant	I>C	-8.893	-3.067	0.103	0.175	0.004	0.007	0.003	0.005	8.615
44	15	50	R-CS	MAR	10	Constant	C>I	-1.205	0.713	0.163	0.299	0.006	0.012	0.005	0.008	0.462
44	15	50	R-CS	MAR	20	Constant	C>I	-5.506	-2.656	0.131	0.238	0.005	0.009	0.004	0.007	0.154
44	15	50	R-CS	MAR	40	Constant	C>I	-17.076	-16.764	0.103	0.169	0.004	0.007	0.003	0.005	2.615
44	15	50	R-CS	MAR	10	Constant	I>C	-0.363	2.118	0.165	0.311	0.006	0.012	0.005	0.009	0.615
44	15	50	R-CS	MAR	20	Constant	I>C	-5.196	-2.206	0.13	0.234	0.005	0.009	0.004	0.007	0.462
44	15	50	R-CS	MAR	40	Constant	I>C	-11.934	-5	0.103	0.164	0.004	0.007	0.003	0.005	3.385
44	15	50	R-CS	MAR	10	Non-constant	C>I	0.364	0.575	0.166	0.307	0.006	0.012	0.005	0.009	0
44	15	50	R-CS	MAR	20	Non-constant	C>I	-6.645	-7.489	0.132	0.229	0.005	0.009	0.004	0.006	0.462
44	15	50	R-CS	MAR	40	Non-constant	C>I	-20.243	-19.564	0.106	0.171	0.004	0.007	0.003	0.005	3.692
44	15	50	R-CS	MAR	10	Non-constant	I>C	-2.514	-8.872	0.168	0.315	0.007	0.012	0.005	0.009	0.154
44	15	50	R-CS	MAR	20	Non-constant	I>C	-3.608	3.675	0.139	0.248	0.005	0.01	0.004	0.007	0.308
44	15	50	R-CS	MAR	40	Non-constant	I>C	-14.016	-10.672	0.099	0.163	0.004	0.007	0.003	0.005	2.769

Table B.24: Results for the complications continuous model with two timescales and MAR missing data mechanism.

No. clusters	m	M	Design	MD	Turnover (%)	Shape	Intvn. effects	OC-52-I bias (%)	OC-26-I bias (%)	OC-52-I empSE	OC-26-I empSE	OC-52-I MCSE (bias)	OC-26-I MCSE (bias)	OC-52-I MCSE (empSE)	OC-26-I MCSE (empSE)	Non-convergence (%)
44	15	15	R-CS	MNAR	10	Constant	C>I	-1.605	-4.757	0.17	0.323	0.007	0.013	0.005	0.009	7.385
44	15	15	OC	MNAR	10	Constant	C>I	-0.222	-1.67	0.168	0.316	0.007	0.013	0.005	0.009	1.846
44	15	15	R-CS	MNAR	20	Constant	C>I	-9.275	-10.154	0.134	0.247	0.005	0.01	0.004	0.007	2.154
44	15	15	OC	MNAR	20	Constant	C>I	-6.55	-9.492	0.13	0.241	0.005	0.01	0.004	0.007	2.462
44	15	15	R-CS	MNAR	40	Constant	C>I	-18.035	-17.338	0.106	0.183	0.004	0.007	0.003	0.005	1.538
44	15	15	OC	MNAR	40	Constant	C>I	-12.4	-13.601	0.102	0.179	0.004	0.007	0.003	0.005	10.769
44	15	15	R-CS	MNAR	10	Constant	I>C	-1.066	-1.028	0.174	0.331	0.007	0.013	0.005	0.009	6.308
44	15	15	OC	MNAR	10	Constant	I>C	-0.529	-0.721	0.17	0.324	0.007	0.013	0.005	0.009	1.077
44	15	15	R-CS	MNAR	20	Constant	I>C	-5.938	-1.095	0.137	0.253	0.005	0.01	0.004	0.007	2
44	15	15	OC	MNAR	20	Constant	I>C	-3.2	-1.314	0.135	0.248	0.005	0.01	0.004	0.007	1.692
44	15	15	R-CS	MNAR	40	Constant	I>C	-16.062	-12.618	0.108	0.184	0.004	0.007	0.003	0.005	1.846
44	15	15	OC	MNAR	40	Constant	I>C	-9.868	-6.887	0.105	0.178	0.004	0.007	0.003	0.005	10.769
44	15	15	R-CS	MNAR	10	Non-constant	C>I	-3.632	-6.291	0.16	0.305	0.006	0.012	0.005	0.009	3.077
44	15	15	OC	MNAR	10	Non-constant	C>I	-0.757	-0.271	0.159	0.3	0.006	0.012	0.004	0.008	0.923
44	15	15	R-CS	MNAR	20	Non-constant	C>I	-9.296	-6.485	0.142	0.256	0.006	0.01	0.004	0.007	1.077
44	15	15	OC	MNAR	20	Non-constant	C>I	-3.595	-0.602	0.138	0.25	0.006	0.01	0.004	0.007	3.385
44	15	15	R-CS	MNAR	40	Non-constant	C>I	-22.617	-21.737	0.11	0.193	0.004	0.008	0.003	0.005	2
44	15	15	OC	MNAR	40	Non-constant	C>I	-15.712	-16.176	0.104	0.183	0.004	0.008	0.003	0.005	9.385
44	15	15	R-CS	MNAR	10	Non-constant	I>C	3.207	11.919	0.172	0.328	0.007	0.013	0.005	0.009	6.308
44	15	15	OC	MNAR	10	Non-constant	I>C	5.111	15.775	0.165	0.313	0.007	0.012	0.005	0.009	1.692
44	15	15	R-CS	MNAR	20	Non-constant	I>C	-12.166	-16.227	0.143	0.267	0.006	0.011	0.004	0.008	2.308
44	15	15	OC	MNAR	20	Non-constant	I>C	-8.142	-13.584	0.136	0.256	0.005	0.01	0.004	0.007	2.308
44	15	15	R-CS	MNAR	40	Non-constant	I>C	-15.122	-9.442	0.107	0.181	0.004	0.007	0.003	0.005	0.615
44	15	15	OC	MNAR	40	Non-constant	I>C	-7.771	-2.073	0.104	0.174	0.004	0.007	0.003	0.005	8
44	15	50	OC	MNAR	10	Constant	C>I	-1.504	-3.264	0.153	0.296	0.006	0.012	0.004	0.008	1.538
44	15	50	OC	MNAR	20	Constant	C>I	-5.586	-8.005	0.138	0.255	0.005	0.01	0.004	0.007	2.308
44	15	50	OC	MNAR	40	Constant	C>I	-11.686	-12.251	0.104	0.183	0.004	0.008	0.003	0.005	12.308
44	15	50	OC	MNAR	10	Constant	I>C	0.693	7.298	0.163	0.309	0.006	0.012	0.005	0.009	0.615
44	15	50	OC	MNAR	20	Constant	I>C	-0.571	5.259	0.142	0.252	0.006	0.01	0.004	0.007	2.615
44	15	50	OC	MNAR	40	Constant	I>C	-7.963	-2.899	0.11	0.188	0.005	0.008	0.003	0.006	10.308
44	15	50	OC	MNAR	10	Non-constant	C>I	2.758	4.988	0.164	0.303	0.007	0.012	0.005	0.008	2
44	15	50	OC	MNAR	20	Non-constant	C>I	-5.172	-4.209	0.127	0.228	0.005	0.009	0.004	0.006	2.923
44	15	50	OC	MNAR	40	Non-constant	C>I	-15.615	-16.267	0.105	0.175	0.004	0.007	0.003	0.005	10.308
44	15	50	OC	MNAR	10	Non-constant	I>C	-1.243	-3.25	0.165	0.314	0.007	0.012	0.005	0.009	2.308
44	15	50	OC	MNAR	20	Non-constant	I>C	-3.278	-3.621	0.135	0.251	0.005	0.01	0.004	0.007	2.154
44	15	50	OC	MNAR	40	Non-constant	I>C	-10.483	-7.107	0.102	0.179	0.004	0.007	0.003	0.005	10.154
44	15	50	R-CS	MNAR	10	Constant	C>I	-0.031	0.287	0.162	0.302	0.006	0.012	0.005	0.008	0.615
44	15	50	R-CS	MNAR	20	Constant	C>I	-6.999	-5.799	0.132	0.239	0.005	0.009	0.004	0.007	0.308
44	15	50	R-CS	MNAR	40	Constant	C>I	-19.54	-22.212	0.105	0.181	0.004	0.007	0.003	0.005	4.769
44	15	50	R-CS	MNAR	10	Constant	I>C	-0.715	1.838	0.162	0.303	0.006	0.012	0.004	0.008	0
44	15	50	R-CS	MNAR	20	Constant	I>C	-6.293	-2.725	0.135	0.24	0.005	0.009	0.004	0.007	0.462
44	15	50	R-CS	MNAR	40	Constant	I>C	-14.162	-8.668	0.106	0.184	0.004	0.007	0.003	0.005	4.308
44	15	50	R-CS	MNAR	10	Non-constant	C>I	0.03	-3.85	0.174	0.323	0.007	0.013	0.005	0.009	0.154
44	15	50	R-CS	MNAR	20	Non-constant	C>I	-8.489	-8.108	0.129	0.236	0.005	0.009	0.004	0.007	0.615
44	15	50	R-CS	MNAR	40	Non-constant	C>I	-22.087	-21.445	0.104	0.175	0.004	0.007	0.003	0.005	2.615
44	15	50	R-CS	MNAR	10	Non-constant	I>C	-3.851	-9.204	0.16	0.297	0.006	0.012	0.004	0.008	0.462
44	15	50	R-CS	MNAR	20	Non-constant	I>C	-8.937	-8.173	0.128	0.238	0.005	0.009	0.004	0.007	0.462
44	15	50	R-CS	MNAR	40	Non-constant	I>C	-16.507	-12.627	0.106	0.183	0.004	0.007	0.003	0.005	4.308

Table B.25: Results for the complications continuous model with two timescales and MNAR missing data mechanism.

Appendix C

Appendix for Chapter 6

C.1 Figures

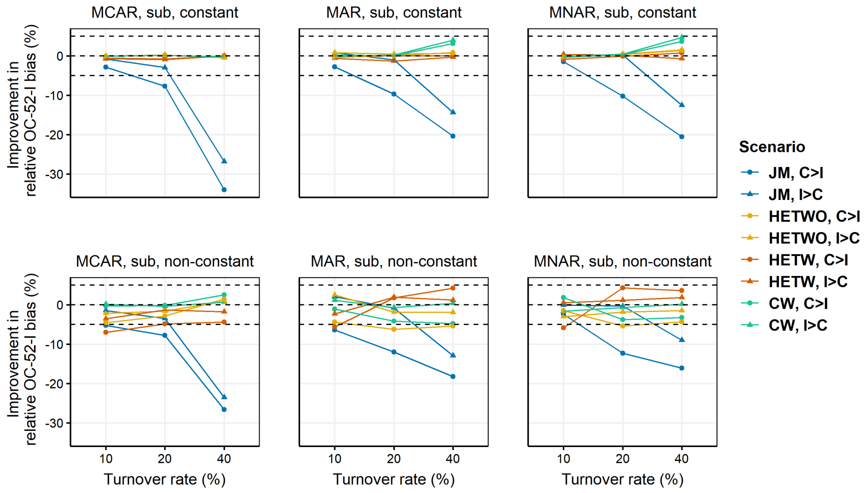


Figure C.1: Improvement in relative bias (%) for all models for the OC-52-I estimand in sub-samples in comparison to the Kasza model. 'Sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

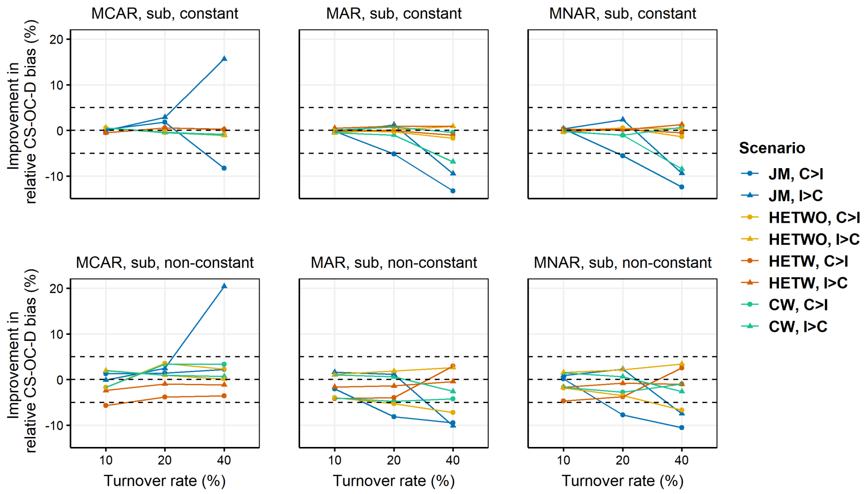


Figure C.2: Improvement in relative bias (%) for all models for the CS-OC-D estimand in sub-samples with 2 timescales in comparison to the CC-I. 'Sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

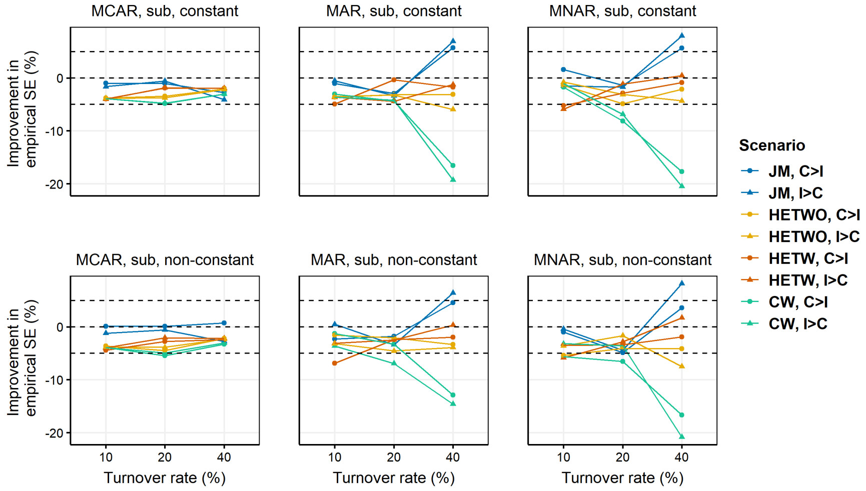


Figure C.3: Improvement in empirical SE (%) for all models for the CS-OC-D estimand in sub-samples with 2 timescales in comparison to the Kasza model. 'Sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

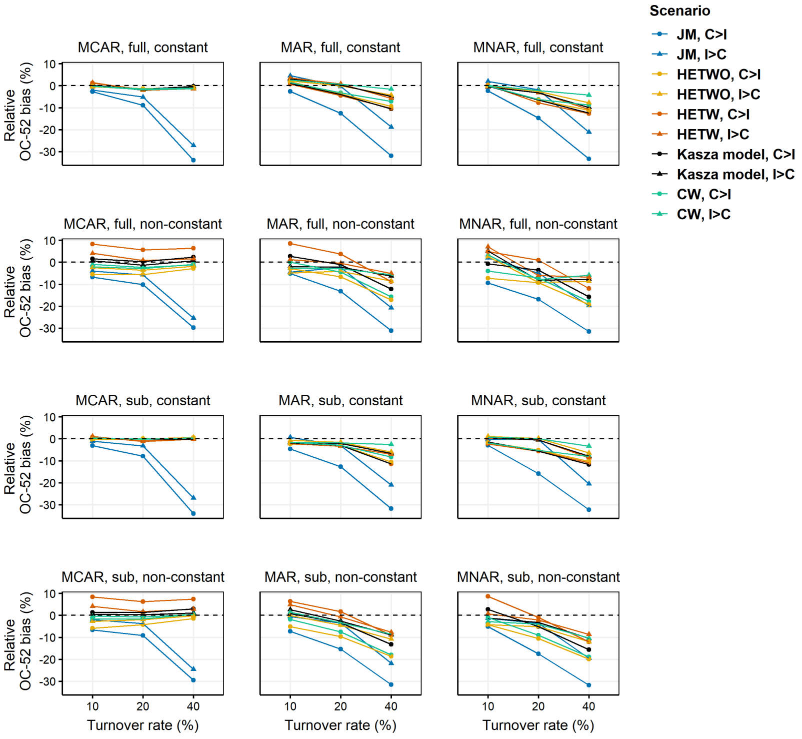


Figure C.4: Relative bias (%) for all models for the OC-52-I estimand with 2 timescales. 'Full' denotes samples of 15 taken from cluster populations of 15; 'sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

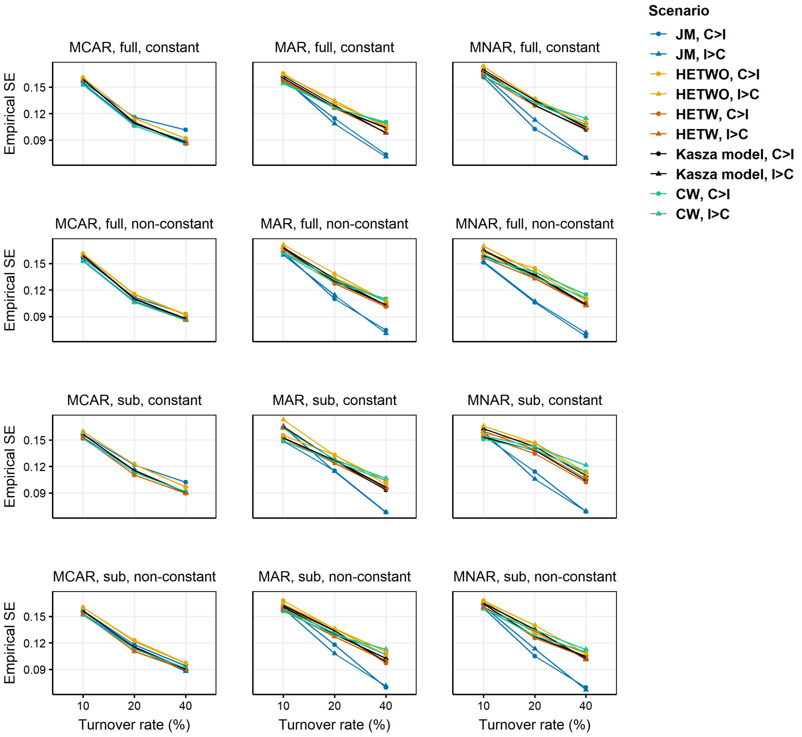


Figure C.5: Empirical SE for all models for the OC-52-I estimand with 2 timescales. 'Full' denotes samples of 15 taken from cluster populations of 15; 'sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

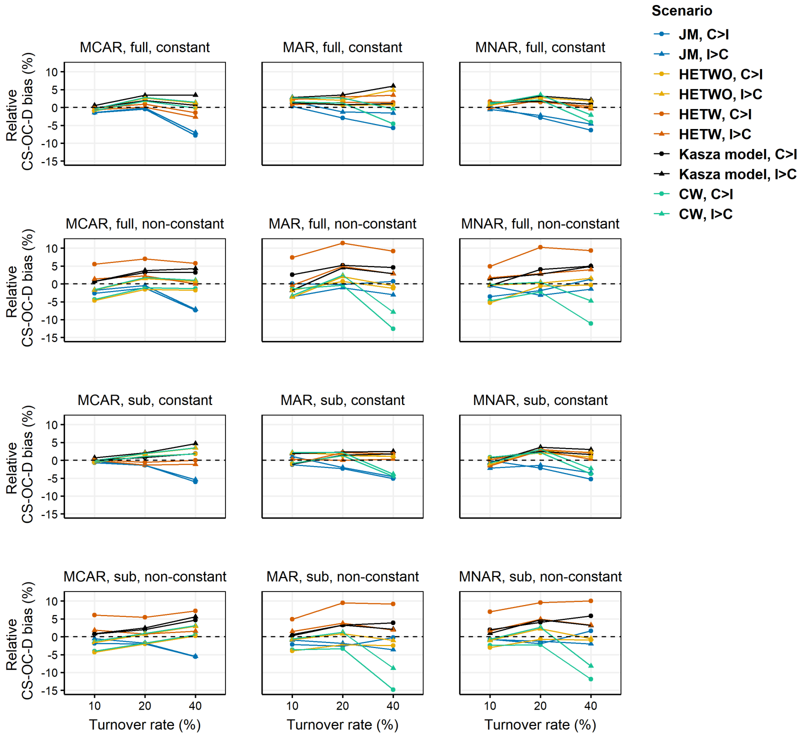


Figure C.6: Relative bias (%) for all models for the CS-OC-D estimand with 1 timescale. 'Full' denotes samples of 15 taken from cluster populations of 15; 'sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

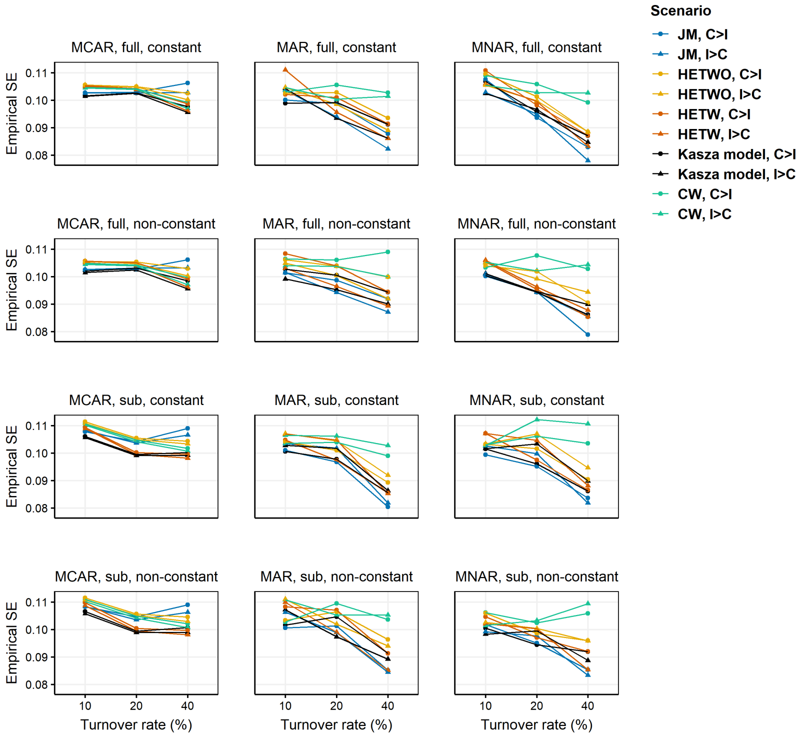


Figure C.7: Empirical SE for all models for the CS-OC-D estimand with 1 timescale. 'Full' denotes samples of 15 taken from cluster populations of 15; 'sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

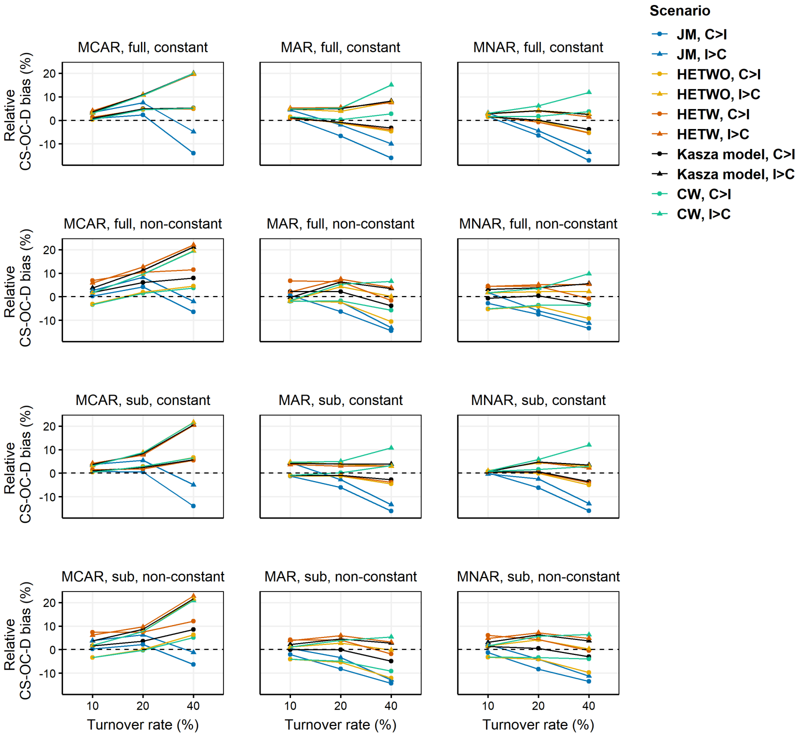


Figure C.8: Relative bias (%) for all models for the CS-OC-D estimand with 2 timescales. 'Full' denotes samples of 15 taken from cluster populations of 15; 'sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

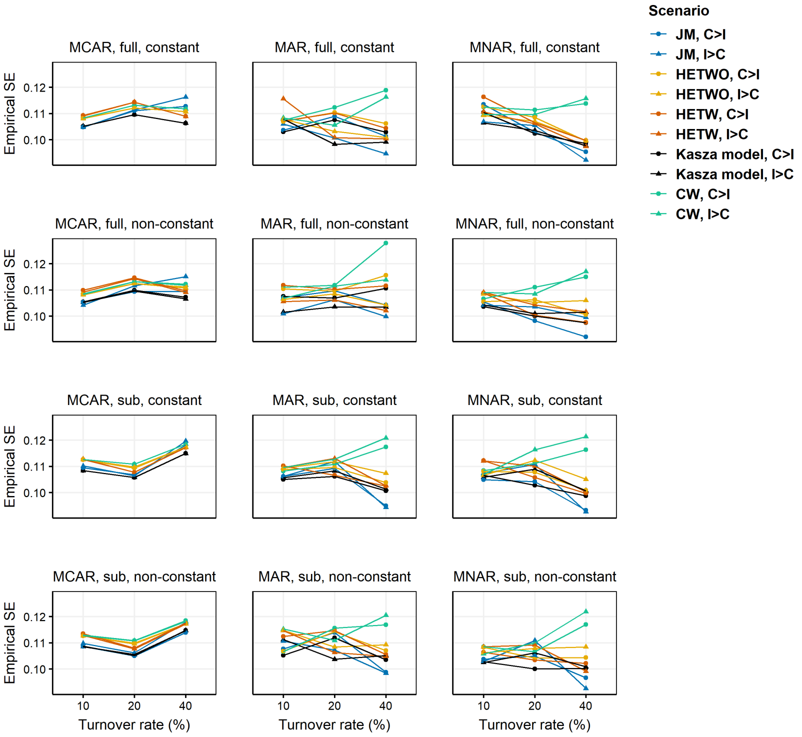


Figure C.9: Empirical SE for all models for the CS-OC-D estimand with 2 timescales. 'Full' denotes samples of 15 taken from cluster populations of 15; 'sub' denotes samples of 15 taken from cluster populations of 50; 'constant' ('non-constant') denote a constant (non-constant) intervention effect rate.

C.2 Stata code and technical details for joint models

The following Stata code uses merlin [238] to fit a joint model. Submodels are specified within parentheses. The first submodel is for the drop-out times, with the response given by evtime. Other effects in the model follow, which here includes the treatment group trt, a fixed effect for cluster, and the expected value association structure is specified using EV[y]. The distribution family is then specified, in this case exponential, and the variable which contains the event indicator, exit. The time variable in this model is then specified using timevar(evtime).

The second model is the longitudinal model starting with the response y. The following two terms are the fixed effects for cr_time and the interaction between cr_time and trt. As I am using EV[y], the association structure is dependent on time due to the fixed effects of time in the longitudinal model. As time needs to be integrated out in the survival model, either fp (fractional polynomials) or rcs (restricted cubic splines) must be used for these time variables to tell Stata to integrate out these terms. The time variable is also specified in each model using timevar, and this need not be the same variable across models. The issue of currently only being able to specify one time variable in merlin is discussed earlier in Section 6.2.4.7. A random effect for cluster is then created using M1[cluster], with the @1 giving this term a coefficient of 1. The family of the longitudinal model is then given as Gaussian and again the time variable involved in this model is given using timevar(cr_time). The difficult and iter(30) terms are explained earlier in Section 6.2.4.7.

```
merlin (evtime trt cluster EV[y], family(exp, failure(exit)) timevar(evtime))
(y fp(cr_time,pow(1)) fp(cr_time,pow(1))#trt M1[cluster]@1, family(gaussian) timevar(cr_time)),
difficult iter(30)
```

When using survival data, merlin requires that in long format datasets where an individual has a row per observation, that the event indicator exit and time evtime have only one entry for each individual.

In initial attempts, before running a joint model I first ran a simple longitudinal model only to get an estimate of the cluster intercept variance, and then used this estimate in the ‘restartvalues’ option of merlin. Sometimes the default starting values for random effect variances (set to 1) lead to problems but in my case, omitting the restartvalues line actually resulted in improved convergence, so this is now not specified in the final code.

There are estimation options that can be amended in Stata for all functions which use maximum likelihood estimation [289], which in this case includes merlin for the joint models and mixed for the heteroscedastic models. The 'difficult' option can be used if the likelihood function contains non-concave regions and is therefore more difficult to maximise, and tells Stata to use a different stepping algorithm. I have included the difficult option because in initial runs there were a lot of non-concave warnings. I also reduced the number of iterations to 30 (from a default of 300) using iter(30). This ensures that runs that are not going to converge are stopped after an initial attempt and do not run for a long time, thus reducing computational time. When convergence is achieved, this usually happens in less than 10 iterations.

C.3 R and Stata code for heteroscedastic model

The R code below creates simple dummy variables pres0, pres26 and pres78 which are equal to 1 if an individual is present at time 0, 26 or 78 respectively, and 0 if they are not. The existing open cohort dataset is given by temp, leavetime is the time an individual leaves the cluster, and entrytime when they enter the cluster. This could be easily adapted for any number of timepoints.

```
temp$pres0[temp$entrytime==0] <- 1
    temp$pres0[temp$entrytime>0] <- 0
    temp$pres26 <- 1
    temp$pres26[temp$entrytime>26 | temp$leavetime<26] <- 0
    temp$pres78 <- 1
    temp$pres78[temp$leavetime<78] <- 0
```

The following Stata code then fits the heteroscedastic (HETW) model using the mixed function. The response is given by y with other fixed effects following this before the double vertical bars. This example includes only the single timescale fixed effects of cr_time and the interaction between cr_time and treatment group, trt. Following the first set of vertical double bars, a random effect for cluster at level 2 is specified which is cross-classified with time. Following the second set of vertical bars there is a random effect for subject at level 1 which has its variance partitioned according to the dummy variables for the three timepoints. Following the comma the estimation method of restricted maximum likelihood is specified using reml. By default, an independent variance-covariance matrix at level 1 is specified such that variances are estimated but covariances are constrained to zero. By specifying nocons the default constant term that is included at level 1 along with the dummy variables is omitted. The final options of difficult and iter(30) have been explained earlier in Section C.2.

```
mixed y cr_time cr_time#trt || cluster: R.cr_time || subject:pres0 pres26 pres78,
reml nocons difficult iter(30)
```

This second version of the heteroscedastic model (HETWO) does not include the cross-classification with time at level 2, by simplying dropping the R.cr_time term:

```
mixed y cr_time cr_time#trt || cluster: || subject:pres0 pres26 pres78,
reml nocons difficult iter(30)
```

C.4 R and Stata code for cluster-weighted model

The R code below calculates the weights for each individual in the dataset based on their length of stay in the cluster. It then assigns these weights to a matrix for all individuals in all clusters and attaches this matrix to the existing dataset in long format. The existing open cohort dataset is given by temp, leavetime is the time an individual leaves the cluster, entrytime when they enter the cluster, and cluster is the cluster ID. The outputted dataset is given by out.

```
temp$weight <- (temp$leavetime-temp$entrytime)/78
n <- max(temp$cluster)
cnames <- c()
for(j in 1:n){
     cnames[j] <- paste("c",j,sep="")
            }

out <- cbind.data.frame(matrix(0, ncol=n, nrow=nrow(temp), dimnames=list(NULL,cnames)), temp)
for(m in 1:nrow(out)){
     c <- out$cluster[m]
     out[m,c] <- test$weight[m]
                   }
```

The following Stata code can be used to fit the CW model with a constant term with the function xtmixed [241]. The response is given by y with other fixed effects following this before the double vertical bars. In this example only the single timescale fixed effects of cr_time and the interaction between cr_time and treatment group, trt, are included. An 'artificial supercluster' is then specified at level 2 using _all and the weights generated above which act as predictor variables are given by c1 to c44 (there are 44 clusters in our example dataset). By specifying covariance(identity), the variance-covariance matrix at level 2 is set as a multiple of the identity matrix, that is, with equal variances and covariances equal to 0. Finally, the random effect at level 1 for `subject` is included after the second set of double bars, and general options at the end for maximum likelihood estimation and outputting variances instead of standard deviations. The code below includes as default a constant term at the same level as the `c1-c44` predictor variables.

```
xtmixed y cr_time cr_time#trt || _all: c1-c44, covariance(identity) || subject:,
mle variance
```

To fit the CW model without the constant term, `nocons` can be added in the level 2 part of the model:

```
xtmixed y cr_time cr_time#trt || _all: c1-c44, nocons covariance(identity) || subject:,
mle variance
```

C.5 Tables

Model	**Turnover (%)**	**Intvn. effects**	**CS-OC-D^1 bias (%)**	**CS-OC-D^1 EmpSE**	**CS-OC-D^1 MCSE (bias)**	**CS-OC-D^1 MCSE (empSE)**	**CS-OC-D^2 bias (%)**	**CS-OC-D^2 EmpSE**	**CS-OC-D^2 MCSE (bias)**	**CS-OC-D^2 MCSE (empSE)**	**OC-52-I bias (%)**	**OC-52-I empSE**	**OC-52-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**Non-convergence^1 (%)**	**Non-convergence^2 (%)**
Kasza	10	C>I	-0.116	0.102	0.004	0.003	0.878	0.105	0.004	0.003	0.215	0.159	0.006	0.004	0.923	0.769
Kasza	10	I>C	0.543	0.101	0.004	0.003	3.557	0.105	0.004	0.003	0.205	0.159	0.006	0.004	0.923	0.615
JM	10	C>I	-1.504	0.103	0.004	0.003	0.732	0.105	0.004	0.003	-2.756	0.156	0.006	0.005	2.615	11.077
JM	10	I>C	-1.417	0.103	0.004	0.003	3.316	0.105	0.004	0.003	-1.942	0.154	0.006	0.005	2	10.462
CW	10	C>I	-0.836	0.105	0.004	0.003	0.298	0.108	0.004	0.003	-0.162	0.153	0.006	0.004	0	0
CW	10	I>C	-0.399	0.104	0.004	0.003	2.999	0.108	0.004	0.003	-0.198	0.153	0.006	0.004	0	0
HETW	10	C>I	-0.161	0.105	0.004	0.003	1.391	0.109	0.004	0.003	1.241	0.159	0.006	0.004	0.308	0.462
HETW	10	I>C	-0.601	0.105	0.004	0.003	4.099	0.109	0.004	0.003	1.391	0.159	0.006	0.004	0.154	0.154
HETWO	10	C>I	-1.002	0.105	0.004	0.003	0.302	0.108	0.004	0.003	-0.253	0.161	0.006	0.004	0.923	0.615
HETWO	10	I>C	-0.459	0.105	0.004	0.003	3.138	0.108	0.004	0.003	-0.473	0.161	0.006	0.004	1.538	0.308
Kasza	20	C>I	1.951	0.103	0.004	0.003	4.838	0.11	0.004	0.003	-1.598	0.11	0.004	0.003	2.462	3.077
Kasza	20	I>C	3.394	0.103	0.004	0.003	10.902	0.11	0.004	0.003	-1.987	0.11	0.004	0.003	2.769	2.923
JM	20	C>I	-0.437	0.103	0.004	0.003	2.3	0.111	0.005	0.003	-8.863	0.116	0.005	0.003	2.923	8
JM	20	I>C	-0.178	0.103	0.004	0.003	7.475	0.111	0.005	0.003	-5.174	0.108	0.004	0.003	2.769	6.615
CW	20	C>I	1.773	0.104	0.004	0.003	4.569	0.113	0.004	0.003	-1.505	0.106	0.004	0.003	0	0
CW	20	I>C	2.759	0.104	0.004	0.003	10.669	0.113	0.004	0.003	-1.839	0.106	0.004	0.003	0	0
HETW	20	C>I	0.91	0.104	0.004	0.003	4.674	0.114	0.004	0.003	-1.672	0.11	0.004	0.003	0.462	0.308
HETW	20	I>C	0.076	0.104	0.004	0.003	10.779	0.114	0.004	0.003	-2.044	0.11	0.004	0.003	0.308	0.308
HETWO	20	C>I	1.823	0.105	0.004	0.003	4.61	0.112	0.004	0.003	-1.344	0.115	0.005	0.003	0.462	0.154
HETWO	20	I>C	2.652	0.105	0.004	0.003	10.711	0.112	0.004	0.003	-1.643	0.115	0.005	0.003	0.615	0.154
Kasza	40	C>I	0.603	0.098	0.004	0.003	5.267	0.106	0.004	0.003	-0.331	0.087	0.004	0.002	5.077	5.385
Kasza	40	I>C	3.412	0.096	0.004	0.003	20.017	0.106	0.004	0.003	-0.427	0.087	0.004	0.002	4.769	5.231
JM	40	C>I	-7.808	0.106	0.004	0.003	-13.98	0.113	0.005	0.003	-33.901	0.102	0.004	0.003	0.769	8.769
JM	40	I>C	-7.048	0.103	0.004	0.003	-4.799	0.116	0.005	0.003	-27.231	0.089	0.004	0.003	1.385	6.769
CW	40	C>I	-0.146	0.099	0.004	0.003	5.153	0.112	0.004	0.003	-1.049	0.086	0.003	0.002	0	0
CW	40	I>C	1.445	0.097	0.004	0.003	19.935	0.112	0.004	0.003	-1.282	0.086	0.003	0.002	0	0
HETW	40	C>I	-1.459	0.099	0.004	0.003	4.843	0.109	0.004	0.003	-1.132	0.086	0.003	0.002	0	0
HETW	40	I>C	-2.698	0.096	0.004	0.003	19.58	0.109	0.004	0.003	-1.384	0.086	0.003	0.002	0	0
HETWO	40	C>I	-0.147	0.103	0.004	0.003	4.96	0.111	0.004	0.003	-0.829	0.092	0.004	0.003	0.154	0
HETWO	40	I>C	1.263	0.1	0.004	0.003	19.715	0.111	0.004	0.003	-1.013	0.092	0.004	0.003	0.308	0

Table C.1: Results for the five models with full-samples, MCAR missing data mechanism and constant intervention effect rate. ^{1} = single timescale model, ^{2} = two timescale model.

Model	**Turnover** (%)	**Intvn. effects**	**CS-OC-D^1 bias (%)**	**CS-OC-D^1 EmpSE**	**CS-OC-D^1 MCSE (bias)**	**CS-OC-D^1 MCSE (empSE)**	**CS-OC-D^2 bias (%)**	**CS-OC-D^2 EmpSE**	**CS-OC-D^2 MCSE (bias)**	**CS-OC-D^2 MCSE (empSE)**	**OC-52-I bias (%)**	**OC-52-I empSE**	**OC-52-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**Non-convergence^1 (%)**	**Non-convergence^2 (%)**
Kasza	10	C>I	0.679	0.102	0.004	0.003	1.778	0.105	0.004	0.003	1.598	0.159	0.006	0.004	0.923	0.923
Kasza	10	I>C	0.733	0.101	0.004	0.003	3.747	0.105	0.004	0.003	0.642	0.159	0.006	0.004	0.923	0.615
JM	10	C>I	-2.627	0.103	0.004	0.003	0.279	0.105	0.004	0.003	-6.816	0.156	0.007	0.005	2.923	11.231
JM	10	I>C	-1.783	0.103	0.004	0.003	2.707	0.104	0.004	0.003	-3.948	0.154	0.006	0.005	2.615	10.615
CW	10	C>I	-4.341	0.105	0.004	0.003	-3.377	0.108	0.004	0.003	-2.134	0.153	0.006	0.004	0	0
CW	10	I>C	-1.593	0.104	0.004	0.003	1.744	0.108	0.004	0.003	-0.995	0.153	0.006	0.004	0	0
HETW	10	C>I	5.491	0.106	0.004	0.003	6.911	0.11	0.004	0.003	8.271	0.159	0.006	0.004	0.923	0.462
HETW	10	I>C	1.362	0.105	0.004	0.003	5.85	0.109	0.004	0.003	4.037	0.159	0.006	0.004	0.462	0.615
HETWO	10	C>I	-4.619	0.105	0.004	0.003	-3.127	0.108	0.004	0.003	-5.506	0.161	0.006	0.004	1.385	0.154
HETWO	10	I>C	-1.72	0.106	0.004	0.003	1.737	0.108	0.004	0.003	-2.435	0.161	0.006	0.004	0.923	0.615
Kasza	20	C>I	3.201	0.103	0.004	0.003	6.084	0.11	0.004	0.003	0.251	0.11	0.004	0.003	2.923	3.077
Kasza	20	I>C	3.756	0.102	0.004	0.003	11.307	0.11	0.004	0.003	-1.379	0.11	0.004	0.003	2.615	3.077
JM	20	C>I	-1.185	0.103	0.004	0.003	4.172	0.109	0.004	0.003	-10.203	0.112	0.005	0.003	3.538	8.308
JM	20	I>C	-0.448	0.103	0.004	0.003	8.278	0.112	0.005	0.003	-5.716	0.107	0.004	0.003	3.231	6.154
CW	20	C>I	-1.063	0.104	0.004	0.003	1.494	0.113	0.004	0.003	-2.898	0.106	0.004	0.003	0	0
CW	20	I>C	1.772	0.104	0.004	0.003	9.591	0.113	0.004	0.003	-2.397	0.106	0.004	0.003	0	0
HETW	20	C>I	6.946	0.105	0.004	0.003	10.417	0.115	0.005	0.003	5.579	0.111	0.004	0.003	0.308	0.308
HETW	20	I>C	2.234	0.104	0.004	0.003	12.718	0.114	0.004	0.003	0.856	0.11	0.004	0.003	0.308	0.308
HETWO	20	C>I	-1.528	0.105	0.004	0.003	1.89	0.112	0.004	0.003	-5.651	0.115	0.005	0.003	0.462	0.462
HETWO	20	I>C	1.491	0.105	0.004	0.003	9.645	0.112	0.004	0.003	-3.619	0.115	0.005	0.003	0	0.462
Kasza	40	C>I	3.152	0.098	0.004	0.003	7.94	0.107	0.004	0.003	2.389	0.088	0.004	0.003	6.308	7.846
Kasza	40	I>C	4.298	0.096	0.004	0.003	21.107	0.107	0.004	0.003	0.702	0.087	0.004	0.002	4.923	5.692
JM	40	C>I	-7.382	0.106	0.004	0.003	-6.397	0.109	0.004	0.003	-29.733	0.093	0.004	0.003	1.692	8
JM	40	I>C	-7.108	0.103	0.004	0.003	-1.985	0.115	0.005	0.003	-25.436	0.087	0.004	0.003	1.077	7.077
CW	40	C>I	-1.304	0.1	0.004	0.003	3.73	0.112	0.004	0.003	-0.995	0.086	0.003	0.002	0	0
CW	40	I>C	1.049	0.097	0.004	0.003	19.397	0.112	0.004	0.003	-1.254	0.086	0.003	0.002	0	0
HETW	40	C>I	5.749	0.099	0.004	0.003	11.445	0.11	0.004	0.003	6.325	0.087	0.003	0.002	0	0.154
HETW	40	I>C	-0.012	0.096	0.004	0.003	22.028	0.109	0.004	0.003	1.594	0.087	0.003	0.002	0	0
HETWO	40	C>I	-1.785	0.103	0.004	0.003	4.514	0.111	0.004	0.003	-2.848	0.092	0.004	0.003	0	0.154
HETWO	40	I>C	0.708	0.1	0.004	0.003	19.541	0.111	0.004	0.003	-1.838	0.092	0.004	0.003	0.154	0

Table C.2: Results for the five models with full-samples, MCAR missing data mechanism and non-constant intervention effect rate. ^{1} = single timescale model, ^{2} = two timescale model.

Model	**Turnover** (%)	**Intvn. effects**	**CS-OC-D[1] bias (%)**	**CS-OC-D[1] EmpSE**	**CS-OC-D[1] MCSE (bias)**	**CS-OC-D[1] MCSE (empSE)**	**CS-OC-D[2] bias (%)**	**CS-OC-D[2] EmpSE**	**CS-OC-D[2] MCSE (bias)**	**CS-OC-D[2] MCSE (empSE)**	**OC-52-I bias (%)**	**OC-52-I empSE**	**OC-52-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**Non-con-vergence[1] (%)**	**Non-con-vergence[2] (%)**
Kasza	10	C>I	0.059	0.106	0.004	0.003	0.963	0.108	0.004	0.003	0.266	0.156	0.006	0.004	0.923	1.077
Kasza	10	I>C	0.724	0.106	0.004	0.003	3.681	0.108	0.004	0.003	0.325	0.156	0.006	0.004	0.923	1.077
JM	10	C>I	-0.659	0.108	0.004	0.003	0.752	0.109	0.004	0.003	-3.147	0.156	0.006	0.005	2.462	8.615
JM	10	I>C	-0.342	0.108	0.004	0.003	3.76	0.11	0.005	0.003	-1.093	0.153	0.006	0.004	2.308	8.615
CW	10	C>I	-0.482	0.111	0.004	0.003	0.417	0.113	0.004	0.003	0.366	0.152	0.006	0.004	0	0
CW	10	I>C	-0.036	0.11	0.004	0.003	3.121	0.113	0.004	0.003	0.447	0.152	0.006	0.004	0	0
HETW	10	C>I	0.174	0.109	0.004	0.003	1.473	0.113	0.004	0.003	0.855	0.152	0.006	0.004	0.462	0.769
HETW	10	I>C	-0.029	0.109	0.004	0.003	4.188	0.113	0.004	0.003	1.057	0.153	0.006	0.004	1.077	0.923
HETWO	10	C>I	-0.636	0.111	0.004	0.003	0.392	0.112	0.004	0.003	-0.376	0.159	0.006	0.004	0.923	0.615
HETWO	10	I>C	-0.258	0.111	0.004	0.003	3.132	0.113	0.004	0.003	-0.316	0.159	0.006	0.004	0.769	0.769
Kasza	20	C>I	0.794	0.099	0.004	0.003	2.36	0.106	0.004	0.003	-0.214	0.115	0.005	0.003	2	2.154
Kasza	20	I>C	2.113	0.099	0.004	0.003	8.311	0.106	0.004	0.003	-0.254	0.116	0.005	0.003	2.154	2.308
JM	20	C>I	-1.37	0.104	0.004	0.003	0.585	0.107	0.004	0.003	-7.919	0.122	0.005	0.004	1.231	9.077
JM	20	I>C	-1.367	0.104	0.004	0.003	5.443	0.106	0.004	0.003	-3.216	0.115	0.005	0.003	1.231	7.538
CW	20	C>I	1.095	0.105	0.004	0.003	2.837	0.111	0.004	0.003	-0.113	0.113	0.004	0.003	0	0
CW	20	I>C	1.975	0.104	0.004	0.003	8.835	0.111	0.004	0.003	-0.138	0.113	0.004	0.003	0	0
HETW	20	C>I	-0.554	0.1	0.004	0.003	1.801	0.108	0.004	0.003	-0.963	0.11	0.004	0.003	0.154	0
HETW	20	I>C	-1.341	0.099	0.004	0.003	7.739	0.108	0.004	0.003	-1.177	0.11	0.004	0.003	0.462	0
HETWO	20	C>I	1.191	0.105	0.004	0.003	2.732	0.109	0.004	0.003	0.018	0.122	0.005	0.003	0.769	0.462
HETWO	20	I>C	1.896	0.105	0.004	0.003	8.762	0.11	0.004	0.003	-0.017	0.122	0.005	0.003	0.308	0.154
Kasza	40	C>I	1.902	0.1	0.004	0.003	5.703	0.115	0.005	0.003	-0.099	0.091	0.004	0.003	4.154	5.385
Kasza	40	I>C	4.654	0.099	0.004	0.003	20.562	0.115	0.005	0.003	-0.121	0.091	0.004	0.003	4.154	5.385
JM	40	C>I	-6.045	0.109	0.004	0.003	-13.989	0.118	0.005	0.003	-34.092	0.102	0.004	0.003	0.769	7.231
JM	40	I>C	-5.436	0.107	0.004	0.003	-4.916	0.12	0.005	0.003	-26.965	0.09	0.004	0.003	0.923	8.154
CW	40	C>I	1.794	0.102	0.004	0.003	6.56	0.118	0.005	0.003	0.322	0.092	0.004	0.003	0	0
CW	40	I>C	3.525	0.101	0.004	0.003	21.539	0.118	0.005	0.003	0.393	0.092	0.004	0.003	0	0
HETW	40	C>I	-0.016	0.1	0.004	0.003	5.483	0.117	0.005	0.003	-0.114	0.09	0.004	0.002	0	0
HETW	40	I>C	-1.121	0.098	0.004	0.003	20.311	0.117	0.005	0.003	-0.139	0.09	0.004	0.002	0	0
HETWO	40	C>I	1.818	0.104	0.004	0.003	6.607	0.117	0.005	0.003	0.512	0.097	0.004	0.003	0	0
HETWO	40	I>C	3.453	0.103	0.004	0.003	21.673	0.117	0.005	0.003	0.582	0.097	0.004	0.003	0.154	0.308

Table C.3: Results for the five models with sub-samples, MCAR missing data mechanism and constant intervention effect rate. [1] = single timescale model, [2] = two timescale model.

Model	**Turnover (%)**	**Intvn. effects**	**CS-OC-D^1 bias (%)**	**CS-OC-D^1 EmpSE**	**CS-OC-D^1 MCSE (bias)**	**CS-OC-D^1 MCSE (empSE)**	**CS-OC-D^2 bias (%)**	**CS-OC-D^2 EmpSE**	**CS-OC-D^2 MCSE (bias)**	**CS-OC-D^2 MCSE (empSE)**	**OC-52-I bias (%)**	**OC-52-I empSE**	**OC-52-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**Non-convergence^1 (%)**	**Non-convergence^2 (%)**
Kasza	10	C>I	0.774	0.106	0.004	0.003	1.614	0.109	0.004	0.003	1.346	0.157	0.006	0.004	0.923	1.231
Kasza	10	I>C	0.886	0.106	0.004	0.003	3.765	0.108	0.004	0.003	0.546	0.156	0.006	0.004	0.923	1.385
JM	10	C>I	-1.873	0.108	0.004	0.003	0.34	0.109	0.004	0.003	-6.622	0.157	0.006	0.005	3.692	9.385
JM	10	I>C	-0.524	0.108	0.004	0.003	3.86	0.11	0.004	0.003	-2.02	0.153	0.006	0.004	3.231	8.154
CW	10	C>I	-4.047	0.111	0.004	0.003	-3.328	0.113	0.004	0.003	-1.668	0.152	0.006	0.004	0	0
CW	10	I>C	-1.251	0.11	0.004	0.003	1.84	0.113	0.004	0.003	-0.378	0.152	0.006	0.004	0	0
HETW	10	C>I	6.063	0.11	0.004	0.003	7.317	0.113	0.004	0.003	8.339	0.153	0.006	0.004	0.308	0.923
HETW	10	I>C	1.841	0.109	0.004	0.003	6.164	0.113	0.004	0.003	4.095	0.153	0.006	0.004	0.769	0.462
HETWO	10	C>I	-4.304	0.111	0.004	0.003	-3.305	0.113	0.004	0.003	-5.861	0.16	0.006	0.004	0.615	0.923
HETWO	10	I>C	-1.627	0.111	0.004	0.003	1.785	0.113	0.004	0.003	-2.725	0.16	0.006	0.004	0.923	0.923
Kasza	20	C>I	1.994	0.099	0.004	0.003	3.641	0.105	0.004	0.003	1.34	0.115	0.005	0.003	2.308	2.615
Kasza	20	I>C	2.466	0.099	0.004	0.003	8.701	0.106	0.004	0.003	0.327	0.115	0.005	0.003	2	2.154
JM	20	C>I	-2.027	0.105	0.004	0.003	2.218	0.105	0.004	0.003	-9.125	0.118	0.005	0.003	2	9.231
JM	20	I>C	-1.742	0.104	0.004	0.003	6.264	0.106	0.004	0.003	-3.789	0.115	0.005	0.003	1.385	6.462
CW	20	C>I	-1.797	0.105	0.004	0.003	-0.28	0.111	0.004	0.003	-1.54	0.113	0.004	0.003	0	0
CW	20	I>C	0.964	0.104	0.004	0.003	7.734	0.111	0.004	0.003	-0.718	0.113	0.004	0.003	0	0
HETW	20	C>I	5.462	0.1	0.004	0.003	7.494	0.108	0.004	0.003	6.203	0.111	0.004	0.003	0.462	0.154
HETW	20	I>C	0.748	0.099	0.004	0.003	9.692	0.108	0.004	0.003	1.66	0.11	0.004	0.003	0.308	0
HETWO	20	C>I	-1.982	0.106	0.004	0.003	0.084	0.11	0.004	0.003	-4.223	0.123	0.005	0.003	0.308	0.462
HETWO	20	I>C	0.714	0.105	0.004	0.003	7.746	0.11	0.004	0.003	-1.923	0.122	0.005	0.003	0.308	0.308
Kasza	40	C>I	4.69	0.101	0.004	0.003	8.496	0.115	0.005	0.003	2.914	0.09	0.004	0.003	5.231	6.462
Kasza	40	I>C	5.603	0.099	0.004	0.003	21.574	0.115	0.005	0.003	1.067	0.091	0.004	0.003	3.846	5.385
JM	40	C>I	-5.616	0.109	0.004	0.003	-6.298	0.114	0.005	0.003	-29.546	0.094	0.004	0.003	1.231	8.154
JM	40	I>C	-5.485	0.106	0.004	0.003	-1.171	0.118	0.005	0.003	-24.574	0.088	0.004	0.003	0.923	9.385
CW	40	C>I	0.575	0.102	0.004	0.003	5.143	0.118	0.005	0.003	0.379	0.092	0.004	0.003	0	0
CW	40	I>C	3.101	0.101	0.004	0.003	20.891	0.118	0.005	0.003	0.377	0.092	0.004	0.003	0	0.154
HETW	40	C>I	7.189	0.1	0.004	0.003	12.068	0.117	0.005	0.003	7.334	0.09	0.004	0.002	0	0
HETW	40	I>C	1.57	0.098	0.004	0.003	22.764	0.117	0.005	0.003	2.842	0.09	0.004	0.002	0	0
HETWO	40	C>I	0.2	0.105	0.004	0.003	6.248	0.117	0.005	0.003	-1.443	0.097	0.004	0.003	0.154	0
HETWO	40	I>C	2.913	0.103	0.004	0.003	21.482	0.117	0.005	0.003	-0.156	0.097	0.004	0.003	0.154	0.154

Table C.4: Results for the five models with sub-samples, MCAR missing data mechanism and non-constant intervention effect rate. ^{1} = single timescale model, ^{2} = two timescale model.

Model	Turnover (%)	Intvn. effects	CS-OC-D[1] bias (%)	CS-OC-D[1] EmpSE	CS-OC-D[1] MCSE (bias)	CS-OC-D[1] MCSE (empSE)	CS-OC-D[2] bias (%)	CS-OC-D[2] EmpSE	CS-OC-D[2] MCSE (bias)	CS-OC-D[2] MCSE (empSE)	OC-52-I bias (%)	OC-52-I empSE	OC-52-I MCSE (bias)	OC-52-I MCSE (empSE)	Non-convergence[1] (%)	Non-convergence[2] (%)
Kasza	10	C>I	1.216	0.099	0.004	0.003	1.156	0.103	0.004	0.003	1.032	0.16	0.006	0.004	1.231	1.231
Kasza	10	I>C	2.832	0.104	0.004	0.003	4.677	0.108	0.004	0.003	3.142	0.163	0.006	0.005	0.769	1.385
JM	10	C>I	0.288	0.1	0.004	0.003	1.264	0.104	0.004	0.003	-2.6	0.158	0.007	0.005	2.615	13.692
JM	10	I>C	1.598	0.104	0.004	0.003	4.423	0.106	0.004	0.003	4.57	0.16	0.007	0.005	3.692	12.462
CW	10	C>I	1.585	0.103	0.004	0.003	1.499	0.107	0.004	0.003	1.411	0.156	0.006	0.004	0	0
CW	10	I>C	2.714	0.105	0.004	0.003	4.776	0.108	0.004	0.003	2.521	0.154	0.006	0.004	0	0
HETW	10	C>I	0.77	0.102	0.004	0.003	0.771	0.107	0.004	0.003	0.713	0.157	0.006	0.004	0.462	0.769
HETW	10	I>C	2.222	0.111	0.004	0.003	5.22	0.116	0.005	0.003	3.571	0.159	0.006	0.004	0.615	0.462
HETWO	10	C>I	1.568	0.102	0.004	0.003	1.422	0.107	0.004	0.003	1.874	0.165	0.007	0.005	1.231	0.769
HETWO	10	I>C	2.505	0.105	0.004	0.003	4.755	0.108	0.004	0.003	2.194	0.164	0.006	0.005	0.462	0.308
Kasza	20	C>I	0.786	0.099	0.004	0.003	-0.926	0.108	0.004	0.003	-3.989	0.127	0.005	0.004	1.538	1.538
Kasza	20	I>C	3.525	0.093	0.004	0.003	5.015	0.098	0.004	0.003	0.324	0.129	0.005	0.004	2.308	2.615
JM	20	C>I	-2.946	0.099	0.004	0.003	-6.704	0.109	0.005	0.003	-12.601	0.115	0.005	0.004	0.308	20
JM	20	I>C	-1.275	0.094	0.004	0.003	-1.96	0.101	0.004	0.003	-0.372	0.108	0.005	0.003	0.308	16.923
CW	20	C>I	1.234	0.106	0.004	0.003	0.24	0.112	0.004	0.003	-3.235	0.126	0.005	0.004	0	0
CW	20	I>C	2.659	0.1	0.004	0.003	5.264	0.106	0.004	0.003	0.628	0.127	0.005	0.004	0	0
HETW	20	C>I	1.377	0.101	0.004	0.003	-0.907	0.11	0.004	0.003	-4.427	0.126	0.005	0.004	0.462	0.615
HETW	20	I>C	2.924	0.096	0.004	0.003	5.461	0.101	0.004	0.003	0.811	0.126	0.005	0.004	0.462	0.769
HETWO	20	C>I	0.619	0.103	0.004	0.003	-1.243	0.11	0.004	0.003	-3.728	0.132	0.005	0.004	0.154	0.462
HETWO	20	I>C	2.051	0.098	0.004	0.003	3.887	0.103	0.004	0.003	-0.116	0.135	0.005	0.004	0.154	0.923
Kasza	40	C>I	1.069	0.091	0.004	0.003	-3.25	0.103	0.004	0.003	-10.562	0.104	0.004	0.003	6.923	8.615
Kasza	40	I>C	5.975	0.086	0.003	0.002	8.204	0.099	0.004	0.003	-4.912	0.098	0.004	0.003	5.077	6.615
JM	40	C>I	-5.733	0.088	0.004	0.003	-15.991	0.101	0.004	0.003	-31.893	0.074	0.003	0.002	5.846	3.231
JM	40	I>C	-1.562	0.082	0.003	0.002	-9.985	0.095	0.004	0.003	-18.804	0.071	0.003	0.002	6.154	3.385
CW	40	C>I	-4.575	0.103	0.004	0.003	2.716	0.119	0.005	0.003	-7.193	0.11	0.004	0.003	0	0
CW	40	I>C	-0.394	0.101	0.004	0.003	15.02	0.116	0.005	0.003	-1.579	0.109	0.004	0.003	0	0
HETW	40	C>I	1.404	0.091	0.004	0.003	-4.073	0.104	0.004	0.003	-10.349	0.105	0.004	0.003	0.308	0
HETW	40	I>C	3.413	0.086	0.003	0.002	7.511	0.1	0.004	0.003	-5.756	0.099	0.004	0.003	0.769	0.308
HETWO	40	C>I	0.758	0.094	0.004	0.003	-4.596	0.106	0.004	0.003	-9.417	0.107	0.004	0.003	0.462	0.769
HETWO	40	I>C	4.862	0.089	0.003	0.002	7.835	0.101	0.004	0.003	-4.273	0.105	0.004	0.003	0.154	0.923

Table C.5: Results for the five models with full-samples, MAR missing data mechanism and constant intervention effect rate. [1] = single timescale model, [2] = two timescale model.

Model	**Turnover (%)**	**Intvn. effects**	**CS-OC-D^1 bias (%)**	**CS-OC-D^1 EmpSE**	**CS-OC-D^1 MCSE (bias)**	**CS-OC-D^1 MCSE (empSE)**	**CS-OC-D^2 bias (%)**	**CS-OC-D^2 EmpSE**	**CS-OC-D^2 MCSE (bias)**	**CS-OC-D^2 MCSE (empSE)**	**OC-52-I bias (%)**	**OC-52-I empSE**	**OC-52-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**Non-convergence^1 (%)**	**Non-convergence^2 (%)**
Kasza	10	C>I	2.588	0.103	0.004	0.003	2.191	0.108	0.004	0.003	2.73	0.167	0.007	0.005	2.154	2.154
Kasza	10	I>C	-1.797	0.099	0.004	0.003	-0.266	0.102	0.004	0.003	-1.948	0.168	0.007	0.005	0.769	1.077
JM	10	C>I	0.117	0.101	0.004	0.003	0.663	0.107	0.005	0.003	-5.062	0.163	0.007	0.005	3.231	17.692
JM	10	I>C	-3.559	0.101	0.004	0.003	-1.872	0.101	0.004	0.003	-4.509	0.16	0.007	0.005	2.615	14
CW	10	C>I	-1.408	0.106	0.004	0.003	-1.905	0.111	0.004	0.003	0.117	0.162	0.006	0.004	0	0
CW	10	I>C	-3.278	0.104	0.004	0.003	-1.563	0.107	0.004	0.003	-2.637	0.163	0.006	0.005	0	0
HETW	10	C>I	7.342	0.108	0.004	0.003	6.717	0.112	0.004	0.003	8.457	0.169	0.007	0.005	0.462	0.462
HETW	10	I>C	-0.359	0.103	0.004	0.003	2.086	0.106	0.004	0.003	1.216	0.165	0.007	0.005	0.462	1.385
HETWO	10	C>I	-1.774	0.106	0.004	0.003	-1.766	0.11	0.004	0.003	-2.808	0.167	0.007	0.005	1.231	1.077
HETWO	10	I>C	-3.68	0.105	0.004	0.003	-1.75	0.106	0.004	0.003	-4.883	0.171	0.007	0.005	0.154	0.462
Kasza	20	C>I	5.234	0.1	0.004	0.003	2.186	0.107	0.004	0.003	-1.032	0.13	0.005	0.004	2	2.308
Kasza	20	I>C	4.533	0.095	0.004	0.003	6.276	0.104	0.004	0.003	-2.243	0.133	0.005	0.004	2.154	2.462
JM	20	C>I	-0.114	0.099	0.004	0.003	-6.247	0.11	0.005	0.003	-13.15	0.11	0.005	0.003	0.615	16.923
JM	20	I>C	-1.047	0.094	0.004	0.003	-2.223	0.106	0.005	0.003	-2.323	0.114	0.005	0.003	1.077	17.538
CW	20	C>I	-0.356	0.106	0.004	0.003	-1.725	0.112	0.004	0.003	-4.506	0.128	0.005	0.004	0	0
CW	20	I>C	2.382	0.104	0.004	0.003	5.412	0.112	0.004	0.003	-2.938	0.132	0.005	0.004	0	0
HETW	20	C>I	11.381	0.104	0.004	0.003	6.396	0.11	0.004	0.003	3.707	0.127	0.005	0.004	0.308	0.615
HETW	20	I>C	4.928	0.096	0.004	0.003	7.509	0.106	0.004	0.003	-0.486	0.131	0.005	0.004	0.154	0
HETWO	20	C>I	0.677	0.104	0.004	0.003	-2.255	0.11	0.004	0.003	-6.665	0.134	0.005	0.004	0.923	0.462
HETWO	20	I>C	2.047	0.1	0.004	0.003	4.327	0.108	0.004	0.003	-4.082	0.139	0.005	0.004	0.308	0.154
Kasza	40	C>I	4.546	0.094	0.004	0.003	-3.865	0.111	0.005	0.003	-12.226	0.103	0.004	0.003	6.615	8.154
Kasza	40	I>C	2.9	0.09	0.004	0.003	3.522	0.103	0.004	0.003	-6.227	0.103	0.004	0.003	6.923	7.077
JM	40	C>I	0.708	0.092	0.004	0.003	-14.382	0.104	0.004	0.003	-31.199	0.075	0.003	0.002	5.538	6.308
JM	40	I>C	-3.046	0.087	0.004	0.002	-13.161	0.1	0.004	0.003	-20.687	0.071	0.003	0.002	6.154	3.385
CW	40	C>I	-12.583	0.109	0.004	0.003	-5.768	0.128	0.005	0.004	-15.708	0.11	0.004	0.003	0	0
CW	40	I>C	-7.878	0.1	0.004	0.003	6.526	0.114	0.004	0.003	-5.619	0.108	0.004	0.003	0	0
HETW	40	C>I	9.159	0.094	0.004	0.003	-1.631	0.112	0.004	0.003	-8.78	0.101	0.004	0.003	0.154	0.462
HETW	40	I>C	2.831	0.089	0.004	0.002	3.917	0.102	0.004	0.003	-5.095	0.103	0.004	0.003	0.462	0.462
HETWO	40	C>I	-1.237	0.1	0.004	0.003	-10.592	0.116	0.005	0.003	-17.105	0.108	0.004	0.003	0.615	0.769
HETWO	40	I>C	-0.571	0.092	0.004	0.003	0.135	0.104	0.004	0.003	-8.56	0.107	0.004	0.003	0.462	0

Table C.6: Results for the five models with full-samples, MAR missing data mechanism and non-constant intervention effect rate. ^{1} = single timescale model, ^{2} = two timescale model.

Model	**Turnover (%)**	**Intvn. effects**	**CS-OC-D[1] bias (%)**	**CS-OC-D[1] EmpSE**	**CS-OC-D[1] MCSE (bias)**	**CS-OC-D[1] MCSE (empSE)**	**CS-OC-D[2] bias (%)**	**CS-OC-D[2] EmpSE**	**CS-OC-D[2] MCSE (bias)**	**CS-OC-D[2] MCSE (empSE)**	**OC-52-I bias (%)**	**OC-52-I empSE**	**OC-52-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**Non-convergence[1] (%)**	**Non-convergence[2] (%)**
Kasza	10	C>I	-1.05	0.101	0.004	0.003	-1.027	0.105	0.004	0.003	-1.862	0.153	0.006	0.004	1.538	2.154
Kasza	10	I>C	1.884	0.103	0.004	0.003	4.169	0.106	0.004	0.003	-1.573	0.165	0.007	0.005	1.077	1.231
JM	10	C>I	-1.259	0.101	0.004	0.003	-1.223	0.106	0.004	0.003	-4.614	0.149	0.006	0.004	1.692	13.846
JM	10	I>C	1.056	0.103	0.004	0.003	4.59	0.106	0.005	0.003	0.714	0.165	0.007	0.005	2.154	15.231
CW	10	C>I	-0.819	0.103	0.004	0.003	-0.923	0.108	0.004	0.003	-1.615	0.149	0.006	0.004	0	0
CW	10	I>C	2.228	0.106	0.004	0.003	4.687	0.11	0.004	0.003	-1.635	0.163	0.006	0.005	0	0
HETW	10	C>I	-1.071	0.105	0.004	0.003	-1.016	0.11	0.004	0.003	-1.847	0.152	0.006	0.004	0.923	0.615
HETW	10	I>C	0.315	0.107	0.004	0.003	3.673	0.11	0.004	0.003	-2.212	0.164	0.006	0.005	0.308	0.308
HETWO	10	C>I	-0.969	0.104	0.004	0.003	-1.073	0.109	0.004	0.003	-2.282	0.155	0.006	0.004	0.769	0.769
HETWO	10	I>C	2.08	0.107	0.004	0.003	4.706	0.109	0.004	0.003	-0.729	0.173	0.007	0.005	0.308	0.462
Kasza	20	C>I	1.451	0.098	0.004	0.003	-0.911	0.106	0.004	0.003	-2.966	0.127	0.005	0.004	2.308	3.077
Kasza	20	I>C	2.296	0.102	0.004	0.003	3.964	0.108	0.004	0.003	-2.055	0.127	0.005	0.004	1.538	2.154
JM	20	C>I	-2.328	0.097	0.004	0.003	-6.086	0.109	0.005	0.003	-12.636	0.115	0.005	0.004	0.615	17.692
JM	20	I>C	-1.996	0.102	0.004	0.003	-2.732	0.112	0.005	0.003	-3.091	0.114	0.005	0.004	0.923	18.615
CW	20	C>I	1.334	0.104	0.004	0.003	0.144	0.111	0.004	0.003	-2.936	0.126	0.005	0.003	0	0
CW	20	I>C	2.156	0.106	0.004	0.003	5.002	0.113	0.004	0.003	-1.817	0.128	0.005	0.004	0	0
HETW	20	C>I	2.141	0.097	0.004	0.003	-0.945	0.107	0.004	0.003	-3.002	0.124	0.005	0.003	0.154	0
HETW	20	I>C	0.019	0.105	0.004	0.003	3.054	0.113	0.004	0.003	-3.382	0.128	0.005	0.004	0.615	0.462
HETWO	20	C>I	1.306	0.101	0.004	0.003	-1.202	0.11	0.004	0.003	-2.924	0.133	0.005	0.004	0.615	0.308
HETWO	20	I>C	1.594	0.104	0.004	0.003	3.811	0.112	0.004	0.003	-1.585	0.133	0.005	0.004	0.462	0.308
Kasza	40	C>I	1.781	0.086	0.003	0.002	-2.79	0.101	0.004	0.003	-11.382	0.093	0.004	0.003	6	8.154
Kasza	40	I>C	2.482	0.086	0.004	0.002	3.876	0.101	0.004	0.003	-6.629	0.096	0.004	0.003	8	9.692
JM	40	C>I	-5.127	0.08	0.003	0.002	-16.044	0.095	0.004	0.003	-31.773	0.068	0.003	0.002	5.385	4
JM	40	I>C	-4.605	0.082	0.003	0.002	-13.356	0.094	0.004	0.003	-21.016	0.068	0.003	0.002	5.692	4.308
CW	40	C>I	-4.5	0.099	0.004	0.003	3.262	0.117	0.005	0.003	-8.218	0.104	0.004	0.003	0	0
CW	40	I>C	-3.86	0.103	0.004	0.003	10.732	0.121	0.005	0.003	-2.613	0.107	0.004	0.003	0	0
HETW	40	C>I	1.766	0.085	0.003	0.002	-3.857	0.102	0.004	0.003	-11.403	0.095	0.004	0.003	0.462	0.154
HETW	40	I>C	0.323	0.085	0.003	0.002	2.993	0.103	0.004	0.003	-6.959	0.097	0.004	0.003	0.462	0.615
HETWO	40	C>I	1.06	0.089	0.004	0.002	-4.493	0.104	0.004	0.003	-10.574	0.1	0.004	0.003	0.615	0.615
HETWO	40	I>C	1.084	0.092	0.004	0.003	2.976	0.107	0.004	0.003	-5.968	0.104	0.004	0.003	0.615	0.308

Table C.7: Results for the five models with sub-samples, MAR missing data mechanism and constant intervention effect rate. [1] = single timescale model, [2] = two timescale model.

Model	**Turnover** (%)	**Intvn. effects**	**CS-OC-D^1 bias (%)**	**CS-OC-D^1 EmpSE**	**CS-OC-D^1 MCSE (bias)**	**CS-OC-D^1 MCSE (empSE)**	**CS-OC-D^2 bias (%)**	**CS-OC-D^2 EmpSE**	**CS-OC-D^2 MCSE (bias)**	**CS-OC-D^2 MCSE (empSE)**	**OC-52-I bias (%)**	**OC-52-I empSE**	**OC-52-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**Non-convergence^1 (%)**	**Non-convergence^2 (%)**
Kasza	10	C>I	0.289	0.102	0.004	0.003	-0.063	0.105	0.004	0.003	0.806	0.163	0.006	0.005	1.538	1.692
Kasza	10	I>C	0.66	0.107	0.004	0.003	2.199	0.111	0.004	0.003	2.569	0.161	0.006	0.004	1.077	1.231
JM	10	C>I	-2.206	0.101	0.004	0.003	-2.105	0.108	0.005	0.003	-7.216	0.163	0.007	0.005	3.231	14
JM	10	I>C	-0.945	0.106	0.004	0.003	0.587	0.111	0.005	0.003	-0.547	0.159	0.007	0.005	2.923	13.077
CW	10	C>I	-3.674	0.103	0.004	0.003	-4.087	0.107	0.004	0.003	-1.87	0.159	0.006	0.004	0	0
CW	10	I>C	-0.59	0.111	0.004	0.003	1.162	0.115	0.005	0.003	1.429	0.156	0.006	0.004	0	0
HETW	10	C>I	4.875	0.108	0.004	0.003	4.217	0.112	0.004	0.003	6.398	0.161	0.006	0.004	0.308	0.308
HETW	10	I>C	1.497	0.11	0.004	0.003	3.849	0.115	0.005	0.003	4.893	0.158	0.006	0.004	0.462	0.308
HETWO	10	C>I	-3.958	0.103	0.004	0.003	-4.03	0.107	0.004	0.003	-5.177	0.168	0.007	0.005	0.769	1.077
HETWO	10	I>C	-0.852	0.111	0.004	0.003	1.074	0.115	0.005	0.003	-0.07	0.164	0.006	0.005	0.462	0.769
Kasza	20	C>I	3.243	0.105	0.004	0.003	-0.06	0.112	0.004	0.003	-3.397	0.134	0.005	0.004	2.769	3.385
Kasza	20	I>C	3.26	0.097	0.004	0.003	4.57	0.104	0.004	0.003	-2.599	0.13	0.005	0.004	2.154	2.308
JM	20	C>I	-2.622	0.101	0.004	0.003	-8.233	0.114	0.005	0.003	-15.41	0.118	0.005	0.004	0.308	17.538
JM	20	I>C	-1.829	0.099	0.004	0.003	-3.407	0.107	0.005	0.003	-3.423	0.108	0.005	0.003	0.615	18.462
CW	20	C>I	-3.305	0.109	0.004	0.003	-4.835	0.116	0.005	0.003	-7.54	0.131	0.005	0.004	0	0
CW	20	I>C	1.249	0.105	0.004	0.003	3.981	0.111	0.004	0.003	-3.309	0.129	0.005	0.004	0	0
HETW	20	C>I	9.427	0.107	0.004	0.003	4.013	0.115	0.004	0.003	1.657	0.134	0.005	0.004	0.462	0.154
HETW	20	I>C	3.774	0.099	0.004	0.003	5.981	0.106	0.004	0.003	-0.688	0.127	0.005	0.004	0.462	0.769
HETWO	20	C>I	-2.237	0.106	0.004	0.003	-5.353	0.114	0.004	0.003	-9.633	0.136	0.005	0.004	0.308	0.615
HETWO	20	I>C	0.839	0.102	0.004	0.003	2.729	0.108	0.004	0.003	-4.466	0.136	0.005	0.004	0.462	0.923
Kasza	40	C>I	3.874	0.091	0.004	0.003	-4.837	0.104	0.004	0.003	-13.262	0.099	0.004	0.003	6.308	9.231
Kasza	40	I>C	2.089	0.089	0.004	0.003	2.785	0.105	0.004	0.003	-8.893	0.103	0.004	0.003	7.077	8.615
JM	40	C>I	-0.263	0.085	0.003	0.002	-14.309	0.099	0.004	0.003	-31.528	0.07	0.003	0.002	6.154	5.846
JM	40	I>C	-3.614	0.084	0.003	0.002	-12.921	0.098	0.004	0.003	-21.83	0.071	0.003	0.002	5.231	2.923
CW	40	C>I	-14.839	0.104	0.004	0.003	-9.085	0.117	0.005	0.003	-18.049	0.107	0.004	0.003	0	0
CW	40	I>C	-8.775	0.105	0.004	0.003	5.388	0.121	0.005	0.003	-8.52	0.113	0.004	0.003	0	0
HETW	40	C>I	9.179	0.091	0.004	0.003	-1.875	0.106	0.004	0.003	-9.069	0.097	0.004	0.003	0.154	0.154
HETW	40	I>C	1.872	0.085	0.003	0.002	3.204	0.105	0.004	0.003	-7.71	0.1	0.004	0.003	0.154	0.308
HETWO	40	C>I	-2.472	0.096	0.004	0.003	-12.05	0.107	0.004	0.003	-18.672	0.107	0.004	0.003	0.154	0.923
HETWO	40	I>C	-1.026	0.094	0.004	0.003	-0.185	0.109	0.004	0.003	-10.819	0.11	0.004	0.003	0.308	0.923

Table C.8: Results for the five models with sub-samples, MAR missing data mechanism and nonconstant intervention effect rate. ^{1} = single timescale model, ^{2} = two timescale model.

Model	Turnover (%)	Intvn. effects	CS-OC-D[1] bias (%)	CS-OC-D[1] EmpSE	CS-OC-D[1] MCSE (bias)	CS-OC-D[1] MCSE (empSE)	CS-OC-D[2] bias (%)	CS-OC-D[2] EmpSE	CS-OC-D[2] MCSE (bias)	CS-OC-D[2] MCSE (empSE)	OC-52-I bias (%)	OC-52-I empSE	OC-52-I MCSE (bias)	OC-52-I MCSE (empSE)	Non-convergence[1] (%)	Non-convergence[2] (%)
Kasza	10	C>I	1.393	0.107	0.004	0.003	1.61	0.11	0.004	0.003	-0.222	0.168	0.007	0.005	1.231	1.846
Kasza	10	I>C	0.927	0.102	0.004	0.003	2.918	0.106	0.004	0.003	-0.529	0.17	0.007	0.005	1.077	1.077
JM	10	C>I	0.239	0.108	0.004	0.003	1.593	0.113	0.005	0.003	-2.327	0.161	0.007	0.005	2.615	14.615
JM	10	I>C	-0.562	0.103	0.004	0.003	3.18	0.107	0.004	0.003	1.96	0.164	0.007	0.005	2.615	12.769
CW	10	C>I	1.291	0.109	0.004	0.003	1.61	0.112	0.004	0.003	-0.057	0.163	0.006	0.005	0	0
CW	10	I>C	0.947	0.105	0.004	0.003	3.061	0.11	0.004	0.003	0.214	0.165	0.006	0.005	0	0
HETW	10	C>I	1.648	0.111	0.004	0.003	1.87	0.116	0.005	0.003	0.331	0.163	0.006	0.005	1.231	0.308
HETW	10	I>C	-0.289	0.106	0.004	0.003	2.826	0.111	0.004	0.003	-0.351	0.167	0.007	0.005	0.308	0.615
HETWO	10	C>I	1.13	0.11	0.004	0.003	1.431	0.113	0.004	0.003	-0.246	0.173	0.007	0.005	0.462	0.462
HETWO	10	I>C	0.657	0.105	0.004	0.003	2.942	0.109	0.004	0.003	0.037	0.173	0.007	0.005	0.462	0.462
Kasza	20	C>I	1.7	0.096	0.004	0.003	0.038	0.102	0.004	0.003	-6.55	0.13	0.005	0.004	2	2.462
Kasza	20	I>C	3.151	0.097	0.004	0.003	4.095	0.104	0.004	0.003	-3.2	0.135	0.005	0.004	1.077	1.692
JM	20	C>I	-2.847	0.094	0.004	0.003	-6.429	0.103	0.005	0.003	-14.771	0.103	0.005	0.003	0.769	24.615
JM	20	I>C	-2.252	0.095	0.004	0.003	-4.476	0.105	0.005	0.003	-1.784	0.113	0.005	0.004	0.769	24.154
CW	20	C>I	2.128	0.106	0.004	0.003	1.739	0.111	0.004	0.003	-6.32	0.132	0.005	0.004	0	0
CW	20	I>C	3.491	0.103	0.004	0.003	6.129	0.11	0.004	0.003	-2.205	0.133	0.005	0.004	0	0
HETW	20	C>I	1.494	0.098	0.004	0.003	-0.76	0.107	0.004	0.003	-7.712	0.129	0.005	0.004	0.308	0.308
HETW	20	I>C	2.112	0.1	0.004	0.003	4.038	0.106	0.004	0.003	-3.229	0.136	0.005	0.004	0.308	0.615
HETWO	20	C>I	1.744	0.101	0.004	0.003	-0.09	0.109	0.004	0.003	-6.488	0.137	0.005	0.004	0.308	0.615
HETWO	20	I>C	2.76	0.1	0.004	0.003	4.189	0.107	0.004	0.003	-2.585	0.136	0.005	0.004	0.462	0.154
Kasza	40	C>I	0.969	0.087	0.004	0.003	-3.793	0.099	0.004	0.003	-12.4	0.102	0.004	0.003	8.462	10.769
Kasza	40	I>C	2.148	0.085	0.003	0.002	2.521	0.098	0.004	0.003	-9.868	0.105	0.004	0.003	9.077	10.769
JM	40	C>I	-6.365	0.083	0.003	0.002	-17.071	0.095	0.004	0.003	-33.321	0.07	0.003	0.002	3.231	7.231
JM	40	I>C	-4.617	0.078	0.003	0.002	-13.562	0.092	0.004	0.003	-21.141	0.07	0.003	0.002	3.846	5.231
CW	40	C>I	-4.072	0.099	0.004	0.003	3.725	0.114	0.004	0.003	-8.796	0.107	0.004	0.003	0	0
CW	40	I>C	-2.11	0.103	0.004	0.003	11.845	0.116	0.005	0.003	-4.283	0.115	0.004	0.003	0	0
HETW	40	C>I	0.28	0.087	0.003	0.002	-5.401	0.1	0.004	0.003	-12.675	0.103	0.004	0.003	0.462	0.462
HETW	40	I>C	-0.381	0.083	0.003	0.002	1.416	0.097	0.004	0.003	-10.855	0.103	0.004	0.003	0.769	0.308
HETWO	40	C>I	0.271	0.088	0.003	0.002	-5.24	0.1	0.004	0.003	-11.202	0.105	0.004	0.003	0.462	1.077
HETWO	40	I>C	1.804	0.088	0.003	0.002	2.673	0.1	0.004	0.003	-7.77	0.11	0.004	0.003	0.769	0.923

Table C.9: Results for the five models with full-samples, MNAR missing data mechanism and constant intervention effect rate. [1] = single timescale model, [2] = two timescale model.

Model	Turnover (%)	Intvn. effects	CS-OC-D^1 bias (%)	CS-OC-D^1 EmpSE	CS-OC-D^1 MCSE (bias)	CS-OC-D^1 MCSE (empSE)	CS-OC-D^2 bias (%)	CS-OC-D^2 EmpSE	CS-OC-D^2 MCSE (bias)	CS-OC-D^2 MCSE (empSE)	OC-52-I bias (%)	OC-52-I empSE	OC-52-I MCSE (bias)	OC-52-I MCSE (empSE)	Non-convergence^1 (%)	Non-convergence^2 (%)
Kasza	10	C>I	-0.47	0.1	0.004	0.003	-0.686	0.104	0.004	0.003	-0.757	0.159	0.006	0.004	1.077	0.923
Kasza	10	I>C	1.4	0.101	0.004	0.003	3.105	0.104	0.004	0.003	5.111	0.165	0.007	0.005	1.231	1.692
JM	10	C>I	-3.536	0.1	0.004	0.003	-2.811	0.105	0.004	0.003	-9.404	0.151	0.006	0.005	2	15.538
JM	10	I>C	-0.52	0.101	0.004	0.003	1.811	0.104	0.004	0.003	1.984	0.152	0.006	0.005	2.615	14.923
CW	10	C>I	-4.74	0.103	0.004	0.003	-5.121	0.107	0.004	0.003	-3.978	0.156	0.006	0.004	0	0
CW	10	I>C	-0.168	0.105	0.004	0.003	1.73	0.109	0.004	0.003	3.336	0.16	0.006	0.004	0	0
HETW	10	C>I	4.92	0.105	0.004	0.003	4.506	0.109	0.004	0.003	4.84	0.157	0.006	0.004	0.154	0.615
HETW	10	I>C	1.745	0.106	0.004	0.003	4.384	0.109	0.004	0.003	7.014	0.165	0.006	0.005	0.154	0.308
HETWO	10	C>I	-5.306	0.104	0.004	0.003	-5.294	0.106	0.004	0.003	-7.319	0.163	0.006	0.005	0.769	1.077
HETWO	10	I>C	-0.349	0.105	0.004	0.003	1.704	0.108	0.004	0.003	2.529	0.17	0.007	0.005	0.462	0.615
Kasza	20	C>I	4.015	0.094	0.004	0.003	0.39	0.1	0.004	0.003	-3.595	0.138	0.006	0.004	2.769	3.385
Kasza	20	I>C	2.705	0.095	0.004	0.003	3.808	0.101	0.004	0.003	-8.142	0.136	0.005	0.004	2	2.308
JM	20	C>I	-1.846	0.094	0.004	0.003	-7.508	0.098	0.004	0.003	-16.818	0.106	0.005	0.003	0.462	26.615
JM	20	I>C	-3.178	0.094	0.004	0.003	-5.912	0.104	0.005	0.003	-4.934	0.107	0.005	0.003	0.769	23.846
CW	20	C>I	-2.221	0.108	0.004	0.003	-3.546	0.111	0.004	0.003	-7.167	0.141	0.006	0.004	0	0
CW	20	I>C	0.472	0.102	0.004	0.003	3.635	0.108	0.004	0.003	-7.933	0.135	0.005	0.004	0	0
HETW	20	C>I	10.209	0.095	0.004	0.003	4.393	0.1	0.004	0.003	0.919	0.133	0.005	0.004	0.462	0.308
HETW	20	I>C	2.91	0.096	0.004	0.003	5.113	0.104	0.004	0.003	-5.969	0.133	0.005	0.004	0.462	0.308
HETWO	20	C>I	-0.627	0.102	0.004	0.003	-4.07	0.106	0.004	0.003	-9.289	0.145	0.006	0.004	0.615	0.615
HETWO	20	I>C	0.225	0.099	0.004	0.003	2.13	0.105	0.004	0.003	-9.113	0.14	0.006	0.004	0.462	0.462
Kasza	40	C>I	4.996	0.086	0.004	0.002	-3.274	0.098	0.004	0.003	-15.712	0.104	0.004	0.003	7.538	9.385
Kasza	40	I>C	4.918	0.09	0.004	0.003	5.581	0.101	0.004	0.003	-7.771	0.104	0.004	0.003	7.692	8
JM	40	C>I	1.354	0.079	0.003	0.002	-13.366	0.092	0.004	0.003	-31.498	0.068	0.003	0.002	4.462	4.769
JM	40	I>C	-1.453	0.085	0.003	0.002	-11.21	0.1	0.004	0.003	-19.752	0.072	0.003	0.002	3.077	3.846
CW	40	C>I	-11.06	0.103	0.004	0.003	-3.461	0.115	0.005	0.003	-17.841	0.115	0.004	0.003	0	0
CW	40	I>C	-4.742	0.104	0.004	0.003	9.792	0.117	0.005	0.003	-5.834	0.111	0.004	0.003	0	0
HETW	40	C>I	9.338	0.085	0.003	0.002	-0.796	0.098	0.004	0.003	-11.999	0.103	0.004	0.003	0.615	0.923
HETW	40	I>C	3.97	0.088	0.003	0.002	5.281	0.102	0.004	0.003	-6.815	0.102	0.004	0.003	0.769	0.615
HETWO	40	C>I	-0.34	0.091	0.004	0.003	-9.305	0.101	0.004	0.003	-19.219	0.111	0.004	0.003	0.769	0.308
HETWO	40	I>C	1.531	0.094	0.004	0.003	2.204	0.106	0.004	0.003	-8.694	0.108	0.004	0.003	0.615	0.769

Table C.10: Results for the five models with full-samples, MNAR missing data mechanism and non-constant intervention effect rate. ^{1} = single timescale model, ^{2} = two timescale model.

Model	Turnover (%)	Intvn. effects	CS-OC-D[1] bias (%)	CS-OC-D[1] EmpSE	CS-OC-D[1] MCSE (bias)	CS-OC-D[1] MCSE (empSE)	CS-OC-D[2] bias (%)	CS-OC-D[2] EmpSE	CS-OC-D[2] MCSE (bias)	CS-OC-D[2] MCSE (empSE)	OC-52-I bias (%)	OC-52-I empSE	OC-52-I MCSE (bias)	OC-52-I MCSE (empSE)	Non-con-vergence[1] (%)	Non-con-vergence[2] (%)
Kasza	10	C>I	0.71	0.101	0.004	0.003	0.583	0.107	0.004	0.003	-1.504	0.153	0.006	0.004	1.538	1.538
Kasza	10	I>C	-0.703	0.102	0.004	0.003	0.669	0.106	0.004	0.003	0.693	0.163	0.006	0.005	0.308	0.615
JM	10	C>I	-0.111	0.099	0.004	0.003	0.234	0.105	0.004	0.003	-3.004	0.155	0.007	0.005	1.538	13.846
JM	10	I>C	-2.189	0.103	0.004	0.003	-0.333	0.107	0.005	0.003	-0.288	0.159	0.007	0.005	3.538	13.385
CW	10	C>I	0.844	0.103	0.004	0.003	0.809	0.108	0.004	0.003	-1.838	0.151	0.006	0.004	0	0
CW	10	I>C	-0.795	0.103	0.004	0.003	0.832	0.107	0.004	0.003	0.86	0.155	0.006	0.004	0	0
HETW	10	C>I	0.116	0.107	0.004	0.003	0.302	0.112	0.004	0.003	-2.362	0.155	0.006	0.004	0.769	0.615
HETW	10	I>C	-1.558	0.107	0.004	0.003	0.816	0.112	0.004	0.003	0.359	0.16	0.006	0.004	0	0.615
HETWO	10	C>I	0.658	0.103	0.004	0.003	0.673	0.108	0.004	0.003	-1.724	0.158	0.006	0.004	0.154	0.154
HETWO	10	I>C	-1.133	0.103	0.004	0.003	1.078	0.107	0.004	0.003	1.081	0.166	0.007	0.005	0.615	1.385
Kasza	20	C>I	2.466	0.096	0.004	0.003	0.589	0.103	0.004	0.003	-5.586	0.138	0.005	0.004	1.846	2.308
Kasza	20	I>C	3.654	0.103	0.004	0.003	4.808	0.109	0.004	0.003	-0.571	0.142	0.006	0.004	2.154	2.615
JM	20	C>I	-2.183	0.095	0.004	0.003	-6.171	0.104	0.005	0.003	-15.829	0.114	0.005	0.004	0.769	25.231
JM	20	I>C	-1.384	0.1	0.004	0.003	-2.451	0.111	0.005	0.003	-0.369	0.106	0.005	0.003	0.462	20.769
CW	20	C>I	2.174	0.106	0.004	0.003	1.624	0.111	0.004	0.003	-5.265	0.139	0.005	0.004	0	0
CW	20	I>C	3.068	0.112	0.004	0.003	5.847	0.116	0.005	0.003	0.116	0.143	0.006	0.004	0	0
HETW	20	C>I	3.047	0.097	0.004	0.003	0.348	0.106	0.004	0.003	-5.68	0.135	0.005	0.004	0.154	0.615
HETW	20	I>C	2.732	0.105	0.004	0.003	4.667	0.11	0.004	0.003	-0.435	0.139	0.005	0.004	0.308	1.231
HETWO	20	C>I	2.07	0.102	0.004	0.003	-0.038	0.108	0.004	0.003	-5.183	0.146	0.006	0.004	0.769	1.077
HETWO	20	I>C	2.744	0.107	0.004	0.003	4.418	0.112	0.004	0.003	0.163	0.147	0.006	0.004	0.769	0.154
Kasza	40	C>I	1.607	0.086	0.004	0.003	-3.593	0.099	0.004	0.003	-11.686	0.104	0.004	0.003	9.692	12.308
Kasza	40	I>C	3.002	0.09	0.004	0.003	3.52	0.101	0.004	0.003	-7.963	0.11	0.005	0.003	8.462	10.308
JM	40	C>I	-5.261	0.084	0.003	0.002	-16.015	0.093	0.004	0.003	-32.251	0.069	0.003	0.002	2.923	6.308
JM	40	I>C	-3.458	0.082	0.003	0.002	-12.924	0.093	0.004	0.003	-20.476	0.069	0.003	0.002	3.538	5.385
CW	40	C>I	-3.829	0.104	0.004	0.003	2.887	0.116	0.005	0.003	-8.006	0.114	0.004	0.003	0	0
CW	40	I>C	-2.308	0.111	0.004	0.003	11.956	0.121	0.005	0.003	-3.297	0.122	0.005	0.003	0	0
HETW	40	C>I	2.069	0.087	0.003	0.002	-4.086	0.1	0.004	0.003	-10.926	0.102	0.004	0.003	0.462	0.462
HETW	40	I>C	0.343	0.088	0.003	0.002	2.224	0.1	0.004	0.003	-8.663	0.106	0.004	0.003	0.615	1.231
HETWO	40	C>I	0.937	0.09	0.004	0.003	-4.962	0.101	0.004	0.003	-10.353	0.111	0.004	0.003	0.462	0.769
HETWO	40	I>C	2.252	0.095	0.004	0.003	3.069	0.105	0.004	0.003	-6.426	0.114	0.004	0.003	0.769	0.462

Table C.11: Results for the five models with sub-samples, MNAR missing data mechanism and constant intervention effect rate. [1] = single timescale model, [2] = two timescale model.

Model	**Turnover (%)**	**Intvn. effects**	**CS-OC-D[1] bias (%)**	**CS-OC-D[1] EmpSE**	**CS-OC-D[1] MCSE (bias)**	**CS-OC-D[1] MCSE (empSE)**	**CS-OC-D[2] bias (%)**	**CS-OC-D[2] EmpSE**	**CS-OC-D[2] MCSE (bias)**	**CS-OC-D[2] MCSE (empSE)**	**OC-52-I bias (%)**	**OC-52-I empSE**	**OC-52-I MCSE (bias)**	**OC-52-I MCSE (empSE)**	**Non-convergence[1] (%)**	**Non-convergence[2] (%)**
Kasza	10	C>I	1.916	0.1	0.004	0.003	1.385	0.103	0.004	0.003	2.758	0.164	0.007	0.005	2	2
Kasza	10	I>C	0.949	0.098	0.004	0.003	3.085	0.102	0.004	0.003	-1.243	0.165	0.007	0.005	2	2.308
JM	10	C>I	-3.536	0.1	0.004	0.003	-2.811	0.105	0.004	0.003	-9.404	0.151	0.006	0.005	2	15.538
JM	10	I>C	-0.52	0.101	0.004	0.003	1.811	0.104	0.004	0.003	1.984	0.152	0.006	0.005	2.615	14.923
CW	10	C>I	-2.433	0.106	0.004	0.003	-3.118	0.109	0.004	0.003	-0.959	0.16	0.006	0.004	0	0
CW	10	I>C	-0.664	0.101	0.004	0.003	1.618	0.106	0.004	0.003	-2.837	0.159	0.006	0.004	0	0
HETW	10	C>I	6.966	0.105	0.004	0.003	6.073	0.106	0.004	0.003	8.609	0.165	0.006	0.005	0.462	0.462
HETW	10	I>C	1.662	0.102	0.004	0.003	4.78	0.108	0.004	0.003	0.709	0.161	0.006	0.004	0.154	0.615
HETWO	10	C>I	-2.991	0.106	0.004	0.003	-3.222	0.108	0.004	0.003	-4.224	0.168	0.007	0.005	1.077	0.923
HETWO	10	I>C	-1.022	0.103	0.004	0.003	1.534	0.106	0.004	0.003	-4.261	0.168	0.007	0.005	0.923	0.308
Kasza	20	C>I	4.004	0.094	0.004	0.003	0.513	0.1	0.004	0.003	-5.172	0.127	0.005	0.004	2.615	2.923
Kasza	20	I>C	4.734	0.099	0.004	0.003	6.348	0.106	0.004	0.003	-3.278	0.135	0.005	0.004	1.846	2.154
JM	20	C>I	-1.846	0.094	0.004	0.003	-7.508	0.098	0.004	0.003	-16.818	0.106	0.005	0.003	0.462	26.615
JM	20	I>C	-3.178	0.094	0.004	0.003	-5.912	0.104	0.005	0.003	-4.934	0.107	0.005	0.003	0.769	23.846
CW	20	C>I	-2.259	0.102	0.004	0.003	-3.312	0.107	0.004	0.003	-8.988	0.128	0.005	0.004	0	0
CW	20	I>C	2.65	0.103	0.004	0.003	5.709	0.11	0.004	0.003	-3.966	0.134	0.005	0.004	0	0
HETW	20	C>I	9.543	0.097	0.004	0.003	4.278	0.103	0.004	0.003	-0.899	0.126	0.005	0.003	0.154	0.769
HETW	20	I>C	4.918	0.1	0.004	0.003	7.155	0.109	0.004	0.003	-2.127	0.133	0.005	0.004	0.154	0.462
HETWO	20	C>I	-0.714	0.099	0.004	0.003	-4.037	0.104	0.004	0.003	-10.567	0.131	0.005	0.004	0.769	0.923
HETWO	20	I>C	2.229	0.1	0.004	0.003	4.203	0.108	0.004	0.003	-5.153	0.14	0.006	0.004	0.615	0.923
Kasza	40	C>I	5.846	0.092	0.004	0.003	-2.985	0.1	0.004	0.003	-15.615	0.105	0.004	0.003	8.154	10.308
Kasza	40	I>C	3.16	0.089	0.004	0.003	3.748	0.101	0.004	0.003	-10.483	0.102	0.004	0.003	8.615	10.154
JM	40	C>I	1.354	0.079	0.003	0.002	-13.366	0.092	0.004	0.003	-31.498	0.068	0.003	0.002	4.462	4.769
JM	40	I>C	-1.453	0.085	0.003	0.002	-11.21	0.1	0.004	0.003	-19.752	0.072	0.003	0.002	3.077	3.846
CW	40	C>I	-11.828	0.106	0.004	0.003	-3.993	0.117	0.005	0.003	-18.83	0.11	0.004	0.003	0	0
CW	40	I>C	-8.196	0.109	0.004	0.003	6.363	0.122	0.005	0.003	-10.349	0.113	0.004	0.003	0	0
HETW	40	C>I	10.044	0.092	0.004	0.003	-0.456	0.102	0.004	0.003	-12.055	0.104	0.004	0.003	0.615	0.769
HETW	40	I>C	3.278	0.085	0.003	0.002	4.799	0.099	0.004	0.003	-8.666	0.101	0.004	0.003	0.154	0.615
HETWO	40	C>I	-0.916	0.096	0.004	0.003	-9.686	0.104	0.004	0.003	-19.933	0.108	0.004	0.003	0.615	0.769
HETWO	40	I>C	-0.508	0.096	0.004	0.003	0.331	0.108	0.004	0.003	-11.965	0.108	0.004	0.003	0.154	0.308

Table C.12: Results for the five models with sub-samples, MNAR missing data mechanism and non-constant intervention effect rate. [1] = single timescale model, [2] = two timescale model.

Bibliography

[1] H. Goldstein. *Multilevel statistical models*, volume 922. John Wiley & Sons, 2011.

[2] S. N. Beretvas. Cross-classified and multiple membership models. *Handbook of Advanced Multilevel Analysis*, pages 313–334, 2011.

[3] A. Caille, S. Kerry, E. Tavernier, C. Leyrat, S. Eldridge, and B. Giraudeau. Timeline cluster: a graphical tool to identify risk of bias in cluster randomised trials. *BMJ*, 354, 2016.

[4] T. P. Morris, I. R. White, and M. J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019.

[5] R. J. Hayes and L. H. Moulton. *Cluster randomised trials.* CRC press, 2017.

[6] M. Porta. *A dictionary of epidemiology.* Oxford University Press, 2014.

[7] K. J. Rothman, S. Greenland, and T. L. Lash. *Modern epidemiology.* Lippincott Williams & Wilkins, 2008.

[8] J. P. Vandenbroucke and N. Pearce. Incidence rates in dynamic populations. *International Journal of Epidemiology*, 41(5):1472–1479, 2012.

[9] M. Szklo and F. J. Nieto. *Epidemiology: beyond the basics.* Jones & Bartlett Publishers, 2014.

[10] U. Trivellato. Issues in the design and analysis of panel studies: A cursory review. *Quality and Quantity*, 33(3):339–351, 1999.

[11] S. Eldridge and S. Kerry. *A practical guide to cluster randomised trials in health services research*, volume 120. John Wiley & Sons, 2012.

[12] A. Donner and N. Klar. *Design and analysis of cluster randomization trials in health research.* 2010.

[13] D. M. Murray and P. J. Hannan. Planning for the appropriate analysis in school-based drug-use prevention studies. *Journal of Consulting and Clinical Psychology*, 58(4):458, 1990.

[14] H. A. Feldman and S. M. McKinlay. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. *Statistics in Medicine*, 13(1):61–78, 1994.

[15] P. Diehr, D. C. Martin, T. Koepsell, A. Cheadle, B. M. Psaty, and E. H. Wagner. Optimal survey design for community intervention evaluations: cohort or cross-sectional? *Journal of Clinical Epidemiology*, 48(12):1461–1472, 1995.

[16] P. S. Levy and S. Lemeshow. *Sampling of populations: methods and applications.* John Wiley & Sons, 2013.

[17] R. Simon. Length biased sampling in etiologic studies. *American Journal of Epidemiology*, 111(4):444–452, 1980.

[18] R. Hooper and L. Bourke. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ*, 350:h2925, 2015.

[19] T. D. Koepsell, D. C. Martin, P. H. Diehr, B. M. Psaty, E. H. Wagner, E. B. Perrin, and A. Cheadle. Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis of variance approach. *Journal of Clinical Epidemiology*, 44(7):701–713, 1991.

[20] J. Salonen, T. Kottke, D. Jacobs Jr, and P. Hannan. Analysis of community-based cardiovascular disease prevention studies - evaluation issues in the North Karelia Project and the Minnesota Heart Health Program. *International Journal of Epidemiology*, 15(2):176–182, 1986.

[21] A. A. Atienza and A. C. King. Community-based health intervention trials: an overview of methodological issues. *Epidemiologic Reviews*, 24(1):72–79, 2002.

[22] D. M. Zucker, E. Lakatos, L. S. Webber, D. M. Murray, S. M. McKinlay, H. A. Feldman, S. H. Kelder, P. R. Nader, C. S. Group, et al. Statistical design of the child and adolescent trial for cardiovascular health (CATCH): implications of cluster randomization. *Controlled Clinical Trials*, 16(2):96–118, 1995.

[23] M. H. Gail, S. D. Mark, R. J. Carroll, S. B. Green, and D. Pee. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15(11):1069–1092, 1996.

[24] J. M. Simpson, N. Klar, and A. Donnor. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *American Journal of Public Health*, 85(10):1378–1383, 1995.

[25] S. M. Mckinlay. Cost-efficient designs of cluster unit trials. *Preventive Medicine*, 23(5):606–611, 1994.

[26] R. A. Carleton, T. M. Lasater, A. R. Assaf, H. A. Feldman, and S. McKinlay. The Pawtucket Heart Health Program: community changes in cardiovascular risk factors and projected disease risk. *American Journal of Public Health*, 85(6):777–785, 1995.

[27] D. R. Jacobs Jr, R. V. Luepker, M. B. Mittelmark, A. R. Folsom, P. L. Pirie, S. R. Mascioli, P. J. Hannan, T. F. Pechacek, N. F. Bracht, R. W. Carlaw, et al. Community-wide prevention strategies: evaluation design of the Minnesota Heart Health Program. *Journal of Chronic Diseases*, 39(10):775–788, 1986.

[28] J. W. Farquhar, S. P. Fortmann, N. Maccoby, W. L. Haskell, P. T. Williams, J. A. Flora, C. B. Taylor, B. W. Brown Jr, D. S. Solomon, and S. B. Hulley. The Stanford five-city project: design and methods. *American Journal of Epidemiology*, 122(2):323–334, 1985.

[29] E. H. Wagner, T. D. Koepsell, C. Anderman, A. Cheadle, S. G. Curry, B. M. Psaty, M. Von Korff, T. M. Wickizer, W. L. Beery, P. K. Diehr, et al. The evaluation of the Henry J. Kaiser Family Foundation's community health promotion grant program: design. *Journal of Clinical Epidemiology*, 44(7):685–699, 1991.

[30] S. N. Blair, P. V. Piserchia, C. S. Wilbur, and J. H. Crowder. A public health intervention model for work-site health promotion: impact on exercise and physical fitness in a health promotion plan after 24 months. *JAMA*, 255(7):921–926, 1986.

[31] K. Hemming, M. Taljaard, J. E. McKenzie, R. Hooper, A. Copas, J. A. Thompson, M. Dixon-Woods, A. Aldcroft, A. Doussau, M. Grayling, et al. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. *BMJ*, 363:k1614, 2018.

[32] A. J. Copas, J. J. Lewis, J. A. Thompson, C. Davey, G. Baio, and J. R. Hargreaves. Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials*, 16(1):352, 2015.

[33] S. J. Arnup, J. E. McKenzie, K. Hemming, D. Pilcher, and A. B. Forbes. Understanding the cluster randomised crossover design. 2017.

[34] C. Rietbergen and M. Moerbeek. The design of cluster randomized crossover trials. *Journal of Educational and Behavioral Statistics*, 36(4):472–490, 2011.

[35] T. D. Koepsell, E. H. Wagner, A. Cheadle, D. L. Patrick, D. Martin, P. H. Diehr, E. B. Perrin, A. Kristal, C. Allan-Andrilla, and L. Dey. Selected methodological issues in evaluating community-based health promotion and disease prevention programs. *Annual Review of Public Health*, 13(1):31–57, 1992.

[36] J. Kasza, R. Hooper, A. Copas, and A. B. Forbes. Sample size and power calculations for open cohort longitudinal cluster randomized trials. *Statistics in Medicine*, 39(13):1871–1883, 2020.

[37] A. Rafferty, P. Walthery, and S. King-Hele. Analysing change over time: repeated cross sectional and longitudinal survey data (UK Data Service). `https://www.ukdataservice.ac.uk/media/455362/changeovertime.pdf`, 2015.

[38] P. Lugtig and P. A. Smith. The choice between a panel and cohort study design. 2019.

[39] Y. Deng, D. S. Hillygus, J. P. Reiter, Y. Si, S. Zheng, et al. Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science*, 28(2):238–256, 2013.

[40] D. Steel and C. McLaren. Design and analysis of repeated surveys. 2008.

[41] J. D. Singer, J. B. Willett, J. B. Willett, et al. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press, 2003.

[42] S. W. Raudenbush and A. S. Bryk. *Hierarchical linear models: Applications and data analysis methods*, volume 1. Sage, 2002.

[43] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[44] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[45] R. A. Bailey. *Design of comparative experiments*, volume 25. Cambridge University Press, 2008.

[46] R. Hooper, A. Forbes, K. Hemming, A. Takeda, and L. Beresford. Analysis of cluster randomised trials with an assessment of outcome at baseline. *BMJ*, 360:k1121, 2018.

[47] L. M. Friedman, C. D. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger. *Fundamentals of clinical trials*. Springer, 2015.

[48] C. J. Coffman, D. Edelman, and R. F. Woolson. To condition or not condition? Analysing 'change' in longitudinal randomised controlled trials. *BMJ Open*, 6(12), 2016.

[49] D. M. Murray et al. *Design and analysis of group-randomized trials*, volume 29. Oxford University Press, USA, 1998.

[50] R. Walwyn and C. Roberts. Therapist variation within randomised trials of psychotherapy: implications for precision, internal and external validity. *Statistical Methods in Medical Research*, 19(3):291–315, 2010.

[51] R. Hooper, S. Teerenstra, E. de Hoop, and S. Eldridge. Sample size calculation for stepped wedge and other longitudinal cluster randomised trials. *Statistics in Medicine*, 35(26):4718–4728, 2016.

[52] S. Teerenstra, S. Eldridge, M. Graff, E. de Hoop, and G. F. Borm. A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in Medicine*, 31(20):2169–2178, 2012.

[53] M. A. Hussey and J. P. Hughes. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28(2):182–191, 2007.

[54] C. A. Surr, R. E. Walwyn, A. Lilley-Kelly, R. Cicero, D. Meads, C. Ballard, K. Burton, L. Chenoweth, A. Corbett, B. Creese, et al. Evaluating the effectiveness and cost-effectiveness of Dementia Care Mapping to enable person-centred care for people with dementia and their carers (DCM-EPIC) in care homes: study protocol for a randomised controlled trial. *Trials*, 17(1):300, 2016.

[55] K. Diaz-Ordaz, R. Froud, B. Sheehan, and S. Eldridge. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. *BMC Medical Research Methodology*, 13(1):127, 2013.

[56] A. Farrin, I. Russell, D. Torgerson, and M. Underwood. Differential recruitment in a cluster randomized trial in primary care: the experience of the UK back pain, exercise, active management and manipulation (UK BEAM) feasibility study. *Clinical Trials*, 2(2):119–124, 2005.

[57] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[58] S. Van Buuren. *Flexible imputation of missing data.* Chapman and Hall/CRC, 2018.

[59] M. Jones, G. D. Mishra, and A. Dobson. Analytical results in longitudinal studies depended on target of inference and assumed mechanism of attrition. *Journal of Clinical Epidemiology*, 68(10):1165–1175, 2015.

[60] B. F. Kurland, L. L. Johnson, B. L. Egleston, and P. H. Diehr. Longitudinal data with follow-up truncated by death: match the analysis method to research aims. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 24(2):211, 2009.

[61] S. R. Seaman, M. Pavlou, and A. J. Copas. Methods for observed-cluster inference when cluster size is informative: A review and clarifications. *Biometrics*, 70(2):449–456, 2014.

[62] H. Arksey and L. O'Malley. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 8(1):19–32, 2005.

[63] M. D. Peters, C. M. Godfrey, H. Khalil, P. McInerney, D. Parker, and C. B. Soares. Guidance for conducting systematic scoping reviews. *International Journal of Evidence-based Healthcare*, 13(3):141–146, 2015.

[64] M. K. Campbell, G. Piaggio, D. R. Elbourne, and D. G. Altman. CONSORT 2010 statement: extension to cluster randomised trials. *BMJ*, 345:e5661, 2012.

[65] D. Giustini and M. N. K. Boulos. Google scholar is not enough to be used alone for systematic reviews. *Online Journal of Public Health Informatics*, 5(2):214, 2013.

[66] N. R. Haddaway, A. M. Collins, D. Coughlin, and S. Kirk. The role of google scholar in evidence reviews and its applicability to grey literature searching. *PloS One*, 10(9):e0138237, 2015.

[67] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, P. Group, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*, 6(7):e1000097, 2009.

[68] L. Chenoweth, I. Forbes, R. Fleming, M. King, J. Stein-Parbury, G. Luscombe, P. Kenny, Y.-H. Jeon, M. Haas, and H. Brodaty. PerCEN: a cluster randomized

controlled trial of person-centered residential care and environment for people with dementia. 2014.

[69] J. A. Weintraub, S. Zimmerman, K. Ward, C. J. Wretman, P. D. Sloane, S. C. Stearns, P. Poole, and J. S. Preisser. Improving nursing home residents' oral hygiene: Results of a cluster randomized intervention trial. *Journal of the American Medical Directors Association*, 19(12):1086–1091, 2018.

[70] M. Underwood, S. E. Lamb, S. Eldridge, B. Sheehan, A.-M. Slowther, A. Spencer, M. Thorogood, N. Atherton, S. A. Bremner, A. Devine, et al. Exercise for depression in elderly residents of care homes: a cluster-randomised controlled trial. *The Lancet*, 382(9886):41–49, 2013.

[71] C. M. Sackley, M. F. Walker, C. R. Burton, C. L. Watkins, J. Mant, A. K. Roalfe, K. Wheatley, B. Sheehan, L. Sharp, K. E. Stant, et al. An occupational therapy intervention for residents with stroke related disabilities in UK care homes (OTCH): cluster randomised controlled trial. *BMJ*, 350, 2015.

[72] C. Ballard, A. Corbett, M. Orrell, G. Williams, E. Moniz-Cook, R. Romeo, B. Woods, L. Garrod, I. Testad, B. Woodward-Carlton, et al. Impact of person-centred care training and person-centred activities on quality of life, agitation, and antipsychotic use in people with dementia living in nursing homes: A cluster-randomised controlled trial. *PLoS Medicine*, 15(2):e1002500, 2018.

[73] Y. Kon, Y. Ichikawa-Shigeta, T. Iuchi, Y. Nakajima, G. Nakagami, K. Tabata, H. Sanada, and J. Sugama. Effects of a skin barrier cream on management of incontinence-associated dermatitis in older women. *Journal of Wound, Ostomy and Continence Nursing*, 44(5):481–486, 2017.

[74] R. M. Daly, S. L. OConnell, N. L. Mundell, C. A. Grimes, D. W. Dunstan, and C. A. Nowson. Protein-enriched diet, with the use of lean red meat, combined with progressive resistance training enhances lean tissue mass and muscle strength and reduces circulating il-6 concentrations in elderly women: a cluster randomized controlled trial. *The American Journal of Clinical Nutrition*, 99(4):899–910, 2014.

[75] J. Jancey, A.-M. Holt, A. Lee, D. Kerr, S. Robinson, L. Tang, A. Anderson, A. P. Hills, and P. Howat. Effects of a physical activity and nutrition program in retirement villages: a cluster randomised controlled trial. *International Journal of Behavioral Nutrition and Physical Activity*, 14(1):92, 2017.

[76] H. Wouters, J. Scheper, H. Koning, C. Brouwer, J. W. Twisk, H. van der Meer, F. Boersma, S. U. Zuidema, and K. Taxis. Discontinuing inappropriate medication use in nursing home residents: a cluster randomized controlled trial. *Annals of Internal Medicine*, 167(9):609–617, 2017.

[77] G. Arendts, P. Deans, K. OBrien, C. Etherton-Beer, K. Howard, G. Lewin, and M. Sim. A clinical trial of nurse practitioner care in residential aged care facilities. *Archives of Gerontology and Geriatrics*, 77:129–132, 2018.

[78] B. S. Husebø, C. Ballard, D. Aarsland, G. Selbaek, D. D. Slettebo, C. Gulla, I. Aasmul, T. Habiger, T. Elvegaard, I. Testad, et al. The effect of a multicomponent intervention on quality of life in residents of nursing homes: a randomized controlled trial (COSMOS). *Journal of the American Medical Directors Association*, 20(3):330–339, 2019.

[79] P. Sambrook, I. Cameron, J. Chen, R. Cumming, S. Durvasula, M. Herrmann, C. Kok, S. Lord, M. Macara, L. March, et al. Does increased sunlight exposure work as a strategy to improve vitamin D status in the elderly: a cluster randomised controlled trial. *Osteoporosis International*, 23(2):615–624, 2012.

[80] M. Boorsma, D. H. Frijters, D. L. Knol, M. E. Ribbe, G. Nijpels, and H. P. van Hout. Effects of multidisciplinary integrated care on quality of care in residential care facilities for elderly people: a cluster randomized trial. *CMAJ*, 183(11):E724–E732, 2011.

[81] P. Koczy, C. Becker, K. Rapp, T. Klie, D. Beische, G. Büchele, A. Kleiner, V. Guerra, U. Rißmann, S. Kurrle, et al. Effectiveness of a multifactorial intervention to reduce physical restraints in nursing home residents. *Journal of the American Geriatrics Society*, 59(2):333–339, 2011.

[82] C. Roos, M. Silén, B. Skytt, and M. Engström. An intervention targeting fundamental values among caregivers at residential facilities: effects of a cluster-randomized controlled trial on residents self-reported empowerment, person-centered climate and life satisfaction. *BMC Geriatrics*, 16(1):130, 2016.

[83] E. O'Shea, D. Devane, A. Cooney, D. Casey, F. Jordan, A. Hunter, E. Murphy, J. Newell, S. Connolly, and K. Murphy. The impact of reminiscence on the quality of life of residents with dementia in long-stay care. *International Journal of Geriatric Psychiatry*, 29(10):1062–1070, 2014.

[84] K. Adane, M. Spigt, B. Winkens, and G.-J. Dinant. Tuberculosis case detection by trained inmate peer educators in a resource-limited prison setting in Ethiopia: a cluster-randomised trial. *The Lancet Global Health*, 7(4):e482–e491, 2019.

[85] M. Sleed, T. Baradon, and P. Fonagy. New beginnings for mothers and babies in prison: A cluster randomized controlled trial. *Attachment & Human Development*, 15(4):349–367, 2013.

[86] R. Umbach, A. Raine, and N. R. Leonard. Cognitive decline as a result of incarceration and the effects of a CBT/MT intervention: A cluster-randomized controlled trial. *Criminal Justice and Behavior*, 45(1):31–55, 2018.

[87] G. Donenberg, E. Emerson, and A. D. Kendall. HIV-risk reduction intervention for juvenile offenders on probation: The PHAT Life group randomized controlled trial. *Health Psychology*, 37(4):364, 2018.

[88] K. Beernaert, T. Smets, J. Cohen, R. Verhofstede, M. Costantini, K. Eecloo, N. Van Den Noortgate, and L. Deliens. Improving comfort around dying in elderly people: a cluster randomised controlled trial. *The Lancet*, 390(10090):125–134, 2017.

[89] R. McCorkle, S. Jeon, E. Ercolano, M. Lazenby, A. Reid, M. Davies, D. Viveiros, and S. Gettinger. An advanced practice nurse coordinated multidisciplinary intervention for patients with late-stage cancer: a cluster randomized trial. *Journal of Palliative Medicine*, 18(11):962–969, 2015.

[90] C. Zimmermann, N. Swami, M. Krzyzanowska, B. Hannon, N. Leighl, A. Oza, M. Moore, A. Rydall, G. Rodin, I. Tannock, et al. Early palliative care for patients with advanced cancer: a cluster-randomised controlled trial. *The Lancet*, 383(9930):1721–1730, 2014.

[91] A. P. Abernethy, D. C. Currow, T. Shelby-James, D. Rowett, F. May, G. P. Samsa, R. Hunt, H. Williams, A. Esterman, and P. A. Phillips. Delivery strategies to optimize resource utilization and performance status for patients with advanced life-limiting illness: results from the "palliative care trial". *Journal of Pain and Symptom Management*, 45(3):488–505, 2013.

[92] M. Vermandere, F. Warmenhoven, E. Van Severen, J. De Lepeleire, and B. Aertgeerts. Spiritual history taking in palliative home care: A cluster randomized controlled trial. *Palliative Medicine*, 30(4):338–350, 2016.

[93] M. Sánchez-López, I. Cavero-Redondo, C. Álvarez-Bueno, A. Ruiz-Hermosa, D. P. Pozuelo-Carrascosa, A. Díez-Fernández, D. Gutierrez-Díaz del Campo, M. J. Pardo-Guijarro, and V. Martínez-Vizcaíno. Impact of a multicomponent physical activity intervention on cognitive performance: The MOVI-KIDS study. *Scandinavian Journal of Medicine & Science in Sports*, 29(5):766–775, 2019.

[94] S. Low, K. Smolkowski, C. Cook, and D. Desfosses. Two-year impact of a universal social-emotional learning curriculum: Group differences from developmentally sensitive trends over time. *Developmental Psychology*, 55(2):415, 2019.

[95] K. D. Martinsen, L. M. P. Rasmussen, T. Wentzel-Larsen, S. Holen, A. M. Sund, M. E. S. Løvaas, J. Patras, P. C. Kendall, T. Waaktaar, and S.-P. Neumer. Prevention of anxiety and depression in school children: Effectiveness of the transdiagnostic emotion program. *Journal of Consulting and Clinical Psychology*, 87(2):212, 2019.

[96] L. Karimli, F. M. Ssewamala, T. B. Neilands, C. R. Wells, and L. G. Bermudez. Poverty, economic strengthening, and mental health among aids orphaned children in Uganda: mediation model in a randomized clinical trial. *Social Science & Medicine*, 228:17–24, 2019.

[97] C. Sundgot-Borgen, O. Friborg, E. Kolle, K. M. Engen, J. Sundgot-Borgen, J. H. Rosenvinge, G. Pettersen, M. K. Torstveit, N. Piran, and S. Bratland-Sanda. The healthy body image (HBI) intervention: Effects of a school-based cluster-randomized controlled trial with 12-months follow-up. *Body Image*, 29:122–131, 2019.

[98] S. Flanagan, J. Kunkel, V. Appleby, S. E. Eldridge, S. Ismail, S. Moreea, C. Griffiths, R. Walton, M. Pitt, A. Salmon, et al. Case finding and therapy for chronic viral hepatitis in primary care (HepFREE): a cluster-randomised controlled trial. *The Lancet Gastroenterology & Hepatology*, 4(1):32–44, 2019.

[99] B. K. Kristiansen, B. Andersen, F. Bro, H. Svanholm, and P. Vedsted. Direct notification of cervical cytology results to women improves follow-up in cervical cancer screening-a cluster-randomised trial. *Preventive Medicine Reports*, 13:118–125, 2019.

[100] A. Siebenhofer, L.-R. Ulrich, K. Mergenthal, A. Berghold, G. Pregartner, B. Kemperdick, S. Schulz-Rothe, S. Rauck, S. Harder, F. M. Gerlach, et al. Primary care management for patients receiving long-term antithrombotic treatment: a cluster-randomized controlled trial. *PloS One*, 14(1):e0209366, 2019.

[101] A. Beratarrechea, S. Abrahams-Gessel, V. Irazola, L. Gutierrez, D. Moyano, and T. A. Gaziano. Using mHealth Tools to Improve Access and Coverage of People With Public Health Insurance and High Cardiovascular Disease Risk in Argentina: A Pragmatic Cluster Randomized Trial. *Journal of the American Heart Association*, 8(8):e011799, 2019.

[102] V. Ferreira, C. Sangalli, P. Leffa, F. Rauber, and M. Vitolo. The impact of a primary health care intervention on infant feeding practices: a cluster randomised controlled trial in Brazil. *Journal of Human Nutrition and Dietetics*, 32(1):21–30, 2019.

[103] NHS. Going into hospital as a patient. `https://www.nhs.uk/nhs-services/hospitals/going-into-hospital/going-into-hospital-as-a-patient/`, 2022.

[104] P. E. Beeler, E. Eschmann, M. Schneemann, and J. Blaser. Negligible impact of highly patient-specific decision support for potassium-increasing drug-drug interactions–a cluster-randomised controlled trial. *Swiss Medical Weekly*, 149:w20035, 2019.

[105] C. C.-H. Chen, Y.-T. Yang, I.-R. Lai, B.-R. Lin, C.-Y. Yang, J. Huang, Y.-W. Tien, C.-N. Chen, M.-T. Lin, J.-T. Liang, et al. Three nurse-administered protocols reduce nutritional decline and frailty in older gastrointestinal surgery patients: A cluster randomized trial. *Journal of the American Medical Directors Association*, 20(5):524–529, 2019.

[106] T. Y. Wang, L. A. Kaltenbach, C. P. Cannon, G. C. Fonarow, N. K. Choudhry, T. D. Henry, D. J. Cohen, D. Bhandary, N. D. Khan, K. J. Anstrom, et al. Effect of medication co-payment vouchers on P2Y12 inhibitor use and major adverse cardiovascular events among patients with myocardial infarction: the ARTEMIS randomized clinical trial. *JAMA*, 321(1):44–55, 2019.

[107] S. Bernitz, R. Dalbye, J. Zhang, T. M. Eggebø, K. F. Frøslie, I. C. Olsen, E. Blix, and P. Øian. The frequency of intrapartum caesarean section use with the WHO partograph versus Zhang's guideline in the Labour Progression Study (LaPS): a multicentre, cluster-randomised controlled trial. *The Lancet*, 393(10169):340–348, 2019.

[108] M. Cuypers, R. E. Lamers, P. J. Kil, L. V. van de Poll-Franse, and M. de Vries. Longitudinal regret and information satisfaction after deciding on treatment for localized prostate cancer with or without a decision aid. results at one-year follow-up in the PCPCC trial. *Patient Education and Counseling*, 102(3):424–428, 2019.

[109] A. Diallo, O. M. Diop, D. Diop, M. N. Niang, J. D. Sugimoto, J. R. Ortiz, E. H. A. Faye, B. Diarra, D. Goudiaby, K. D. Lewis, et al. Effectiveness of seasonal influenza vaccination in children in Senegal during a year of vaccine mismatch: a cluster-randomized trial. *Clinical Infectious Diseases*, 69(10):1780–1788, 2019.

[110] E. Loha, W. Deressa, T. Gari, M. Balkew, O. Kenea, T. Solomon, A. Hailu, B. Robberstad, M. Assegid, H. J. Overgaard, et al. Long-lasting insecticidal nets and indoor residual spraying may not be sufficient to eliminate malaria in a low malaria incidence area: results from a cluster randomized controlled trial in Ethiopia. *Malaria Journal*, 18(1):141, 2019.

[111] L. Von Seidlein, T. J. Peto, J. Landier, T.-N. Nguyen, R. Tripura, K. Phommasone, T. Pongvongsa, K. M. Lwin, L. Keereecharoen, L. Kajeechiwa, et al. The impact of targeted malaria elimination with mass drug administrations on falciparum malaria in Southeast Asia: a cluster randomised trial. *PLoS Medicine*, 16(2):e1002745, 2019.

[112] A. C. Lee, L. C. Mullany, M. Quaiyum, D. K. Mitra, A. Labrique, P. Christian, P. Ahmed, J. Uddin, I. Rafiqullah, S. DasGupta, et al. Effect of population-based antenatal screening and treatment of genitourinary tract infections on birth outcomes in Sylhet, Bangladesh (MIST): a cluster-randomised clinical trial. *The Lancet Global Health*, 7(1):e148–e159, 2019.

[113] A. Rahman, M. N. Khan, S. U. Hamdani, A. Chiumento, P. Akhtar, H. Nazir, A. Nisar, A. Masood, I. U. Din, N. A. Khan, et al. Effectiveness of a brief group psychological intervention for women in a post-conflict setting in Pakistan: a single-blind, cluster, randomised controlled trial. *The Lancet*, 393(10182):1733–1744, 2019.

[114] A. J. Pickering, Y. Crider, S. Sultana, J. Swarthout, F. G. Goddard, S. A. Islam, S. Sen, R. Ayyagari, and S. P. Luby. Effect of in-line drinking water chlorination at the point of collection on child diarrhoea in urban Bangladesh: a double-blind, cluster-randomised controlled trial. *The Lancet Global Health*, 7(9):e1247–e1256, 2019.

[115] J. L. Bell, J. W. Collins, and S. Chiou. Effectiveness of a no-cost-to-workers, slip-resistant footwear program for reducing slipping-related injuries in food service workers: a cluster randomized trial. *Scandinavian Journal of Work, Environment & Health*, 45(2):194–202, 2019.

[116] M. P. J. Hermans, J. Kooistra, S. C. Cannegieter, F. R. Rosendaal, D. O. Mook-Kanamori, and B. Nemeth. Healthcare and disease burden among refugees in long-

stay refugee camps at Lesbos, Greece. *European Journal of Epidemiology*, 32(9):851–854, 2017.

[117] C. Klein, P. Luig, T. Henke, and P. Platen. Injury burden differs considerably between single teams from german professional male football (soccer): surveillance of three consecutive seasons. *Knee Surgery, Sports Traumatology, Arthroscopy*, 28(5):1656–1664, 2020.

[118] N. M. Ivers, L. Desveaux, J. Presseau, C. Reis, H. O. Witteman, M. K. Taljaard, N. McCleary, K. Thavorn, and J. M. Grimshaw. Testing feedback message framing and comparators to address prescribing of high-risk medications in nursing homes: protocol for a pragmatic, factorial, cluster-randomized trial. *Implementation Science*, 12(1):86, 2017.

[119] S. A. Lippman, A. Pettifor, D. Rebombo, A. Julien, R. G. Wagner, M.-S. K. Dufour, C. W. Kabudula, T. B. Neilands, R. Twine, A. Gottert, et al. Evaluation of the Tsima community mobilization intervention to improve engagement in HIV testing and care in South Africa: study protocol for a cluster randomized trial. *Implementation Science*, 12(1):9, 2017.

[120] A. Piotrowski, M. Meyer, I. Burkholder, D. Renaud, M. A. Müller, T. Lehr, S. Laag, J. Meiser, L. Manderscheid, and J. Köberlein-Neu. Effect of an interprofessional care concept on the hospitalization of nursing home residents: study protocol for a cluster-randomized controlled trial. *Trials*, 21:1–11, 2020.

[121] G. A. Pape, J. S. Hunt, K. L. Butler, J. Siemienczuk, B. H. LeBlanc, W. Gillanders, Y. Rozenfeld, and K. Bonin. Team-based care approach to cholesterol management in diabetes mellitus: two-year cluster randomized controlled trial. *Archives of Internal Medicine*, 171(16):1480–1486, 2011.

[122] A. d'Arminio Monforte, H. Diaz-Cuervo, A. De Luca, F. Maggiolo, A. Cingolani, S. Bonora, A. Castagna, E. Girardi, A. Antinori, S. Lo Caputo, et al. Evolution of major non-HIV-related comorbidities in HIV-infected patients in the Italian Cohort of Individuals, Naïve for Antiretrovirals (ICONA) Foundation Study cohort in the period 2004–2014. *HIV Medicine*, 20(2):99–109, 2019.

[123] K. Azad, S. Barnett, B. Banerjee, S. Shaha, K. Khan, A. R. Rego, S. Barua, D. Flatman, C. Pagel, A. Prost, et al. Effect of scaling up women's groups on birth outcomes in three rural districts in Bangladesh: a cluster-randomised controlled trial. *The Lancet*, 375(9721):1193–1202, 2010.

[124] T. Clasen, S. Boisson, P. Routray, O. Cumming, M. Jenkins, J. H. Ensink, M. Bell, M. C. Freeman, S. Peppin, and W.-P. Schmidt. The effect of improved rural sanitation on diarrhoea and helminth infection: design of a cluster-randomized trial in Orissa, India. *Emerging Themes in Epidemiology*, 9(1):7, 2012.

[125] T. A. Houweling, K. Azad, L. Younes, A. Kuddus, S. Shaha, B. Haq, T. Nahar, J. Beard, E. F. Fottrell, A. Prost, et al. The effect of participatory women's groups on birth outcomes in Bangladesh: does coverage matter? Study protocol for a randomized controlled trial. *Trials*, 12(1):208, 2011.

[126] P. Tripathy, N. Nair, S. Barnett, R. Mahapatra, J. Borghi, S. Rath, S. Rath, R. Gope, D. Mahto, R. Sinha, et al. Effect of a participatory intervention with women's groups on birth outcomes and maternal depression in Jharkhand and Orissa, India: a cluster-randomised controlled trial. *The Lancet*, 375(9721):1182–1192, 2010.

[127] J. A. Finkelstein, S. S. Huang, K. Kleinman, S. L. Rifas-Shiman, C. J. Stille, J. Daniel, N. Schiff, R. Steingard, S. B. Soumerai, D. Ross-Degnan, et al. Impact of a 16-community trial to promote judicious antibiotic use in Massachusetts. *Pediatrics*, 121(1):e15–e23, 2008.

[128] M. Greiver, S. Dahrouge, P. OBrien, D. Manca, M. Lussier, J. Wang, F. Burge, M. Grandy, A. Singer, M. Twohig, et al. Improving care for elderly patients living with polypharmacy: protocol for a pragmatic cluster randomized trial in community-based primary care practices in Canada. *Implementation Science*, 14(1):55, 2019.

[129] W. Chaboyer, T. Bucknall, J. Webster, E. McInnes, B. M. Gillespie, M. Banks, J. A. Whitty, L. Thalib, S. Roberts, M. Tallott, et al. The effect of a patient centred care bundle intervention on pressure ulcer incidence (INTACT): a cluster randomised trial. *International Journal of Nursing Studies*, 64:63–71, 2016.

[130] M. Baiocchi, B. Omondi, N. Langat, D. B. Boothroyd, J. Sinclair, L. Pavia, M. Mulinge, O. Githua, N. H. Golden, and C. Sarnquist. A behavior-based intervention that prevents sexual assault: the results of a matched-pairs, cluster-randomized study in Nairobi, Kenya. *Prevention Science*, 18(7):818–827, 2017.

[131] G. Agarwal, M. Girard, R. Angeles, M. Pirrie, M.-T. Lussier, F. Marzanek, L. Dolovich, J. M. Paterson, L. Thabane, and J. Kaczorowski. Design and rationale for a pragmatic cluster randomized trial of the Cardiovascular Health Awareness Program (CHAP) for social housing residents in Ontario and Quebec, Canada. *Trials*, 20(1):760, 2019.

[132] S. G. Staedke, C. Maiteki-Sebuguzi, D. D. DiLiberto, E. L. Webb, L. Mugenyi, E. Mbabazi, S. Gonahasa, S. P. Kigozi, B. A. Willey, G. Dorsey, et al. The impact of an intervention to improve malaria care in public health centers on health indicators of children in Tororo, Uganda (PRIME): a cluster-randomized trial. *The American Journal of Tropical Medicine and Hygiene*, 95(2):358, 2016.

[133] N. Bavarian, K. M. Lewis, D. L. DuBois, A. Acock, S. Vuchinich, N. Silverthorn, F. J. Snyder, J. Day, P. Ji, and B. R. Flay. Using social-emotional and character development to improve academic outcomes: A matched-pair, cluster-randomized controlled trial in low-income, urban schools. *Journal of School Health*, 83(11):771–779, 2013.

[134] A. Thompson, A. K. Wright, D. M. Ashcroft, T. P. van Staa, and M. Pirmohamed. Epidemiology of alcohol dependence in UK primary care: results from a large observational study using the clinical practice research Datalink. *PloS One*, 12(3):e0174818, 2017.

[135] O. A. Sawicki, A. Mueller, R. Klaaßen-Mielke, A. Glushan, F. M. Gerlach, M. Beyer, M. Wensing, and K. Karimova. Strong and sustainable primary healthcare is associated with a lower risk of hospitalization in high risk patients. *Scientific Reports*, 11(1):1–9, 2021.

[136] G. Agarwal, R. N. Angeles, L. Dolovich, J. Kaczorowski, J. Gaber, D. Guenter, F. D. Arnuco, H. Y. Lam, L. Thabane, D. OReilly, et al. The Community Health Assessment Program in the Philippines (CHAP-P) diabetes health promotion program for low-to middle-income countries: study protocol for a cluster randomized controlled trial. *BMC Public Health*, 19(1):682, 2019.

[137] R. G. Tobe, S. E. Haque, K. Ikegami, and R. Mori. Mobile-health tool to improve maternal and neonatal health care in Bangladesh: a cluster randomized controlled trial. *BMC Pregnancy and Childbirth*, 18(1):102, 2018.

[138] H. J. Overgaard, N. Alexander, M. I. Mátiz, J. F. Jaramillo, V. A. Olano, S. Vargas, D. Sarmiento, A. Lenhart, R. Seidu, and T. A. Stenström. Diarrhea and dengue control in rural primary schools in Colombia: study protocol for a randomized controlled trial. *Trials*, 13(1):182, 2012.

[139] J. Cairney, S. Veldhuizen, S. King-Dowling, B. E. Faught, and J. Hay. Tracking cardiorespiratory fitness and physical activity in children with and without motor

coordination problems. *Journal of Science and Medicine in Sport*, 20(4):380–385, 2017.

[140] J. Hippisley-Cox, C. Coupland, Y. Vinogradova, J. Robson, R. Minhas, A. Sheikh, and P. Brindle. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*, 336(7659):1475–1482, 2008.

[141] A. Ntouva, K. A. Toulis, D. Keerthy, N. J. Adderley, W. Hanif, R. Thayakaran, K. Gokhale, G. N. Thomas, K. Khunti, A. A. Tahrani, et al. Hypoglycaemia is associated with increased risk of fractures in patients with type 2 diabetes mellitus: a cohort study. *European Journal of Endocrinology*, 180(1):51–58, 2019.

[142] K. A. Toulis, B. H. Willis, T. Marshall, B. Kumarendran, K. Gokhale, S. Ghosh, G. N. Thomas, K. K. Cheng, P. Narendran, W. Hanif, et al. All-cause mortality in patients with diabetes under treatment with dapagliflozin: a population-based, open-cohort study in the health improvement network database. *The Journal of Clinical Endocrinology & Metabolism*, 102(5):1719–1725, 2017.

[143] F. Atzeni, A. Carletto, R. Foti, M. Sebastiani, V. Panetta, F. Salaffi, G. Bonitta, F. Iannone, E. Gremese, M. Govoni, A. Marchesoni, E. G. Favalli, R. Gorla, R. Ramonda, P. Sarzi-Puttini, G. Ferraccioli, and G. Lapadula. Incidence of cancer in patients with spondyloarthritis treated with anti-TNF drugs. *Joint Bone Spine*, 85(4):455–459, 2018.

[144] P. S. Fontela, R. W. Platt, I. Rocher, C. Frenette, D. Moore, E. Fortin, D. Buckeridge, M. Pai, and C. Quach. Epidemiology of central line-associated bloodstream infections in Quebec intensive care units: A 6-year review. *American Journal of Infection Control*, 2012.

[145] J. Robson, I. Dostal, V. Madurasinghe, A. Sheikh, S. Hull, K. Boomla, H. Page, C. Griffiths, and S. Eldridge. The NHS Health Check programme: implementation in east London 2009-2011. *BMJ Open*, 5(4):e007578, 2015.

[146] B. Feakins, J. Oke, E. McFadden, J. Aronson, D. Lasserson, C. OCallaghan, C. Taylor, N. Hill, R. Stevens, and R. Perera. Trends in kidney function testing in UK primary care since the introduction of the quality and outcomes framework: a retrospective cohort study using CPRD. *BMJ Open*, 9(6):e028062, 2019.

[147] Y. Xie, C. Song, H. Cheng, C. Xu, Z. Zhang, J. Wang, L. Huo, Q. Du, J. Xu, Y. Chen, et al. Long-term follow-up of helicobacter pylori reinfection and its risk

factors after initial eradication: a large-scale multicentre, prospective open cohort, observational study. *Emerging Microbes & Infections*, 9(1):548–557, 2020.

[148] R. C. hechter, C. R. Chao, M. A. Sidell, L. S. Sy, B. K. Ackerson, J. M. Slezak, N. J. Patel, H. F. Tseng, and S. J. Jacobsen. Quadrivalent Human Papillomavirus Vaccine Initiation in Boys Before and Since Routine Use: Southern California, 2009-2013. *American Journal of Public Health*, 105(12):2549–2556, 2015.

[149] J. B. Cannata-Andía, J. L. Fernández-Martín, J. B. Cannata-Andia, J. L. Fernandez-Martin, C. Zoccali, G. M. London, F. Locatelli, M. Ketteler, A. Ferreira, A. Covic, et al. Current management of secondary hyperparathyroidism: a multicenter observational study (COSMOS). *Journal of Nephrology*, 21(3):290–298, 2008.

[150] M. Taljaard, J. McGowan, J. M. Grimshaw, J. C. Brehaut, A. McRae, M. P. Eccles, and A. Donner. Electronic search strategies to identify reports of cluster randomized trials in MEDLINE: low precision will improve with adherence to reporting standards. *BMC Medical Research Methodology*, 10(1):1–8, 2010.

[151] A.-W. Chan, J. M. Tetzlaff, D. G. Altman, A. Laupacis, P. C. Gøtzsche, K. Krleža-Jerić, A. Hróbjartsson, H. Mann, K. Dickersin, J. A. Berlin, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. *Annals of Internal Medicine*, 158(3):200–207, 2013.

[152] M. K. Campbell, D. R. Elbourne, and D. G. Altman. CONSORT statement: extension to cluster randomised trials. *BMJ*, 328(7441):702–708, 2004.

[153] K. Suresh, S. V. Thomas, and G. Suresh. Design, data analysis and sampling techniques for clinical research. *Annals of Indian Academy of Neurology*, 14(4):287, 2011.

[154] P. Ravaud, R. Flipo, I. Boutron, C. Roy, A. Mahmoudi, B. Giraudeau, and T. Pham. ARTIST (osteoarthritis intervention standardized) study of standardised consultation versus usual care for patients with osteoarthritis of the knee in primary care in France: pragmatic randomised controlled trial. *BMJ*, 338, 2009.

[155] C. R. Ramsay, A. M. Grant, S. A. Wallace, P. Garthwaite, A. Monk, and I. Russell. Statistical assessment of the learning curves of health technologies. 2001.

[156] O. Papachristofi, D. Jenkins, and L. D. Sharples. Assessment of learning curves in complex surgical interventions: a consecutive case-series study. *Trials*, 17(1):266, 2016.

[157] S. Eldridge, S. Kerry, and D. J. Torgerson. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *BMJ*, 339, 2009.

[158] C. Easter, J. A. Thompson, S. Eldridge, M. Taljaard, and K. Hemming. Cluster randomised trials of individual-level interventions were at high risk of bias. *Journal of Clinical Epidemiology*, 2021.

[159] J. F. Bobb, H. Qiu, A. G. Matthews, J. McCormack, and K. A. Bradley. Addressing identification bias in the design and analysis of cluster-randomized pragmatic trials: a case study. *Trials*, 21(1):1–12, 2020.

[160] P. Logan, K. McCartney, S. Armstrong, A. Clarke, S. Conroy, J. Darby, J. Gladman, M. Godfrey, A. Gordon, L. Irvine, et al. Evaluation of the guide to action care home fall prevention programme in care homes for older people: protocol for a multi-centre, single blinded, cluster randomised controlled trial (finch). 2019.

[161] S. Hartley, R. Foy, R. E. Walwyn, R. Cicero, A. J. Farrin, J. J. Francis, F. Lorencatto, N. J. Gould, J. Grant-Casey, J. M. Grimshaw, et al. The evaluation of enhanced feedback interventions to reduce unnecessary blood transfusions (AFFINITIE): protocol for two linked cluster randomised factorial controlled trials. *Implementation Science*, 12(1):1–11, 2017.

[162] S. Shinde, H. A. Weiss, B. Varghese, P. Khandeparkar, B. Pereira, A. Sharma, R. Gupta, D. A. Ross, G. Patton, and V. Patel. Promoting school climate and health outcomes with the SEHER multi-component secondary school intervention in Bihar, India: a cluster-randomised controlled trial. *The Lancet*, 392(10163):2465–2477, 2018.

[163] J. Young, J. Green, A. Farrin, M. Collinson, S. Hartley, J. Smith, E. Teale, N. Siddiqi, and S. K. Inouye. A multicentre, pragmatic, cluster randomised, controlled feasibility trial of the pod system of care. *Age and Ageing*, 49(4):640–647, 2020.

[164] D. J. Torgerson. Contamination in trials: is cluster randomisation the answer? *BMJ*, 322(7282):355–357, 2001.

[165] K. Hemming, M. Taljaard, M. Moerbeek, and A. Forbes. Contamination: How much can an individually randomized trial tolerate? *Statistics in Medicine*, 40(14):3329–3351, 2021.

[166] J. Young, J. Green, M. Godfrey, J. Smith, F. Cheater, C. Hulme, M. Collinson, S. Hartley, S. Anwar, M. Fletcher, et al. The Prevention of Delirium system of

care for older patients admitted to hospital for emergency care: the POD research programme including feasibility RCT. *Programme Grants for Applied Research*, 9(4), 2021.

[167] T. E. Davison, M. P. McCabe, L. Busija, A. Graham, V. Camões-Costa, J. Kelly, and J. Byers. The effectiveness of the Program to Enhance Adjustment to Residential Living (PEARL) in reducing depression in newly admitted nursing home residents. *Journal of Affective Disorders*, 282:1067–1075, 2021.

[168] L. Graham, A. Ellwood, K. Hull, J. Fisher, B. Cundill, M. Holland, M. Goodwin, D. Clarke, R. Hawkins, C. Hulme, et al. A posture and mobility training package for care home staff: results of a cluster randomised controlled feasibility trial (the PATCH trial). *Age and Ageing*, 49(5):821–828, 2020.

[169] M. Halek, S. Reuther, R. Müller-Widmer, D. Trutschel, and D. Holle. Dealing with the behaviour of residents with dementia that challenges: A stepped-wedge cluster randomized trial of two types of dementia-specific case conferences in nursing homes (FallDem). *International Journal of Nursing Studies*, 104:103435, 2020.

[170] J. Hewitt, K. M. Refshauge, S. Goodall, T. Henwood, and L. Clemson. Does progressive resistance and balance exercise reduce falls in residential aged care? Randomized controlled trial protocol for the sunbeam program. *Clinical Interventions in Aging*, 9:369, 2014.

[171] B. Lichtwarck, G. Selbaek, Ø. Kirkevold, A. M. M. Rokstad, J. Š. Benth, J. Myhre, S. Nybakken, and S. Bergh. Time–targeted interdisciplinary model for evaluation and treatment of neuropsychiatric symptoms: protocol for an effectiveness-implementation cluster randomized hybrid trial. *BMC Psychiatry*, 16(1):1–12, 2016.

[172] D. K. Pasay, M. S. Guirguis, R. C. Shkrobot, J. P. Slobodan, A. S. Wagg, C. A. Sadowski, J. M. Conly, L. M. Saxinger, and L. C. Bresee. Antimicrobial stewardship in rural nursing homes: impact of interprofessional education and clinical decision tool implementation on urinary tract infection treatment in a cluster randomized trial. *Infection Control & Hospital Epidemiology*, 40(4):432–437, 2019.

[173] A. M. M. Rokstad, J. Røsvik, Ø. Kirkevold, G. Selbaek, J. S. Benth, and K. Engedal. The effect of person-centred dementia care to prevent agitation and other neuropsychiatric symptoms and enhance quality of life in nursing home patients: a 10-month randomized controlled trial. *Dementia and Geriatric Cognitive Disorders*, 36(5-6):340–353, 2013.

[174] E. L. Sampson, A. Feast, A. Blighe, K. Froggatt, R. Hunter, L. Marston, B. McCormack, S. Nurock, M. Panca, C. Powell, et al. Pilot cluster randomised trial of an evidence-based intervention to reduce avoidable hospital admissions in nursing home residents (Better Health in Residents of Care Homes with Nursing-BHiRCH-NH Study). *BMJ Open*, 10(12):e040732, 2020.

[175] J. K. Sluggett, E. Y. Chen, J. Ilomäki, M. Corlis, J. Van Emden, M. Hogan, T. Caporale, C. Keen, R. Hopkins, C. E. Ooi, et al. Reducing the burden of complex medication regimens: SImplification of Medications Prescribed to Long-tErm care Residents (SIMPLER) cluster randomized controlled trial. *Journal of the American Medical Directors Association*, 21(8):1114–1120, 2020.

[176] J. L. Campbell, N. Britten, C. Green, T. A. Holt, V. Lattimer, S. H. Richards, D. A. Richards, C. Salisbury, R. S. Taylor, and E. Fletcher. The effectiveness and cost-effectiveness of telephone triage of patients requesting same day consultations in general practice: study protocol for a cluster randomised controlled trial comparing nurse-led and GP-led management systems (ESTEEM). *Trials*, 14(1):1–19, 2013.

[177] T. Harris, S. M. Kerry, E. S. Limb, C. R. Victor, S. Iliffe, M. Ussher, P. H. Whincup, U. Ekelund, J. Fox-Rushby, C. Furness, et al. Effect of a Primary Care Walking Intervention with and without Nurse Support on Physical Activity Levels in 45-to 75-Year-Olds: The P edometer A nd C onsultation E valuation (PACE-UP) Cluster Randomised Clinical Trial. *PLoS Medicine*, 14(1):e1002210, 2017.

[178] G. I. Harrison, K. Murray, R. Gore, P. Lee, A. Sreedharan, P. Richardson, A. J. Hughes, M. Wiselka, W. Gelson, E. Unitt, et al. The hepatitis C awareness through to treatment (HepCATT) study: improving the cascade of care for hepatitis C virus-infected people who inject drugs in England. *Addiction*, 114(6):1113–1122, 2019.

[179] A. Heaven, P. Bower, B. Cundill, A. Farrin, M. Foster, R. Foy, S. Hartley, R. Hawkins, C. Hulme, S. Humphrey, et al. Study protocol for a cluster randomised controlled feasibility trial evaluating personalised care planning for older people with frailty: PROSPER V2 27/11/18. *Pilot and Feasibility Studies*, 6(1):1–11, 2020.

[180] H. Moore, C. D. Summerbell, D. C. Greenwood, P. Tovey, J. Griffiths, M. Henderson, K. Hesketh, S. Woolgar, and A. J. Adamson. Improving management of obesity in primary care: cluster randomised trial. *BMJ*, 327(7423):1085, 2003.

[181] R. Mullis, M. R. J. R. Aquino, S. N. Dawson, V. Johnson, S. Jowett, E. Kreit, and J. Mant. Improving primary care after stroke (IPCAS) trial: protocol of a

randomised controlled trial to evaluate a novel model of care for stroke survivors living in the community. *BMJ Open*, 9(8):e030285, 2019.

[182] T. A. Willis, M. Collinson, L. Glidewell, A. J. Farrin, M. Holland, D. Meads, C. Hulme, D. Petty, S. Alderson, S. Hartley, et al. An adaptable implementation package targeting evidence-based indicators in primary care: a pragmatic cluster-randomised evaluation. *PLoS Medicine*, 17(2):e1003045, 2020.

[183] K. E. Duhig, J. Myers, P. T. Seed, J. Sparkes, J. Lowe, R. M. Hunter, A. H. Shennan, L. C. Chappell, R. Bahl, G. Bambridge, et al. Placental growth factor testing to assess women with suspected pre-eclampsia: a multicentre, pragmatic, stepped-wedge cluster-randomised controlled trial. *The Lancet*, 393(10183):1807–1818, 2019.

[184] K. E. Harding, D. A. Snowdon, L. Prendergast, A. K. Lewis, B. Kent, S. F. Leggat, and N. F. Taylor. Sustainable waiting time reductions after introducing the STAT model for access and triage: 12-month follow up of a stepped wedge cluster randomised controlled trial. *BMC Health Services Research*, 20(1):1–9, 2020.

[185] R. E. Hopkins, T. Bui, A. H. Konstantatos, C. Arnold, D. J. Magliano, D. Liew, and M. J. Dooley. Educating junior doctors and pharmacists to reduce discharge prescribing of opioids for surgical patients: a cluster randomised controlled trial. *Medical Journal of Australia*, 213(9):417–423, 2020.

[186] X. Pourrat, C. Roux, B. Bouzige, V. Garnier, A. Develay, B. Allenet, M. Fraysse, J.-M. Halimi, J. Grassin, and B. Giraudeau. Impact of drug reconciliation at discharge and communication between hospital and community pharmacists on drug-related problems: study protocol for a randomized controlled trial. *Trials*, 15(1):1–7, 2014.

[187] S. P. Singh, H. Tuomainen, G. De Girolamo, A. Maras, P. Santosh, F. McNicholas, U. Schulze, D. Purper-Ouakil, S. Tremmery, T. Franić, et al. Protocol for a cohort study of adolescent mental health service users with a nested cluster randomised controlled trial to assess the clinical and cost-effectiveness of managed transition in improving transitions from child to adult mental health services (the MILESTONE study). *BMJ Open*, 7(10):e016055, 2017.

[188] S. A. Clemes, D. D. Bingham, N. Pearson, Y.-L. Chen, C. L. Edwardson, R. R. McEachan, K. Tolfrey, L. Cale, G. Richardson, M. Fray, et al. Stand out in class: Restructuring the classroom environment to reduce sitting time–findings from a pilot cluster randomised controlled trial. *International Journal of Behavioral Nutrition and Physical Activity*, 17(1):1–16, 2020.

[189] C. E. Evans, J. K. Ransley, M. S. Christian, D. C. Greenwood, J. D. Thomas, and J. E. Cade. A cluster-randomised controlled trial of a school-based fruit and vegetable intervention: Project Tomato. *Public Health Nutrition*, 16(6):1073–1081, 2013.

[190] R. Pechey, G. J. Hollands, and T. M. Marteau. Are meat options preferred to comparable vegetarian options? An experimental study. *BMC Research Notes*, 14(1):1–5, 2021.

[191] S. A. Clemes, V. V. Mato, F. Munir, C. L. Edwardson, Y.-L. Chen, M. Hamer, L. J. Gray, N. B. Jaicim, G. Richardson, V. Johnson, et al. Cluster randomised controlled trial to investigate the effectiveness and cost-effectiveness of a Structured Health Intervention For Truckers (the SHIFT study): a study protocol. *BMJ Open*, 9(11):e030175, 2019.

[192] M. Vasiljevic, E. Cartwright, M. Pilling, M.-M. Lee, G. Bignardi, R. Pechey, G. J. Hollands, S. A. Jebb, and T. M. Marteau. Impact of calorie labelling in worksite cafeterias: a stepped wedge randomised controlled pilot trial. *International Journal of Behavioral Nutrition and Physical Activity*, 15(1):1–12, 2018.

[193] N. Fairhall, S. E. Kurrle, C. Sherrington, S. R. Lord, K. Lockwood, B. John, N. Monaghan, K. Howard, and I. D. Cameron. Effectiveness of a multifactorial intervention on preventing development of frailty in pre-frail older people: study protocol for a randomised controlled trial. *BMJ Open*, 5(2):e007091, 2015.

[194] M. Neuman, P. Indravudh, R. Chilongosi, M. dElbée, N. Desmond, K. Fielding, B. Hensen, C. Johnson, P. Mkandawire, A. Mwinga, et al. The effectiveness and cost-effectiveness of community-based lay distribution of HIV self-tests in increasing uptake of HIV testing among adults in rural Malawi and rural and peri-urban Zambia: protocol for STAR (self-testing for Africa) cluster randomized evaluations. *BMC Public Health*, 18(1):1–12, 2018.

[195] S. Cox, A. Ford, J. Li, C. Best, A. Tyler, D. Robson, L. Bauld, P. Hajek, I. Uny, S. Parrott, et al. Exploring the uptake and use of electronic cigarettes provided to smokers accessing homeless centres: a four-centre cluster feasibility trial. *Public Health Research*, 2021.

[196] S. E. Lamb, S. Gates, M. A. Williams, E. M. Williamson, S. Mt-Isa, E. J. Withers, E. Castelnuovo, J. Smith, D. Ashby, M. W. Cooke, et al. Emergency department

treatments and physiotherapy for acute whiplash: a pragmatic, two-step, randomised controlled trial. *The Lancet*, 381(9866):546–556, 2013.

[197] A. R. Martineau, Y. Hanifa, K. D. Witt, N. C. Barnes, R. L. Hooper, M. Patel, N. Stevens, Z. Enayat, Z. Balayah, A. Syed, et al. Double-blind randomised controlled trial of vitamin D3 supplementation for the prevention of acute respiratory infection in older adults and their carers (ViDiFlu). *Thorax*, 70(10):953–960, 2015.

[198] L. Rodríguez-Mañas, A. J. Bayer, M. Kelly, A. Zeyfang, M. Izquierdo, O. Laosa, T. C. Hardman, and A. J. Sinclair. An evaluation of the effectiveness of a multimodal intervention in frail and pre-frail older people with type 2 diabetes-the MID-Frail study: study protocol for a randomised controlled trial. *Trials*, 15(1):1–9, 2014.

[199] R. Kipping, R. Jago, C. Metcalfe, J. White, A. Papadaki, R. Campbell, W. Hollingworth, D. Ward, S. Wells, R. Brockman, et al. NAP SACC UK: protocol for a feasibility cluster randomised controlled trial in nurseries and at home to increase physical activity and healthy eating in children aged 2–4 years. *BMJ Open*, 6(4), 2016.

[200] C. Bonell, E. Beaumont, M. Dodd, D. R. Elbourne, L. Bevilacqua, A. Mathiot, J. McGowan, J. Sturgess, E. Warren, R. M. Viner, et al. Effects of school environments on student risk-behaviours: evidence from a longitudinal study of secondary schools in England. *J Epidemiol Community Health*, 73(6):502–508, 2019.

[201] N. Siddiqi, F. Cheater, M. Collinson, A. Farrin, A. Forster, D. George, M. Godfrey, E. Graham, J. Harrison, A. Heaven, et al. The PiTSTOP study: a feasibility cluster randomized trial of delirium prevention in care homes for older people. *Age and Ageing*, 45(5):652–661, 2016.

[202] C. A. Surr, I. Holloway, R. E. Walwyn, A. W. Griffiths, D. Meads, A. Martin, R. Kelley, C. Ballard, J. Fossey, N. Burnley, et al. Effectiveness of dementia care mapping to reduce agitation in care home residents with dementia: an open-cohort cluster randomised controlled trial. *Aging & Mental Health*, pages 1–14, 2020.

[203] M. Tadrous, K. Fung, L. Desveaux, T. Gomes, M. Taljaard, J. M. Grimshaw, C. M. Bell, and N. M. Ivers. Effect of academic detailing on promoting appropriate prescribing of antipsychotic medication in nursing homes: a cluster randomized clinical trial. *JAMA Network Open*, 3(5):e205724–e205724, 2020.

[204] C. Bonell, E. Allen, E. Warren, J. McGowan, L. Bevilacqua, F. Jamal, Z. Sadique, R. Legood, M. Wiggins, C. Opondo, et al. Modifying the secondary school environment to reduce bullying and aggression: the INCLUSIVE cluster RCT. *Public Health Research*, 7(18):1–164, 2019.

[205] L. Desveaux, T. Gomes, M. Tadrous, L. Jeffs, M. Taljaard, J. Rogers, C. M. Bell, and N. M. Ivers. Appropriate prescribing in nursing homes demonstration project (APDP) study protocol: pragmatic, cluster-randomized trial and mixed methods process evaluation of an Ontario policy-maker initiative to improve appropriate prescribing of antipsychotics. *Implementation Science*, 11(1):1–10, 2015.

[206] J. A. Sterne, J. Savović, M. J. Page, R. G. Elbers, N. S. Blencowe, I. Boutron, C. J. Cates, H.-Y. Cheng, M. S. Corbett, S. M. Eldridge, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*, 366, 2019.

[207] R. Hooper and A. Copas. Stepped wedge trials with continuous recruitment require new ways of thinking. *Journal of Clinical Epidemiology*, 116:161–166, 2019.

[208] C. Bova, C. Jaffarian, S. Crawford, J. B. Quintos, M. Lee, and S. Sullivan-Bolyai. Intervention fidelity: Monitoring drift, providing feedback and assessing the control condition. *Nursing Research*, 66(1):54, 2017.

[209] T. C. Hoffmann, P. P. Glasziou, I. Boutron, R. Milne, R. Perera, D. Moher, D. G. Altman, V. Barbour, H. Macdonald, M. Johnston, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*, 348, 2014.

[210] S. M. Eldridge, D. Ashby, and S. Kerry. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology*, 35(5):1292–1300, 2006.

[211] F. Steele. Module 5 (Concepts): Introduction to Multilevel Modelling. LEMMA VLE, Centre for Multilevel Modelling. University of Bristol. `https://www.cmm.bris.ac.uk/lemma/`, 2008.

[212] C. Thomadakis, L. Meligkotsidou, N. Pantazis, and G. Touloumi. Longitudinal and time-to-drop-out joint models can lead to seriously biased estimates when the drop-out mechanism is at random. *Biometrics*, 75(1):58–68, 2019.

[213] C. for Human Medicinal Products. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials, Step 5. 2020.

[214] L. Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

[215] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[216] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021.

[217] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team. nlme: Linear and nonlinear mixed effects models. `https://CRAN.R-project.org/package=nlme`, 2020. R package version 3.1-147.

[218] M. Matsumoto and T. Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30, 1998.

[219] B. C. Kahan, T. P. Morris, I. R. White, C. D. Tweed, S. Cro, D. Dahly, T. M. Pham, H. Esmail, A. Babiker, and J. R. Carpenter. Treatment estimands in clinical trials of patients hospitalised for COVID-19: ensuring trials ask the right questions. *BMC Medicine*, 18(1):1–8, 2020.

[220] I. R. White, J. Carpenter, and N. J. Horton. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clinical Trials*, 9(4):396–407, 2012.

[221] J. G. Ibrahim and G. Molenberghs. Missing data methods in longitudinal studies: a review. *Test*, 18(1):1–43, 2009.

[222] A. Lawrence Gould, M. E. Boye, M. J. Crowther, J. G. Ibrahim, G. Quartey, S. Micallef, and F. Y. Bois. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine*, 34(14):2181–2195, 2015.

[223] R. M. Elashoff, G. Li, and N. Li. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64(3):762–771, 2008.

[224] A. A. Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

[225] M. J. Sweeting and S. G. Thompson. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53(5):750–763, 2011.

[226] A. A. Tsiatis, V. Degruttola, and M. S. Wulfsohn. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, 90(429):27–37, 1995.

[227] D. Rizopoulos. *Joint models for longitudinal and time-to-event data: With applications in R.* CRC press, 2012.

[228] I. Sousa. A review on joint modelling of longitudinal measurements and time-to-event. *Revstat Stat J*, 9:57–81, 2011.

[229] C. Roberts and R. Walwyn. Design and analysis of non-pharmacological treatment trials with multiple therapists per patient. *Statistics in Medicine*, 32(1):81–98, 2013.

[230] H. Goldstein, S. Burgess, and B. McConnell. Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4):941–954, 2007.

[231] S. L. Brilleman, M. J. Crowther, M. Moreno-Betancur, J. Buros Novik, J. Dunyak, N. Al-Huniti, R. Fox, J. Hammerbacher, and R. Wolfe. Joint longitudinal and time-to-event models for multilevel hierarchical data. *Statistical Methods in Medical Research*, 28(12):3502–3515, 2019.

[232] J. G. Ibrahim, H. Chu, and L. M. Chen. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796, 2010.

[233] M. Sudell, R. Kolamunnage-Dona, and C. Tudur-Smith. Joint models for longitudinal and time-to-event data: a review of reporting quality with a view to meta-analysis. *BMC Medical Research Methodology*, 16(1):1–11, 2016.

[234] D. Rizopoulos. JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33, 2010.

[235] P. Philipson, I. Sousa, P. J. Diggle, P. Williamson, R. Kolamunnage-Dona, R. Henderson, and G. L. Hickey. *joineR: Joint Modelling of Repeated Measurements and Time-to-Event Data*, 2018. R package version 1.2.6.

[236] S. Brilleman, M. Crowther, M. Moreno-Betancur, J. Buros Novik, and R. Wolfe. Joint longitudinal and time-to-event models via Stan. *Proceedings of StanCon 2018*, 2018.

[237] M. J. Crowther. STJM: Stata module to fit shared parameter joint models of longitudinal and survival data. `https://EconPapers.repec.org/RePEc:boc:bocode:s457502`, 2013. Boston College Department of Economics.

[238] M. J. Crowther. merlin - A unified modeling framework for data analysis and methods development in Stata. *The Stata Journal*, 20(4):763–784, 2020.

[239] StataCorp. *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC, 2019.

[240] G. Cafri, D. Hedeker, and G. A. Aarons. An introduction and integration of cross-classified, multiple membership, and dynamic group random-effects models. *Psychological Methods*, 20(4):407, 2015.

[241] G. Leckie. Multiple Membership Multilevel Models - Stata Practical. LEMMA VLE Module 13, 1-42. Centre for multilevel modelling. University of Bristol. `http://www.bristol.ac.uk/cmm/learning/course.html`, 2013.

[242] M. J. Crowther, T. M.-L. Andersson, P. C. Lambert, K. R. Abrams, and K. Humphreys. Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification. *Statistics in Medicine*, 35(7):1193–1209, 2016.

[243] A. García-Hernandez, T. Pérez, M. d. C. Pardo, and D. Rizopoulos. Mmrm vs joint modeling of longitudinal responses and time to study drug discontinuation in clinical trials using a “de jure” estimand. *Pharmaceutical Statistics*, 19(6):909–927, 2020.

[244] M. W. Arisido, L. Antolini, D. P. Bernasconi, M. G. Valsecchi, and P. Rebora. Joint model robustness compared with the time-varying covariate Cox model to evaluate the association between a longitudinal marker and a time-to-event endpoint. *BMC Medical Research Methodology*, 19(1):1–13, 2019.

[245] D. C. Miller, S. MaWhinney, J. L. Patnaik, K. L. Christopher, A. M. Lynch, and B. D. Wagner. Predictors of refraction prediction error after cataract surgery: a shared parameter model to account for missing post-operative measurements. *Statistical Methods & Applications*, pages 1–22, 2021.

[246] C. Leyrat, A. Caille, A. Donner, and B. Giraudeau. Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study. *Statistics in Medicine*, 32(19):3357–3372, 2013.

[247] J. Moberg and M. Kramer. A brief history of the cluster randomised trial design. *Journal of the Royal Society of Medicine*, 108(5):192–198, 2015.

[248] R. Booy, R. I. Lindley, D. E. Dwyer, J. K. Yin, L. G. Heron, C. R. Moffatt, C. K. Chiu, A. E. Rosewell, A. S. Dean, T. Dobbins, et al. Treating and preventing influenza in aged care facilities: a cluster randomised controlled trial. 2012.

[249] K. Chami, G. Gavazzi, A. Bar-Hen, F. Carrat, B. de Wazières, B. Lejeune, N. Armand, M. Rainfray, J. Hajjar, F. Piette, et al. A short-term, multicomponent infection control program in nursing homes: a cluster randomized controlled trial. *Journal of the American Medical Directors Association*, 13(6):569–e9, 2012.

[250] Y.-H. Chen and L.-C. Lin. Ability of the pain recognition and treatment (PRT) protocol to reduce expressions of pain among institutionalized residents with dementia: a cluster randomized controlled trial. *Pain Management Nursing*, 17(1):14–24, 2016.

[251] S.-T. Cheng, P. K. Chow, C. Edwin, and A. C. Chan. Leisure activities alleviate depressive symptoms in nursing home residents with very mild or mild dementia. *The American Journal of Geriatric Psychiatry*, 20(10):904–908, 2012.

[252] C. S. Colón-Emeric, K. Corazzini, E. S. McConnell, W. Pan, M. Toles, R. Hall, M. P. Cary, M. Batchelor-Murphy, T. Yap, A. L. Anderson, et al. Effect of promoting high-quality staff interactions on fall prevention in nursing homes: a cluster-randomized trial. *JAMA Internal Medicine*, 177(11):1634–1641, 2017.

[253] M. J. Connolly, M. Boyd, J. B. Broad, N. Kerse, T. Lumley, N. Whitehead, and S. Foster. The Aged Residential Care Healthcare Utilization Study (ARCHUS): a multidisciplinary, cluster randomized controlled trial designed to reduce acute avoidable hospitalizations from long-term care facilities. *Journal of the American Medical Directors Association*, 16(1):49–55, 2015.

[254] T. E. Davison, G. Karantzas, D. Mellor, M. P. McCabe, and D. Mrkic. Staff-focused interventions to increase referrals for depression in aged care facilities: A cluster randomized controlled trial. *Aging & Mental Health*, 17(4):449–455, 2013.

[255] L. De Visschere, J. Schols, G.-J. van der Putten, C. de Baat, and J. Vanobbergen. Effect evaluation of a supervised versus non-supervised implementation of an oral health care guideline in nursing homes: a cluster randomised controlled clinical trial. *Gerodontology*, 29(2):e96–e106, 2012.

[256] M. Ersek, M. B. Neradilek, K. Herr, A. Jablonski, N. Polissar, and A. Du Pen. Pain management algorithms for implementing best practices in nursing homes: Results of a randomized controlled trial. *Journal of the American Medical Directors Association*, 17(4):348–356, 2016.

[257] E. Galik, B. Resnick, N. Lerner, M. Hammersla, and A. L. Gruber-Baldini. Function focused care for assisted living residents with dementia. *The Gerontologist*, 55(Suppl_1):S13–S26, 2015.

[258] S. Gravenstein, H. E. Davidson, M. Taljaard, J. Ogarek, P. Gozalo, L. Han, and V. Mor. Comparative effectiveness of high-dose versus standard-dose influenza vaccination on numbers of US nursing home residents admitted to hospital: a cluster-randomised trial. *The Lancet Respiratory Medicine*, 5(9):738–746, 2017.

[259] J. Hewitt, S. Goodall, L. Clemson, T. Henwood, and K. Refshauge. Progressive resistance and balance training for falls prevention in long-term residential aged care: a cluster randomized trial of the sunbeam program. *Journal of the American Medical Directors Association*, 19(4):361–369, 2018.

[260] M. Hödl, R. J. Halfens, and C. Lohrmann. Effectiveness of conservative urinary incontinence management among female nursing home residents - A cluster RCT. *Archives of Gerontology and Geriatrics*, 81:245–251, 2019.

[261] N. Jøranson, I. Pedersen, A. M. M. Rokstad, and C. Ihlebaek. Effects on symptoms of agitation and depression in persons with dementia participating in robot-assisted activity: a cluster-randomized controlled trial. *Journal of the American Medical Directors Association*, 16(10):867–873, 2015.

[262] J. Kerr, D. Rosenberg, R. A. Millstein, K. Bolling, K. Crist, M. Takemoto, S. Godbole, K. Moran, L. Natarajan, C. Castro-Sweet, et al. Cluster randomized controlled

trial of a multilevel physical activity intervention for older adults. *International Journal of Behavioral Nutrition and Physical Activity*, 15(1):1–9, 2018.

[263] F. Könner, A. Budnick, R. Kuhnert, I. Wulff, S. Kalinowski, P. Martus, D. Dräger, and R. Kreutz. Interventions to address deficits of pharmacological pain management in nursing home residents–a cluster-randomized trial. *European Journal of Pain*, 19(9):1331–1341, 2015.

[264] S. Köpke, I. Mühlhauser, A. Gerlach, A. Haut, B. Haastert, R. Möhler, and G. Meyer. Effect of a guideline-based multicomponent intervention on use of physical restraints in nursing homes: a randomized controlled trial. *JAMA*, 307(20):2177–2184, 2012.

[265] J. Kuck, M. Pantke, and U. Flick. Effects of social activation and physical mobilization on sleep in nursing home residents. *Geriatric Nursing*, 35(6):455–461, 2014.

[266] K. L. Lapane, C. M. Hughes, L. A. Daiello, K. A. Cameron, and J. Feinberg. Effect of a pharmacist-led multicomponent intervention focusing on the medication monitoring phase to prevent potential adverse drug events in nursing homes. *Journal of the American Geriatrics Society*, 59(7):1238–1245, 2011.

[267] W. Leslie, M. Woodward, M. Lean, H. Theobald, L. Watson, and C. Hankey. Improving the dietary intake of under nourished older people in residential care homes using an energy-enriching food approach: a cluster randomised controlled study. *Journal of Human Nutrition and Dietetics*, 26(4):387–394, 2013.

[268] L.-F. Low, H. Brodaty, B. Goodenough, P. Spitzer, J.-P. Bell, R. Fleming, A.-N. Casey, Z. Liu, and L. Chenoweth. The Sydney Multisite Intervention of LaughterBosses and ElderClowns (SMILE) study: cluster randomised trial of humour therapy in nursing homes. *BMJ Open*, 3(1):e002072, 2013.

[269] A.-G. Mamhidir, B.-M. Sjölund, B. Fläckman, A. Wimo, A. Sköldunger, and M. Engström. Systematic pain assessment in nursing homes: a cluster-randomized trial using mixed-methods approach. *BMC Geriatrics*, 17(1):1–16, 2017.

[270] S. Meeks, K. Van Haitsma, B. Schoenbachler, and S. W. Looney. BE-ACTIV for depression in nursing homes: primary outcomes of a randomized clinical trial. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 70(1):13–23, 2015.

[271] S. L. Mitchell, M. L. Shaffer, S. Cohen, L. C. Hanson, D. Habtemariam, and A. E. Volandes. An advance care planning video decision support tool for nursing home residents with advanced dementia: a cluster randomized clinical trial. *JAMA Internal Medicine*, 178(7):961–969, 2018.

[272] W. Moyle, C. J. Jones, J. E. Murfield, L. Thalib, E. R. Beattie, D. K. Shum, S. T. O'Dwyer, M. C. Mervin, and B. M. Draper. Use of a robotic seal as a therapeutic tool to improve dementia symptoms: a cluster-randomized controlled trial. *Journal of the American Medical Directors Association*, 18(9):766–773, 2017.

[273] C. Olsen, I. Pedersen, A. Bergland, M.-J. Enders-Slegers, G. Patil, and C. Ihlebæk. Effect of animal-assisted interventions on depression, agitation and quality of life in nursing home residents suffering from cognitive impairment or dementia: A cluster randomized controlled trial. *International Journal of Geriatric Psychiatry*, 31(12):1312–1321, 2016.

[274] E. O'Shea, D. Devane, A. Cooney, D. Casey, F. Jordan, A. Hunter, E. Murphy, J. Newell, S. Connolly, and K. Murphy. The impact of reminiscence on the quality of life of residents with dementia in long-stay care. *International Journal of Geriatric Psychiatry*, 29(10):1062–1070, 2014.

[275] M. A. Rapp, T. Mell, T. Majic, Y. Treusch, J. Nordheim, M. Niemann-Mirmehdi, H. Gutzmann, and A. Heinz. Agitation in nursing home residents with dementia (VIDEANT trial): effects of a cluster-randomized, controlled, guideline implementation trial. *Journal of the American Medical Directors Association*, 14(9):690–695, 2013.

[276] L. M. Roets-Merken, S. U. Zuidema, M. J. Vernooij-Dassen, S. Teerenstra, P. G. Hermsen, G. I. Kempen, and M. J. Graff. Effectiveness of a nurse-supported self-management programme for dual sensory impaired older adults in long-term care: a cluster randomised controlled trial. *BMJ Open*, 8(1):e016674, 2018.

[277] A. J. Sinclair, A. J. Girling, R. Gadsby, I. Bourdel-Marchasson, and A. J. Bayer. Diabetes in care homes: a cluster randomised controlled trial of resident education. *The British Journal of Diabetes & Vascular Disease*, 12(5):238–242, 2012.

[278] I. Testad, T. E. Mekki, O. Førland, C. Øye, E. M. Tveit, F. Jacobsen, and Ø. Kirkevold. Modeling and evaluating evidence-based continuing education program in nursing home dementia care (MEDCED) - training of care home staff to

reduce use of restraint in care home residents with dementia. A cluster randomized controlled trial. *International Journal of Geriatric Psychiatry*, 31(1):24–32, 2016.

[279] G.-J. van der Putten, J. Mulder, C. de Baat, L. M. De Visschere, J. N. Vanobbergen, and J. M. Schols. Effectiveness of supervised implementation of an oral health care guideline in care homes; a single-blinded cluster randomized controlled trial. *Clinical Oral Investigations*, 17(4):1143–1153, 2013.

[280] K. N. Williams, Y. Perkhounkova, R. Herman, and A. Bossen. A communication intervention to reduce resistiveness in dementia care: A cluster randomized controlled trial. *The Gerontologist*, 57(4):707–718, 2017.

[281] K. Yokoi, K. Yoshimasu, S. Takemura, J. Fukumoto, S. Kurasawa, and K. Miyashita. Short stick exercises for fall prevention among older adults: a cluster randomized trial. *Disability and Rehabilitation*, 37(14):1268–1276, 2015.

[282] S. Black-Tiong, D. Gonzalez-Chica, and N. Stocks. Trends in long-term opioid prescriptions for musculoskeletal conditions in Australian general practice: a national longitudinal study using MedicineInsight, 2012–2018. *BMJ Open*, 11(4):e045418, 2021.

[283] C. de Burgos-Lunar, I. del Cura-Gonzalez, M. A. Salinero-Fort, P. Gomez-Campelo, L. Perez de Isla, and R. Jimenez-Garcia. Delayed diagnosis of hypertension in diabetic patients monitored in primary care. *Revista Espanola de Cardiologia*, 66(9):700–6, 2013.

[284] J. L. Fernandez-Martin, P. Martinez-Camblor, M. P. Dionisi, J. Floege, M. Ketteler, G. London, F. Locatelli, J. L. Gorriz, B. Rutkowski, A. Ferreira, W. J. Bos, A. Covic, M. Rodriguez-Garcia, J. E. Sanchez, D. Rodriguez-Puyol, J. B. Cannata-Andia, and C. group. Improvement of mineral and bone metabolism markers is associated with better survival in haemodialysis patients: the COSMOS study. *Nephrology Dialysis Transplantation*, 30(9):1542–51, 2015.

[285] K. Islam, T. Anggondowati, P. Deviany, J. Ryan, A. Fetrick, D. Bagenda, M. Copur, A. Tolentino, I. Vaziri, H. McKean, et al. Patient preferences of chemotherapy treatment options and tolerance of chemotherapy side effects in advanced stage lung cancer. *BMC Cancer*, 19(1):1–9, 2019.

[286] A. Ntouva, K. A. Toulis, D. Keerthy, N. J. Adderley, W. Hanif, R. Thayakaran, K. Gokhale, G. N. Thomas, K. Khunti, A. A. Tahrani, and K. Nirantharakumar.

Hypoglycaemia is associated with increased risk of fractures in patients with type 2 diabetes mellitus: A cohort study. *European Journal of Endocrinology*, 180(1):51–58, 2019.

[287] S. S. Oguz, H. G. Kanmaz, and U. Dilmen. Off-label and unlicensed drug use in neonatal intensive care units in Turkey: the old-inn study. *International Journal of Clinical Pharmacy*, 34(1):136–141, 2012.

[288] R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

[289] W. Gould, J. Pitblado, and W. Sribney. *Maximum likelihood estimation with Stata.* Stata press, 2006.