Sample size calculations for cluster randomised trials, with a focus on ordinal outcomes.

Robinson, Clare Marie

Sample size calculations for cluster randomised trials, with a focus on ordinal outcomes

Clare Marie Robinson (née Rutterford)
School of Medicine and Dentistry,
Queen Mary University of London,

Thesis submitted in partial fulfilment of the requirements of,
the Degree of Doctor of Philosophy (PhD)

I, Clare Marie Robinson (née Rutterford), confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Date: 19th April 2016

Details of collaboration and publications:

Some aspects of the work described in Chapter Three were conducted collaboratively with Monica Taljaard (MT) and Stephanie Dixon (SDX) who were based at the Ottawa Hospital Research Institute at Ottawa Hospital and the Schulich School of Medicine and Dentistry at the University of Western Ontario in Canada respectively. My role in the collaboration is clearly described within Chapter Three. I took the lead role in the analysis of this data and preparing and submitting the work for the following publication.

Rutterford C, Taljaard M, Dixon S, Copas A and Eldridge S. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *J Clin Epidemiol.* 2015; 68: 716-23.

My literature review of sample size methods for cluster randomised trials conducted in Chapter Two has been published.

Rutterford, C. and Copas, A. and Eldridge, S. Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology.* 2015; 1051-1067

Abstract

Background

A common approach to sample size calculation for cluster randomised trials (CRTs) is to calculate the sample size assuming individual randomisation and multiply it by an inflation factor, the design effect. This approach is well established for binary and continuous outcomes, but less so for ordinal. As the variety in trial design increases alternative or more complex methods are required. There is currently no single resource that provides a comprehensive summary of methods. This thesis aims to provide a unique contribution towards the review and development of sample size methods for CRTs, with a focus on ordinal outcomes.

Methods

I provide a comprehensive review of sample size methods for CRTs and summarise the methodological gaps that remain. Through simulation I evaluate the power performance, under realistic trial scenarios, of the design effect for ordinal outcomes calculated using a kappa-type intracluster correlation coefficient (ICC), the ICC on an assumed underlying variable and an ANOVA ICC. I provide practical guidance for sample size calculation for ordinal outcomes in CRTs.

Results

Simulation results showed when the number of clusters was large the ANOVA and kappa-type estimates were equivalent, and smaller than the latent variable ICC. Use of the ANOVA ICC in the design effect produced adequately powered trials and power was marginally reduced under a minor deviation from the common assumption of proportional odds used in ordered regression.

Conclusions

For outcomes with three to five categories the ANOVA ICC, calculated by assigning numerical equally spaced scores to the ordinal categories, can be used in the simple design effect to produce an adequately powered trial. The method assumes an analysis by random effects ordered regression with proportional odds, a reasonable number of clusters, and clusters of the same size.

Acknowledgments

I am extremely thankful to the support and guidance I have received from my supervisors Sandra Eldridge and Andrew Copas. I am also grateful to Sally Kerry and Graham Foster who acted as examiners on my progression panel throughout the entire PhD and offered encouragement and input into the direction of this research.

This research was made possible by the many researchers involved in cluster randomised trials whose valuable input and comments shaped the content of this thesis. I thank Allan Donner, Obi Ukoumunne, Mike Campbell, Steven Teerenstra and Gerard van Breukelen for their feedback on the coverage of my review of sample size methods and to Sally Kerry who provided valuable feedback on an early draft of the publication of this review. In 2013 I was lucky enough to visit Stephen Teerenstra and Esther de Hoop at Radbound University Medical Centre, Nijmegen in the Netherlands. This provided a great opportunity to learn more about the use of stepped wedge designs and be involved in some collaborative work to look at sample size development in this area. Also in 2013 I met with Monica Taljaard and her colleagues based at the Ottawa Hospital Research Institute at Ottawa Hospital in Ottawa Canada followed by a visit to meet Allan Donner and colleagues at the Schulich School of Medicine and Dentistry at the University of Western Ontario. I would like to thank Monica for her hospitality during my trip and for her invaluable guidance, enthusiasm and encouragement during our collaboration. I would also like to thank all those who I met and discussed cluster randomised trials with over the course of this research for their contributions.

I am grateful to the responses I received from researchers I contacted about their trials or methodology. In particular Alan Agresti for his input regarding the generation of clustered ordinal data, Hae-Young Kim who kindly shared his SAS code for implementing his method of sample size calculation for clustered ordinal outcomes and Hakan Demirtas who corresponded with me about his method for generating clustered ordinal data.

Finally I would like to thank all the statisticians at the Pragmatic Clinical Trials Unit at Queen Mary University of London who have provided valuable input and feedback on my work. In particular Stephen Bremner who reviewed the first two chapters of this thesis and Neil Wright who reviewed my simulation code. I am also grateful to the Medical Research Council who funded this work.

I have always been a keen long distance runner and I see many similarities between this hobby and undertaking this research. Both require perseverance, dedication, sacrifice and a large amount of endurance. I dedicate this thesis to all those who have helped me bring these attributes to my PhD and reach the finish line. In particular my parents Anne and Frederick Rutterford for showing unfaltering faith in my abilities throughout my life and my dad for providing the inspiration to undertake a PhD. I thank my sister Sarah De Pian for her support and many words of encouragement along the way. I am forever grateful to my husband Jonathan whose love and support made this PhD both possible and enjoyable.

Contents

List of Figures

Page

1 Introduction

2 Sample size methods for CRTs

4 Analysis of ordinal outcomes

5 Simulation studies

List of Tables

Page

1 Introduction

2 Sample size methods for CRTs

3 Ordinal outcomes in CRTs

4 Analysis of ordinal outcomes

5 Simulation studies

6 Remaining methodological gaps in sample size methods for CRTs

Abbreviations

AC Andrew Copas

ANOVA Analysis Of Variance

CACE Complier Average Causal Effect

CI Confidence Interval

COMET Core Outcome Measures in Effectiveness Trials

CONSORT CONsolidated Standards Of Reporting Trials

CR Clare Robinson

CRT Cluster Randomised Trial

CS Cluster Specific

DE Design Effect

GEE Generalised Estimating Equation

GP General Practitioner

HPC High Performance Computing

ICC Intracluster Correlation Coefficient

ITT Intention-to-treat

MAR Missing at random

MCAR Missing completely at random

MeSH Medical Subject Heading

MT Monica Taljaard

NHS National Health Service

OR Odds Ratio

PA Population Averaged

PCTU Pragmatic Clinical Trials Unit

PP Per-Protocol

PROM Patient Reported Outcome Measure

QMUL Queen Mary University of London

RCT Randomised Controlled Trial

RE Relative efficiency

SD Standard Deviation

SDX Stephanie Dixon

SE Sandra Eldridge

VIF Variance Inflation Factor

Notation

Subscripts

i Represents clusters, $i = 1, 2, \dots C$

j Represents individuals, $j = 1, 2, \dots N$

1 Represents intervention arm

2 Represents control arm

Clusters

n Cluster size, when assumed constant

n_i The size of cluster i, $i = 1, \dots, C$

$\bar{n}$ The arithmetic mean of cluster size

s_n The standard deviation of cluster sizes

CV The coefficient of variation in cluster size,$s_n/\bar{n}$

C The total number of clusters

c The number of clusters in each treatment group, under equal allocation

c_1, c_2 The number of clusters in the intervention and control arms

Statistics and parameters

m The number of individuals required per treatment group

δ The minimally clinically important difference between treatments

Δ Standardised effect size δ/σ

σ^2 The total variance in the outcome

s^2 The sample variance in the outcome

π_1, π_2 The probability of an event in the intervention and control arms

$\bar{x}_1, \bar{x}_2$ Sample means of a continuous outcome in the intervention and control arms

μ_1, μ_2 Population means of a continuous outcome in the intervention and control arms

θ Hazard ratio

σ_w^2 Within-cluster variance

Measures of between-cluster variability

ρ Intracluster correlation coefficient

k The coefficient of variation in outcome

k_m The coefficient of variation in outcome within matched pairs

σ_b^2 Between-cluster variance

Standard Normal Distribution

z_x The x'th percentage point of the standard Normal distribution

Hypothesis testing

H_0, H_1 The null and alternative hypotheses

β Probability of a Type II error

α Probability of a Type I error

Ordinal outcomes

k The number of ordinal categories with each ordinal category $q = 1, 2, \ldots, k$

π_q The proportion expected in each ordinal category $q = 1, 2, \ldots, k$

Glossary

Cluster Randomisation The process by which groups of individuals are randomised to each arm of the trial

Cluster size The sample of the cluster that is to be included in the analysis, which may or may not be the entire cluster

Cluster-level analysis A summary statistic is calculated for each cluster and then standard statistical analysis methods applied

Coefficient of variation in outcome An alternative measure to the ICC for estimating the between-cluster variation

Cohort design The individuals sampled from the cluster at baseline are then measured again at subsequent time points

CONSORT The CONSORT statement consists of a 25-item checklist for improving and standardising the reporting of clinical trials

Cross sectional design A different sample within the cluster is selected for measurement at each time point

Design Effect An inflation factor by which the sample size calculated under individual randomisation is inflated by to account for randomisation by cluster

Individual-level analysis Analysis is conducted at the level of the individual suitably adjusted to account for clustering

Intracluster correlation coefficient The proportion of the overall variation in outcome that can be accounted for by the between-cluster variation

Matched pair design Clusters are paired based on similarities such as cluster size, one cluster from the matched pair is allocated to control and the other to intervention

Ordinal outcome An outcome which consists of a set of categories which can be ordered or ranked

Sample size The total number of participants required that will provide sufficient probability of detecting a clinically important treatment difference if such a difference exists

SAS Data analysis and statistical software

Stata Data analysis and statistical software

Stepped wedge design A design where all clusters receive the control at baseline. At points in the trial one of more clusters will cross-over to receive the intervention, with all clusters receiving the treatment intervention by the end of the trial

Chapter 1

Introduction

The methodology for conducting sample size calculations for individually randomised trials is well established. Methods for cluster randomised trials are less developed and can be more complex due to the clustering inherent to the design. As I show in this chapter the use of cluster randomised trials has increased in recent years and hence there is now more methodological development around the design, conduct and analysis of these trials. The focus of this thesis is to provide a unique contribution towards the review and development of sample size methods for cluster randomised trials.

Many methods of sample size calculation for cluster randomised trials are extensions of methods used for individually randomised trials. Hence this chapter starts with a description of sample size formulae for individually randomised trials with explanation of how these are derived in general.

Cluster randomised trials are the focus of the remainder of the chapter, including the rationale behind their use and a description of the added complexity they present for sample size calculation. The most common approach to their sample size calculation is presented and its limitations described.

To place my research in context I provide a brief review of the historical development of sample size calculations for cluster randomised trials. The chapter concludes with a detailed outline of the aims, objectives and structure of the thesis.

1.1 Randomised Controlled Trials (RCTs)

Well designed and conducted Randomised Controlled Trials (RCTs) are seen as the gold standard in research design for the evaluation of medical interventions to improve participant outcomes. In the most common design individuals are randomly allocated, with pre-defined probability, to treatment group and their outcomes compared at a pre-defined time point post-randomisation. The design usually involves two groups; those that receive the intervention under investigation and those that receive a control (either the currently accepted standard treatment, no treatment, or a placebo). This is known as a standard two-arm parallel group design.

1.2 Sample size calculations for RCTs

The aim of an RCT is to provide reliable evidence of whether a treatment is safe and efficacious. In designing an RCT we must calculate in advance the number of participants required that will provide sufficient probability of detecting a clinically important treatment difference if such a difference exists, while also providing reasonable evidence to conclude no treatment difference if none is seen. An appropriately sized trial allows one to reliably conclude that any difference seen between the treatment groups can be attributed to a real effect of treatment rather than to chance.

At a minimum calculations require knowledge about 1) the primary outcome measure and a description of its variability, 2) an estimate of the minimally important treatment difference, 3) the analysis to be performed i.e. the null and alternative hypotheses with associated test statistic, 4) the Type I error to be tolerated and, 5) the amount of statistical power required.[1]

The following formulae can be used for sample size calculation for different data types in a two-arm, parallel group individually randomised trial. The number of individuals required per treatment group is denoted m.

1.2.1 Continuous outcomes

$$m = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 2\sigma^2}{\delta^2} \tag{1.1}$$

Where Z_x is the x'th percentage point of the standard normal distribution, δ represents the minimally clinically important difference in treatment means, $\mu_1 - \mu_2$, and σ^2 the variance in the outcome.

1.2.2 Binary outcomes

$$m = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{\delta^2} \tag{1.2}$$

Where π_1 is the probability of an event in the experimental group and π_2 the probability of an event in the control group and δ represents the minimally clinically important difference in treatment proportions, $\pi_1 - \pi_2$.

1.2.3 Ordinal outcomes

For an ordered categorical outcome with k levels, $q = 1, 2, \dots k$ the formula for m is given by Whitehead[2]

$$m = \frac{6(z_{1-\alpha/2} + z_{1-\beta})^2/(logOR)^2}{1 - \sum_{q=1}^{k} \bar{\pi}_q^3} \tag{1.3}$$

$\bar{\pi}_q$ is the mean proportion expected in category q and OR is the odds ratio of a patient being in a given category or less, in one group compared to the other. Whitehead assumes the lower the category the better the outcome.

1.2.4 Time-to-event outcomes

For time-to-event outcomes there are two commonly used formulae. The first, as defined by Schoenfeld is:[3]

$$m = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{log_e^2\theta(1-\pi_c)} \tag{1.4}$$

Where θ denotes the hazard ratio of group 1 over group 2 and π_c the probability of being censored.

The second formula, defined by Freedman,[4] provides an estimate of the total number of events required, E

$$E = (z_{1-\alpha/2} + z_{1-\beta})^2 \frac{(1+\theta)^2}{(1-\theta)^2} \tag{1.5}$$

1.2.5 Count outcomes

$$m = \frac{[z_{\alpha/2}\sqrt{2} + z_{\beta}\sqrt{[1+e^{-\tilde{b}}]}]^2}{e^{\beta_0}\tilde{b}^2} \tag{1.6}$$

Where β_0 represents the event rate in the control group and b is the model parameter related to treatment effect.

1.2.6 Rate outcomes

The number of person years required in each arm is given by

$$m = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2(\lambda_1 + \lambda_2)}{(\lambda_1 - \lambda_2)^2} \tag{1.7}$$

Where λ_1 and λ_2 are the expected rates in each arm.

1.2.7 Derivation of sample size formula

Regardless of the trial design the process of deriving a sample size formula can be described in four steps. These steps are illustrated here, assuming a normally distributed continuous outcome and a standard two-arm, parallel group design.

Step 1: Define the treatment difference and variance
For continuous normally distributed outcomes the mean provides a useful summary statistic to describe the response in each treatment group. μ_1 and μ_2 represent the population mean of the treatment and control groups with associated variance σ^2 (assumed to be the same in each group).

These population means will be estimated by sample means $\bar{x}_1$ and $\bar{x}_2$ with standard deviation s at the end of the trial. For the purpose of sample size calculation estimates of $\bar{x}_2$ and s might be estimated from previous similar studies and the value of $\bar{x}_1$ is usually chosen on the basis of clinical expertise. The difference between the means $d = \bar{x}_1 - \bar{x}_2$ provides an estimate of the minimal clinically important treatment effect. The standard error of this difference is $\sqrt{\frac{2\sigma^2}{m}}$ and can be estimated using the square of the sample standard deviation, s^2 in place of σ^2.

Step 2: Define the analysis method

In the context of clinical trials the Null hypothesis, denoted H_0 is usually that there is no difference between the two treatment groups.

H_0: $\bar{x}_1 - \bar{x}_2 = d = 0$

The alternative hypothesis (our hypothesis of interest), denoted H_1 is that there is a difference between treatment groups, usually a difference in either direction is considered (a two-sided test).

H_1: $\bar{x}_1 - \bar{x}_2 = d \neq 0$ (2-sided)

Step 3: Define the test statistic

In hypothesis testing we assume that the null hypothesis is true and we use the data collected from the trial, to find evidence against it. A probability value (or P-value) is calculated to say how likely it is that we would have obtained the observed trial data, or something more extreme, if in fact the null hypothesis were true. A P-value of less than 5% is usually taken as providing enough evidence to reject the null hypothesis. However, it has been argued by some that the results of trials should not be categorised into significant or non-significant on the basis of a P-value cut-off. Instead they should be interpreted in the context of the type of study and other available evidence.[5]

In order to derive the P-value we must first calculate a test statistic. The test statistic is a function of our data and reduces the observed data to a single value. For continuous outcomes this is usually the z-test:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d}{\sigma\sqrt{\frac{2}{m}}} \tag{1.8}$$

An important property of the test statistic is that the distribution under the null hypothesis is known. In this case, under the null hypothesis the test statistic follows a standard normal distribution $Z \sim N(0, 1)$.

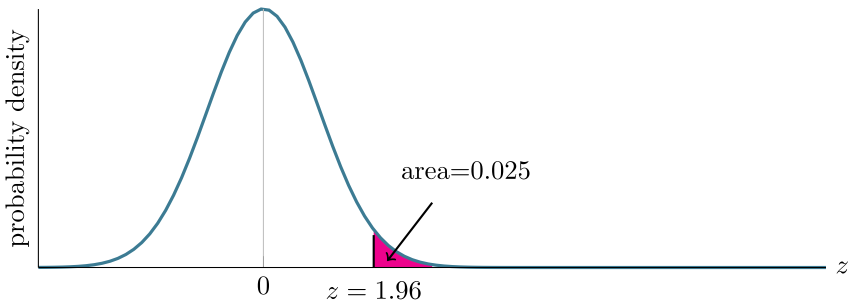


Figure 1.1: The probability density function (pdf) of a standard Normal distribution

When the value for z is 1.96, we know that for a standard normal distribution 0.975 of the distribution lies below this and 0.025 above. Therefore if our test statistic gives a value of 1.96 or greater there is a 2.5% chance or less that we would have observed our data under the Null hypothesis, and hence we would reject the null hypothesis, see Figure 1.1. Due to the symmetry of the normal distribution 2.5% of the distribution also lies below -1.96. Therefore for a two-sided hypothesis test at the 5% significance level a z value less than -1.96 or greater than 1.96 would provide significant evidence to reject the Null hypothesis.

Step 4: Evaluate the test statistic under the null and alternative hypotheses

When deciding whether to reject the null hypothesis we can potentially make two types of errors (1) Type I error: We reject the null hypothesis when it is in fact true (2) Type II error: We fail to reject the null hypothesis when it is false. The acceptable levels for these errors are controlled by the researcher at the design stage. The convention is to set the probability of a Type I error to 5% or less (see step 3), referred to as α and the probability of a Type II error, referred to as β to 20% or less.

Under the Null hypothesis, the hypothesised treatment effect is 0 and the test statistic, Equation 1.8, can be written as:

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{\frac{2}{m}}} \tag{1.9}$$

Under the null hypothesis if the value of this test statistic is $z_{1-\alpha}$ or greater we are incorrectly rejecting the null hypothesis i.e. making a Type I error, see Part A of figure 1.2.

$$z_{1-\alpha} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{\frac{2}{m}}}$$

$$(\bar{x}_1 - \bar{x}_2) = z_{1-\alpha}[\sigma\sqrt{\frac{2}{m}}]$$

Under the Alternative hypothesis, where the hypothesised treatment effect is δ, if the value of the test statistic is $-z_{1-\beta}$ or less we make a Type II error, see Part B of figure 1.2.

$$-z_{1-\beta} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sigma\sqrt{\frac{2}{m}}}$$

$$(\bar{x}_1 - \bar{x}_2) = \delta - z_{1-\beta}[\sigma\sqrt{\frac{2}{m}}]$$

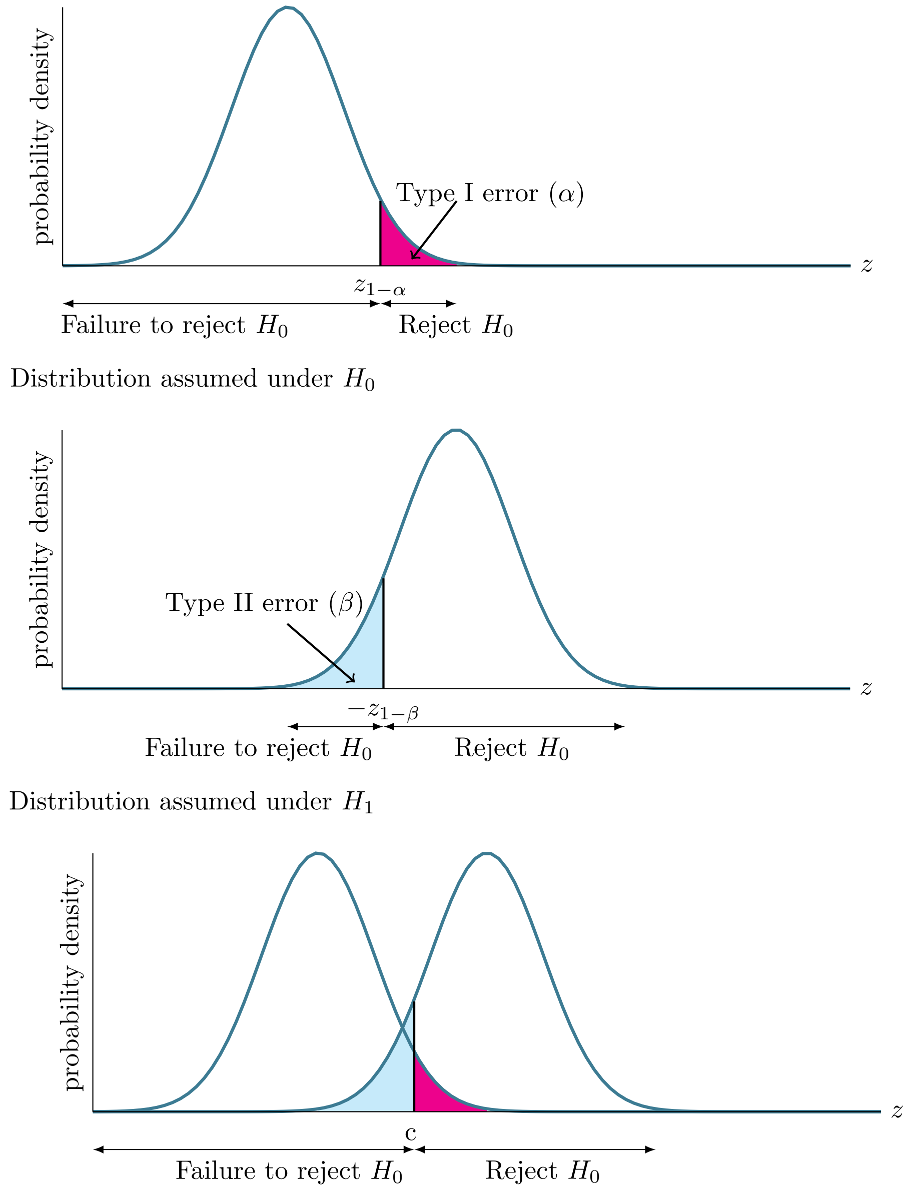
Equating these two expressions

$$z_{1-\alpha}[\sigma\sqrt{\frac{2}{m}}] = \delta - z_{1-\beta}[\sigma\sqrt{\frac{2}{m}}]$$

upon rearranging we are left with m, the required number of individuals per treatment group for a continuous outcome and standard parallel group design

$$m = \tfrac{2(z_{1-\alpha}+z_{1-\beta})^2}{\Delta^2}$$

Where $\Delta = \frac{\delta}{\sigma}$. For a two-sided test α should be replaced with $\alpha/2$.



A: Distribution assumed under H_0

B: Distribution assumed under H_1

C: Distributions under both H_0 and H_1

Figure 1.2: Graphical illustration of hypothesis testing errors

1.3 Cluster randomised trials (CRTs)

In some circumstances groups (or clusters) of individuals are randomised together, as a unit, rather than individually. These clusters are formed by some shared characteristic; they may be members of the same family or people within a particular geographical location. The cluster may be large, such as a hospital or General Practice surgery, or smaller like a family. These trials are referred to as Cluster Randomised Trials (CRTs) and are the focus of this thesis.

1.3.1 Rationale for cluster based randomisation

Broadly speaking the rationale for cluster based randomisation is driven by the nature of the intervention, the minimisation of costs, and/or the logistics of implementing the intervention.[6–8]

Some interventions are naturally applied at the level of the cluster and hence lend themselves to cluster-based randomisation. For example a water filtration system to prevent diseases from contaminated water must be implemented at the water source either for a community or household and therefore would affect all those living within that community or household.

It may be more cost efficient to implement an intervention at the cluster level. The IRIS trial assessed the effectiveness of a training intervention to improve the recognition of domestic violence by General Practitioners (GPs) working in different areas across the UK.[9] Under randomisation of general practice half of the GP practices were visited and trained in the intervention by trial staff. This was more efficient than to have visited all practices and trained only half of the GPs within each practice (as would occur under randomisation of individual GPs).

Avoidance of contamination is one of the advantages often cited for cluster-based randomisation. Contamination refers to the process by which within at least one of the trial arms there may be an unwanted presence of another trial arm. Contamination of treatment groups can result in a dilution of the true treatment effect. In a trial of pain management to reduce behavioural disturbances among dementia sufferers the care home was the unit of randomisation.[10] The authors felt that it would be difficult to individually randomise dementia suffers. Care staff receiving training in the assessment and treatment of pain would be required to provide specific support to some participants and to, in effect, “unlearn” this training when providing treatment to control participants. Therefore an effect of the training would likely be present in the control arm. Care staff may also struggle with the ethical implications of having to do this. Randomisation by cluster in the IRIS trial also reduced the risk of contamination among health professionals, whereby GPs would likely share aspects of their training and new knowledge with other GPs in their practice had they been randomised individually.

Randomisation of clusters can also provide advantages in terms of treatment compliance. For example interventions aimed at smoking cessation are more easily implemented at the level of the general practice. Participants attending the same practice may know each other or come into contact when attending the practice. The fact that participants are receiving the same intervention and able to discuss the intervention and support each other, may actually lead to greater compliance to the treatment than would be obtained in an individually randomised trial, and hence maximize the effect of the treatment. Along similar lines, cluster randomised trials are often conducted in infectious diseases or vaccine trials where the impact of the intervention is maximized when a large proportion receives the intervention.[11]

1.3.2 Between-cluster variability

The outcomes among individuals within the same cluster are likely to be more similar than those from a random sample of individuals. For example individuals attending the same general practice will live within the practice catchment area and likely share a common environment, education and economic status as well as being treated by the same health care professionals. These factors will influence their health outcomes in a similar way, leading to similarity within the cluster.

The intracluster correlation coefficient

The magnitude of this similarity, or clustering, is most commonly quantified by a parameter known as the intracluster correlation coefficient (ICC), usually denoted by ρ. When $\rho = 0$ we have, in effect, statistical independence between members of a cluster. When $\rho = 1$ the opposite is true and we have total dependence among members of a cluster.

For continuous outcomes the ICC can be expressed as

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \tag{1.10}$$

Where σ_w^2 is the within-cluster variance and σ_b^2 the between-cluster variance.

Historically ICCs were not explicitly published in reports of CRT's. Nowadays, however, there are many sources of ICC estimates[8, 12, 13] as well as publications which describe the general behaviour and patterns in ICCs across different therapeutic areas, outcomes and clusters.[14–18] For three types of cluster that vary in size: the spouse pair; general practice and country Donner calculated the within-cluster correlation for lifestyle outcomes: hypertension; smoking status; alcohol consumption; and body fat. For all outcomes the within-cluster correlation decreased with the corresponding size of the cluster.[19]

The coefficient of variation in outcome

An alternative measure of clustering to the ICC is the coefficient of variation in the outcome. This is the between-cluster standard deviation divided by the parameter of interest i.e. the proportion, rate or mean within each cluster.[20]

$$k = \frac{\sigma_b}{\mu} \tag{1.11}$$

This measure is particularly useful when the primary outcome variable is a rate as an ICC cannot be calculated in that situation.

The relationship between the ICC and coefficient of variation

The clustering of outcomes must be accounted for in the sample size calculation using one of the two measures of correlation, which will be described in more detail later (Section 1.4). For binary outcomes a simple relationship exists between the two correlation measures but they make different assumptions about how the between-cluster variation differs across treatment groups and so will produce different sample size requirements. The use of the ICC is recommended for sample size calculations of binary outcomes unless the proportion is very small, where both methods give similar results.[21]

1.3.3 Prevalence of cluster based randomisation

Several reviews of CRTs have been published and are summarised in Table 1.1. Many of these reviews note that the prevalence of CRTs is increasing over time.[22–30] In the review by Eldridge et al the number of cluster trials identified in primary care in 1997 had almost doubled by 1999.[25]

This increased use of CRTs is reflected in the increase in the methodological literature. The paper by Cornfield in 1978[31] was the first to highlight the fact that randomisation by cluster produces a less efficient design. Since then five textbooks have been published on cluster trials by: Murray[32] in 1998; Donner and Klar[6] in 2000; Hayes and Moulton[7] in 2009; Eldridge and Kerry[21] in 2012; and Campbell and Walters[33] in 2014. The journals Statistics in Medicine, Clinical Trials, Statistical Methods in Medical Research, and the American Journal of Epidemiology have each had a special issue dedicated to cluster randomised trials. Most recently, in 2015, the Trials Journal published a collection of articles on stepped-wedge designs, a type of CRT. Cluster randomised trials is a very exciting, and rapidly expanding, area of research to work in.

Table 1.1: Characteristics of 20 published reviews of cluster randomised trials

Author	Dates	Area	Main sources	N	Review topic
Donner (1990)[34]	1979-1989	Non therapeutic intervention trials	Lancet, NEJM, AJE, IJE	16	Quality of methods
Simpson (1995)[35]	1990-1993	Primary prevention trials	American Journal of Public Health, Preventive Medicine	21	Quality of methods
Chuang (2000)[36]	1974-1998	Computer-based clinical decision support systems	Medline, Embase, and INSPEC	24	Quality of methods
Hayes (2000)[11]	1986-1999	infectious diseases	Medline database	21	Evaluation of design features
Isaakidis (2003)[23]	1973-2001	Trials in Sub-Saharan Africa	Medline, Cochrane Controlled Trials Register, African Published Trials Register	51	Quality of methods and reporting
Puffer (2003)[37]	1997-2002	General medicine	BMJ, Lancet, NEJM	36	Risk of bias
Eldridge (2004)[25]	1997-2000	Primary care	Cochrane Controlled Trials register , UK National Research Register and conference proceedings	152	Quality of methods
Varnell (2004)[22]	1998-2002	General medicine	American Journal of Public Health, Preventative Medicine	60	Quality of methods

Continued on next page

Table 1.1 – continued from previous page

Author	Dates	Area	Main sources	N	Review topic
Bland (2004)[26]	1983-2003	General medicine	BMJ	18	Quality of methods
Murray(2008)[38]	2002-2006	Oncology	Medline, PubMed	75	Quality of methods
Eldridge (2008)[39]	2004-2005	Primary care	BMJ, BJGP, FP, preventive medicine, Annals of internal medicine, journal of general internal medicine, Paediatrics	34	Internal and external validity
Bowater (2009)[40]	1998-2007	Tropical parasitic disease	Medline database	35	Quality of methods
Handlos (2009)[27]	1998-2008	Maternal and child health	Pubmed, SCOPUS, Cochrane library	35	Quality of methods
Mdege (2010)[24]	To Jan 2010	Stepped wedge trial	Medline, Embase, PsycINFO, HMIC, CINAHL, Cochrane library, Web of Knowledge, current controlled trials register, google scholar	25	Areas of application
Ivers (2011)[30]	2000-2008	General Medicine	Medline database	300	Quality of methods and reporting
Walleser (2011)[41]	2004-2010	Trials in children	Medline, CINAHL, Embase, cochrane central register	106	Quality of reporting
Crespi (2011)[42]	1995-2010	Cancer screening interventions	PubMed, Web of Science	50	quality of analysis/outcome reporting
Brierley (2012)[43]	2008	General Medicine	BMJ, Lancet, JAMA, NEJM	24	Bias in recruitment

Continued on next page

Table 1.1 – continued from previous page

Author	**Dates**	**Area**	**Main sources**	**N**	**Review topic**
Giraudeau (2012)[44]	2008	General Medicine	Medline database	173	Reporting of informed consent
Froud(2012)[45]	2005-2009	Oral Health	PubMed and experts in the field	23	Quality of methods and reporting
Diaz-Ordaz (2013)[28]	up to 2010	Trials in old age residential facilities	Medline database	73	Quality of methods and reporting
Diaz-Ordaz (2013)[29]	up to 2010	Trials in old age residential facilities	Medline database	73	Reporting of informed consent
Sutton (2013)[46]	2003-2011	Stroke	PubMed	15	Quality of reporting

1.3.4 Guidelines for best practice

Many of the fundamental principles of RCTs such as informed consent, randomisation, sample size and analysis are more complex with randomisation by cluster. Methodology and guidelines for best practice are not as well developed as those for individually randomised trials. The guidance that is available is described in the following sections.

Design, analysis and conduct

In addition to the published textbooks on CRTs there are some additional sources of guidance on the design, analysis and conduct of cluster randomised trials. In 1999 the methodological literature around cluster randomised trials was reviewed and synthesized into 12 methodological recommendations designed to aid investigators designing and conducting these trials. The review covers study design, measures of between-cluster variation, sample size and analysis.[8] With regards to sample size, the focus of my research, the advice given was to avoid designing studies with less than four clusters per group and to calculate a sample size appropriate for cluster randomisation. In 2002 the Medical Research Council (MRC) produced a brief guidance booklet on the methodological and ethical aspects of cluster randomised trials which provides a basic introduction to the main issues to be aware of when randomising clusters. Their advice on sample size was for a minimum of five clusters per group. (http://www.cebma.org/wp-content/uploads/Cluster-randomised-trials-Methodological-and-ethical-considerations.pdf). Guidance has also been developed for specific aspects of the cluster randomised trial such as consent procedures for trials conducted in residential facilities.[29]

Reporting

The CONSORT statement consists of a 25-item checklist for improving and standardising the reporting of clinical trials. The statement was first published in 1996[47] and has since gone through two further revisions.[48, 49] In 2004 the statement was extended for cluster randomised trials[50] and this extension was updated in 2012 following the update to the main statement in 2010.[51]

The item which relates to describing the sample size calculation for individually randomised trials formally recommends the following descriptive elements to be reported (i) the estimated outcomes in each group (which implies the minimum important treatment effect); (ii) the level of significance (or the α (type I) error level); (iii) the statistical power (or the β (type II) error level); and (iv) for continuous outcomes, the assumed standard deviation of the outcome. For cluster randomised trials the 2004 CONSORT extension additionally recommends the reporting of two further descriptive elements (v) the number of clusters or the cluster size(s) and (vi) the intracluster correlation coefficient (ICC) or coefficient of variation (k), along with a measure of its uncertainty. The 2012 revision additionally recommends specification of whether equal or unequal cluster sizes are assumed.

1.4 Sample size calculations for cluster randomised trials

For cluster randomised trials there may be several combinations of the number of clusters and cluster size that produce designs with equivalent power. In these situations, to determine the optimal design, one may additionally consider the efficiency of these designs in terms of the costs involved in recruiting and measuring clusters and individuals within clusters. These scenarios are briefly considered in Chapter Six but the focus of this thesis are scenarios where the cluster size or number of clusters is fixed, or constrained, at the point of sample size calculation and the trial is designed to produce a specified level of power.

In this section I briefly describe the most common approach to sample size calculation for cluster randomised trials and describe recent developments prior to the start of my research and unresolved issues to place my research in context. In Chapter Six I will return to look in more detail at the developments and unresolved issues that remain at the end of my research.

1.4.1 Early work

Cornfield[31] recognised that randomisation by cluster resulted in a less efficient design and so the sample size assuming individual randomisation must be inflated to achieve adequate power under cluster randomisation. Cornfield's work was followed in 1981 by Donner[52] who quantified this inflation factor and described it as the Design effect. Despite being over 30 years old these two papers still remain highly cited and use of the design effect for sample size calculation remains the most common approach.

1.4.2 The Design Effect(DE)

For continuous and binary outcomes the sample size calculated assuming individual randomisation (Equations 1.1 or 1.2) is multiplied by the design effect to account for randomisation by cluster. This design effect is given by

$$DE = 1 + (n - 1)\rho \tag{1.12}$$

Where n is the number of individuals per cluster (assumed constant) and ρ is the intracluster correlation coefficient. When we conduct a CRT we may sample an entire cluster such as all participants registered at a General Practice, or take a sub-sample for inclusion into the trial. Throughout this thesis, when I refer to cluster size, I am specifically referring to the sample of the cluster that is to be included in the analysis, which may or may not be the entire cluster.

1.4.3 Recent developments

The design effect proposed by Donner was derived for continuous or binary data analysed at the cluster level assuming a fixed cluster size. In some trials such as those conducted in ophthalmology where a subject is the cluster and measurements are taken on each eye fixed cluster size may be a reasonable assumption to make. However, to have variable cluster sizes is more common. In trials where the cluster size is very variable use of the average cluster size in the design effect will likely underestimate the required sample size. For cluster-level analysis simple methods are available that provide an appropriate sample size using the harmonic mean of the sample size in each cluster.[7] Recent reviews have shown developments in sample size calculations including methods allowing for: variable cluster sizes, matched designs, re-estimation using internal pilots, attrition, incorporation of covariates or multiple time points, time-to-event outcomes, and incorporation of imprecision in the ICC.[8, 53–55]

1.4.4 Unresolved issues

The methodological reviews indicate that although many sample size methods are being developed to deal with variations and complexities in trial design the majority are still only applicable to binary or continuous outcomes. Methodology for alternative outcomes such as ordinal, count or time-to-event is in the minority, especially for variations to the standard parallel group trial. However, these reviews present only selected methods and do not provide a comprehensive review of all available sample size methods. In Chapter Two I conduct a systematic review to provide a comprehensive description of published sample size methods for cluster randomised trials.

1.4.5 Quality of methodology and reporting

Despite the simplicity of the design effect many trialists are still unaware of the need to adjust sample size calculations to account for clustering in cluster randomised trials, or perhaps unaware that the design they are using induces such clustering. Many of the reviews in Table 1.1 examined the quality of both the trial methodology and reporting. The proportion of trials that reported a sample size calculation and the proportion that reported an appropriate calculation can be seen in Table 1.2. Despite the introduction of the CONSORT Statement many of the reviews showed that the reporting was inadequate. In the largest review by Ivers et al whose sample is perhaps the most representative of the health research field just over half of the trials reported a sample size calculation.[30]

Given that the CONSORT extension for cluster randomised trials was first published in 2004 it is clear from Table 1.2 that sample sizes that appropriately account for clustering are reported with low frequency. Much improvement to reporting is needed, poor reporting can make it difficult for those designing trials to obtain the estimates they need for sample size calculations.

In this section I have described the use of the design effect as the most common method for sample size calculation in cluster randomised trials. However, as the variety in trial designs increase such as variable cluster sizes, attrition or repeated measurements the simple design effect may not always be appropriate. The focus of my research is sample size calculations for ordinal outcomes. In the following section I provide the definition of an ordinal outcome used throughout this research and describe some simple and intuitive approaches to sample size calculation and why they may not always be adequate.

Table 1.2: Description of the quality of sample size reporting identified in published reviews of cluster randomised trials

Author	Reported sample size calculation	%	Reported appropriate sample size calculation	%
Donner (1990)[34]	NA		3/16	19%
Simpson (1995)[35]	5/21	24%	4/21	19%
Chuang (2000)[36]	1/24	4%	0/24	0%
Isaakidis (2003)[23]	47/51	92%	10/51	20%
Puffer (2003)[37]	NA		20/36	56%
Eldridge (2004)[25]	68/152	45%	21/152	14%
Varnell (2004)[22]	NA		9/60	15%
Murray(2008)[38]	40/75	53%	18/75	24%
Eldridge (2008)[39]	29/34	85%	21/34	62%
Bowater (2009)[40]	17/35	49%	10/35	29%
Handlos (2009)[27]	33/35	94%	25/35	71%
Mdege (2010)[24]	8/15	53%	3/15	20%
Ivers (2011)[30]	164/300	55%	100/300	33%
Walleser (2011)[41]	87/106	82%	63/106	59%
Froud(2012)[45]	21/23	91%	15/23	65%
Diaz-Ordaz (2013)[28]	43/73	59%	20/73	27%
Sutton (2013)[46]	NA		12/15	80%

1.5 Ordinal outcomes

1.5.1 Definition

An ordinal variable is one which consists of a set of categories which can be ordered or ranked. Disease severity (mild, moderate, severe) or measures of agreement (completely agree, agree, do not agree, disagree, do not agree at all) are examples of ordinal variables. The difference between participants in adjacent categories may not be the same, and are often unmeasurable. For example the difference in disease severity between moderate and severe could be much greater than the difference between mild and moderate.

Where the outcome consist of categories which cannot be ordered, therefore each level does not differ in magnitude for example marital status (single, married, divorced) the variable is referred to as categorical, or nominal, within this thesis.

1.5.2 Example of a trial with an ordinal outcome

Encouraging lifestyle changes such as smoking cessation, decreasing fat intake and increasing regular physical activity can help prevent cardiovascular disease. Visits to primary healthcare can be an opportunity for those at high risk of coronary heart disease to receive advice about changing their lifestyle. In a trial by Steptoe et al 20 general practices were randomised to lifestyle counselling or usual health promotion.[56] Patients with one or more risk factors for coronary heart disease were included in the study. Each patient completed a questionnaire prior to their physician visit and four and twelve months after. This questionnaire measured a patient's stage of motivation concerning regular exercise and patients were categorised into one of five stages of change: Pre-contemplation (patients are not eating a low-fat diet or currently exercising or are smokers, and they are not seriously considering changing behaviour), contemplation (patients are considering a change in behaviour but are not confident they will carry this out within the next month), preparation (patients are seriously planning to change behaviour and are confident that they will make changes within the next month), action (patients have changed behaviour within the last 6 months) and maintenance (patients have maintained the change for at least 6 months).

1.5.3 Sample size approaches

In the Steptoe study the authors chose to dichotomise the ordinal outcome. Those in the action and maintenance stages were combined and those in the remaining categories were combined. The benefits of using a binary outcome are: sample size and analysis methods are well established; parameter estimates are likely to be available; the dichotomised version may be more clinically relevant; and it avoids problems in the analysis caused by a small number of observations in one of the ordinal categories.

Had the authors analysed the outcome in its ordinal form they would have likely increased the power of their study, as dichotomisation of the outcome results in a loss of information. For individually randomised trials using a sample size calculation appropriate for the ordinal version of the outcome can result in trials being on average 28% smaller compared to those powered on the dichotomous version.[57] It is unknown how conservative the dichotomous approach would be for the clustered case. For a cluster randomised trial there is likely to be a large cost associated with recruiting an additional cluster as compared to recruiting an additional subject in an individually randomised trial. Therefore to calculate a conservative estimate of sample size could be considered wasteful.

In addition to the dichotomisation approach there are two other valid approaches that can be taken to arrive at a sample size estimate. The first is to choose an alternative primary outcome for which sample size can be calculated easily. In some situations this may be a reasonable approach if several outcome measures all of clinical relevance are being considered. The Core Outcome Measures in Effectiveness Trials (COMET) initiative was set up in 2010 to develop a set of standardised core outcomes that should be measured and reported in all trials of a certain condition. The consistent use of these core outcomes in trials will ensure that more trials can be included in meta-analyses and, most importantly, as each set of proposed core outcomes were chosen to be relevant to patients, clinicians and policy makers the findings from the trial are likely to influence current practice.[58] These are important justifications for the choice of outcome measure, convenience for the sample size calculation is not.

The second method is to calculate the sample size via simulation methods. It can be computationally intensive but provides a lot of flexibility, allowing full control of all parameters and so giving a closer representation of real life. The procedure involves simulating a large number of data sets, each one to represent a potential data set of results from the trial. For each simulated data set the planned analysis is conducted and the empirical power calculated as the percentage of tests where the null hypothesis is rejected. Changes in the input parameters can be made until adequate power is achieved. This is a valid approach to use for ordinal outcomes and user written commands are available in the statistical computer package Stata to aid the implementation.[59] However, this approach has disadvantages in both the time taken to compile the simulation and its potential complexity. The aim of my research is to recommend an approach that is simpler to implement, even if its use is only to provide an initial benchmark estimate that may be further refined with more complex procedures.

1.5.4 The importance of this research

The literature around sample size calculations for cluster randomised trials has focused heavily on continuous and binary outcomes. There is little guidance around sample size methods for ordinal outcomes. Simple approaches such as calculating a sample size based on the dichotomisation of the outcome can be used but may be overly conservative and simulation methods require programming knowledge to implement. Although less common than binary and continuous outcomes the use of ordinal outcomes is not rare and hence there is a need for appropriate sample size calculations.

In my opinion the use of ordinal outcomes will become more prevalent in the future, particularly for trials conducted in the UK setting. This is because the National Health Service in the UK is becoming more patient centred and many Patient Reported Outcome Measures (PROMs) such as quality of life tend to be ordinal.

1.6 The research aim

The aim of this research is to comprehensively review the existing state of knowledge of sample size calculations for cluster randomised trials and to focus on the development of methods for ordinal outcomes. These aims are covered in the following chapters:

- **Chapter Two** A comprehensive summary of sample size methods available for cluster randomised trials. The review has two main aims: To identify the prevalence of sample size methods for ordinal outcomes which will provide evidence for or against the need for development in this area; and to identify the methods that have been used to derive sample size calculations for cluster randomised trials, it may be possible to adapt or emulate one of these approaches to derive a sample size formula for ordinal outcomes.
- **Chapter Three** A review of 300 published cluster randomised trials to determine the prevalence and characteristics of trials with ordinal outcomes, providing additional motivation and context for the development of sample size methods for ordinal outcomes.
- **Chapter Four** Analysis methods available for clustered ordinal outcomes: their assumptions, advantages and disadvantages. Several proposed estimates for measuring between-cluster variability in ordinal data that may be used in sample size calculation are also presented.
- **Chapter Five** Monte Carlo simulation studies are conducted to explore the relationship between ICC estimators for ordinal outcomes and to assess the performance of using each of these ICCs in sample size calculations.

- **Chapter Six** A return to the review of sample size methods for cluster randomised trials with the aim of identifying where there is scope for further development beyond this thesis.
- **Chapter Seven** A summary and discussion of the main findings from this research.

Chapter 2

Sample size methods for CRTs

At the time of this research there is no single, up-to-date, published, resource that provides a comprehensive summary of sample size methods available for cluster randomised trials. In this chapter I undertake such a review to meet three main aims:

- The first is to identify the current availability of sample size methods for ordinal outcomes, the results of which will inform the direction of my research.

- The second aim is to identify the different approaches that have been used to derive sample size calculations for cluster randomised trials. It may be possible to adapt or emulate one of these approaches, if necessary, to derive new sample size formulae for ordinal outcomes.

- The final aim is to provide researchers with a comprehensive summary of sample size methods for cluster randomised trials that allows them to easily identify the formula to use for a given design, outcome and analysis method.

This chapter describes the methods I used to conduct the review followed by the results relevant to the first two aims: a description of the methods available for ordinal outcomes and a summary of the approaches that have been used to derive sample size formulae. This chapter concludes with a discussion of how the findings impact upon the research plan. The results of the third aim will be provided in Chapter Six where the focus will be on summarising where the methodological gaps remain at the end of my research to inform the direction of future work. My review has been published and can be found in appendix (ix).[60]

2.1 Review methods

The methods of the review were specified in advance and documented in a protocol (version 1.0 31/03/2011) to be found in appendix (i). SE and AC reviewed the protocol prior to implementation. The following sections describe the actual conduct of the review, section 2.1.7 provides the detail and explanation for any differences from the planned protocol.

2.1.1 Data sources

Cluster randomised trials are found within the educational literature as well as within health research. However, the health research literature is the focus of my review as this is my area of expertise. In addition CRTs are commonly used in health research and much of their methodological development is focused around health research.

I conducted the review using electronic online databases, a personal collection of 41 articles on sample size in CRT's provided by SE, key text books on cluster randomised trials,[6–8, 32] and special issue journals on cluster randomised trials.[61–64] For the electronic search I used the online databases PubMed and Web of Science.

The PubMed database is a free online database developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM). It contains over 20 million citations from the biomedical literature. The MEDLINE database is the largest component of the PubMed database. Searches conducted in the MEDLINE database cover biomedicine and health articles dating from 1946 onwards. Articles in the MEDLINE database are indexed using Medical Subject Heading (MeSH) terms. MeSH terms are used by indexers to provide a consistent way to index articles that may have used different terminology to describe the same concepts. In addition to the MEDLINE database, PubMed also contains additional references such as those which are yet to be indexed with MeSH and citations that precede the date that a journal was selected for MEDLINE indexing.

The Web of Science online database contains seven citation databases. The search was conducted using only the most relevant, the “Science citation index expanded database”. This database contains articles across 150 scientific disciplines from the year 1970.

2.1.2 Inclusion and exclusion criteria

I included an article in the review if it provided a method of sample size calculation for cluster randomised trials, via formula, simulation or other approach. The first paper to report a particular methodology was included in the review; subsequent papers describing the same approach were excluded. The two electronic databases searched, PubMed and Web of Science contained articles from 1946 and 1970 onwards respectively, no further date restrictions were applied. I excluded those papers written in a language other than English. To have conducted this search in other languages would not have been feasible due to the cost involved in translation of manuscripts.

The following types of paper were excluded from the review as they were considered to be: too general for the review; unlikely to contain new methodology or sufficient detail on sample size calculation; or contain information irrelevant to the review aims

- Those reporting a trial protocol or trial results
- Those that provide a very general discussion on the effects of clustering or correlated data
- Papers focusing on non-randomised observational studies
- Those that only discuss the calculation of the within-cluster correlation component of sample size estimation
- Those describing correlated data but individually randomised, for example:
    - Clustering induced by the treatment itself, for example several participants attending a group therapy intervention, but individually randomised to treatment group
    - Clustering present due to centre effects
    - Correlated data due to recurrent events occurring to the same subject
    - Correlation due to multivariate outcomes

Development of the inclusion and exclusion criteria

I developed the inclusion and exclusion criteria from a preliminary basic search I conducted of the sample size literature for CRTs in the PubMed database. This search revealed articles covering a broad range of topics:

1. Those which derive some or all methodological aspects of sample size formulae, using either a Bayesian or frequentist approach.
2. Those which provide suggested adaptations or extensions to previously derived formulae.
3. Papers with a broader focus that discuss design features in general for cluster randomised trials.
4. Overviews or summaries of the current methodology in a particular area.
5. Papers which focus on the calculation/estimation of the intracluster correlation or coefficient of variation of outcome.
6. Papers evaluating or comparing alternative methods.
7. Papers with a general discussion on the effects of clustering upon sample size and power.
8. Papers describing application tools for implementation of the methodology.
9. Protocols or design papers for specific cluster randomised trials, describing the sample size calculation, as applied to a specific trial.
10. Reports of results from cluster randomised trials.

To fulfil the objectives of this review I developed the search criteria to target articles of type 1-4. Papers of type 5-10 were to be excluded due to reasons described at the beginning of this section.

2.1.3 Search terms

To improve the relevancy of results retrieved I chose to base my search terms on only the article title and MeSH term used to index the paper.

The list below contains the search terms for the PubMed database. This list includes all changes which were made after validation of the search terms (see section 2.1.6)

1. cluster analysis[MeSH] AND sample size[MeSH]
2. "sample size" [Title]
3. "design effect"[Title] OR "design effects"[Title] OR "variance inflation factor"[Title]
4. (design*[Title] OR plan*[Title] OR siz*[Title]) AND cluster*[Title] [1]
5. power [title] AND cluster*[Title]
6. "intraclass correlation*"[Title] OR "interclass correlation*"[Title] OR "intracluster correlation*"[Title] OR "coefficient of variation"[Title] OR "between cluster"[Title]
7. coefficient[Title] AND variation[Title]
8. (design[Title] OR matching[Title]) AND community[Title]
9. power[Title] AND correlated[Title]
10. number[Title] AND clusters[Title]
11. 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10

The Web of Science database does not index articles with MeSH so I did not perform this component of the search for this database. I additionally refined the searches within Web of Science by limiting the search to articles only, thereby excluding conference proceedings and letters etc. which are not listed in the PubMed database.

All articles retrieved using these search terms were assessed for eligibility following the procedure described in section 2.1.4.

[1] The "*" is used as a wildcard to search on all words that are formed from the truncated word entered. For example siz* will find all words formed from siz such as "size" and "sizing".

Development of search terms

I developed the search terms using the personal collection of 41 articles from SE to determine (1) the most common MeSH terms used to index these papers and (2) the terminology commonly used in the title description.

Thirty five of the papers were available via PubMed and had 199 associated MeSH terms. The most frequent being "randomized controlled trials as Topic" (n=28), "cluster analysis" (n=20), "research design" (n=16), and "sample size" (n=20). However, the definition of the MeSH term "Cluster analysis" is not strictly applicable to cluster randomised trials. MeSH description of cluster analysis is

> "A set of statistical methods used to group variables or observations into strongly interrelated subgroups. In epidemiology it may be used to analyse a closely grouped series of events or cases of disease or other health related phenomenon with well-defined distribution patterns in relation to time or place or both"

The term cluster analysis is unreliable due to its incorrect and inconsistent use for indexing papers. Therefore its use was combined with the term sample size to reduce retrieval of irrelevant results. The term "randomized controlled trials as topic" and "research design" are very broad and their use would retrieve a vast amount of literature irrelevant to this review, including literature related to individually randomised trials. With this in mind I did not use the MeSH term for randomised controlled trials or research design. A review of all MeSH terms available for indexing did not identify any further terms that might be useful to this review.

Due to the inconsistent use of MeSH terms and the lack of terms specific to the needs of this review the use of MeSH terms alone would likely retrieve a large number of irrelevant results and may miss some that are relevant. Therefore I supplemented the search by a search for specific text within the article title only. In the 41 papers of the personal collection the term "sample size" was the most frequently used. However, the terms "Design effect" and "power" were also frequently used. As the term power has a large number of meanings, I restricted its search to those papers which additionally mentioned clustering in some form within the title in order to reduce the number of irrelevant articles retrieved.

Within PubMed the word “of” is a stopword and is ignored when searching for a phrase. Therefore rather than search on the phrase “coefficient of variation” I chose to search on the separate components, similarly with the phrase “number of clusters”.

Web of Science treats the word “of” as a placeholder when contained in the search for a phrase. Therefore a search with the phrase “coefficient of variation” will retrieve all results where the two terms are separated by one word, such as “coefficient and variation” and “coefficient of variation”. I used this search rather than the search on the two separate components which was required when searching PubMed. The word “between” is similarly treated as a placeholder in Web of Science. This means that a search on ”“between cluster” will retrieve any article with the word cluster in the title. This retrieves a vast number of articles, with many irrelevant articles relating to disease clusters and laboratory data. Therefore I excluded the search term “between cluster” from the Web of Science search.

2.1.4 Search strategy

I applied the search terms first to the online databases, this was followed by a hand search of personal collections, textbooks and special issue journals.

An article was first assessed for inclusion based upon the title. Where eligibility was unclear I then reviewed the abstract, and if still unclear I reviewed the full text. Any paper where eligibility was still unclear after examination of the full text was discussed and eligibility agreed with SE and AC.

The references of all papers identified as eligible were searched. Any new eligible articles were added to the review and the process continued until it was agreed with SE and AC that sufficient saturation of new methodology or concepts had been achieved.

I sent the final list of included papers to experts who have made significant contribution in the field for their thoughts upon its coverage: G.J.P. van Breukelen, Mike Campbell, Allan Donner, Steven Tereenstra, and Obi Ukoumunne. Additional papers suggested by these experts for inclusion or review were assessed using the process described in this section.

2.1.5 Data collection and management

All details of the articles retrieved from the search of the electronic databases were imported into a Microsoft Access 2010 database for storage. I designed and created this database and screen prints can be found in the protocol (see appendix i). Each article was stored with a unique identification number and the information imported from the electronic databases included authors, year, title, journal, volume, issue, start page, end page, and the abstract. Within the Access database I categorised each retrieved result for inclusion or exclusion, with associated reason.

For each paper that was identified as eligible I extracted the following information 1) the authors, title, journal, date of publication and database id 2) the trial design features 3) the formulae described for each design 4) the correlation measure used 5) the analysis method assumed 6) the assumptions underpinning the methodology 7) the simulation procedures used to evaluate the methodology 8) the strengths and weaknesses as stated by the authors and 9) any extensions of the methodology. This extracted information was collected on a paper based data extraction form which I designed and can be found in the protocol (appendix i).

Where I made assumptions with regard to information that was not explicitly stated within a paper I clearly identified as "assumed" on the data extraction form. For example it was not always explicit that fixed cluster sizes were assumed but it could be reasonably assumed if the formula clearly included only an estimate for the average cluster size or a fixed cluster size. Where reasonable assumptions could not be made the information was recorded as unclear on the data extraction form. In papers where much of the paper and its methodology was unclear I collected only basic information about the trial design and relevant sample size formulae, as agreed with SE and AC.

I performed all data extraction and articles with ambiguity were discussed with SE and AC.

2.1.6 Validation

I implemented validation methods at various stages of the review process in order to provide confidence in the quality of the review and its ability to fulfil its intended objectives.

Inclusion/exclusion criteria

I piloted the inclusion/exclusion criteria on 20 search results in order to identify any need for revision. No revision was deemed necessary. However, during the review some of the criteria were more explicitly defined.

Search terms and strategy

At the design stage the appropriateness of the search terms and general strategy to retrieve the targeted papers was validated in two ways.

First I assessed the titles of the papers included in the personal collection of 41 papers to determine if they would be identified from the proposed search terms, and in the majority of cases they were. Those which were not identified were cited by eligible papers and so would have been identified as eligible for the review from a search of references, alongside the context of the citation in the included paper.

Secondly I compared the search terms with the retrieved results from a search that could be considered to be a "gold standard". A computer based search of the Statistics in Medicine journal was used as the "gold standard" search as this journal was considered most likely to contain articles on methodology of sample size calculations. Using PubMed I retrieved all articles published in Statistics and Medicine from 1982 to March 2011 and included reference to clustered data or cluster randomised trials. This search produced fewer than 200 results. Based upon an individual assessment of each title these results were reduced to those which would warrant further examination and compared to the proposed search terms. Of these 200 results many of the early articles made reference to community intervention trials rather than cluster randomised trials, as they are now more commonly known. The term community was therefore added to the search terms in combination with design or matching to pick up these articles in the main review. "Number of clusters" was also a common phrase used to describe sample size and was added to the search terms.

Data extraction

Before its implementation I piloted the data extraction form on five papers from the personal collection of articles. The test papers were selected to cover different styles of article, from those of a very practical nature to those with detailed information on the mathematical derivations. The extraction form was then refined based on this pilot.

The data extraction process was validated by an independent double data extraction for ten articles. Five were additionally extracted by SE and five by AC. Discussions took place to agree any differences in interpretation. In the protocol I stated that these ten articles would be randomly selected. However, during the process of the review it became clear that the ease of data extraction varied among the papers and that it would be most useful to perform double data extraction on the more challenging papers.

In addition during the process of preparing the review for publication many of the papers were discussed with SE and AC which provided further reassurance of my interpretation and understanding of the methods described and the assumptions made.

2.1.7 Changes in conduct from the planned protocol

In the protocol I stated that monthly email notifications based on the search terms would be set up within the electronic databases for the duration of the project in order to keep informed of new developments. However, the email notifications were not received, possibly due to anti-virus or spam technology. Therefore to identify any new methodological developments since the main search date I re-ran the search on the two online databases near the end of my research, 27th August 2015. The additionally identified methods are described in Chapter Six where I provide a comprehensive summary of sample size methods.

During the course of the review several special issue journals on cluster randomised trials came to light. All special issue journals were then additionally included as data sources as they contained a high concentration of articles with potential for eligibility in the review.

I had planned an electronic search based on first authors of papers identified as eligible. After completion of the electronic searches and reference searches of eligible papers it was agreed with SE and AC that a search on first authors would not be conducted. Due to the large number of distinct contributors in the review a search based on first authors would likely retrieve a very large number of potential results for assessment and, given the time needed to review the results, was unlikely to identify a significant number of methods that would not have already been identified through the other data sources.

In the protocol I stated that

> "articles will be eligible for the review if they discuss any aspect relating to the methodology of sample size or power. This includes those papers which may discuss or simulate only components of the sample size."

During the process of the review I decided to exclude papers that reported only on the calculation or estimation of the ICC for two reasons. The first being that the definitions of ICCs relevant for a sample size calculation would be most likely found in the paper that described the sample size calculation. ICCs calculations without reference to a sample size formula were not likely to be useful. Secondly the ICC has other uses in addition to its place in CRTs so irrelevant results would be retrieved. For example it can be used in quantifying the inter-rater reliability of instruments, often psychometric scales. Inter-rater reliability assesses the degree to which two raters give consistent observations, if raters do not agree either the scale is defective or the raters need to be trained in its use. In this context the ICC is high when there is little variability among the raters.

In the protocol sample size calculations for correlated data due to repeated measurements on the same individual were to be excluded. However, during the review one paper that reported a sample size method for longitudinal data was included. The reason for its inclusion was that it was one of very few papers to consider ordinal outcomes and I considered it important to assess its potential to be applied to cluster randomised trials.

2.2 Results

2.2.1 Description of included papers

I applied the search terms to the online databases on the 31st of March 2011. From 8393 retrieved records 77 papers were identified as eligible. A further 4 were identified as eligible from the personal collection of articles on sample size in CRTs provided by SE and special issue journals and 4 were identified as eligible ad-hoc during the review and collation of results (identified: via colleagues; from feedback from experts in the field; or published after the initial search). In total 85 papers describing sample size methodology were included in the review. A list of the included papers can be found in appendix (ii). Figure 2.1 shows the flow of articles through the review process.

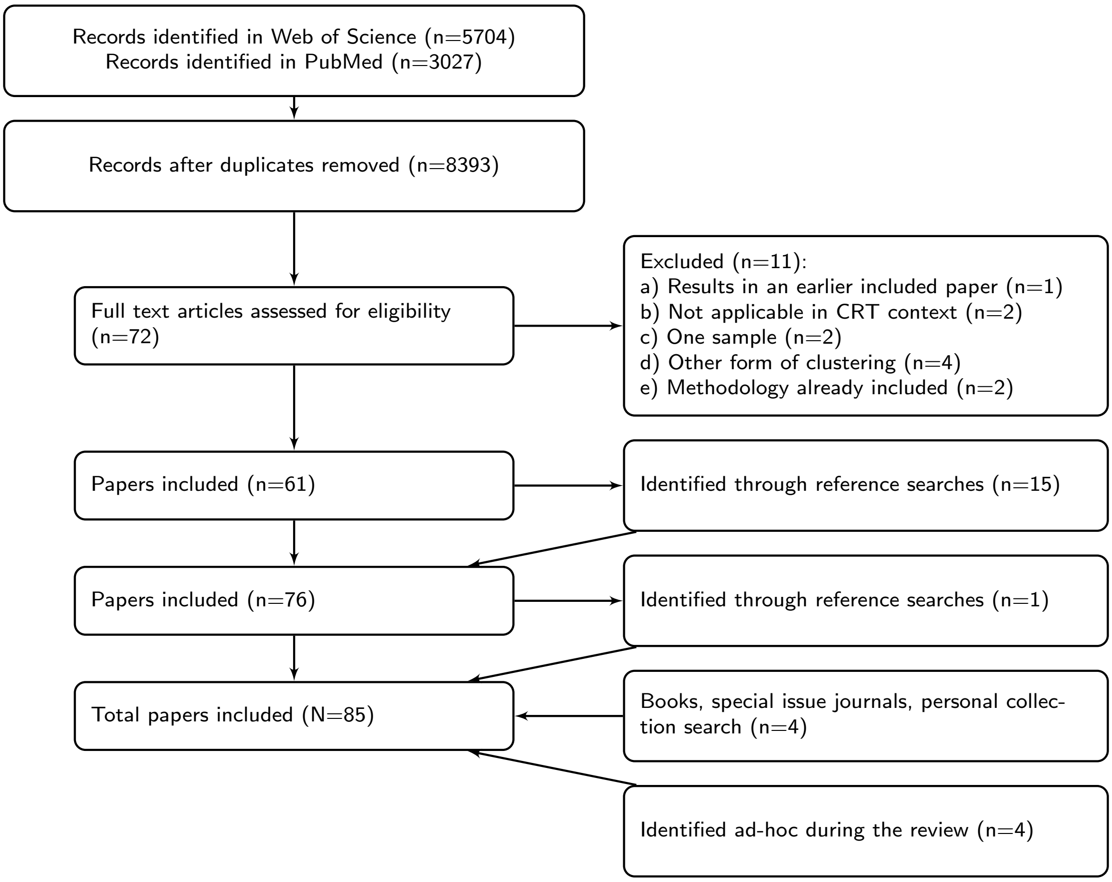

Records identified in Web of Science (n=5704)
Records identified in PubMed (n=3027)
Records after duplicates removed (n=8393)
Full text articles assessed for eligibility (n=72)
Excluded (n=11):
a) Results in an earlier included paper (n=1)
b) Not applicable in CRT context (n=2)
c) One sample (n=2)
d) Other form of clustering (n=4)
e) Methodology already included (n=2)
Papers included (n=61)
Identified through reference searches (n=15)
Papers included (n=76)
Identified through reference searches (n=1)
Total papers included (N=85)
Books, special issue journals, personal collection search (n=4)
Identified ad-hoc during the review (n=4)

Figure 2.1: Systematic review sample size methods selection: Flow diagram describing number of papers screened, assessed for eligibility and included

The two journals where the largest number of the included papers were published were Statistics in Medicine 29 (34%) and Biometrics 7 (8%). The number of sample size methods has increased over time, with 58 (68%) of papers published since 2000. In terms of the trial characteristics the majority of sample size methodology was aimed at binary or continuous outcomes, completely randomised, parallel group superiority designs, see Table 2.1.

The results of the review are now presented according to the initial objectives:

- To identify and describe sample size methods available for ordinal outcomes.
- To identify the approaches used to derive sample size formulae.

Table 2.1: Characteristics of the 85 published sample size methods for CRTs included in the systematic review

		N	(%)
Journal	Statistics in Medicine	29	(34%)
	Biometrics	7	(8%)
	International Journal of Epidemiology	4	(5%)
	American Journal of Epidemiology	2	(2%)
	Journal of Clinical Epidemiology	3	(4%)
	Controlled Clinical Trials	4	(5%)
	Clinical trials	3	(4%)
	Journal of Biopharmaceutical Statistics	3	(4%)
	Journal of Educational and Behavioral Statistics	3	(4%)
	Journal of Educational Statistics	1	(1%)
	Journal of the Royal Statistical Society	3	(4%)
	Other Journals/books	23	(27%)
Year of publication	pre 1990	6	(7%)
	1990-2000	21	(25%)
	2000-2010	43	(51%)
	2010 onwards	15	(18%)
Type of outcome	*Binary/Continuous	74	(87%)
	Time-to-event	5	(6%)
	Rates	2	(2%)
	Count	2	(2%)
	Ordinal	2	(2%)
Randomisation	Simple	78	(92%)
	Matched	6	(7%)
	Stratified	1	(1%)
	Minimisation	0	(0%)
Design	Parallel group	80	(94%)
	Cross-over	3	(3%)
	Stepped wedge	2	(2%)
Objective	Superiority	83	(98%)
	Non-inferiority	1	(1%)
	equivalence	1	(1%)

* Notes. Binary and continuous outcomes were often contained within the same paper. They have been combined here so the denominator is the 85 papers included in the review.

2.2.2 Objective 1: Sample size methods for ordinal outcomes

Of the 85 sample size methods included there were two that were applicable to clustered ordinal data. The first method was proposed by Kim et al in 2005.[65] It deals with clustered ordinal data occurring in repeated measurement designs where the individual can be thought of as a cluster and measurements are taken on each individual at several time points. The second method was proposed by Campbell and Walters in their textbook on the design, analysis and reporting of cluster randomised trials published in 2014.[33] A standard parallel group design is assumed where measurements are taken on individuals within a cluster.

In the following sections I present the details of each method with a worked example. I start with the method proposed by Campbell and Walters which is more readily applied to the CRT context and the simpler of the two methods.

Campbell and Walters sample size method for ordinal outcomes

Notation

Campbell and Walters[33] state that the sample size formula for ordinal outcomes in individually randomised trials, derived by Whitehead,[2] can be inflated by the standard design effect to account for randomisation by cluster. For an ordered categorical outcome with k levels, $q = 1, 2, \ldots k$ where a higher category implies a worse outcome the sample size per group, m, is calculated as:

$$m = \frac{6[z_{1-\alpha/2} + z_{1-\beta}]^2/\theta^2}{[1 - \sum_{q=1}^{k} \bar{\pi}_q^3]}[1 + (n-1)\rho] \tag{2.1}$$

Where α and β are the overall Type I and Type II errors respectively, $z_{(1-\alpha/2)}$ and $z_{(1-\beta)}$ the corresponding percentage points of a standard normal distribution. $\bar{\pi}_q$ is the mean proportion expected in ordinal category q calculated as $\bar{\pi}_q = (\pi_{q1} + \pi_{q2})/2$, where π_{q1} and π_{q2} are the proportions in category q for the intervention and control groups. The number of individuals per cluster, n, is assumed constant and ρ is the intracluster correlation coefficient. The log-odds-ratio, θ_q is the probability of a response q or better (i.e. in a lower category) for a subject in the experimental group relative to a subject in the control group and is calculated as

$$\theta_q = log \left(\frac{P_{q1}(1 - P_{q2})}{P_{q2}(1 - P_{q1})} \right) \tag{2.2}$$

Where P_{q1} and P_{q2} are the cumulative probabilities of an individual in the experimental and control groups respectively giving a response in category q or better.

The underlying assumption of Whitehead's method is that of proportional odds, i.e. the log-odds-ratios for being in each category or better, θ_q share a common value, θ, which is used in the sample size calculation. If the observed probabilities do not satisfy the assumption of proportional odds power may be substantially reduced. Whitehead does not suggest an alternative sample size calculation if the proportional odds assumption is violated.

The method by Whitehead assumes an eventual analysis by a Mann-Whitney test which is equivalent to an ordinal regression when only treatment group is fitted in the model.[66] Ordinal regression methods can be extended to deal with clustered data and are more available in statistical packages than their non-parametric counterparts. It is assumed in this thesis that use of the design effect for ordinal outcomes assumes a random effects ordinal regression with proportional odds, although this is not explicitly stated by Campbell and Walters. More detail on the analysis of ordinal outcomes will be considered in Chapter Four.

Worked example

The use of the formula is illustrated by extending the example provided in the Whitehead paper to a CRT design.

In a two-arm clinical trial patient response after three months of treatment is measured as a four-level ordinal outcome (very good, good, moderate, poor). The proportions, π_q, and cumulative proportions, P_q, expected in each category are provided in Table 2.2 for the control (2) and experimental groups (1).

As can be seen in the table the log odds ratio does appear to be reasonably equal across the response categories, indicating the assumption of proportional odds is reasonable. In this example the reference

Table 2.2: Sample size calculation for an ordinal outcome example: Expected proportions in each outcome category for each treatment group of a trial*

	Very Good	Good	Moderate	Poor
	$(q = 1)$	$(q = 2)$	$(q = 3)$	$(q = 4)$
π_{q2}	0.2	0.5	0.2	0.1
P_{q2}	0.2	0.7	0.9	1
π_{q1}	0.378	0.472	0.106	0.044
P_{q1}	0.378	0.85	0.956	1
$\bar{\pi}_q = \frac{1}{2}(p_{q1} + p_{q2})$	0.289	0.486	0.153	0.072
$\bar{\pi}_q^3 =$	0.024	0.115	0.00358	0.000373
θ_q	0.888	0.887	0.881	

* (Source: Whitehead[2])

improvement for the sample size calculation was chosen by Whitehead as 0.887, power was set at 90% and a two-sided alpha of 0.05 chosen. For illustration I assume an ICC of 0.05 (a typical value for outcomes and clusters in health services research[21]) and cluster size of 5 in the design effect. Using equation 2.1 the number of individuals required per group for the clustered design is

$$m = \frac{6[z_{1-\alpha/2}+z_{1-\beta}]^2/\theta^2}{[1-\sum_{q=1}^{k}\bar{\pi}_q^3]}[1 + (n-1)\rho]$$

$$= \frac{6[1.96+1.282]^2/(0.887)^2}{[0.857]}[1 + (5-1) \times 0.05]$$

$$= 93.53 \times 1.2 = 115 \text{ people per arm}$$

This corresponds to 23 clusters of size 5 in each treatment group.

Areas for further development

The use of the design effect for ordinal outcomes has been suggested only recently and therefore is not as well established as it is for binary or continuous outcomes. This means that the reporting of the relevant estimates required may not be standard practice and it may be difficult for researchers to implement this formula if they have no good estimates to base their calculations upon. Examples of the parameter estimates calculated from real data may aid researchers in implementing this method. It would also be useful to use real datasets to consider how realistic the assumption of proportional odds might be. I present some real-life estimates in Chapter Four.

As described in Chapter One there are several papers that describe the calculation of the ICC for binary or continuous outcomes. The calculation for ordinal outcomes has received much less attention. For ordinal outcomes there are no recommendations about which ICC estimator to use in the design effect. Therefore evaluating the performance of this method using several different estimators of the ICC and under deviations from the underlying assumption of proportional odds is needed in order to provide clear recommendations for its use. I undertake this work in Chapter Five.

Kim et al sample size method for ordinal outcomes

Notation and model

Kim et al[65] propose a sample size calculation based upon an analysis by GEE, with inference based on the Wald test. Their sample size method is an extension of the method proposed by Rochon[67] for binary outcomes and assumes the analysis method of Lipsitz et al.[68] The method by Kim assumes a longitudinal design where measurements are taken at several time points for each individual. In the following notation I refer to sub-units within a cluster, in this example these subunits are time points but within the usual CRT context these would be individuals. At the end of this example I discuss the application of this method in the CRT context.

For each sub-unit we observe the response on a k-level ordinal outcome with categories $q = 1, 2, \ldots, k$. A higher category here is used to indicate a better outcome. Let Z_{ij} denote the ordinal response of the j'th sub-unit in the i'th cluster, $j = 1, 2, \ldots, n$ and $i = 1, 2, \ldots C$. The size of the cluster is assumed constant and denoted by n.

We form k indicator variables Y_{ijq}, where $Y_{ijq} = 1$ if sub-unit j has response q and $Y_{ijq} = 0$ otherwise. For each sub-unit we form a k-1 response vector $\mathbf{Y_{ij}} = [Y_{ij1}, \ldots, Y_{ij(k-1)}]'$ formed from the indicator variables, and for each cluster $\mathbf{Y_i} = [\mathbf{Y'_{ij}}, \ldots, \mathbf{Y'_{in}}]'$

The marginal probability of a response in category q is denoted by $Pr[Z_{ij} = q] = E[Y_{ijq}] = Pr[Y_{ijq} = 1] = \pi_{ijq}$ and the corresponding marginal cumulative probabilities by $Pr[Z_{ij} \leq q] = P_{ijq}$

Lipsitz et al analyse the data using a marginal model based on cumulative logits

$$logit[P_{ijq}] = \mathbf{X}\beta \tag{2.3}$$

Where $\mathbf{X}$ denotes a $(k-1) \times k$ design matrix for the $j'th$ sub-unit of the $i'th$ cluster and $\beta = [\alpha_1, \ldots, \alpha_{k-1}, \beta]'$ denotes a $k \times 1$ parameter vector. Where the α_q correspond to the $q'th$ cumulative logit and β denotes the effect of the intervention.

The GEE method assumed for the analysis is taken from Lipsitz et al,[68] an extension of Liang and Zegar's method for binary data.[69] The estimate of the treatment effect, β, is found by solution of the generalised estimating equation proposed by Lipsitz.[68] Calculation requires specification of the correlation matrix and the working covariance matrix. The significance of the treatment effect is then determined using the Wald test statistic. The variance of the treatment effect used in the Wald test can be calculated in two ways: via the model-based estimate of variance or a robust calculation which is less sensitive to miss-specification of the working covariance matrix. The details of these calculations will be seen in the worked example that follows and discussed in Chapter Four. The sample size method is described as being conservative if a maximum likelihood approach is taken.

Worked example

Here I work through the example described in section four of the original paper by Kim. I include additional detail providing the explicit step by step calculations to implement the method.

A clinical trial was conducted to assess the efficacy of auranofin compared to placebo for rheumatoid arthritis. Patients were randomly assigned to one of the two groups with equal allocation and assessed at 1 month, 3 months and 5 months post randomisation. The outcome measurement was a three-level ordinal outcome measuring self-assessment of arthritis (1) poor, (2) fair, (3) good. In the paper the power calculation was performed post-hoc with the correlation parameters and observed proportions in each ordinal category at each time point taken directly from the analysis, power was identified to be only 68%. In order to replicate the result given in the paper and for illustration of using this method to calculate a sample size at the design stage of a trial I set the power in the following worked example to be 68% and assume the observed proportions are those that were assumed for the trial design , with a two-sided 5% significance level. In this example the authors plan a test for an overall difference between the two groups for the three timepoints i.e. they assume the treatment effect is the same at every time point. This may not be the case for all longitudinal studies.

There are twelve steps in calculating the sample size.

Step 1: Specify the error rates

$\alpha = 0.05$

Type II error=β=0.32

$\therefore power = 0.68$

Step 2: Specify the number of treatment groups and their relative sizes

The number of treatment groups, T=2 [t=1 (intervention), t=2 (control)]. The relative size of each treatment group=1 i.e. equal allocation

Step 3: Specify the number of categories in the ordinal response

Number of categories in the ordinal response, k=3 (poor, fair, good)

Step 4: Specify the cluster size

In this example 3 measurements are taken on each individual at 1 month, 3 month and 5 month follow-up time points. Therefore the cluster size n=3.

Step 5: For each sub-unit specify the probabilities of a response for each ordinal category, by treatment group

The probabilities are described in Table 2.3.

Table 2.3: Sample size calculation for an ordinal outcome example: Observed proportions in each outcome category for each time-point and treatment group of a trial*

Treatment	Response	1 month	3 months	5 months
Placebo	Poor	29.7% (π_{i11})	27.7% (π_{i21})	25.2% (π_{i31})
	Fair	33.8% (π_{i12})	42.6% (π_{i22})	35.4% (π_{i32})
	Good	36.5% (π_{i13})	29.7% (π_{i23})	39.5% (π_{i33})
Auranofin	Poor	11.9% (π_{i11})	20.3% (π_{i21})	15.1% (π_{i31})
	Fair	51.0% (π_{i12})	35.1% (π_{i22})	34.9% (π_{i32})
	Good	37.1% (π_{i13})	44.6% (π_{i23})	50.0% (π_{i33})

* (Source: Kim[65])

Step 6: Specify the link function

The link function assumed is the cumulative logit link (clogit)

Step 7: Specify the design matrix

Assuming a proportional odds model the $[Tn(k-1) \times k]$ design matrix is given as:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

This can be defined separately for each treatment group.

$$X_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} ,X_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

Step 8: Determine the correlation structure, R

The correlation matrix R is a matrix of $n(k-1) \times n(k-1)$ dimension. The n diagonal blocks of R are defined by the $(k-1) \times (k-1)$ standard correlation matrix of a multinomial variable, the diagonal elements of which are

$$corr(Y_{ijq}, Y_{ijq}) = 1$$

and the off-diagonal elements are

$$corr(Y_{ijq}, Y_{ijq'}) = \frac{-\pi_{ijq}\pi_{ijq'}}{\sqrt{(\pi_{ijq}(1-\pi_{ijq})\pi_{ijq'}(1-\pi_{ijq'}))}}$$

where q and q' refer to levels of the response variable and $q \neq q'$

The remaining unknown elements of R are the elements of the $(k-1) \times (k-1)$ correlation matrix of two sub-units from the same cluster. The diagonal elements represent the correlation for pairs of indicators between two sub-units for the same category (either 'poor' or 'fair'). The off-diagonal elements represent the correlation between two sub-units for two adjacent categories. In this example the matrix used is that provided in Table V directly from the analysis of this dataset by Lipsitz et al, as the calculation is being performed post-hoc.[68] Exchangeability is assumed among members of a cluster.

$$\Phi = Corr(\mathbf{Y}_{ij}, \mathbf{Y}_{ij'}) = \begin{bmatrix} Corr(Y_{ij1}, Y_{ij'1}) & Corr(Y_{ij1}, Y_{ij'2}) \\ Corr(Y_{ij2}, Y_{ij'1}) & Corr(Y_{ij2}, Y_{ij'2}) \end{bmatrix} = \begin{bmatrix} 0.392 & 0.151 \\ 0.151 & 0.208 \end{bmatrix}$$

Therefore, the correlation matrices for the treatment and control group are defined as:

$$R_1 = \begin{bmatrix} 1 & -0.37 & 0.392 & 0.151 & 0.392 & 0.151 \\ -0.37 & 1 & 0.151 & 0.208 & 0.151 & 0.208 \\ 0.392 & 0.151 & 1 & -0.371 & 0.392 & 0.151 \\ 0.151 & 0.208 & -0.371 & 1 & 0.151 & 0.208 \\ 0.392 & 0.151 & 0.392 & 0.151 & 1 & -0.308 \\ 0.151 & 0.208 & 0.151 & 0.208 & -0.308 & 1 \end{bmatrix}$$

$$R_2 = \begin{bmatrix} 1 & -0.462 & 0.392 & 0.151 & 0.392 & 0.151 \\ -0.462 & 1 & 0.151 & 0.208 & 0.151 & 0.208 \\ 0.392 & 0.151 & 1 & -0.533 & 0.392 & 0.151 \\ 0.151 & 0.208 & -0.533 & 1 & 0.151 & 0.208 \\ 0.392 & 0.151 & 0.392 & 0.151 & 1 & -0.428 \\ 0.151 & 0.208 & 0.151 & 0.208 & -0.428 & 1 \end{bmatrix}$$

Step 9: Derive the working covariance matrix V and the matrix D

The matrix $\mathbf{V_i}$ is the working covariance matrix of $\mathbf{Y_i}$, the cluster response vector. Since the marginal distribution of $\mathbf{Y_{ij}}$ is multinomial the $(k-1) \times (k-1)$ diagonal blocks of $\mathbf{V_i}$ are the multinomial covariance matrices. That is each diagonal entry is the variance $\pi_{ijq}(1-\pi_{ijq})$ with covariance $-\pi_{ijq}\pi_{ijq'}$ on the off diagonal.

The covariance matrix V can be defined for each treatment group as

$$\mathbf{V}_t = \mathbf{A}_t^{1/2}\mathbf{R}_t\mathbf{A}_t^{1/2}$$

Where A_{ij} is a diagonal matrix with the binary variances of the indicator variable, $var(Y_{ijq})$, on the diagonal $A_{ij} = Diag[\pi_{ij1}(1-\pi_{ij1}), \ldots \pi_{ij(k-1)}(1-\pi_{ij(k-1)})]$.
$\therefore A_{it}^{1/2} = Diag[\pi_{it1}(1-\pi_{it1})^{1/2}, \ldots \pi_{it(k-1)}(1-\pi_{it(k-1)})^{1/2}]$
and the matrix $A_t = Diag[A_{i1}, A_{i2}, \ldots, A_{i(k-1)}]$

Using the matrices R and A the matrix V for the intervention group is calculated as:

$$\mathbf{V}_1 = \begin{bmatrix} 0.104839 & -0.06069 & 0.0510534 & 0.0233353 & 0.0454454 & 0.0233046 \\ -0.06069 & 0.2499 & 0.0303625 & 0.0496275 & 0.0270273 & 0.0495621 \\ 0.0510534 & 0.0303625 & 0.161791 & -0.071253 & 0.0564554 & 0.0289506 \\ 0.0233353 & 0.0496275 & -0.071253 & 0.227799 & 0.0258045 & 0.0473198 \\ 0.0454454 & 0.0270273 & 0.0564554 & 0.0258045 & 0.128199 & -0.052699 \\ 0.0233046 & 0.0495621 & 0.0289506 & 0.0473198 & -0.052699 & 0.227199 \end{bmatrix}$$

$$\Rightarrow \mathbf{V}_1^{-1} = \begin{bmatrix} 25.366113 & 10.094366 & -9.066519 & -5.892443 & -7.66796 & -4.19997 \\ 10.094366 & 8.5726621 & -4.535907 & -3.380871 & -3.730698 & -2.488696 \\ -9.066519 & -4.535907 & 16.285962 & 8.3823852 & -6.047153 & -3.304235 \\ -5.892443 & -3.380871 & 8.3823852 & 9.313599 & -3.812866 & -2.550373 \\ -7.66796 & -3.730698 & -6.047153 & -3.812866 & 17.726874 & 7.2767969 \\ -4.19997 & -2.488696 & -3.304235 & -2.550373 & 7.2767969 & 8.0152027 \end{bmatrix}$$

and the control group

$$\mathbf{V}_2 = \begin{bmatrix} 0.208791 & -0.100386 & 0.0801588 & 0.0341188 & 0.0777666 & 0.0329952 \\ -0.100386 & 0.223756 & 0.0319649 & 0.0486532 & 0.031011 & 0.047051 \\ 0.0801588 & 0.0319649 & 0.200271 & -0.118002 & 0.0761634 & 0.032315 \\ 0.0341188 & 0.0486532 & -0.118002 & 0.244524 & 0.0324182 & 0.0491861 \\ 0.0777666 & 0.031011 & 0.0761634 & 0.0324182 & 0.188496 & -0.089208 \\ 0.0329952 & 0.047051 & 0.032315 & 0.0491861 & -0.089208 & 0.228684 \end{bmatrix}$$

$$\Rightarrow \mathbf{V}_2^{-1} = \begin{bmatrix} 33.168681 & 24.883259 & -29.4581 & -23.04584 & -2.800597 & -1.878367 \\ 24.883259 & 23.892531 & -24.02663 & -19.19547 & -2.02613 & -1.77263 \\ -29.4581 & -24.02663 & 56.39634 & 42.262828 & -21.67354 & -16.32029 \\ -23.04584 & -19.19547 & 42.262828 & 36.393116 & -16.87288 & -13.1071 \\ -2.800597 & -2.02613 & -21.67354 & -16.87288 & 26.991706 & 18.041944 \\ -1.878367 & -1.77263 & -16.32029 & -13.1071 & 18.041944 & 17.171919 \end{bmatrix}$$

$D = \frac{\partial \pi}{\partial \beta}$ is a matrix of partial derivatives of the mean of the outcome with respect to the regression parameters. If we assume η represents the linear predictor using the chain rule Rochon derives:[67]
$\frac{\partial \pi}{\partial \beta} = \frac{\partial \pi}{\partial \eta} \times \frac{\partial \eta}{\partial \beta} = \frac{\partial \pi}{\partial \eta}\mathbf{X}$
Therefore the matrix D for each treatment group is defined as:

$$\mathbf{D}_t = \mathbf{\Delta}_t \mathbf{X}_t$$

For each treatment group the D matrices are calculated as:

$$\mathbf{D}_1 = \begin{bmatrix} 0.208791 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0229840 & 0 & 0 & 0 & \\ 0 & 0 & 0.200271 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.00852 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.188496 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.050268 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$\therefore \mathbf{D}_1 = \begin{bmatrix} 0.208791 & 0 & 0 \\ 0 & 0.022984 & 0 \\ 0.200271 & 0 & 0 \\ 0 & 0.00852 & 0 \\ 0.188496 & 0 & 0 \\ 0 & 0.050268 & 0 \end{bmatrix}$$

$$\mathbf{D}_2 = \begin{bmatrix} 0.104839 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.12852 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.161791 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.085293 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.128199 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.121801 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$\therefore \mathbf{D}_2 = \begin{bmatrix} 0.104839 & 0 & 0.104839 \\ 0 & 0.12852 & 0.12852 \\ 0.161791 & 0 & 0.161791 \\ 0 & 0.085293 & 0.085293 \\ 0.128199 & 0 & 0.128199 \\ 0 & 0.121801 & 0.121801 \end{bmatrix}$$

Step 10: Calculate $\hat{\beta}$

The vector of parameters in the model can now be estimated by solution of the following generalised estimating equation.

$\hat{\beta} = [\sum_t \mathbf{D}_t' \mathbf{V}_t^{-1} \mathbf{D}_t]^{-1} [\sum_t \mathbf{X}_t' \mathbf{W}_t h(\theta_t)]$

Where $\mathbf{W}_t = \Delta_t V_t^{-1} \Delta_t$ and $h(\theta_t)$ is a vector of cumulative logits i.e. ln(cumulative probability/(1-cumulative probability). For the treatment and placebo groups these vectors are:

$$h(\theta_1) = \begin{bmatrix} -2.001934 \\ 0.5279292 \\ -1.367649 \\ 0.2168457 \\ -1.726779 \\ 0 \end{bmatrix} ,h(\theta_2) = \begin{bmatrix} -0.861625 \\ 0.5537276 \\ -0.959392 \\ 0.8616248 \\ -1.087974 \\ 0.4305291 \end{bmatrix}$$

Using the matrices as previously defined

$$[\sum_t \mathbf{D}_t' \mathbf{V}_t^{-1} \mathbf{D}_t]^{-1} = \begin{bmatrix} 2.7910582 & 2.1564694 & -2.572864 \\ 2.1564694 & 10.146841 & -4.90384 \\ -2.572864 & -4.90384 & 6.2484071 \end{bmatrix}$$

$$\mathbf{W}_1 = \begin{bmatrix} 1.4459449 & 0.1194111 & -1.231784 & -0.040996 & -0.110221 & -0.019714 \\ 0.1194111 & 0.0126216 & -0.110595 & -0.003759 & -0.008778 & -0.002048 \\ -1.231784 & -0.110595 & 2.2619711 & 0.0721134 & -0.818182 & -0.1643 \\ -0.040996 & -0.003759 & 0.0721134 & 0.0026418 & -0.027098 & -0.005614 \\ -0.110221 & -0.008778 & -0.818182 & -0.027098 & 0.9590353 & 0.1709531 \\ -0.019714 & -0.002048 & -0.1643 & -0.005614 & 0.1709531 & 0.0433912 \end{bmatrix}$$

and in the other treatment group

$$\mathbf{W}_2 = \begin{bmatrix} 0.2788044 & 0.1360106 & -0.153786 & -0.05269 & -0.103059 & -0.053631 \\ 0.1360106 & 0.141598 & -0.094317 & -0.037061 & -0.061467 & -0.038958 \\ -0.153786 & -0.094317 & 0.4263067 & 0.1156739 & -0.125427 & -0.065114 \\ -0.05269 & -0.037061 & 0.1156739 & 0.0677555 & -0.041692 & -0.026495 \\ -0.103059 & -0.061467 & -0.125427 & -0.041692 & 0.2913409 & 0.1136255 \\ -0.053631 & -0.038958 & -0.065114 & -0.026495 & 0.1136255 & 0.1189094 \end{bmatrix}$$

$$\therefore \beta = \begin{bmatrix} -0.9802 \\ 0.4783 \\ -0.5032 \end{bmatrix}$$

Therefore the GEE estimate of the treatment effect is -0.5032.

Step 11: Identify the specific hypothesis of interest

The Null and Alternative hypotheses are defined as:

$H_0 : \mathbf{H}\beta = h_o$ versus $H_1 : \mathbf{H}\beta \neq h_0$, In this example $h_0 = 0$

$\mathbf{H} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$

Step 12: Calculate the sample size per group (m)

The sample size calculations are based on the Wald test statistic which is asymptotically distributed as a $\chi^2_{(h),\lambda}$ random variable with non-centrality parameter λ. In this example, as in most cluster randomised trials, we conduct a one degree of freedom hypothesis test on the difference between the two treatments. Therefore h=1.[70]

The sample size is found by solving the following equation for the minimum number of subjects, m, required in each treatment group.

$$1 - \beta = \int_{\chi^2_{(h),1-\alpha}}^{\infty} f(x; h, \lambda)dx \tag{2.4}$$

Where β is the type II error rate, $\chi^2_{(h),1-\alpha}$ is the critical value from the central $\chi^2_{(h)}$ distribution with h degrees of freedom and $f(x; h, \lambda)dx$ is the probability density function of the $\chi^2_{(h),\lambda}$ distribution. In this example from tables of the chi-squared distribution on 1 degree of freedom at the 5% 2-sided level, the critical value from the central chi-squared distribution is 3.84.

The non-centrality parameter is calculated as

$\lambda \approx m(\mathbf{H}\hat{\beta} - \mathbf{h}_0)'[\mathbf{H}[\sum_t \mathbf{D}_t'\mathbf{V}_t^{-1}\mathbf{D}_t]^{-1}\mathbf{H}']^{-1}(\mathbf{H}\hat{\beta} - \mathbf{h}_0)$

Using previously defined estimates

$\lambda \approx m[-0.5032 \times 0.1600407722 \times -0.5032] = m[0.04052]$

Given h, α and β we can use the CNONCT function in SAS to return the non-negative non-centrality parameter from a non-central chi-square distribution

$$\lambda\text{=CNONCT(3.84, 1, 0.32)=5.8259921.}$$

Therefore, m per group=5.8259921/0.04052=144. For 90% power m per group increases to 260.

Adaptation to cluster randomised trials

Methods designed for longitudinal data are not always generalizable to CRTs. Often a treatment interaction with time is the main parameter of interest, the correlation structure is often assumed to be autoregressive and the cluster sizes tend to be small. In some respects the methodology by Kim might be transferable to the CRT scenario. The form of the model used in the example presented here does not include an effect for time and the correlation matrix for two individuals from the same cluster is assumed to be exchangeable, a common assumption in the CRT context too. However, there are a number of practical limitations that limit its usefulness in calculating sample size for a CRT.

In the longitudinal context, for which the method was designed, the number of observations per person (or cluster size) is usually small. In the cluster randomised trial context the cluster size is much larger and the definition of matrices and matrix algebra involved in using this method become increasingly complex with increasing cluster size and the number of ordinal categories. This may substantially increase the computational time of using this method in the CRT context. Currently the GEE approach to the analysis of clustered ordinal outcomes assumed in this approach is a less popular approach and not available in all statistical software packages. The complexity involved in using this method also means that in practice a statistician would be required to implement it.

Areas for development

Further work could be done on the methodology by Kim to: look at whether this method could be adapted to be suitable to the CRT design; consider its performance under situations that are more reflective of a CRT, such as larger cluster sizes; and find appropriate estimates to use in its calculation. I chose not to look at developing this method further given that it did not seem to offer any advantages over the design effect approach, the calculations are more complex and the GEE is not a common approach in the analysis of ordinal outcomes.

2.2.3 Objective 2: Methods used in the derivation of sample size formulae

The second objective of the review was to identify the different approaches used in the derivation of sample size formula for cluster randomised trials. In a minority of papers (n=9, 11%) a sample size method was presented and evaluated but the description of how it was derived was unclear. The approaches used in the remaining 76 papers could be broadly categorised into one of the following six methods: calculation of the power function from first principles (n=44, 52%); derivation of the inflation factor as a ratio of variances of the treatment effect (n=8, 10%); sample size subject to optimality or cost constraint criteria (n=11, 14%); sample size by simulation (n=5, 6%); Bayesian methods (n=4, 5%); or adaptation of a pre-existing method (n=3, 4%).

I now describe each of these approaches in more detail.

Calculation of a power function from first principles

The majority of papers, n=44 (52%) took a general approach to sample size derivation, mirroring the derivation method described in Chapter One for individually randomised trials. The steps in this approach are as follows:

1. Describe the estimate of the intervention effect
2. Select a test statistic for the intervention effect
3. Express the variance of the intervention effect under the null (and possibly alternative hypothesis)
4. Determine the distribution of the test statistic under the null hypothesis
5. Generate a statement about power or sample size required

In 18 (41%) of these papers the variance of the intervention effect calculated in step 3 simplified to the variance under individual randomisation multiplied by a factor, termed the design effect. The sample size formula then simplified to the formula for an individually randomised trial multiplied by the design effect. The first of these papers was published in 1982 by Donner, Birkett and Buck for continuous and binary data analysed at the cluster level. Since then design effects have been established for: time-to-event outcomes;[71,72] count outcomes;[73] variable cluster sizes;[74–76] cross-over, equivalence, three-level and stepped-wedge designs;[77–82]analysis by GEE[75,83] or analysis of covariance,[84] and adjusting for attrition.[85] In the majority of cases the design effect methods were derived for cluster-level analysis. For continuous outcomes with fixed cluster sizes this is equivalent to an individual-level analysis adjusted for clustering.

In the remaining 26 (59%) papers there was no such simplification of the variance into the two components. These methods were more likely to assume an individual-level analysis where the estimate of the treatment effect can be more complex.

For derivation based upon a GEE (or marginal model) the estimate of treatment effect became more complex with the inclusion of covariates,[86] more complex correlation structures,[87] ordinal outcomes,[65] adaptive designs[88] and time-to-event outcomes.[89]

For a random effects model the treatment effect, β, is usually estimated via maximization of the likelihood function and the significance of the treatment effect assessed using the likelihood-ratio, Score or Wald statistics. The use of random effects models appears to have been constrained to continuous outcomes, where estimation of the treatment effect is more straightforward. For discrete responses, in particular ordinal outcomes, except in rare cases (the complementary log-log function with the log of a gamma or inverse Gaussian distribution for the random effects), the likelihood function cannot be written in closed form and therefore must be approximated.[90] The recommended approximation is via Gauss-Hermite quadrature, a procedure for performing numerical integration of the likelihood function using a series expansion evaluated at certain quadrature points. An alternative method is to take a quasi-likelihood approach. A quasi-likelihood approach starts by linearizing and approximating the nonlinear response model using a Taylor series expansion before maximising the likelihood. If the Taylor series expansion is based on the fixed parameters only it is referred to as a marginal quasi-likelihood (MQL), if it is also based on the random effects it is called penalized quasi-likelihood (PQL) and depending upon the extent of the Taylor approximation it may be first- or second-order quasi-likelihood.

Ratio of variances

Rather than generating a specific statement about power or sample size requirement some papers derived inflation factors directly by specification and comparison of the variance of the treatment effect estimate under two alternative designs.

These inflation factors are referred to as design effects or measures of relative efficiency. The difference between the definitions of these two inflation factors is not always clear. In this thesis I use the following definitions, taken from the Cambridge Dictionary of Statistics[91]

The design effect is defined as 'the ratio of the variance of an estimator under the particular sampling design used in a study to its variance at equivalent sample size under simple random sampling without replacement'. Therefore, in the context of this thesis, the design effect is a comparison of the cluster randomised design with the individually randomised design. The sample size required for the individually randomised design is multiplied by the design effect to calculate the sample size required for the cluster randomised case.

The relative efficiency is defined as 'the ratio of the variances of two possible estimates of a parameter, or the ratio of the sample sizes required by two statistical procedures to achieve the same power'. Therefore, the relative efficiency can be a comparison of the cluster randomised and individually randomised designs but can also be used to compare two alternative clustered designs. In the comparison of cluster and individually randomised trials the reciprocal of the relative efficiency is equivalent to the design effect. The reciprocal of relative efficiency provides the multiplication factor by which the sample size must be inflated. Many of the papers identified in the review resulted in a sample size calculation involving a design effect. However, in only one paper was this calculated directly as the variance of the cluster summary statistic assuming variable cluster sizes divided by the equivalent statistic ignoring clustering.[92]

In the literature relative efficiency inflation factors have been defined to compare: unequal versus equal cluster sizes;[93–95] cross-sectional versus cohort samples;;[96] matched versus unmatched designs;[97] and a cross-over versus parallel group design.[98]

Defining optimality criteria

In cluster randomised trials there may be several combinations of cluster size and number of clusters that produce designs with equivalent power. In these situations the optimal sample size at each level may be constrained by the total budget for the trial. This total cost function is fixed and is made up of the cost of sampling a cluster and the cost of sampling an individual. The optimal allocation of clusters and individuals, subject to cost constraints have been derived for binary and continuous outcomes.[99–109]

Sample size by simulation

Five papers were identified that described sample size calculation through the use of simulation.[110–114] These authors used simulation due to the lack of available simple sample size formula for their design or because of the inherent complexity in their proposed design.

The general process for sample size by simulation involves simulating a large number of data sets, most commonly at the level of the individual, each one to represent a potential dataset of results from the trial. For each simulated data set the planned analysis is conducted and the empirical power calculated as the percentage of tests where the null hypothesis is rejected. Changes in the input parameters, such as number of clusters, can be made until adequate power is achieved. Simulation has the flexibility to incorporate all the complexities of the trial. However, it can be computationally intensive to implement.

From a more technical viewpoint the simulation procedure starts with generating an observation for each cluster alongside a variable to indicate which treatment group the cluster belongs to. Within each cluster an observation is generated for each individual. The number of individuals per cluster can be chosen to be fixed, randomly selected for each cluster with defined probability from one of several user-defined cluster size values, or randomly selected from a user-defined probability distribution. It is also possible to randomly assign covariates at this stage using user-defined probability distributions. A model is then defined for generating the outcome variable for the dataset, with a user defined treatment effect, for example a linear or logistic random effects model for continuous or binary outcomes respectively. The level of the ICC is controlled through the assumed distribution of the cluster-level random effects and the individual-level residuals. This procedure is repeated to generate a large number of datasets.

Bayesian methodology

Four papers took a Bayesian perspective to sample size determination.[115–118] Bayesian statistics is a branch of statistics that expresses uncertainty about unknown parameters in terms of a probability distribution (known as a posterior distribution). This posterior distribution is calculated using both prior knowledge and the observed data. In several of these papers the prior knowledge involved specification of a distribution for the ICC. The probability distribution for the power of the study reflecting ICC uncertainty would then be derived. Posterior distributions for unknown parameters can often be computationally complex and in these situations Markov Chain Monte Carlo (MCMC) simulation methods using specialist Bayesian software such as WinBUGS are needed to find estimates from the distribution, for example the average power. Bayesian methods are considered by many to be more complex to implement than classical methods. The development of statistical methodology for clinical trials and the corresponding statistical software has largely focused on classical, or frequentist, approaches. Many of the papers that implemented a Bayesian approach to sample size did assume that the eventual analysis would take a frequentist approach.

Adaptation of a pre-existing method

There were three papers that adapted or extended a previously derived method. Examples included extending methods used for rates to time-to-event data, methods for re-estimating the sample size at an interim look and dealing with ICC uncertainty.[119–121]

2.3 Discussion

2.3.1 Main findings

This review is the most comprehensive and up-to-date review of sample size calculations for cluster randomised trials. Additional methodology published between the date of this review and the end of my research is summarised in Chapter Six of this thesis.

The results of this review show that there is a large body of literature available on sample size methodology for cluster randomised clinical trials, 85 papers were identified. The literature is dominated by methods which are applicable to binary or continuous outcomes. Methods for alternative outcomes, particularly ordinal outcomes were lacking. In 2005 a paper by Kim discussed sample size for correlated ordinal outcomes. Their method assumes that the analysis will be by GEE with the treatment effect evaluated using a Wald test, and that each cluster is of the same size. As it currently stands it is not obvious how this method can be transferred to the cluster randomised trial situation.

A second method proposed by Campbell and Walters proposes multiplication of the sample size formula for individually randomised trials for ordinal outcomes by the standard design effect.[33] This method offers great benefits of being simple to implement and familiar to researchers. However its performance has not been formally evaluated and no recommendations are provided to guide researchers on an appropriate estimate of the ICC to use.

Of the 85 papers the most popular derivation approach was to derive a sample size formula from first principles. This approach was used in over half of the papers, 52%. For individual-level analyses the GEE method of analysis was assumed for discrete outcomes. The use of random effects models was constrained to continuous outcomes.

2.3.2 Strengths and limitations

This review has some limitations. The search results are biased towards methodology that has been published in the English language. There is also a bias towards methodology published in the medical/healthcare literature with the choice of the electronic databases chosen for searching. However, I felt that this is where the majority of cluster randomised trial methodology is published and health research is my area of expertise. Potential limitations of the search process include the fact that the search was not performed by two independent researchers and decisions about initial inclusion of a paper were made by me alone. This means there is potential that some papers may have been missed during the process. However, the search results were reviewed by experts in the field and I am therefore confident that no key methods have been excluded.

There is limited guidance available for the conduct and reporting of reviews of methodology such as this and different methods from those used in more traditional systematic reviews are required. A strength of this review is that the search procedure was developed to be as systematic as possible and the methods used mirrored those of a more traditional review where possible to ensure a comprehensive summary of the methodology available. The search terms were validated prior to use and data abstraction independently reviewed for a select number of articles. The only previous review of CRT methodology to implement such robust methods was conducted 15 years ago[8]

2.3.3 Implications for this research

Sample size methodology for ordinal outcomes were set out in only two of the 85 papers, of these both have scope for further development.

At the beginning of this research I started, in parallel, two approaches to sample size calculation for ordinal outcomes. The first was to derive a formula from first principles and the second was to calculate the required sample size using simulation based methods and explore whether any patterns in the required sample size emerged.

As the most common approach to the analysis of ordinal outcomes for the non-clustered case is by proportional odds model it seemed most sensible to assume the analysis for the clustered case would be the random effects extension to this model. However, it became clear from my review that there would be great difficulty in using a random effects model as the basis for formula derivation from first principles due to the complexity involved in estimation and inference, an issue that had been identified for binary outcomes. The alternative was a GEE approach but this had already been used by Kim et al and is not a popular analysis approach for ordinal outcomes.[65]

From an initial exploration of sample size by simulation it appeared that some patterns were emerging and the simple design effect might offer an appropriate approach to sample size calculation for ordinal outcomes.

While working on the two approaches above the use of the design effect for sample size calculation with ordinal outcomes was proposed by Campbell and Walters.[33] However, the performance of the method in different scenarios was unknown and there was no guidance around which estimate of the ICC should be used.

The focus of my research therefore was diverted in order to evaluate the design effect method and provide guidance for applied researchers around its use. I made this decision because the design effect method: has the advantage of simplicity that would likely outweigh any newly derived formula, that may be a more complex calculation; is already well established for binary and continuous outcomes and Donner's paper proposing this method still remains highly cited, despite its age; is familiar to researchers; and the calculation is easy to perform.

Chapter 3

Ordinal outcomes in CRTs

In this thesis I concentrate on providing practical guidance for using the sample size method for clustered ordinal outcomes which suggests multiplication of the sample size derived by Whitehead[2] for individual randomisation by the design effect.

In this chapter I establish whether there is sufficient demand for such guidance i.e. a prevalence of cluster randomised trials with primary outcomes that are ordinal. Using an existing sample of 300 published cluster randomised trials, I estimate the prevalence of ordinal outcomes, look at the methodological approaches used in their design and analysis, and describe the design characteristics of these trials. The characteristics of these trials will be used to inform the design of simulation studies presented in Chapter Five that will evaluate the design effect approach for sample size calculation in CRTs with ordinal outcomes.

Some of the work within this chapter was undertaken collaboratively. Aspects that were conducted collaboratively, and my role in the collaboration, are clearly described within the following sections.

3.1 Background

Ivers et al conducted the largest review to date of 300 cluster randomised trials published between 2000 and 2008 with the aim of describing the quality of methods and reporting.[30] I planned to use the same sample of 300 trials to meet the aims of this chapter. My plan was to utilize some of the information the Ivers et al group had already collected on this sample such as whether a sample size was calculated and whether it was appropriately adjusted for clustering. I would then review all 300 papers myself to extract the additional information I required to meet my objectives. However, the dataset of the information extracted in the original review of these 300 trials was not publicly available at the time of my research, as the research team were planning further work using these trials. The statisticians involved in the original review were Monica Taljaard (MT) and Stephanie Dixon (SDX) who were based at the Ottawa Hospital Research Institute at Ottawa Hospital and the Schulich School of Medicine and Dentistry at the University of Western Ontario in Canada respectively. An agreement was reached between us whereby the information previously extracted would be shared with me and we would combine our resources to extract further information from each trial to satisfy the planned objectives of my research and that of the Canadian based research team.

3.2 Methods

3.2.1 Objectives

The review was designed to meet the objectives of my research and that of the Canadian based team. As the focus of this thesis is my research I have described my objectives as primary and those of the Canadian research team as secondary to distinguish the two.

The primary objectives:

- To describe the prevalence of ordinal primary outcomes in cluster randomised trials

and for trials with ordinal outcomes:

- To describe the design features of trials using ordinal primary outcomes
- To describe the prevalence of reporting observed estimates required for future sample size calculations i.e. the proportions in each ordinal category in each treatment group and a measure of the observed ICC.
- To describe the methodological approach used in the sample size calculation

• To describe the quality of sample size reporting in comparison to the recommended elements provided in the CONSORT 2004 extension for cluster randomised trials.[50]

The CONSORT 2004 statement requires five elements to be included in the description of the sample size calculation. (1) the type I error rate, (2) power, (3) estimates of outcomes in each group or minimum important target effect, (4) the number of clusters or average cluster size and, (5) the assumed measure of intracluster correlation, design effect or coefficient of variation. In 2012 the CONSORT 2004 statement for cluster randomised trials was updated.[51] The 2012 version additionally recommends that it be specified whether equal or variable cluster sizes has been assumed. In this sample the trials were published between 2000 and 2008 and therefore not expected to report according to the 2012 extension.

The secondary objectives:

The secondary objectives of this review are restricted to the subset of trials that reported a sample size calculation. These objectives look at the reporting and methodological quality of sample size calculations in cluster randomised trials in more detail than was considered in the original review by Ivers et al. These objectives were generated by the Canadian based research team and jointly agreed by both research groups.

• To describe the methodological quality of sample size reporting in cluster randomised trials
• To establish which sample size descriptive elements, as recommended in the CONSORT extension for cluster randomised trials, are best and worst reported
• To determine whether sample size reporting practices have improved since the introduction of the CONSORT 2004 extension for cluster randomised trials
• To evaluate the accuracy of the a priori estimates used in the sample size calculation by making comparisons with their observed values at the end of the trial

The methods and results of these secondary objectives have been published, and are not repeated within this thesis. I took the lead role in the analysis of this data and preparing and submitting the work for publication.[122]

In the following sections I describe the methods used to meet the objectives of this review.

3.2.2 Data source

This review was conducted using the 300 published reports of cluster randomised trials identified by Ivers et al.[30]

For trials with ordinal outcomes each primary author was contacted by email to see if their trial dataset was available, in order to calculate estimates of the ICC and proportions in each ordinal category where missing from the trial report. If the primary author could not be reached an alternative author was contacted.

3.2.3 Justification of data source

I chose trial reports as the data source for this review as I considered them to provide a more representative sample of clinical trials than other sources, such as published protocols. However, the level of detail and general reporting quality may be less than ideal given the strict length restrictions often imposed by journals at the time these trials were reported.

I did consider published trial protocols as a data source. The biggest advantage of the trial protocol over the trial report is the level of detail is likely to be greater, as they are not subject to the same length restrictions. The trial protocol may contain detailed information about decisions made during the course of design, for example possible justification for dichotomising an ordinal outcome, treating it as continuous, or choosing an alternative outcome. The protocol may provide insight into whether these decisions were driven by clinical relevance or the availability and complexity of statistical methods for ordinal outcomes. However, although the publication of trial protocols appears to be more common in recent years those trials which publish their protocols are more likely to be of high quality and not a representative sample of all cluster randomised trials. Their use was not considered further.

In Chapter one twenty-three reviews of cluster randomised trials were summarised in Table 1.1. The number of trials included in each review ranged from 15 to 300. With the availability of these existing reviews, and due to the time limitations imposed upon my research, I chose to utilize an existing review as my data source rather than conduct a new one. I opted to use the review by Ivers et al for two reasons.[30] Firstly it was the largest review to have been conducted and secondly, unlike some of the other reviews it was less restrictive in the databases searched and health areas included and was more likely to provide a representative sample of published cluster randomised trials across all areas of health research.

3.2.4 Inclusion and exclusion criteria

The search strategy implemented in the review by Ivers et al produced a sample of CRTs that the authors describe to be representative of all Medline publications. A publication was included if it was the main study report and was published in an English language journal between the years 2000 and 2008. Trial protocols, pilot studies, or papers that reported only baseline results or secondary analyses were excluded.

For my research I initially considered all 300 trials and then refined this to the subset reporting ordinal outcomes. The collaboration work with the Canadian group was restricted to those trials that reported a sample size calculation.

3.2.5 Data collection

In the original review by Ivers et al 47% of trials did not identify a primary outcome. In these cases, for the purpose of data extraction an outcome was designated as primary. Where multiple outcomes were specified the primary outcome was chosen as that which was reported first in the abstract or analysis. All data extraction in the original review was related to the primary outcome. The same approach was taken in my review with the exception being those that reported a sample size calculation. For these trials data extraction was based upon the outcome used in the sample size calculation if this was different from the originally defined primary outcome. This was not a frequent occurrence and was done in order to maximise information gained about sample size calculations. Where several follow-up time points were given data extraction was based on the final time point, unless an earlier time-point was identified as primary.

Descriptive information collected in the original review was made available to me by the authors. It included information on year of publication, impact factor, trial design (parallel trial, factorial, cross-over, other), method of randomisation (completely randomised, stratified, pair-matched, other), a description of the primary outcome and whether a sample size calculation had been reported.

The following sections describe the details of what information was additionally extracted and by whom.

Information extracted collaboratively (by CR, MT, and SDX)

For papers that reported a sample size calculation the following information was collected collaboratively: (the full data extraction form can be found in appendix iii):

Study design:

- The outcome for which the sample size calculation was performed for

Sample size:

- The method, or citation, of sample size calculation
- Data type of the primary outcome (binary, categorical, ordinal, continuous, count, time-to-event data, other, unclear)
- Type I, Type II error rate, and whether a one or two-sided test assumed
- The estimator used to describe the correlation within clusters, its value and any justification for its value
- Additional aspects accounted for in the calculation i.e. attrition, variable cluster sizes
- The value of the expected response in the control and treatment arms with a measure of the effect size and any justification provided for these values
- Target sample size: Total number of clusters and individuals

Analysis:

- The achieved sample size used in the analysis: Total number of clusters and individuals
- The method of analysis used
- Data type at the level (cluster or individual) corresponding to the analysis (binary, categorical, ordinal, continuous, count, time-to-event data, other, unclear)
- Data type as used in the analysis (binary, categorical, ordinal, continuous, count, time-to-event data, other, unclear)
- Observed values of: the measure of correlation; the response in the control and treatment arms;

and unadjusted effect size

Information extracted independently (by CR)

The following information was extracted independently on all articles that did not report a sample size calculation:

Study design:

• Data type of the primary outcome (binary, categorical, ordinal, continuous, count, time-to-event data, other, unclear)

For the subset of trials identified with an ordinal primary outcome the following information was extracted :

Study design:

• Description of the outcome

• Number of ordinal categories, with category description

• Description of the intervention

• Description of the cluster

• Description of the sub-units within a cluster

If a sample size calculation was present the following information was included in the information collected collaboratively, else this information was extracted independently:

Analysis:

• The achieved sample size used in the analysis: total number of clusters and sub-units

• The method of analysis used

• Data type as used in the analysis (binary, categorical, ordinal, continuous, count, time-to-event data, other, unclear)

• Observed values of: the measure of correlation; the response in the control and treatment arms; and unadjusted effect size

3.2.6 Data management

Collaborative extraction (by CR, MT, and SDX)

I drafted a data extraction form. This was discussed and finalised with MT, SDX, SE, and AC. The extracted data was transcribed from the data extraction form and stored in an Access database, to which the data collection from the originally review was also imported. I designed and tested the database (screen shots available in appendix iv).

CR, MT, SDX extracted the data in pairs for each article that reported a sample size calculation. The trials were divided into batches of approximately ten trials. Each batch of trials was assigned to an extracting pair. MT was responsible for the allocation and rotation of extraction pairs. Each member of the team had a copy of the database and was responsible for storing the paper and electronic versions for the trials to which they were assigned. After each set of 10 trials had been extracted the electronic database from each extracting pair was emailed to me. I imported each set of data into Stata where I used the cf2 command to compare the two datasets and list the discrepancies. The list of discrepancies was then sent to the extracting pair and were reviewed and resolved by consensus within the pair.

After the discussion of any discrepancies in the data extraction one member of each extracting pair was responsible for updating the database. This was MT for all her batches and CR for the batches marked with SDX. At the end of the project these two datasets were merged to create the final dataset on which data checking took place, described in the next section.

Individual abstraction (by CR)

The majority of the information I extracted independently was for those trials that had an ordinal primary outcome. As the number of trials for which this information was extracted was small an Excel spreadsheet was deemed adequate for its storage.

3.2.7 Validation and data checking

Collaborative extraction

The final version of the data extraction form was tested on five papers selected by MT, who was most familiar with the trials included in the sample, with the aim of testing the form on a variety of trials. No changes to the form were deemed necessary after piloting the form.

After the collection of all the data a data checking plan was agreed to ensure consistent recording between both extraction pairs and extractions over time. This data checking included some of the following aspects:

- Agreeing the categorisation of free text responses such as the method of analysis used
- Ensuring consistency in reporting percentages as decimals rather than whole numbers for example 0.85 versus 85%
- Double checking papers where a large discrepancy was seen between target and actual sample size
- Double checking papers where a large discrepancy was seen between target and actual sample size parameter estimates
- Checking that absolute and relative differences had been calculated correctly
- Part way through data extraction it was agreed that the target total number of clusters required should be left blank if not explicitly reported in the sample size calculation. All papers where this question was not missing were double checked to ensure it had been explicitly reported and not inferred.
- Logical checks, for example, if the sample size does not account for the ICC then no value should be given for the ICC

<u>Individual extraction</u>

All information I extracted on trials with ordinal outcomes was performed twice, a month apart, in order to provide a double check of the extraction. A second data extraction was feasible due to the small number of trials included. As both extractions were performed by me there was some limitation to the validation provided. However, it was thought that the information to be extracted would be a key part of the trial report and should be easily identified. If any ambiguity was present it was discussed with SE and AC.

3.3 Results

3.3.1 Description of included trials

The characteristics of the 300 trials have previously been described in detail.[30] Selected characteristics are presented in Table 3.1. Just over half of the trials reported a sample size calculation, N=166 and of these 61% accounted for clustering. This number is larger than the 164 reported in the original review because that review focused on whether a sample size calculation was present for the variable defined as primary. In this review the outcome used in the sample size calculation was deemed primary.

Despite the willingness of authors no datasets for the eleven trials with ordinal outcomes were available to calculate observed data summaries. The unavailability was due to data being deleted, archived, corrupt or the research team disbanded and the data location unknown.

3.3.2 Prevalence of ordinal outcomes and design characteristics

There were 11 (4%) trials identified as having an ordinal primary outcome. The design of these trials were most often parallel group 10 (91%), two-arm 6 (55%), completely randomised 8 (73%), and used a cohort sample, meaning the same individuals within a cluster are measured at more than one time point, 9 (82%) (Table 3.2). In all but one of these cohort trials the baseline measure or post randomisation measurements have been incorporated into the analysis. Use of the design effect for clustered ordinal data does not allow for the inclusion of baseline or post randomisation measures. The inclusion of such measurements would introduce additional components of correlation or require the ICC to be suitably adjusted. For ordinal outcomes estimates of the ICC are not routinely published and are likely to be difficult to find, estimates of additional or adjusted correlations due to incorporation of baseline or repeated measures will be more so. Therefore, for the purpose of this thesis I focus on sample size methods for an analysis of the final time point only, which should be conservative for analyses that include baseline or repeated measures.

The majority of the interventions being tested within these trials were counselling or skills training interventions aimed at changing various aspects of behaviour (9/11 (82%) of trials). The ran-

Table 3.1: Characteristics of the 300 CRTs included in the published review by Ivers*. Figures are numbers (percentages) of trials unless stated otherwise

Characteristic	N	%
Publication year:		
2000-4	139	(46)
2005-6	93	(31)
2007-8	68	(23)
Journal impact factor (n=294)		
Median (IQR)	2.9	(2.1-5.1)
Range	0.45-50.0	
Setting:		
Clinical	169	(56)
Non-clinical	131	(44)
No. of clusters randomised (n=285)		
Median (IQR)	21.0	(12-52)
Range	2-605	
Average cluster size (n=271)		
Median (IQR)	33.9	(12.5-88.5)
Range	1.7-122 855	
No. of participants per arm (n=290)		
Median (IQR)	329	(143-866)
Range	20-614 275	
Sample size reported	166	(55)
Accounted for clustering in sample size	102	(61)

* Ivers et al[30]

domised units were either health care practices/providers with outcomes measured for each patient (5/11(45%)) or schools/classrooms with outcomes measured for each pupil (4/11(36%)). Over half of these ordinal outcomes had either four or five levels, Table 3.2.

The total number of clusters randomised was small, median 18 with inter-quartile range 9 to 61. Similarly the number of individuals enrolled per trial arm was small, median 128 with interquartile range 71 to 150, Table 3.2.

In two of the trials a random effects ordinal regression was performed. The outcome was dichotomised in four trials, treated as continuous in four trials and categorical in one.

3.3.3 Reported values of ICCs and category proportions

Campbell and Walters' design effect method for sample size calculation for clustered ordinal outcomes requires estimates for the expected proportions in each category within each treatment group at the end of the trial as well as a measure of the ICC. These estimates were not routinely reported, only two of the eleven trials provided information on the observed proportions in each category at the end of the trial. These two trials and their estimates are summarised below.

The trial by Steptoe et al assessed the impact of a behavioural counselling programme, implemented at the level of general practice, in changing the state of change in fat reduction of patients at risk of coronary heart disease.[56] State of change was measured on 5-levels (pre-contemplation, contemplation, preparation, action and maintenance). The authors combined the action and maintenance stages since they felt the distinction between the two may be subject to recall bias. After 4 months the proportions in these categories in the intervention group were 9.9%, 8.9%, 14.1% and 67.1% and in the control group 17.7%, 10.9%, 17.7% and 53.6%. These figures translate to log-odds ratios of -0.67, -0.54 and -0.56. At the 12 month time point the proportions in the intervention group were 14.1%, 5.7%, 11.9%, 68.4% and the control group 16.5%, 7.5%, 16.8% and 59.2%. These figures provide log-odds ratios of -0.19, -0.25, and -0.40. From observation alone the log-odds are not similar for each dichotomisation indicating that the assumption of proportionality may not be met. The assumption of proportional odds, what it means and how it can be formally assessed will be explored further in the next chapter.

Table 3.2: Design features for 11 CRTs with ordinal primary outcomes identified from the 300 published CRTs included in the review by Ivers*. Figures are number (percentage) of trials unless stated otherwise

Trial Design		**N**	**(%)**
Publication year	2000-2004	5	45%
	2005-2008	6	55%
Trial design	Parallel trial	10	91%
	Factorial trial	0	0%
	Cross-over trial	1	9%
	Other	0	0%
Randomisation method	Completely randomised	8	73%
	Stratified	1	9%
	Pair-matched	0	0%
	Other	2	18%
Design at patient level	Cross sectional	2	18%
	Cohort	9	82%
Number of arms	Two	6	55%
	Three	4	36%
	Four	1	9%
Sample size reported		1	9%
Number of ordinal levels	3	2	18%
	4	4	36%
	5	3	27%
	6	0	0%
	7	1	9%
	8	0	0%
	9	1	9%
No. of clusters randomised	median (IQR)	18	(9 to 61)
	Range		7 to 345
No. of participants enrolled per arm	median (IQR)	128	(71 to 150)
	Range		30 to 316

* Ivers et al[30]

In the second trial by Howlin expert training and consultancy was provided to teachers of children with autism. The frequency of communication initiations by children was measured on a 4-level outcome variable.[123] Using figure 2 from the trial publication the proportions in each ordinal category for the two follow up measurements can be estimated. In the immediate treatment group the proportions observed in each category are 7.7%, 57.7%, 11.5%, 23.1% and in the no treatment group these are 17.9%, 46.4%, 32.1%, 3.6%. These translate to log-odds ratios of -0.96, 0.05, and -2.08 . At the third time point the proportions in the immediate treatment group are 16%, 52%, 20% and 12% and in the control group 14.3%, 39.3%, 39.3% and 7.1%. These translate to log-odds ratios 0.13, 0.61, and -0.58. Each of these log-odds are very different and not always in the same direction, again indicating that the proportional odds assumption may not be reasonable.

3.3.4 Reporting quality according to CONSORT guidelines

Of the 11 trials with ordinal primary outcomes only one paper by Howlin et al reported a sample size calculation based on an expected odds ratio.[123] The paper by Brody did state that an a priori power analysis based on effect sizes from a previous study informed the sample size but no details of the calculation were actually reported and the corresponding author did not respond to my e-mail inquiries about this study.[124] The expected proportions in each ordinal category at the design stage were not provided in any of the eleven papers.

In the Howlin trial all five of the elements recommended by CONSORT were reported (1) the type I error rate (2) power (3) estimates of outcomes in each group or minimum important target effect, (4) the number of clusters or average cluster size and (5) the assumed measures of intracluster correlation, design effect or coefficient of variation. No allowances were made for attrition or possible cluster size imbalance.

3.3.5 Assessment of methodological approach to sample size calculation

In the Howlin trial the clustered nature of the data was acknowledged and accounted for in the sample size calculation. The ICC was used as the measure of correlation, but no justifications for any of the estimates used in the calculation were provided.

It was agreed through the collaborative extraction that the Howlin paper assumed a binary outcome and the sample size compares a difference in proportions adjusted for clustering. However, on further consideration the exact calculation of the sample size is slightly ambiguous as to whether it has been calculated on the ordinal outcome, or a dichotomous version. The reference quoted for the method used would allow for both of these options and the description of the method does not include enough information to recreate the formula in order to check either approach. The authors were not contactable by email and the study protocol was not available to check the method used.

3.4 Discussion

3.4.1 Main findings

In a random sample of 300 cluster randomised trials 11 (4%) trials had primary outcomes that were ordinal. Nine (82%) of these outcomes had between 3 and 5 ordinal categories and most often measured an aspect of behaviour 9/11 (82%).

The design of these trials were most often parallel group (10, 91%), two-arm (6, 55%), completely randomised (8, 73%), and used a cohort sample (9, 82%). Both the total number of clusters randomised and the number of enrolled participants per arm tended to be small.

Only one paper reported a sample size calculation which incorporated adjustments for clustering and followed all the CONSORT recommendations for sample size reporting for cluster randomised trials.

Few trials reported the observed proportions in each category at the end of the trial that might be used in future sample size calculations. Where provided these proportions did not appear to fulfil the assumption of proportional odds, a requirement of Whitehead's sample size formula for individually randomised trials to which the design effect is applied.

3.4.2 Strengths and limitations

The sample used in this review is unique due to its size and coverage. With 300 cluster randomised trials included it is the largest review of CRTs that has been conducted to date, with the next largest containing 173.[44] The majority of other reviews have contained less than 40 trials. Many previous reviews have focused on particular areas of health such as stroke, oral health, and primary care or targeted particularly high ranking journals during the search. The way in which the sample of 300 used in this review was selected makes it the most representative of cluster randomised trials across the medical research field. However it should be acknowledged that the sample contains trials published between 2000 and 2008. There is a possibility that the inclusion of more recent publications may have provided a different estimate of ordinal outcome prevalence. The true estimate of demand for ordinal sample size methods might also be larger as this review does not identify those trials that chose an alternative outcome due to a lack of available methods for ordinal outcomes and the added complexity of the analysis.

3.4.3 Comparison with other work

The objectives of previous reviews of cluster randomised trials have mainly focused on assessing the quality of methods used and the quality of reporting. This is the first review of cluster randomised trials to focus on the type of outcome used. Some recent research has been conducted on the analysis of ordinal outcomes for cluster randomised trials by Ruochu Gao, a PhD student of Allan Donner, whose thesis is available online.[125] In this work Gao provides examples of 11 recent cluster randomised trials using ordinal outcomes. The types of outcomes are consistent with those seen in my review: behavioural outcomes related to tobacco, drug and condom use and satisfaction outcomes related to patient and physician. The types of clusters are also similar: schools, medical practices and physicians. In nine (82%) of these trials the number of ordinal categories is between three and five, the same figure as seen in my review. Gao does not describe how her sample of trials was selected but it provides consistent findings about the characteristics of trials that use ordinal outcomes.

3.4.4 Implications for this research

Four percent of trials were identified as having an ordinal primary outcome. Although less common than binary or continuous outcomes ordinal outcomes are not especially rare. In only one of these eleven trials was a sample size calculation reported. It is clear that researchers still need further support and guidance in the design of trials for clustered ordinal outcomes and they need access to appropriate sample size methodology to do this. The development towards practical guidance for sample size calculations with ordinal outcomes follows in the proceeding chapters.

Chapter 4

Analysis of ordinal outcomes

Whitehead's sample size formula for ordinal outcomes in individually randomised trials assumes proportional odds and an analysis by Mann-Whitney test, which is equivalent to ordinal regression when only treatment is included in the model. Ordinal regression methods are more available in statistical software than their non-parametric counterparts and are a popular approach to the analysis of ordinal outcomes. The assumed analysis method for Campbell and Walters' extension of Whitehead's method for clustered data is not explicitly stated. The extension of ordinal regression for clustered data, the random effects ordered logistic regression model, would seem a logical approach to use.

The aim of this chapter is to describe analysis methods for clustered ordinal data with a focus on the random effects ordered logistic regression model and its assumptions. As the analysis methods for clustered data are often extensions of the methods used for non-clustered data the methods for non-clustered data are described first. Methodological issues surrounding each analysis method are discussed, providing justification to the decisions made in the design of the simulation studies to evaluate the design effect approach for sample size calculations for clustered ordinal outcomes, which follow in the next chapter.

Three proposed ICC estimators for ordinal outcomes are presented in this chapter and their use in the design effect will be evaluated in the next chapter.

4.1 Analysis methods for non-clustered data

An ordinal outcome, as defined in chapter one, is a variable which consists of a set of categories that can be ordered or ranked, for example disease severity (mild, moderate or severe). The absolute difference between the categories is often unmeasurable or unknown.

Alan Agresti has made a large contribution to the work on the analysis of ordinal outcomes. The first version of his book on the analysis of ordered categorical data was published in 1984 and a second edition, published in 2010, included developments for clustered data. Agresti has contributed to several reviews of methods available to analyse ordinal data with the earliest in 1989 and the latest in 2005.[90, 126–128] These articles highlight the substantial methodological development that has occurred in recent years in the analysis of ordinal outcomes. In the most recent review paper invited researchers were given the opportunity to discuss the work and give their thoughts on the future direction of ordinal outcome research, power and sample size calculations for clustered data, the focus of this thesis, was raised as an area for development. Lall et al have also reviewed ordinal regression models applied specifically to health-related quality of life outcomes.[129]

With an ordinal outcome one can either ignore or incorporate the ordinality in the data into the analysis. Methods which ignore the ordinal nature of the data are described here for completeness, as they are commonly undertaken, but are not recommended as they will likely give different results to an analysis that incorporates ordinality.[130]

4.1.1 Methods which ignore ordinality

If ignoring the ordinal nature of the data the outcome would alternatively be considered to be nominal, continuous or could be reduced to a binary outcome. Treating the outcome as nominal means treating the categories as if there is no natural ordering among them. The Pearson's Chi-squared test is commonly used to analyse nominal data. However, it has been shown that when used to analyse an ordinal outcome the Chi-squared test can produce different conclusions to those made using analyses that take the ordering into account.[130]

Another simple approach which ignores the ordinality is to combine adjacent categories to reduce the ordinal outcome to a binary variable, to which standard statistical methods such as the chi-squared test or logistic regression can be applied. This produces a valid analysis but dichotomisation of the outcome results in a loss of information, more power can be gained by retaining the full ordinal variable. The power of this dichotomisation approach has been examined using data from the CRASH trial which investigated the efficacy of corticosteroids in traumatic brain injury patients.[131] The primary outcome variable in this trial was the 5-point Glasgow Outcome Scale which was dichotomised as unfavourable (dead, vegetative, severe disability) or favourable (moderate disability, good recovery). The analysis of the binary outcome was non-significant, yet the analysis of the ordinal outcome with proportional odds regression was highly significant. The authors attributed this difference in conclusions to the increased statistical power of the ordinal approach. In previous simulation studies the authors had also explored the effects of non-proportionality, where a significant treatment effect was present in only one category cut-off. They reported that surprisingly the results showed that even when the assumption of proportionality was not met the ordinal analysis assuming proportional odds had more power than the binary approach (dichotomised at the point of the significant treatment effect).

According to the authors trials in traumatic brain injury have traditionally been powered on the dichotomous variable of a favourable versus unfavourable outcome. The increased power from an ordinal analysis implies that powering on the ordinal outcome could reduce sample sizes. However, the authors investigating the CRASH data did not advise going down this route because they argue that trials in critical care medicine are generally under powered due to a systematic overestimate of the treatment effect size during the design. Instead the increased efficiency of an ordinal analysis should aid in the detection of smaller treatment effects for the same sample size. However, I would argue that under powering in the strictest sense refers to a trials inability to recruit and measure the required number of individuals as indicated by the sample size calculation. The issue here seems to be that the sample size calculation, in particular the way that the minimum clinically important difference is chosen, is not adequate. Investigators may be too optimistic about the expected treatment effect or the treatment effects are chosen to provide a sample size requirement that is attainable to recruit.

In contrast to the above recommendations on sample size for critical care medicine the recommendation for stroke trials is to conduct the design and analysis using approaches which utilize the ordinality of the data, as opposed to a dichotomous approach based on stroke/no stroke, as the reduction in sample size can reduce the competition for patients between trials and reduce the cost and complexity of the trial itself.[132] For individually randomised trials Whitehead's method of sample size calculation for ordinal outcomes has been shown to produce sample sizes that are on average 28% smaller than those for a binary version of the outcome.[57] The impact on power when using more than two categories is large. However, additional power gains are marginal once the number of ordinal categories goes beyond five.[2]

The final approach ignoring the ordinal nature of the data assigns scores to the ordinal categories and assumes the variable is continuous and analysed using methods such as ANOVA or linear regression. Most commonly equally spaced scores are used across the ordinal categories, although other scoring systems may be used. Aside from how to assign scores the biggest problem with this approach is that the created variable often violates the normality assumption required for many analyses, more so when the sample size is small. The t-test and ANCOVA however, have been shown to be robust to the normality assumption (i.e. to produce significance levels close to nominal levels) using simulation with three-, four- and five-level ordinal data.[133, 134] To avoid the assumption of normality non-parametric methods such as the Mann-Whitney-U test can be used.

Walters et al have explored methods for sample size and analysis within the context of quality of life data. Their results suggest that when the outcome has a limited number of discrete values (less than 7) and/or the proportion of cases at either of the bounds is high Whiteheads method of sample size performs well. However, where seven or more populated categories are present and the proportion of cases at the bounds is low then sample size and analysis methods based on the simplifying assumption of an assumed continuously distributed variable may be used.[135, 136]

4.1.2 Methods that incorporate ordinality

In this chapter I assume that the ordinal outcome is a discrete measure of an underlying continuous variable and therefore I focus on the model often most appropriate for this situation, the proportional odds model (also referred to as ordered logistic regression) which is described by McCullagh.[137] Lall et al have reviewed ordinal regression models applied to health-related quality of life assessments, and include discussion of the stereotype model that can be useful in situations where the categories are not assumed to be a discrete version of an underlying continuous variable. The ordinality of the response is assessed within the model.[129]

Model formulation

Let us assume an ordinal response variable with k ordered categories $q = 1, 2, \ldots, k$ and Y_j be the categorical response for the $j'th$ individual. Y_j takes the value q if the response is in category q.

The probability of an individual j being in category q is π_{jq} and the cumulative probability of being in category q or below, denoted P_{jq} is given by

$$P(Y_j \leq q) = P_{jq} = \pi_{j1} + \pi_{j2} + \ldots + \pi_{jq}$$

We assume that the ordinal response is a crude measure of some underlying continuous distribution which is unknown and unmeasurable (referred to as a latent response), Y_j^*. The ordinal variable is obtained by chopping Y_j^* into categories using a series of cut points α_q where $q = 1, \ldots, k-1$. Figure 4.1 illustrates an unobserved latent response for a four-level ordinal outcome. A value of $Y^* < \alpha_1$ corresponds to a response in the first category, values between α_1 and α_2 correspond to a response in the second category and so on. The cumulative probability of being in category q or below, denoted P_{jq} is now given by

$$P(Y_j^* \leq \alpha_q) = P(Y_j \leq q) = P_{jq} \qquad (q = 1, 2. \ldots k-1)$$

If we know the distribution of the latent response this cumulative probability can be easily calculated. The most common choice for the distribution of the latent response is a logistic distribution with mean μ and variance, $\frac{\pi^2}{3}$. The cumulative distribution function for the standard logistic distribution is:

$$F(x) = P(X \leq x) = \frac{1}{1+e^{-(x-\mu)}} \text{ or equivalently } F(x) = \frac{e^{(x-\mu)}}{1+e^{(x-\mu)}}$$

Therefore if we assume the underlying latent response Y_j^* follows a logistic distribution

$$F(y_j^*) = P(Y_j^* \leq y_j^*) = \frac{e^{y_j^* - \mu}}{1+e^{y_j^* - \mu}}$$

and so

$$P_{jq} = F(\alpha_q) = \frac{e^{\alpha_q - \mu}}{1+e^{\alpha_q - \mu}}$$

Applying the logistic transformation it follows that

$$logit[P(Y_j \leq q)] = log\left(\frac{P_{jq}}{1-P_{jq}}\right) = \alpha_q - \mu$$

Assuming that μ is a linear combination of the explanatory variables the proportional odds model, or more specifically the proportional odds version of the cumulative logit model is given as:

$$logit[P(Y_j \leq q)] = \alpha_q - \boldsymbol{\beta}\mathbf{x_j} \tag{4.1}$$

Where β represents the treatment effect, assumed the same across all q (the proportional odds assumption).

The existence of a latent response is not required for model interpretation. However, if the latent response can be assumed then the treatment effect is unaffected by the choice of number of categories and cut points. Therefore when the model fits well different trials using different scales to measure the treatment effect should agree in their conclusions.

Alternative link functions

The model can be written more generally as

$$G^{-1}[P(Y_j \leq q)] = \alpha_q - \boldsymbol{\beta}\mathbf{x_j} \tag{4.2}$$

Where G^{-1} is a link function, the inverse of the continuous cumulative distribution G. In the previous section a logit link was applied, other link functions may be used such as the probit, log-log and complementary log-log.

The logit link is the most commonly used followed by the probit link. The probit link assumes the underlying latent response Y_j^* follows a standard normal distribution and the link function G^{-1} is

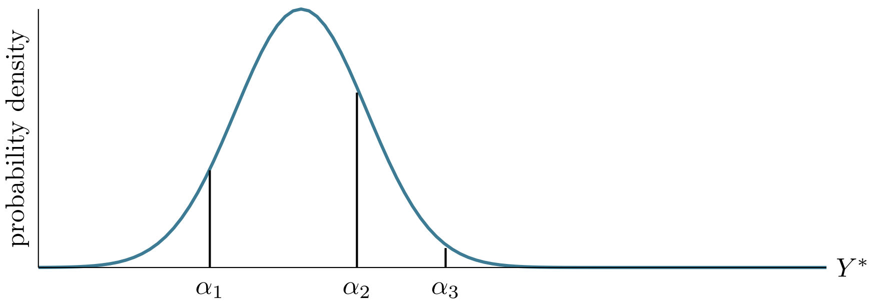


Figure 4.1: Graphical representation of how a 4-level ordinal outcome, Y, can be represented by an assumed underlying continuous latent variable Y^* with cut points α

the inverse of the standard normal distribution. The shape of the normal and logistic distributions is similar so if the model fits well with logit link it should also fit well with a probit link. The advantage of the logit link is that the treatment effect can be interpreted as an odds ratio. Under the probit link the treatment effect relates to an underlying normal latent variable for an ordinary regression model. The log-log and complementary log-log link functions are appropriate when the distribution of the underlying continuous variable is non symmetric. For small sample sizes it may be difficult to assess which link function is most appropriate.[128] In this thesis I consider only the two most frequently used links, the logit and probit links.

Assessment of proportional odds

The proportional odds assumption assumes that the treatment effect is the same for all the possible ways that the k-category response variable might be collapsed to a binary variable.

One approach to checking the assumption of proportional odds is to fit separate treatment effects across the categories, replacing β by β_q and then compare this model to the single effect model using methods such as a likelihood ratio test. Another approach proposed by Brant[138] views the proportional odds model as describing a set of k-1 separate binary logistic regression models. Proportionality is assessed by examining and comparing the fit of these binary logistic models. A goodness-of-fit test statistic, which follows a Chi-squared distribution, formally assesses the proportional odds assumption and is implemented in Stata as part of the user-written *omodel* command. Peterson and Harrell[139] also propose a formal test of proportional odds via a score test.

If the assumption of proportional odds is not met the main approaches to analysis, which still make use of the ordinal nature of the data, include using a more general model which involves fitting a separate effect for each category or a partial proportional odds model as suggested by Peterson and Harrell.[139] The partial proportional odds model allows the treatment effect to be the same for some values of the categories and different for others. This method reduces the number of required estimates in comparison to fitting a separate effect for each category. Partial proportional odds models can be estimated in Stata using the user-written *gologit* command.

Methods that do not make use of the ordinal nature of the data but are sometimes considered include assuming a nominal response or dichotomising the outcome and using logistic regression. However, different conclusions may be made using these methods compared to those for ordinal data.

4.1.3 Significance testing

Under maximum likelihood estimation significance tests for the treatment effect for the proportional odds model are usually conducted using the likelihood-ratio, Wald or score test statistics.

The likelihood ratio test compares two nested models, the null and alternative i.e. in this case with and without a treatment effect fitted. The likelihood ratio statistic follows a Chi-squared distribution and is calculated as twice the difference in the log-likelihoods from each model. The test produces similar results to the Wald test for large sample sizes but performs better under smaller sample sizes, where use of the Wald test can lead to inflated Type I error rates.

The Wald statistic approximates the likelihood ratio test and is used to test the hypothesis that the estimate of the treatment effect, β, is 0. It is the default in most statistical packages and is calculated as:

$$w = \frac{\beta}{se(\beta)}$$

Under the null distribution this statistic follows a normal distribution, and the square of the statistic follows a Chi-squared distribution. The Wald test is advantageous over the likelihood ratio test in that it only requires one model to be fitted, but for small sample sizes the Type I error rate can be inflated meaning a statistically significant treatment effect may be declared when no such difference exists.

The score test (also known as the Lagrange multiplier test) is asymptotically equivalent to the likelihood ratio test. It follows a Chi-squared distribution and is based on the derivative of the log-likelihood and its standard error evaluated at the null hypothesis value. Both the Wald and Likelihood ratio tests are easily implemented in statistical software, but the score test less so.

4.2 Analysis of clustered ordinal outcomes

I now consider the extension of these methods to account for clustering.

4.2.1 Non-parametric methods

Rosner and Grove[140] have extended the Wilcoxon rank sum test to individual-level analysis of clustered ordinal data. The authors present the variance of the test statistic for both the clustered and unclustered case, hence a design effect could be calculated, although not of simple form. The authors propose that where the number of ordinal categories is large their method is more appropriate than parametric methods, such as the proportional odds model. However their method requires estimates of four clustering parameters, values for which may not be readily available. The method does not allow the inclusion of covariates and is not available as standard in statistical packages. Hence, I do not consider its use further in this thesis.

4.2.2 Random effects model

In a cluster randomised trial the ordinal response is Y_{ij} for individual j in cluster i, $i = 1, \ldots, C$, $j = 1, \ldots, n$. The cumulative logit model with random effects, assuming proportional odds, is given as[141]

$$logit[P(Y_{ij} \leq q)] = \alpha_q - \beta x_{ij}\prime + \mu_i, \qquad q = 1, 2, \ldots, k-1 \tag{4.3}$$

Where μ_i represents a random effect of the cluster and distributed $N(0, \sigma_b^2)$.

As in the non-clustered case this model can be motivated by the assumption of an underlying continuously distributed latent response Y_{ij}^* such that

$$Y_{ij}^{*} = x_{ij}\boldsymbol{\beta} + \mu_i + \epsilon_{ij}$$

and

$$y_{ij} = \begin{cases} 1 & \text{if } y_{ij}^{*} \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < y_{ij}^{*} \leq \alpha_2 \\ \vdots & \\ k & \text{if } \alpha_{k-1} < y_{ij}^{*} \end{cases}$$

The error terms ϵ_{ij} are distributed as logistic with mean zero and variance $\pi^2/3$ and are independent of μ_i. Error terms distributed as $N(0,1)$ correspond to a probit link, see Section 4.1.2

As touched upon in Chapter two, Section 2.2.3 random effects models with discrete outcomes can be difficult to fit via maximum likelihood and the likelihood function must be approximated by such methods as Gauss-Hermite quadrature.

4.2.3 Marginal models

In the random effects model (or cluster-specific model) the focus is on modelling a cluster-specific response and the regression coefficient for treatment represents the average effect of treatment if an individual stays in the same cluster but moves from the control to the intervention arm. The random effects model uses maximum likelihood estimation and explicitly models the covariance structure by introducing cluster-specific random effects into the model.

In a population averaged, or marginal, model the regression coefficient for treatment represents the effect of treatment if an individual in the population moves from control to intervention arm. The model does not fully specify the distribution of the population, as in the random effects model, instead the marginal expectations are modelled and a variance-covariance structure (referred to as the working or hypothesised correlation) is chosen to describe the correlation between members of a cluster. The model is not fitted via maximum likelihood; an estimate of the treatment effect can be found by solution of a generalised estimating equation (GEE).

Generalised Estimating Equation methods for the analysis of ordinal outcomes have been described by Lipsitz, Kim and Zhao and are summarised here.[68]

For each individual we observe the response on a k-level ordinal outcome with categories $q = 1, 2, \ldots, k$. To keep with the terminology of Lipsitz et al a higher category here is used to indicate a better outcome. Let Z_{ij} denote the ordinal response of the j'th individual in the i'th cluster, $j = 1, 2, \ldots, n$ and $i = 1, 2, \ldots C$. The size of the cluster is assumed fixed and denoted by n.

We form k indicator variables Y_{ijq}, where $Y_{ijq} = 1$ if subject j has response q and $Y_{ijq} = 0$ otherwise. For each subject we form a k-1 response vector of indicator variables $\mathbf{Y_{ij}} = [Y_{ij1}, \ldots, Y_{ij(k-1)}]'$, and for each cluster $\mathbf{Y_i} = [\mathbf{Y'_{ij}}, \ldots, \mathbf{Y'_{in}}]'$

The marginal probability is denoted by $Pr[Z_{ij} = q] = E[Y_{ijq}] = Pr[Y_{ijq} = 1] = \pi_{ijq}$ and the corresponding marginal cumulative probabilities by $Pr[Z_{ij} \leq q] = P_{ijq}$

Lipsitz et al analyse the data using a marginal model based on cumulative logits

$$logit[P_{ijq}] = \mathbf{X}\beta \tag{4.4}$$

Where $\mathbf{X}$ denotes a $(k-1) \times k$ design matrix for the $j'th$ individual of the $i'th$ cluster and $\beta = [\alpha_1, \ldots, \alpha_{k-1}, \beta]'$ denotes a $k \times 1$ parameter vector. Where the α_q corresponds to the $q'th$ cumulative logit and β denotes the effect of treatment.

In Chapter Two, Section, 2.2.2, I presented a worked example of the sample size methodology proposed by Kim et al which assumed a GEE model. In Step 10 of their sample size process the parameter vector β is found by solution of the GEE equation:

$$\hat{\beta} = [\textstyle\sum_t \mathbf{D}'_t \mathbf{V}_t^{-1} \mathbf{D}_t]^{-1} [\textstyle\sum_t \mathbf{X}'_t \mathbf{W}_t h(\theta_t)]$$

Where t is an index representing treatment group, $\mathbf{D}_t = \mathbf{\Delta}_t \mathbf{X}_t$ where $\mathbf{\Delta}_t = \frac{\partial \pi}{\partial \eta}$ a matrix of partial derivatives of the mean of the outcome with respect to the regression parameters, η is the linear predictor $X\beta$, $\mathbf{X}_t$ is the design matrix and $\mathbf{V}_t$ is the working covariance matrix and $\mathbf{W}_t = \Delta_t V_t^{-1} \Delta_t$ and $h(\theta_t)$ is a vector of cumulative logits i.e. ln(cumulative probability/(1-cumulative probability).

The Wald test statistic for the model with only treatment fitted with $H_0 : \beta = 0$ is

$$W = \tfrac{\beta^2}{var(\beta)} \sim \chi^2_1$$

The variance of the treatment effect can be calculated in two ways. The model-based estimate of the variance provides valid inferences only if the working covariance matrix is correctly specified.

$$var(\beta) = \sum_{t=1}^{T} [D'_t V_t^{-1} D_t]^{-1} \tag{4.5}$$

An alternative estimate that is more robust to miss-specification of the working covariance matrix is calculated by

$$var_r(\beta) = \sum_{t=1}^{T} [D'_t V_t^{-1} D_t]^{-1} \sum_{t=1}^{T} [D'_t V_t^{-1} var(Y_t) V_t^{-1} D_t]^{-1} \sum_{t=1}^{T} [D'_t V_t^{-1} D_t]^{-1} \tag{4.6}$$

4.2.4 Comparison of random effects and marginal models

Of those trials identified in Chapter Three with ordinal outcomes the random effects model was the most popular choice of analysis.

The GEE, or population averaged (PA) method provides consistent estimation even when the correlation structure is miss-specified and is computationally simple compared to the random effects (RE) model, where the likelihood function must be approximated. However, because the method does not specify a full multivariate distribution for the responses the GEE method does not have a likelihood function. No likelihood function means that the likelihood ratio test and other likelihood based methods cannot be used to check model fit, compare models or make inferences about the model parameters. With a GEE model inference about the model parameters must be made via a Wald test, which can give spurious results, particularly for small samples. The empirical-based standard errors calculated from a GEE model may be underestimated unless the sample size is very large.

The two models also differ in the assumptions they make with regard to missing data. Missing data mechanisms have been described by Little and Rubin and I use their terminology here.[142] The GEE model makes the strongest assumption, that the data are Missing Completely At Random (MCAR). In simple terms this means that the probability that the observation is missing does not depend on the value itself, or any other observed measurement. Maximum likelihood based random effects methods make a weaker assumption that the data are Missing At Random (MAR), that is the probability that the observation is missing does not depend upon the value itself but can be explained by other observed measurements. Violations of these assumptions in either model may lead to biased results.

The relationship between the treatment effect estimated from a RE model and that estimated from a PA model have been described by Agresti.[130] These relationships are briefly summarised here as they will be used in the design of the simulation study described in the next chapter.

• The treatment effect from a marginal model will be smaller than that from a random effects model, the difference increases as the level of within-cluster correlation increases.
• When a probit link is used the treatment effect estimate from the RE model and that from the PA model can be directly compared, with the RE estimate being $\sqrt(1+\sigma_w^2)$ times that of the PA effect.
• For the logit link the relationship between the model estimates is only approximate with the RE estimate being $\sqrt{1+0.346\sigma_w^2}$times that of the PA effect.
•Despite the difference in interpretation and magnitude between the random effects and population averaged models the significance of the treatment effect is likely to be similar.[130]
• Using the fact that the standard normal cumulative distribution function (cdf) at a point z is well approximated by the standard logistic cdf at 1.7z the estimates from models with a logit link are approximately 1.7 times those from probit models.[130]

4.2.5 Software

In Stata version 13 the *xtologit* command fits random-effects models via maximum likelihood using adaptive Gauss-Hermite quadrature as the default for approximating the likelihood. The accuracy of this quadrature can be checked using the command *quadchk*. The *xtologit* command assumes that larger values of the ordinal response correspond to better outcomes. The command *xtoprobit* is available for ordered probit models. The *gllamm* command is a user-written command that can also be used to fit these models and pre-dates the inbuilt functions. Currently there is no option available in Stata to analyse ordinal outcomes with a GEE model.

The SAS software can accommodate random effects and GEE models using PROC NLMIXED and PROC GENMOD respectively.

4.2.6 Assessment of proportional odds

For the individually randomised trial formal methods have been proposed to test the assumption of proportional odds and for cases when the assumption of proportional odds is not valid non-proportional or partial proportional odds models have been suggested as alternative analysis methods and have been incorporated into statistical software, see Section 4.1.2.

For the clustered case there has been less development and limited guidance around how to formally assess proportional odds. As in the individually randomised case to test the assumption of proportional odds one may consider fitting a separate effect for each category and comparing this to the proportional odds model via a likelihood ratio test. Hedeker and Mermelstein have extended the random effects proportional odds model to allow for non-proportional or partial proportional odds.[143] Their approach extends Peterson and Harrell's approach for partial proportional odds for the fixed effects model.[139] The partial proportional odds method described by Hedeker and Mermelstein has been implemented in an extension to the MIXOR package available in R for mixed effects ordinal regression.[144]

The partial proportional odds model of Hedeker and Mermelstein was developed within the context of behavioural state of change data, where participants are categorised according to their readiness to change ranging from pre-contemplation to action. The authors considered the assumption of proportional odds to be unreasonable for this type of data. Of the 11 CRTs with ordinal outcomes identified in Chapter Three there were three trials that used such an outcome. Therefore non-proportional odds may be a significant problem in the design and analysis of ordinal outcome trials with behavioural outcomes.

4.2.7 Significance testing

As in the individual case the Null hypothesis that there is no effect of treatment can be assessed via a likelihood ratio, Wald or score test for random effects models or Wald test for GEE. However, the approximation of the Wald test statistic to the normal distribution is worse in the clustered case when the number of clusters is small or the cluster size is variable. This tends to inflate the Type I error rate, so we are likely to see more than 5% of calculated P-values being less than 0.05 under the null hypothesis. A suggested solution is to compare the Wald statistic to a t-distribution, but this may reduce the Type I error to below the nominal value.

4.3 The ICC

Estimators of the ICC have been extensively described for binary and continuous data.[145, 146] The most common interpretation of the ICC is that it represents the proportion of variance due to between-cluster variation, for continuous outcomes this is defined as

$$\rho = \frac{\sigma_b^2}{\sigma_b^2+\sigma_w^2}$$

Where σ_b^2 is the between-cluster variance, σ_w^2 is the within-cluster variance and the sum of the two is the total variance.

For binary outcomes the definition is slightly different

$$\rho = \frac{\sigma_b^2}{\pi(1-\pi)}$$

The ICC is dependent on π, the prevalence in the population. The total variance is calculated as $\pi(1-\pi)$. The prevalence, and therefore the within-cluster variance will likely vary between clusters. Therefore the assumption of constant within-cluster variance does not hold for binary outcomes. This ICC for binary outcomes is referred to as being on the proportions scale and it is this ICC which should be used in the design effect when calculating sample size for binary outcomes.[147]

Like ordinal outcomes the model for binary outcomes can be motivated by the existence of an underlying continuous variable on the logistic scale and an ICC can also be calculated on this logistic scale. This estimate will be different from the ICC on the proportions scale and there is no easy formula to convert one to the other, instead simulation must be used.[148] Eldridge, Ukoumunne and Carlin have compared the two estimates through simulation and provide a useful table that shows the relationship between the ICC on the proportions scale and that on the logistic scale for different levels of overall prevalence. The difference between the two values is greatest when the ICC on the proportions scale is large or the prevalence is further from 50%.

4.3.1 ICC of the latent response

The underlying latent variable, Y^*_{ij} is assumed to be continuous and follows the random effects model

$$Y^*_{ij} = x_{ij}\boldsymbol{\beta} + \mu_i + \epsilon_{ij}$$

with error terms ϵ_{ij} distributed as logistic with mean zero and variance $\pi^2/3$ and are independent of μ_i, the random effects for clusters, which are distributed $N(0, \sigma^2_b)$.

The intracluster correlation coefficient on this underlying (logistic) scale is defined as

$$\rho_{(l)} = \frac{\sigma^2_b}{\sigma^2_b + \pi^2/3} \tag{4.7}$$

This ICC is relatively straight forward to calculate and is often automatically provided by statistical software, such as Stata, after model fitting. However, the fact that this ICC relates to the underlying variable means it is not clear whether its use in the design effect for ordinal outcomes is appropriate.

For the probit link $\pi^2/3$ is replaced by 1 in the denominator.

4.3.2 Analysis of variance ICC

It is possible to assign numerical values to the ordinal categories and calculate an ICC using a one-way analysis of variance.[146]

$$\rho = \tfrac{MSB - MSW}{MSB + (n-1)MSW}$$

Where MSB and MSW are the mean squares between and within clusters estimated from the analysis of variance and n is the cluster size.

The estimated ICC will be dependent upon how the numerical scores have been assigned. For example the simplest approach is to assign equally spaced values such as 1, 2, 3 and 4. Alternatively if it was felt that the gap between categories 3 and 4 was twice that of other adjacent categories the scores could be assigned as 1, 2, 3, and 5. If the categorical variable was formed from the grouping of scores from a health questionnaire, for example, scores may be defined from using mid-points of the categories. In this thesis I assume equally spaced scores. Departures from equally spaced scores are heavily dependent upon the nature of the outcome and the ordinal outcome trials reviewed in Chapter Three provided no evidence against equally spaced scores. Use of equally spaced scores should therefore apply more widely than alternative scoring methods.

4.3.3 Kappa-type ICC

For discrete outcomes the ICC can be interpreted as a version of the kappa statistic. For binary outcomes kappa is a measure of agreement corrected for chance, most often used to measure inter-rater agreement, which is the agreement among a set of raters recording measurements about a binary trait on the same individual. It is calculated as the observed proportion of agreement minus that expected by chance, divided by the maximum agreement over chance:

$$\hat{\rho}_k = \frac{\hat{\pi}_O - \hat{\pi}_E}{1 - \hat{\pi}_E} \tag{4.8}$$

Where π_O and π_E are the proportions of observed and expected agreement, respectively. A value of 1 for the statistic represents perfect agreement.

Gao has proposed a kappa-type ICC estimator for ordinal data in the cluster randomised trial context.[125] It is similarly constructed as the proportion of observed agreement among pairs of observations within clusters minus that expected by chance, divided by the maximum agreement over chance. However, weighting is used to define the distance between ordinal categories. For equally spaced scores the weight, $w_{jj'}$ (termed the square error rate) corresponding to the agreement between two categories q and q' is:

$$w_{qq'} = 1 - \tfrac{(q-q')^2}{(k-1)^2}$$

$w_{qq'} = 1$ if $q = q'$

	Response				
Response	1	2	3	4	Total
1	$Y_{i1}(Y_{i1}{-}1)/2$				
2	$Y_{i2}Y_{i1}$	$Y_{i2}(Y_{i2}{-}1)/2$			
3	$Y_{i3}Y_{i1}$	$Y_{i3}Y_{i2}$	$Y_{i3}(Y_{i3}{-}1)/2$		
4	$Y_{i4}Y_{i1}$	$Y_{i4}Y_{i2}$	$Y_{i4}Y_{i3}$	$Y_{i4}(Y_{i4}{-}1)/2$	
Total					n(n-1)/2

Table 4.1: Intermediate calculations required for estimating the kappa-type ICC for an ordinal outcome: How to define the number of each possible pairwise response across a 4-level ordinal category within a cluster

The weighting increases the closer the categories are to each other. For example, for a 4-level ordinal outcome the two categories furthest apart (1 and 4) are given a weighting of 0. Categories that are two apart (1 and 3, 2 and 4) are given a weighting of 0.55. Categories that are 1 apart (1 and 2, 2 and 3, 3 and 4) are given a weighting of 0.88 and categories that are the same are given a weighting of 1.

Table 4.1 provides a visual representation of all possible pairings among two observations from a cluster for a 4-level ordinal outcome, where Y_{iq} is the number of observations in category q for cluster i. The table summarises the numbers of each possible pair, with a total of $n(n-1)/2$ possible pairs of observations for a cluster of size n. The details of how the number of each possible pairings is calculated is as follows:

For each cluster it is straightforward to calculate the number of observations in category q, Y_{iq}. If we know the number of observations in a given category we can calculate the number of possible pairs in perfect agreement, i.e. the same category, using standard combination rules $Y_{iq}(Y_{iq}-1)/2$. For two different categories q and q' we can calculate the total number of possible pairings ($[(Y_{iq}+Y_{iq'})(Y_{iq}+Y_{iq'}-1)]/2$) and then subtract the number of pairings expected to be in perfect agreement ($[Y_{iq}(Y_{iq}-1)+Y_{iq'}(Y_{iq'}-1)]/2$). This simplifies so that the number of pairs of two different categories q and q' i.e. the off-diagonal observations in Table 4.1 is $Y_{iq}Y_{iq'}$.

The estimated proportion of observed agreement for a cluster, $\hat{\pi}_O$ is a weighted summation of those observations in perfect agreement (on the diagonal in Table 4.1) and the remaining observations on the lower diagonal, divided by the total number of possible observation pairs:

$$\frac{\frac{1}{2}\sum_{q=1}^{k} w_{qq}Y_{iq}(Y_{iq}-1)+\sum_{q=1}^{k}\sum_{q'>q}^{k} w_{qq'}Y_{iq}Y_{iq'}}{\frac{1}{2}n(n-1)}$$

Therefore the average weighted proportion of agreement over all the clusters is:

$$\hat{\pi}_O = \frac{1}{c}\sum_{i=1}^{c} \frac{\frac{1}{2}\sum_{q=1}^{k} w_{qq}Y_{iq}(Y_{iq}-1)+\sum_{q=1}^{k}\sum_{q>q'}^{k} w_{qq'}Y_{iq}Y_{iq'}}{\frac{1}{2}n(n-1)}$$

Using similar calculations the averaged expected (E) proportion of pairwise agreement, $\hat{\pi}_E$ is given by:

$$\hat{\pi}_E = \frac{\frac{1}{2}\sum_{q=1}^{k} w_{qq}Y_{iq}(Y_{iq}-1)+\sum_{q=1}^{k}\sum_{q>q'}^{k} w_{qq'}Y_{iq}Y_{iq'}}{\frac{1}{2}nc(nc-1)}$$

These calculations can be performed separately for each treatment group and the estimates combined to provide an overall value.

4.3.4 Other ICC estimators for ordinal data

In 1979 Rothery proposed a nonparametric measure of intracluster correlation. Rothery illustrates the idea with observations taken on individuals within a family (cluster). Take a pair of randomly selected families, observations $x_{\alpha i}$ and $x_{\alpha j}$ are from two individuals in one family and observation $x_{\beta k}$ from one individual from the other family. The ICC ρ_c is defined to be the complement of the probability that the outsider $x_{\beta k}$ falls between the two observations from the same family, $x_{\alpha i}$ and $x_{\alpha j}$.[149] Therefore ρ_c is largest when the two observations from the same family are close together. When the data is normally distributed this estimate is a monotone function of the parameter from a one-way ANOVA.[150] A similar measure was proposed by Shirahata.[151] As these estimates are not as familiar to interpret or easy to calculate as the ANOVA and kappa-type estimates they are not considered further.

4.3.5 Relationship between ICC estimators

Gao evaluated the relative bias of the ANOVA and kappa-type ICC estimators via simulation with parameters defined as follows:[125]

- Ordinal outcomes with 3 categories

- Trials with 10 or 20 clusters per arm
- Mean cluster size 50 and 120
- Coefficient of variation in cluster size 0.8 and 1
- Treatment effect (OR) 1 and 1.2

Gao generated clustered ordinal data from a marginal model using a Dirichlet-multinomial distribution with corresponding fixed ICCs of 0, 0.005 and 0.01. Relative bias was calculated as (average observed value minus the true value)/true value. Therefore a positive value indicated an overestimate and a negative value an underestimate in comparison to the ICC from the data generating model. ANOVA ICCs were calculated by assigning both equally spaced and midrank scores. Negative ICCs were truncated at zero.

The simulation results showed that both ICC estimates were closer to the true ICC when either cluster size or the number of clusters was large. When the cluster size was small the ANOVA estimate was a much larger overestimate of the true ICC than the kappa-type ICC. Truncating the ICCs at zero may have elevated the resulting average ICC. However, the percentage of negative values was similar for the two estimates for fixed cluster size. When cluster size was variable the ANOVA estimate produced a larger proportion of negative values and was less biased than the kappa-type estimate for variable cluster size.

Although the work by Gao provides an insight into the relationship between the ANOVA and kappa-type ICC estimates it does not provide information about how these relate to the ICC on the continuous latent response assumed to underlie the random effects ordered logistic regression model, the focus of my research, as Gao used a marginal model to generate and analyse her clustered ordinal data. Also the situations explored by Gao do not reflect the targeted scenarios for my research identified in Chapter Three.

4.3.6 Estimates of ICC from real-life data

In the next chapter I will formally explore the relationship between the ICC on the latent variable, the ANOVA ICC and the kappa-type ICC through a simulation study. In order to inform what range of ICC values should be considered in the simulation study, and to gain some understanding of potential patterns in ICC values for different outcomes/trial settings, I had planned to calculate each ICC estimate using the data from the 11 trials with ordinal outcomes that I identified in Chapter Three. However, none of their accompanying datasets were available from the authors. Instead I calculated, where possible, the ANOVA ICC, latent variable (via a random effects ordinal regression) and kappa-type ICC for three datasets that were available publicly or through the Pragmatic Clinical Trials Unit (PCTU). The estimates are presented below for each dataset, separately for each treatment group and pooled where applicable.

The first dataset is the scenario most relevant to this research, a cluster randomised trial, although the total number of clusters recruited (N=47) was moderate. I was the trial statistician for this trial and was involved in the planning and conduct of the analysis.[152] The second dataset is an observational study and so the ICC for the latent variable cannot be calculated as there is no treatment effect to be modelled but this dataset is large with 720 clusters of size two. The final example is a fairly large longitudinal clinical trial in arthritis with 289 individuals followed up over three time points. This example was referenced by both Lipsitz and Kim in their respective methods for the analysis and sample size calculation of clustered ordinal outcomes by GEE.[65,68]

The EPOS trial: A pragmatic CRT in community care

In the EPOS study clinicians within community mental health teams across three London boroughs were randomised to receive the intervention, DIALOG+, or no intervention. DIALOG+ is a computer mediated intervention consisting of a structured assessment of patients concerns, combined with a solution-focused approach to initiate change. Clinicians (n=47) implemented DIALOG+ with their community patients suffering with psychosis (n=147).

All outcomes were measured at baseline, three months, and six months post randomisation. Social outcomes, including employment, accommodation, and living situation were assessed using the Objective Social Outcomes Index (SIX) with a total score ranging from 0 to 6. A higher score indicates a more positive social outcome. To avoid a small number of observations in some categories and for the purpose of illustration I combined responses to form a 3-level ordinal outcome with the categories low, medium and high.

Response	Control N (%)	Intervention N (%)	Log odds
Low	32 (43%)	26 (36%)	-0.29
Medium	32 (43%)	28 (39%)	-0.72
High	10 (14%)	19 (26%)	

Table 4.2: Real life example results from the EPOS trial of community mental health: 3-level ordinal outcome of social functioning at 6 months

At the 6 month time point there were 47 clinicians who each treated between 1 and 6 patients, the average cluster size was 3.13 (SD 1.24) which implies a coefficient of variation in cluster size of 0.40. The total number of individuals was 147 and the ordinal response at 6 months can be seen in Table 4.2.

I calculated each of the three ICC estimates. The ANOVA estimate of the ICC, accounting for variable cluster sizes using the method described by Eldridge and Kerry[21] was -0.01 in the treatment group and 0.009 in the control group. The pooled estimate was -0.0001. The pooled Kappa-type estimate was -0.25 with estimates of -0.163 in the control group and -0.326 in the treatment group.

I fitted a random effects ordinal regression model to the data with treatment included as a covariate. From this model the estimated ICC for the underlying latent variable was approximately zero with both logit link and probit link. All the ICCs for this trial were low but also showed some differences across treatment groups.

An observational study of Diabetic Retinopathy

The Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR) was a population based epidemiological study based in southern Wisconsin, USA designed to identify the risk factors of diabetic retinopathy. The dataset from this study is publicly available,
www.stat.ufl.edu/ aa/ordinal/ord.html. It contains 720 individuals with measurements on the severity of retinopathy taken on both eyes for each individual.[153]

Categories of the diabetic retinopathy scale were combined to create a 4-level ordinal outcome of severity formed of the categories none, mild, moderate and proliferative. The responses from each eye for each individual can be seen in Table 4.3.

		Right eye				
	Response	None	Mild	Moderate	Proliferative	Total
Left eye	None	237	37	1	0	275
	Mild	31	200	35	4	270
	Moderate	0	39	80	9	128
	Proliferative	0	1	11	35	47
	Total	268	277	127	48	720

* (Source: Williamson et al[153])

Table 4.3: Real life example results from an observational study of Diabetic Retinopathy: 4-level ordinal outcome of severity for each eye*

I calculated the ANOVA estimate of the ICC to be 0.841, and the value for the kappa-type ICC to be 0.837, indicating a strong correlation between the measurements recorded on each eye. As this is an observational model there is no treatment estimate to be assessed and therefore it was not possible to fit a random effects model to calculate the ICC on the underlying variable.

A longitudinal design: An Arthritis clinical Trial

This trial was used in Chapter Two to demonstrate the Kim methodology for sample size calculations with ordinal outcomes.[65] Subjects were randomised to receive the drug Auranofin or placebo for the treatment of arthritis. The primary outcome was self-assessment of arthritis, measured as poor, fair or good at baseline, 1, 3, and 5 months post randomisation. The dataset for this trial is publicly available www.stat.ufl.edu/ aa/ordinal/ord.html.

Eighteen observations were missing in total across the follow-up time points. For the purpose of this example individuals with any missing data have been excluded. There were 289 subjects included in the analysis, each with three follow up measurements. Table 4.4 shows the responses over time for each treatment group.

I calculated the pooled ANOVA estimate of the ICC to be 0.517, 0.559 in the placebo group and 0.447 in the treatment group. Values for the pooled kappa-type ICC were 0.518, 0.577 in the placebo group and 0.446 in the treatment group. I fitted a random effects ordinal regression model to the data with treatment included as a covariate. The estimated ICC for the underlying latent variable was 0.60 with logit link and 0.62 with probit link. These ICCs indicate substantial correlations among the repeat responses.

Response	Baseline N (%)	1 month N (%)	3 months N (%)	5 months N (%)
Auranofin				
Poor	45 (31%)	17 (12%)	28 (19%)	20 (14%)
Fair	66 (46%)	73 (51%)	51 (35%)	51 (35%)
Good	33 (23%)	54 (38%)	65 (45%)	73 (51%)
Total	144	144	144	144
Placebo				
Poor	46 (32%)	43 (30%)	40 (28%)	37 (26%)
Fair	67 (46%)	48 (33%)	61 (42%)	50 (34%)
Good	32 (22%)	54 (37%)	44 (30%)	58 (40%)
Total	145	145	145	145

*(Source: Lipsitz et al[68])

Table 4.4: Real life example results from a longitudinal arthritis clinical trial: 4-level ordinal outcome of self-assessed arthritis over time*

4.4 Discussion

4.4.1 Main findings

Analyses that do not account for the ordinality in the data are not recommended as they can be less powerful and produce different results to those methods that take account of the ordinal nature of the data. Two popular approaches to the analysis of clustered ordinal outcomes that account for the ordinal nature of the outcome are the GEE model and the random effects ordered logistic model. The ordered logistic model, is an extension of the proportional odds model commonly used to analyse ordinal outcomes in the non-clustered situation. Both GEE and random effect models allow incorporation of cluster- and individual-level covariates and can be implemented in standard statistical software. The GEE model makes the stronger assumption that any missing data is MCAR, while the random effects model assumes MAR.

Three ICC estimators were considered for ordinal data. The simplest is calculated using the standard ANOVA method, after assigning equally spaced scores to the ordinal categories. The second ICC is a kappa-type ICC which, for a large number of clusters, is equivalent to the ANOVA estimate. The third is the ICC on the latent continuous scale assumed to underlie the ordinal outcome.

4.4.2 Strengths and limitations

The focus of this chapter has been on the two most well established methods for the analysis of clustered ordinal outcomes. The first is the random effects extension to the cumulative logit model with proportional odds proposed by McCullagh in 1980 and the second is the marginal GEE model proposed by Lipsitz in 1994.[68, 137] These two classes of model were the focus of Agresti and Natarajan's 2001 review on methods for modelling clustered ordinal data.[90] Their review also identified some alternative methods such as Bayesian modelling and semi-parametric random effects models as well as raising some areas for future research such as the handling of missing data. In this thesis I did not consider these alternative methods as the random effects method was shown, in Chapter Three, to be the most popular approach used for the analysis of clustered data, can be easily fitted in current software, and is appropriate given the assumptions behind Campbell and Walters' proposed use of the design effect in sample size calculation. The focus on easy to use available analysis methods is a strength of my research. Given ten years has passed since Agresti and Natarajan's review there is scope for it to be updated. This will be beneficial for future work on sample size methods for ordinal outcomes in circumstances where the assumptions underlying the use of the design effect may not be appropriate, for example under extensive deviations from the proportional odds assumption, and when alternative analysis methods are required.

To find papers describing the calculation of potential estimators of the ICC for ordinal outcomes I performed a broad search in PubMed using the search terms "'ordinal"' and "'cluster*"'. I then targeted alternative data sources that I thought were most likely to report ICC calculations. These were: the PhD thesis of Ruochu Gao which looked at the statistical analysis of correlated ordinal outcomes;[125] the PhD thesis by Sandra Eldridge that focused on cluster randomised trials with a chapter on sample size methods and calculating the ICC; and papers identified during the course of my research describing ICC estimators.[145, 147] There may be alternative ICC estimators that were not identified in this review. However, I expect the possibility of this to be small given ICC estimators were investigated in the two PhD theses I searched, the most recent of which was completed in 2012. The ICC estimators that I have considered have the advantages of being simple to calculate and all three are familiar concepts to statisticians.

4.4.3 Comparison with other work

Estimates of ICCs for continuous and binary outcomes have been extensively summarised.[145, 146] The methods available for ordinal outcomes have received little attention. Gao has compared the ANOVA and kappa-type ICCs for trials with a small number of large clusters and an ordinal outcome with three categories. The data was generated and analysed using a marginal model, hence the link between these ICCs and that defined on an assumed underlying continuous variable is unknown. In my research I plan to evaluate the design effect for sample size calculation which assumes a random-effects ordered logistic regression model with proportional odds. There is still a need to evaluate the relationship between the ICC on the assumed underlying variable for this model and the ANOVA and kappa-type estimators. In this chapter I presented a few real-life examples of these ICCs. However, further datasets are required to identify any clear patterns. There has been no work that I am aware of that looks at patterns in ICCs for ordinal data. Future work should focus on providing more real life estimates and exploring patterns in ICCs.

4.4.4 Implications for this research

This chapter has summarised the main approaches to the analysis of clustered ordinal outcomes and discussed their advantages and disadvantages. In this thesis I focus on analysis using the random effects ordered logistic regression, with proportional odds assumption. The random effects model is the assumed method underlying the use of the design effect for sample size calculations for ordinal outcomes. The estimated treatment effect from the model can be interpreted easily as an odds ratio and it is implemented in the majority of statistical software packages. The model is an extension of the most popular method for analysing ordinal data for the non-clustered case and therefore is likely to be more familiar and acceptable to researchers. In Chapter Three I identified 11 papers with ordinal outcomes, of those that accounted for ordinality in the analysis they did so by using a random effects ordered logistic regression model rather than GEE.

The performance of the ANOVA ICC, Kappa-type ICC, and the ICC on the underlying continuous variable in the design effect for sample size calculation will be explored in the next chapter.

Chapter 5

Simulation studies

Although not as common as binary or continuous outcomes the use of ordinal outcomes in CRTs is not particularly rare. In Chapter Three, using a sample of 300 published cluster randomised trials I identified 11/300 (4%) trials with ordinal outcomes. The design of these eleven trials were most often parallel group (10, 91%), two-arm (6, 55%), completely randomised (8, 73%), and used a cohort sample (9, 82%). Both the total number of clusters randomised (median 18, IQR 9 to 61) and the number of enrolled participants per arm (median 128, IQR 71 to 150) tended to be small. Nine (82%) of these trials had outcomes with between 3 and 5 ordinal categories, the remaining trials had between 6 and 9 categories.

A simple approach to sample size calculation in cluster randomised trials with ordinal outcomes is to inflate the sample size formula for individually randomised trials by the standard design effect (described in Chapter Two). However, the performance of this method under different design scenarios has not been studied and no recommendations exist for how the ICC in the design effect should be calculated.

In this chapter Monte Carlo simulation studies are conducted to explore the relationship between the three possible ICC estimators for ordinal outcomes defined in the previous chapter, (ANOVA, kappa-type estimate, and the ICC on the underlying continuous variable) and to assess the performance in terms of the resulting empirical power when using each of these ICCs in the design effect for sample size calculation. The simulation scenarios were chosen to reflect, as much as possible, the design characteristics of the eleven trials identified with ordinal outcomes. However, I chose to exclude situations where the number of clusters was particularly small as the analysis method is unlikely to perform adequately in these situations. The use of a cohort design in 9/11 (82%) of these trials meant that the analysis of these trials incorporated either a baseline measure or post randomisation measurements of the outcome into the analysis. The effect of such an analysis is to reduce the between-cluster variance. For simplicity the methods evaluated within this chapter assume analysis at the final time point only and are expected to provide conservative estimates for designs which incorporate baseline or repeated measurements.

The chapter concludes with recommendations for implementation of the design effect approach for sample size calculation for cluster randomised trials with ordinal outcomes. The practicalities around using this method are discussed in the final chapter of this thesis.

5.1 Aims and objectives

Monte Carlo simulation studies were conducted to meet the following aims:

1. To explore the relationship between three ICC estimators for ordinal data; the ANOVA ICC, kappa-type ICC and the ICC on the latent continuous response.
2. To determine which ICC provides empirical power closest to the expected nominal value when used in the design effect for sample size calculation.
3. To determine the effect that minor deviations from the proportional odds assumption has on power when using an appropriately (identified in objective 2) calculated design effect for sample size calculation.
4.To determine whether empirical power is consistent when alternative analysis methods are used (i.e. random effects ordered logistic or probit regression, assuming proportional odds) when using an appropriately (identified in objective 2) calculated design effect for sample size calculation.

5.2 Simulation procedures

The best known guidance on the design and reporting of simulation studies was published in 2006 by Burton et al.[154] The simulations described within this chapter were designed and reported following this guidance.

The computational time required to run each simulation restricted the number of scenarios I could reasonably evaluate.

5.2.1 Level of dependence between simulated data sets

The simulations were fully independent, in that a completely different set of independent datasets was generated for each scenario considered. This was achieved by using a different random seed for each scenario. For details on selection of the random seed see section 5.2.4.

5.2.2 Treatment of failures

Any simulated data sets for which the analysis model did not converge were discarded and replacement data sets were generated.

5.2.3 Software

The simulations were performed in Stata version 13. The procedure I used was to write a user-defined command which generated a dataset of clustered ordinal data, performed an appropriate analysis, or calculation of ICC, and stored the relevant estimates. I then used the Stata command *simulate* to replicate my command a large number of times. The details of each of these procedures are described in the sections that follow, and the Stata code can be found in appendix v.

The simulations were performed using the Queen Mary University of London High Performance Computing (HPC) cluster, named Apocrita. The aim of the HPC cluster is to enable tasks to be completed quickly by splitting them across processors. The HPC is accessed remotely and therefore has the advantage that jobs may be submitted to it from any location, thus leaving the user's desktop free to complete other tasks. The HPC uses a Linux operating system with commands entered via the command line, therefore some learning and use of Linux commands was required in running these simulations. Using the HPC meant my simulations were run in parallel and hence the computational time was substantially reduced compared to running these directly within Stata from a desktop computer.

5.2.4 Random number generator and starting seeds

Random number generation within Stata is technically pseudo-random in that numbers are not truly random, but generated by a specific algorithm. Number generation can therefore be replicated by specification of a starting value for the algorithm, referred to as the starting seed (a number between 0 and 2,147,483,647 in Stata). In my command I needed to generate variables from a Normal distribution. To do this I used the random number function *rnormal (μ, σ)* in Stata which returns a normal variable with mean μ and standard deviation σ.

For each scenario I used a different starting seed. The associated Stata help guide for use of random number functions states that it does not matter how the seeds are chosen as long as they do not exhibit any patterns. I chose seeds of varying length and without the use of any systematic selection that would exhibit patterns. I checked that all the seeds used were different from each other.

5.3 Methods for generating the datasets

At the time of my research there was no universally recommended method for generating clustered ordinal data. I explored the 85 papers identified in my review of sample size methods to see how simulation data was generated for binary outcomes but no one method was consistently used. I explored the literature to evaluate the best approach to use for generating clustered ordinal data.

The generation of ordinal clustered data has been described in the literature in the multivariate context i.e. several dependent ordinal outcomes that are correlated with each other. To translate these methods to the cluster randomised trial context we could think of each dependent outcome as representing an individual within a cluster.

In 1995 Gange proposed a method for generating clustered ordinal outcomes.[155] This method simulates ordinal data with a specified marginal and pairwise probability structure using an iterative algorithm. A disadvantage of the algorithm, as stated by the author is, that as cluster size and the number of ordinal categories increases the required computing power may limit the feasibility of the approach. With the advances in computing since 1995 it is not clear whether this issue is still relevant today. The method was used to evaluate the GEE-based sample size method proposed by Kim et al.[65] However, a cluster size of three was assumed and the number of ordinal categories was four.

In 2004 Biswas proposed an algorithm for data generation for specific correlation structures, namely first and second order auto-regressive correlations.[156] These correlation structures imply that as the distance between two observations within the same cluster increases their correlation decreases. This structure is most relevant to longitudinal designs where a cluster is an individual and measurements are taken at repeated time points, and hence it may be reasonable to assume that the correlation will decrease as the lag between time points increases. This correlation structure is a less obvious choice for cluster randomised trials and so this method has limited applicability to my research.

In 2006 Demirtas proposed a data generation procedure which in the first step generates binary data, using what the authors describe as well-accepted data generation methods, and in the second step converts these to ordinal outcomes. The authors state that their method is more general than the methods suggested by Gange[155] and Biswas[156] in that there are no restrictions on the marginal distributions and pairwise correlations and that a large number of ordinal categories does not lead to excessive computational complexity. In 2014 the package ”‘MultiOrd”’ was developed in R to implement the methods of Demirtas.[157] The software makes the method more straightforward to implement and ensures that the data are generated as intended by the authors. However, my experience of using this package with large cluster sizes highlighted that the computational time required was still lengthy.

The method by Gange appears to be the most popular method, with 80 citations in Google Scholar at the time of writing. A common disadvantage to all of the above methods is the complexity and computational time involved as the cluster size increases. Given the number of simulations to be undertaken and the values of the parameters to be investigated in my research it was not feasible to use these methods.

Demirtas described an alternative common approach to ordinal data generation; generate the latent continuous variable and convert to an ordinal outcome by chopping up the continuous outcome into categories using appropriate threshold values. However, Demirtas states that this latent variable approach is generally inappropriate as correlations between the ordinal variables are not of simple form or interpretation, the same comment was also made earlier by Biswas, without any further elaboration.[156] I contacted Demirtas by email and he explained that what he meant by this comment was that the correlation of the underlying latent variable will be different to that of the ordinal version of the outcome and that there is no simple formula to convert one to the other. A similar issue exists for binary data and simulations have been used to link values of the two correlations on the different scales.[147]

Although this latent variable approach is described by Demirtas as being common there is no associated source that provides any evidence of this. In fact it is unclear how often any of these data generation methods have been used. The latent variable approach appears to have been used by Jung and Kang in their evaluation of a score test for clustered ordered categorical data, although the published description of their approach lacks sufficient detail to replicate it.[158] There is scope for further research looking at the prevalence and evaluation of these data generation methods for ordinal outcomes.

I chose to simulate clustered ordinal data using the latent variable approach, evaluating the link between the ICC calculated on the latent continuous variable and the ICCs calculated on the ordinal outcome (kappa-type and ANOVA) by simulation. This method: appears to most closely reflect the proportional odds model whose derivation is motivated by the existence of an underlying continuous variable; the methodology can be easily described and hence replicated by others; the generation can be easily implemented in any statistical package; and the computational time required to generate a dataset is reasonable.

5.3.1 Data generating model

Clustered ordinal data was generated using the latent variable approach. Under this approach we think of the ordinal response categories as being a crude measure of some underlying continuous scale. A linear random-intercept model describes this underlying continuous response Y_{ij}^*

$$Y_{ij}^{*} = \beta x_{ij} + \mu_i + \epsilon_{ij}$$

Where μ_i represents a random effect of the cluster and are independent $N(0, \sigma_b^2)$. The distribution of the error term, ϵ_{ij}, informs the link function used in the random effects model for the ordinal response. A standard logistic distribution with mean 0 and variance $\pi^2/3$ corresponds to a logit link and a standard normal distribution to the probit link. I used the probit link and fixed the distribution of the error terms to be Normal with mean 0 and variance 1.

The ordinal variable Y_{ij} was determined from a categorisation of the latent response via the following threshold model

$$Y_{ij} = \begin{cases} 1 & \text{if } Y_{ij}^{*} \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < Y_{ij}^{*} \leq \alpha_2 \\ \vdots & \vdots \\ c & \text{if } \alpha_{c-1} < Y_{ij}^{*} \end{cases}$$

The thresholds were chosen to reflect the proportions expected in each ordinal category in each treatment group (See sections 5.4.4 and 5.4.5). The Stata function *invnormal* was used to find the inverse of the cumulative standard Normal distribution, and hence the appropriate threshold values. To utilise this function the variable Y_{ij}^{*} first had to be standardised (subtraction of the mean and division by the standard deviation) to follow a standard Normal distribution.

Using the probit link meant that the distribution of Y_{ij}^{*} for the control group followed a known Normal distribution $N(0, 1 + \sigma_b^2)$ and hence the calculation of appropriate threshold values to achieve the required proportions in each ordinal category for the ordinal response were much simpler than with a logit link.

5.4 Scenarios to be investigated

The values of the parameters were chosen to reflect as closely as possible the scenarios identified from the eleven real life examples of trials with ordinal outcomes described in Chapter Three.

For all simulations equal allocation to two treatment arms was assumed: 50% of the clusters allocated to intervention and 50% to the control. The cluster size, number of ordinal categories, ICC on the underlying scale and treatment effect were controlled and varied across scenarios. A description of the estimates used are described in the following sections and summarised in Table 5.1

5.4.1 Analysis method

In a cluster randomised trial for an ordered categorical outcome with k levels $q = 1, 2, \ldots, k$ the ordinal response is given by Y_{ij} for individual j in cluster i, $i = 1, \ldots, C$, $j = 1, \ldots, N$. Each simulated dataset was analysed by a random effects ordered regression model, given by

$$G^{-1}[P(Y_{ij} \leq q)] = \alpha_q - \beta x_{ij}\prime + \mu_i, \qquad q = 1, 2, \ldots, k-1 \tag{5.1}$$

Where the μ_i represent a random effect of the cluster and are distributed $N(0, \sigma_b^2)$ and G^{-1} is the probit link. Although use of the probit link provides simpler calculations in the generation of clustered ordinal data use of the logit link is a more popular choice for real life analysis, as it allows the model parameters to be interpreted as odds ratios. The normal and logistic distributions are similar and therefore I expect the simulation results to be similar if either the probit or logit link were used in the analysis. To confirm an analysis via logit link was also investigated in the simulations. The proportional odds assumption for ordinal regression with logit link implies that the value of the treatment effect does not depend on q, the level of the ordinal response, and therefore the treatment effect of being in category q or better can be represented by a single value β for each value of q. For the ordered probit model the treatment effect is also assumed the same for each cumulative probability. However, as the interpretation of the treatment effect does not relate to odds ratios it is not appropriate to refer to the ordered probit model as a proportional odds model.[130]

5.4.2 Number of clusters and cluster size

The relationship between ICC estimators was investigated for both small and large cluster sizes (5 and 50) and a large number of clusters (100). I conducted sensitivity analysis to look at the relationship of ICC estimators when the number of clusters was small (10).

Table 5.1: Summary of the scenarios to be investigated in the Monte Carlo simulations, designed to meet the following aims (i) to explore the relationship between ICC estimators and (ii) to determine the empirical power when using each ICC estimator in the design effect for sample size calculation

Parameters	Values
Exploration of the relationship between ICC estimators (40 scenarios)	
Allocation ratio	50:50
Number of ordinal categories	3,4,5
Cluster size (fixed)	5, 50
Total clusters*	10*, 100
ICC on the underlying scale	0.01, 0.08, 0.16, 0.25, 0.53
Treatment effect (log odds)	0
*investigated only for the 4-level outcome	
Investigation of empirical power using each ICC in the DE (120 scenarios)	
Allocation ratio	50:50
Number of ordinal categories	3,4,5
Cluster size (fixed)	5, 10, 50
ICC on the underlying scale	0.01, 0.08, 0.16, 0.25, 0.53
ICC in DE	ANOVA, underlying ICC*
Treatment effect (log odds)	0.493, 0.887
*investigated only for the 4-level outcome	
Investigation the effect of non-proportional odds (30 scenarios)	
Allocation ratio	50:50
Number of ordinal categories	4
Cluster size (fixed)	5, 10, 50
ICC on the underlying scale	0.01, 0.08, 0.16, 0.25, 0.53
ICC investigated	ANOVA
Treatment effect (log odds)	0.35, 0.45
Investigation of empirical power with ANOVA ICC and logit link (30 scenarios)	
Allocation ratio	50:50
Number of ordinal categories	4
Cluster size (fixed)	5, 10, 50
ICC on the underlying scale	0.01, 0.08, 0.16, 0.25, 0.53
ICC investigated	ANOVA
Treatment effect (log odds)	0.493, 0.887

For all other simulations cluster sizes were fixed at 5, 10 and 50. The number of clusters for each scenario was determined by calculating the sample size required under individual randomisation using Whitehead's formula and multiplying by the appropriate design effect for the scenario.

5.4.3 Number of ordinal categories

The simulation studies focused on ordinal outcomes with four levels, the most common number of levels used in the trials I identified in Chapter Three with ordinal outcomes. The sensitivity of the main results, the assessment of power when using the design effect for sample size calculation, was explored for outcomes with three and five levels.

5.4.4 Estimated probabilities in each category

The control group proportions for the four-level ordinal outcome were chosen to be the same as those used in the example by Whitehead. The four ordinal categories were very good, good, moderate and poor and the corresponding proportions expected in each category were 0.20, 0.50, 0.20, 0.10. These proportions correspond to threshold values for the underlying latent response of $\alpha_1 = -0.84, \alpha_2 = 0.52, \alpha_3 = 1.28$.

For the exploration of relationships amongst ICC estimators I also considered an alternative categorisation of the four-level ordinal outcome of 0.10, 0.30, 0.40 and 0.20 to determine whether the relationships are affected by the proportions expected in each category.

For the sensitivity analyses the control proportions for the three-level version of the outcome were 0.20, 0.70 and 0.10 with corresponding threshold values for the underlying latent response of $\alpha_1 = -0.84, \alpha_2 = 1.28$. For the five-level outcome the control proportions were 0.20, 0.20, 0.30, 0.20, and 0.10 with corresponding threshold values for the underlying latent response of $\alpha_1 = -0.84, \alpha_2 = -0.25, \alpha_3 = 0.52, \alpha_4 = 1.28$. Proportions for the three and five-level versions of the outcome were chosen by expanding or combining categories from the four-level version of the outcome so as to introduce minimal changes and maintain the same underlying proportions to maintain comparability amongst the three-, four-, and five-level versions of the outcome.

The proportional odds assumption implies that the treatment effect can be represented by a single value β for each value of q. Violation of the proportional odds assumption may occur in a variety of ways. The treatment effect may be different for each category q, with the effects being in the same or alternative directions across the categories. Alternatively one or more of the log odds ratios may be different for one or two categories only. The magnitude of any differences in the treatment effect across categories will determine whether proportional odds are still a reasonable assumption. In situations where proportional odds is unreasonable the partial proportional odds model or nominal regression may be more suitable analysis methods.

In this simulation I examined the empirical power of the design effect approach, followed by an analysis by random effects ordinal regression with probit link, under a minor deviation from proportional odds. I considered the deviation where one odds ratio was slightly different from the rest for a four-level outcome. The log odds ratio for the first dichotomisation was 0.35 and 0.493 for the other categories. These log-odds were chosen to ensure a relatively large number of clusters per arm in order to provide reliable empirical power estimates.

To assess the effect of non-proportional odds an initial dataset was generated using the methods described in this section. In order to adjust the observed proportions to reflect non-proportionality I generated a uniformly distributed random variable for every observation using Stata's *runiform()* command. The dataset was then sorted by treatment group, ordinal category and the random variable. In the treatment group the first 10% in the first ordinal category were assigned the category below. Therefore the proportions in the control group remained as 0.20, 0.50, 0.20 and 0.10 and in the treatment group the proportions became 0.26, 0.53, 0.14 and 0.06, which correspond to the required log-odds of 0.35 for the first dichotomisation and 0.493 for the others.

5.4.5 Treatment effect

For the simulations that explored the relationships between ICC estimators I assumed there was no treatment effect i.e. a log-odds of 0. See the discussion section at the end of this chapter for a discussion of the implications of this assumption.

The simulations were designed around the example provided in Whitehead's paper on sample size methods for ordinal outcomes in the individually randomised design. Whitehead's example was used as a starting point in order to be able to compare the results to the clustered case. In Whitehead's example a log-odds of 0.887 (OR 2.43) was assumed, which indicates superiority of the new treatment. I expanded his example to additionally consider log-odds either side of this, 0.493 (OR 1.6), and 1.207 (OR 3.3) in order to explore patterns for small, medium and large effect sizes. However, the larger odds ratio was later removed from the simulations as the analysis model did not perform well on the small number of clusters calculated for the sample size with this treatment effect.

The above estimates of log-odds were based on a proportional odds cumulative logit model. I used an ordered probit model to generate my data as it provided simpler calculations (see section 5.3.1). Therefore in the data generating process the above log-odds needed to be transformed to the corresponding treatment effect that would be expected from an ordered probit regression on this data. Using the fact that the standard normal cumulative distribution function (cdf) at a point z is well approximated by the standard logistic cdf at 1.7z the estimates (log-odds) from the logit model are approximately 1.7 times those from ordered probit models.[130] Therefore, the corresponding treatment estimates for the probit model used in data generation were 0.29 and 0.52. As the treatment effect under a random effects model is proportional to the total variance the treatment effect was multiplied by the value of $\sqrt{(1+\sigma_w^2)}$ for that particular scenario to account for an analysis by random effects ordered probit regression. In the presentation of the results I use the values of the log odds from the logit model, 0.887 and 0.493 as their interpretation is more familiar and simpler than the treatment effect under a probit link.

When exploring the empirical power for the minor deviation from proportionality for the four-level outcome the proportions expected in the control group are 0.20, 0.50, 0.20 and 0.10 and in the treatment group 0.26, 0.53, 0.14 and 0.06. This corresponds to log-odds ratios of 0.35, 0.493 and 0.493 for the three possible dichotomisations of the outcome. Whitehead's sample size method requires all expected proportions to be specified and a single estimate for the treatment effect. I investigated the resulting power under two different scenarios for estimating the treatment effect (i) summarise the log-odds using the mean, 0.45 (ii) use the smallest log odds ratio of 0.35.

5.4.6 Incorporation of clustering

It is only the ICC on the underlying continuous variable that can be fixed during the data generation process. The ICC for the underlying continuous variable, with probit link is defined as

$$\rho_l = \frac{\sigma_b^2}{\sigma_b^2+1}$$

The relationship between this and other ICC estimators is unknown. The values of ICCs likely to be observed in practice for ordinal outcomes are relatively unknown, although they are probably not that dissimilar to those seen for binary and continuous outcomes, particularly if an underlying continuous variable can be assumed.

Several publications describe the general behaviour and patterns in ICCs across different therapeutic areas, outcomes and clusters.[14–18] These patterns have been summarised by Eldridge and Kerry.[21] Process outcomes generally have higher ICC values (median 0.063) than clinical or individual level outcomes (median 0.03); ICCs are higher in secondary care settings (median 0.061) than primary care (median 0.045); ICCs are larger when the natural cluster size is small; ICCs are lower for binary outcomes with extreme prevalence; ICCs over 0.35 are unlikely for binary outcomes. The values of ICC for the underlying continuous variable considered in the simulation (assumed to be the same in both treatment arms) were chosen to cover a large range of ICC estimates, 0.01, 0.08, 0.16, 0.25, and 0.53. The reason these specific values were chosen was that in the paper by Eldridge these ICCs on the underlying latent scale were shown to correspond to ICCs for binary outcomes on the proportions scale of 0.01, 0.05, 0.1, 0.15, 0.3, when overall prevalence is 0.30.[147] Using these values with known relationship to the binary ICC on the proportions scale allowed scope for some validation of my data generation (see section 5.7).

5.5 Estimates to be stored

Summary statistics

For each simulated dataset the proportion in each ordinal category and treatment group were calculated and stored.

Analysis

For each simulated dataset a random effects ordered probit regression model was fitted to the data and the estimated treatment effect was stored. The hypothesis of no treatment effect was tested via the Wald statistic which was calculated as the estimated treatment effect divided by its standard error. The Wald statistic was compared to a Normal distribution and the P-value stored.

Post-estimation

For each simulated dataset the ANOVA estimate of the ICC was calculated by first assigning equally spaced scores to the ordinal outcome. The Stata inbuilt commands for calculating the ANOVA ICC truncate the ICC at zero, inflating the average ANOVA ICC estimate. To avoid this I calculated the ANOVA ICC from first principles using the estimates of the between and within sums of squares (SS_b and SS_w) and associated degrees of freedom (DF_b, DF_w) that are provided by the Stata command *loneway* which performs a one-way analysis of variance . The between-group mean squares MS_b is calculated as SS_b/DF_b and similarly the within-group mean squares, MS_w is calculated as SS_w/DF_w and n is the cluster size. The ANOVA ICC is then calculated as:

$$\rho_a = \frac{MS_b - MS_w}{MS_b + (n-1)MS_w}$$

The observed ICC on the latent scale was calculated using the between-cluster estimate of the variance provided directly from the probit model fitted in Stata, input into the following formula

$$\rho_l = \frac{\sigma_b^2}{\sigma_b^2 + 1}$$

I programmed the calculation of the kappa-type ICC using the formula described in Chapter Four, Section 4.3.

Summary statistics over all simulations

Descriptive statistics for ICC estimates calculated over all simulated datasets within a scenario were mean (SD), median(IQR), minimum and maximum, and the percentage of negative values.

Empirical Power

For each scenario the empirical power was calculated as the proportion of datasets with a P-value of less than 0.05. The sample size was deemed appropriate if the empirical power was close to the calculated power of 90%. In order to determine how others had defined what would be considered to be sufficiently close I reviewed 18 papers that evaluated their sample size method through simulation, from the 85 sample size papers identified in Chapter Two. However, in the majority of papers no formal definition of what constituted sufficiently close was provided. A few papers calculated an absolute difference between empirical and nominal power, reporting the maximum difference seen or those that were at least 2% different. One paper reported the relative bias of power calculated as the calculated power minus empirical power divided by the empirical power. Biases that were greater than 10% were highlighted. In my simulations I report empirical power, the standard error and the absolute difference from the calculated power. I consider the design effect approach appropriate if simulated power is within 2% of nominal power.

5.6 Number and size of simulated datasets

Due to the large number of scenarios studied the number of simulations for each scenario was limited to 1000. In comparison to the 18 papers identified with simulation studies from my sample of 85 sample size papers the number of simulations conducted were 1000 (n=7), 2000 (n=2), 4000 (n=1), 5000 (n=3) and 10,000 (n=5). In only one of these papers was the number of simulations (10,000) justified. However, it is unclear whether this justification was post-hoc. The authors state that the Monte Carlo standard error for the resulting power calculations is no greater than 0.25%.[72] The standard error is presented for each of my simulation scenarios in the results section.

The size of each dataset was dependent upon the scenario being investigated. The sample size was first calculated using Whitehead's formula assuming individual randomisation with 90% power and a 2-sided 5% significance level. Table 5.2 summarises the sample size required under individual randomisation for the 3-,4- and 5-level ordinal responses under each of the possible treatment effects and when the assumption of proportional odds is relaxed. The sample size was multiplied by the standard design effect, $1+(n-1)\rho$ for each combination of cluster size and ICC estimate investigated. Estimates were always rounded up to ensure an equal number of clusters, of required size, per treatment group.

Table 5.2: Sample size requirements under individual randomisation for ordinal outcomes with 3,4, or 5 categories for various treatment effects

Outcome levels	Control proportions	Treatment proportions	Treatment effect (log OR)	Total sample size required
3-level	0.20, 0.70, 0.10	0.29, 0.65, 0.06	0.493	764
		0.38, 0.58, 0.04	0.887	226
4-level	0.20, 0.50, 0.20, 0.10	0.29, 0.50, 0.14, 0.06	0.493	608
		0.38, 0.47, 0.11, 0.04	0.887	188
	non-proportional odds			
	0.20, 0.50, 0.20, 0.10	0.26, 0.53, 0.14, 0.06	0.35	1200
	0.20, 0.50, 0.20, 0.10	0.26, 0.53, 0.14, 0.06	0.45	738
5-level	0.20, 0.20, 0.30, 0.20, 0.10	0.29, 0.23, 0.27, 0.14, 0.06	0.493	550
		0.38, 0.24, 0.23, 0.11, 0.04	0.887	172

5.7 Validation

Data Generating model

After discussion and agreement of the data generating process with SE and AC several strategies were employed to validate the process and ensure confidence in the results.

I sent a written summary of the data generating process to Alan Agresti, one of the leaders in the field of ordinal outcomes. However, he did not have experience of generating clustered data and as he has now retired from academia he did not have any PhD students or colleagues that I might confirm the process with.

My simulation code was checked by a colleague, Neil Wright, from the PCTU who has experience of generating clustered binary data using similar methods to mine. No issues for concern were identified.

From the datasets generated for the four-level ordinal outcome to explore the relationships between ICC estimators I created a binary outcome which combined the first two categories and the last two categories, resulting in an overall prevalence of 0.30. From the work conducted by Eldridge et al it is known that for an overall prevalence of 0.30 ICC values on the underlying scale of 0.01, 0.08, 0.16, 0.25, and 0.53 correspond to ICCs on the proportions scale of 0.01, 0.05, 0.1, 0.15 and 0.3.[147] This result was reassuringly replicated in my dataset.

The analysis method

In order to ensure that my test procedure, the Wald test, performed well under the simulation scenarios I examined the empirical Type I error rate to determine the minimum number of clusters required in the analysis to provide an empirical Type I error rate close to the nominal error rate of 5%.

Datasets of clustered ordinal outcomes were generated assuming the null hypothesis of no difference between treatment groups and analysed with a random effects effects ordered probit regression model. The effect of treatment was tested using a Wald test comparing the test statistic to a normal distribution and using a Wald test comparing the test statistic to a t-distribution with degrees of freedom calculated as the number of clusters minus two. One thousand datasets were generated for a 4-level ordinal outcome with 5, 10, 15, 20, 25 and 40 clusters per arm, cluster sizes of 5, 10 and 50,

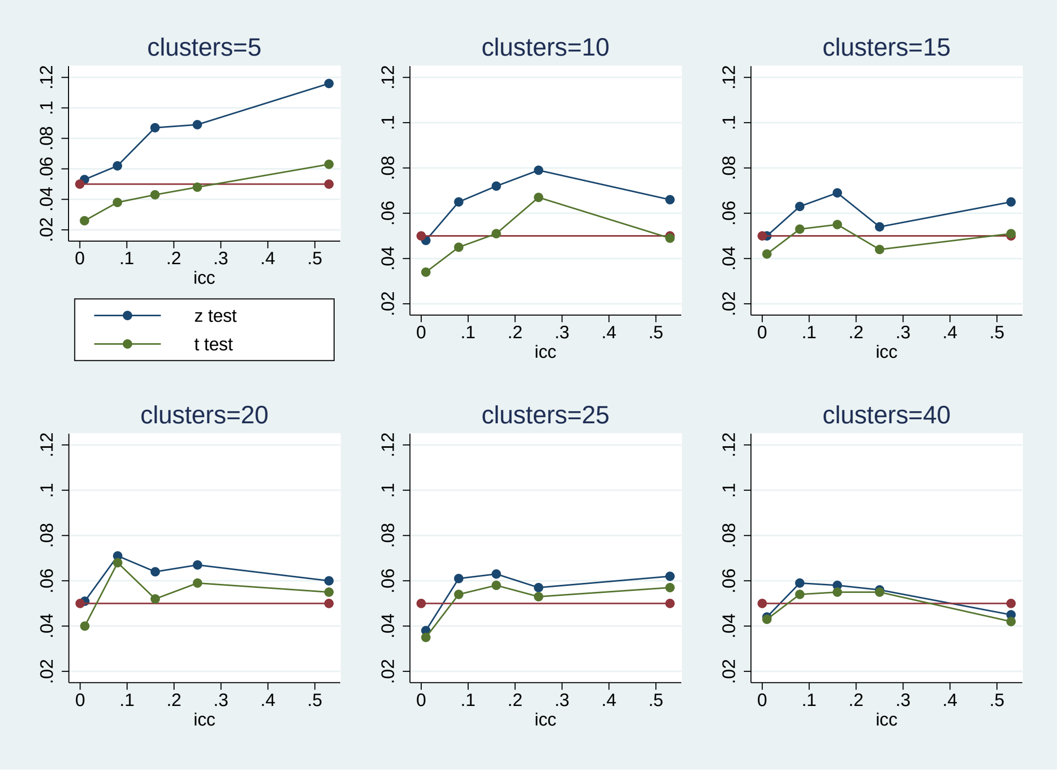


Figure 5.1: Empirical Type I error rates for the Wald test for the treatment effect from a random effects ordered probit model, compared to both the Normal and t-distributions (varying number of clusters of size 5)

and ICCs on the underlying latent variable of 0.01, 0.08, 0.16, 0.25, and 0.53. The empirical Type I error rate was calculated as the proportion of P-values that were less than 0.05.

The results for clusters of size 5, 10 and 50 are presented in figures 5.1, 5.2 and 5.3. These figures show that the type I error for the Wald test compared to a t-distribution was always lower than that compared to the normal distribution. The empirical Type I error rate comes closer to the nominal value as the number of clusters increases, with 40 clusters per arm showing good results.

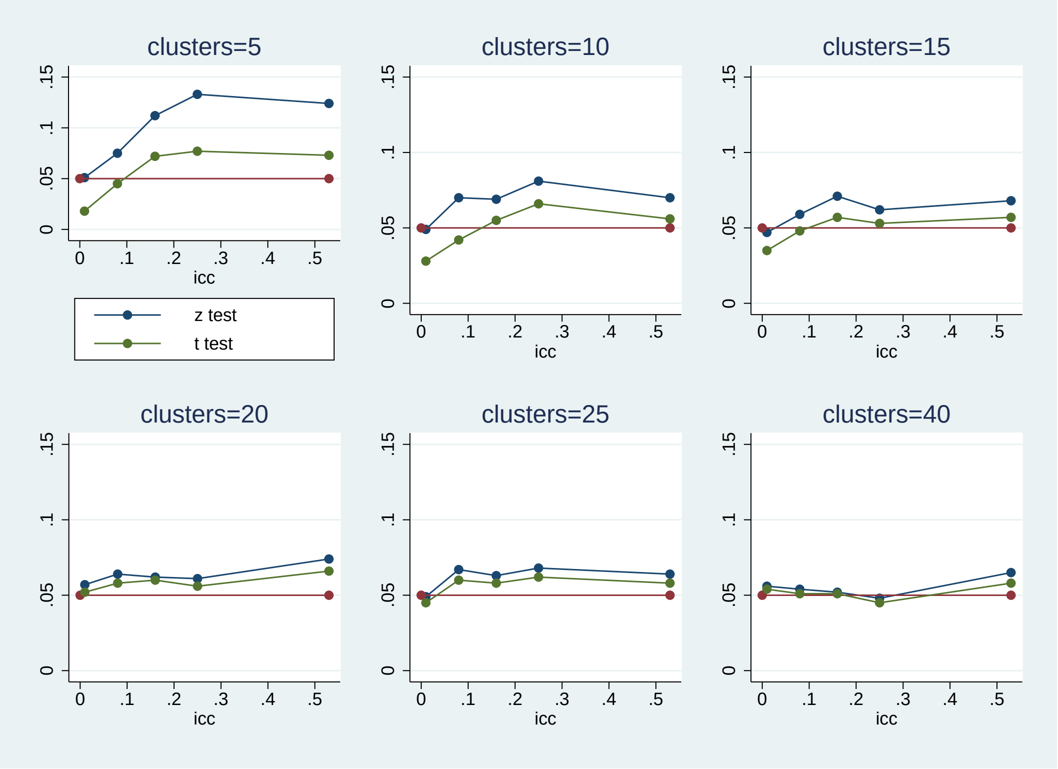


Figure 5.2: Empirical Type I error rates for the Wald test the treatment effect from a random effects ordered probit model, compared to both the Normal and t-distributions (varying number of clusters of size 10)

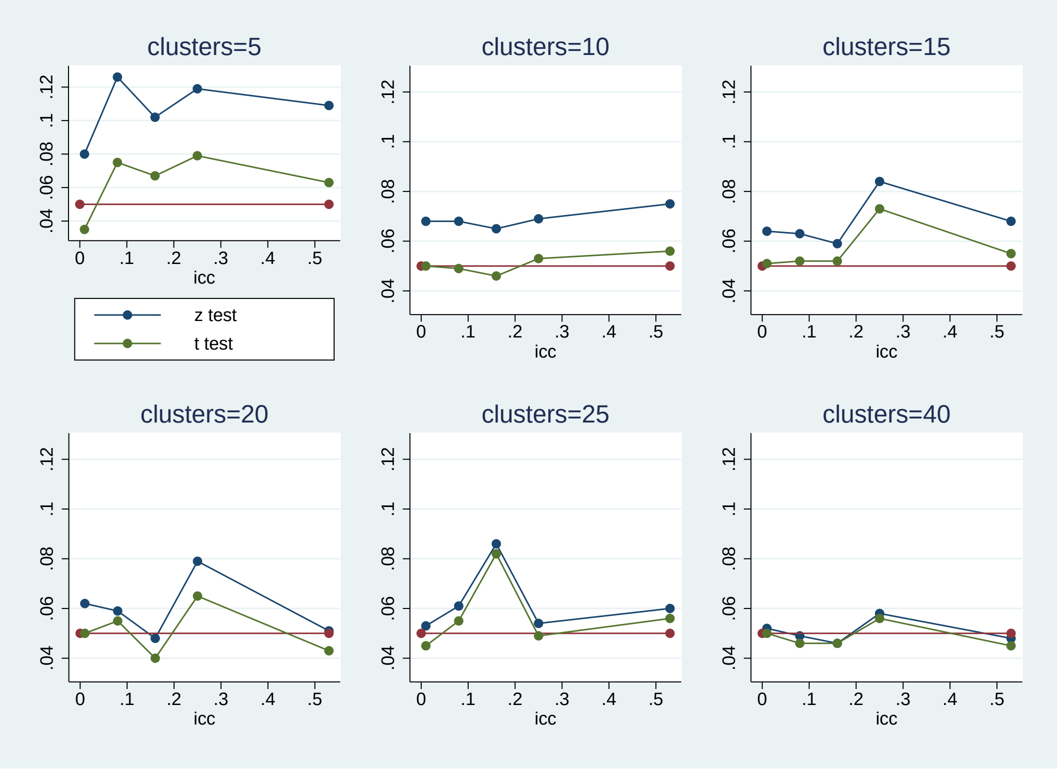


Figure 5.3: Empirical Type I error rates for the Wald test for the treatment effect from a random effects ordered probit model, compared to both the Normal and t-distributions (varying number of clusters of size 50)

5.8 Results

All scenarios were simulated reliably. The number of analysis models that did not converge was less than 10 (1%) for all scenarios considered.

5.8.1 Relationship between ICC estimators

For fixed values of the ICC on the latent continuous variable (ρ_l) 1000 datasets were generated for a total of 100 clusters of size 5 and 50 for ordinal outcomes with 3,4, and 5 levels. For each dataset the ANOVA ICC (ρ_a) and the kappa-type ICC (ρ_k), were calculated. The ANOVA ICC calculated on a dichotomised version of the outcome (ρ_b) was also calculated as validation of the data generating process, as its relationship with the ICC on the latent scale is known.

The relationship between ICCs for the 3-, 4-, and 5-level outcomes are provided in Tables 5.3, 5.4 and 5.5. The results show the ICC on the latent response to be largest, followed by the ANOVA and kappa-type ICCs that were almost identical. Both the ANOVA ICC and kappa-type estimates produced a similar proportion of negative estimates in each scenario. The proportion of negative estimates decreased as the level of clustering and number of clusters increased. As the number of ordinal categories increased so too did the estimated ANOVA and Kappa-type ICCs. These observed patterns were consistent across the 3-, 4- and 5-level outcome variables.

Sensitivity analysis looked at the relationship between ICC estimators for the 4-level outcome when the number of clusters was small (Table 5.6). Results show that the ICC on the latent response tended to be largest followed by the ANOVA and kappa-type ICCs, with the ICC on the dichotomised version of the outcome being the smallest. As the level of clustering increases so too does the difference between the ANOVA and kappa-type estimates with the ANOVA estimate being consistently larger. The relationships between ICC estimators for the 4-level outcome with a small and large number of clusters are displayed in Figures 5.4 and 5.5.

Table 5.7 shows the relationship between ICCs for a four-level ordinal outcome where different proportions from those used in the main simulations are expected in each category. Comparing these to the results seen in Table 5.4 shows that for a fixed ICC on the underlying latent variable the values of the ANOVA and kappa-type ICC are the same even when the proportions in each category change if the level of clustering is low. When the level of clustering is larger the values of the ANOVA and kappa-type ICC are affected by the change in proportions expected in each category.

Table 5.3: Results showing the relationship between the ICC on the underlying continuous latent variable (ρ_l) and ICCs calculated on the 3-level ordinal outcome, ANOVA (ρ_a) and kappa (ρ_k), when the number of clusters is large. The dichotomised version of the ordinal outcome assumes an overall prevalence of 0.2 and its associated ICC is given by (ρ_b). It was assumed there was no effect of treatment.

Parameters (1000 simulations)				Descriptive statistics			
Total clusters	Cluster size	$\rho_l(\sigma_b^2)$	ICC	Mean (SD)	Median (IQR)	min, max	% negative
100	5	0.01 (0.01)	ρ_a	0.01 (0.03)	0.01 (-0.02, 0.03)	-0.10, 0.13	435 (44%)
			ρ_k	0.01 (0.03)	0.01 (-0.02, 0.03)	-0.10, 0.13	435 (44%)
			ρ_b	0.00 (0.03)	0.00 (-0.02, 0.02)	-0.08, 0.12	463 (46%)
		0.08 (0.09)	ρ_a	0.06 (0.04)	0.06 (0.03, 0.08)	-0.06, 0.16	59 (6%)
			ρ_k	0.06 (0.04)	0.06 (0.03, 0.08)	-0.06, 0.16	59 (6%)
			ρ_b	0.04 (0.04)	0.04 (0.01, 0.07)	-0.08, 0.19	138 (14%)
		0.16 (0.19)	ρ_a	0.11 (0.04)	0.11 (0.09, 0.14)	-0.03, 0.25	3(< 1%)
			ρ_k	0.11 (0.04)	0.11 (0.09, 0.14)	-0.03, 0.25	3(< 1%)
			ρ_b	0.08 (0.04)	0.08 (0.05, 0.11)	-0.04, 0.21	18 (2%)
		0.25 (0.33)	ρ_a	0.17 (0.05)	0.17 (0.14, 0.21)	0.05, 0.35	0
			ρ_k	0.17 (0.04)	0.17 (0.14, 0.20)	0.05, 0.34	0
			ρ_b	0.13 (0.05)	0.13(0.10, 0.16)	-0.00, 0.28	1(< 1%)
		0.53 (1.13)	ρ_a	0.38 (0.05)	0.38 (0.34, 0.42)	0.18, 0.54	0
			ρ_k	0.38 (0.05)	0.38 (0.34, 0.41)	0.17, 0.54	0
			ρ_b	0.32 (0.06)	0.32 (0.27, 0.36)	0.11, 0.52	0
	50	0.01 (0.01)	ρ_a	0.01 (0.00)	0.01 (0.00, 0.01)	-0.01, 0.02	31 (3%)
			ρ_k	0.01 (0.00)	0.01 (0.00, 0.01)	-0.01, 0.02	31 (3%)
			ρ_b	0.00 (0.00)	0.00 (0.00, 0.01)	-0.00, 0.02	88 (9%)
		0.08 (0.09)	ρ_a	0.06 (0.01)	0.06 (0.05, 0.07)	0.03, 0.10	0
			ρ_k	0.06 (0.01)	0.06 (0.05, 0.07)	0.03, 0.09	0
			ρ_b	0.04 (0.01)	0.04 (0.04, 0.05)	0.01, 0.07	0
		0.16 (0.19)	ρ_a	0.11 (0.02)	0.11 (0.10, 0.13)	0.07, 0.17	0
			ρ_k	0.11 (0.02)	0.11 (0.10, 0.12)	0.07, 0.17	0
			ρ_b	0.08 (0.01)	0.08 (0.07, 0.09)	0.05, 0.14	0
		0.25 (0.33)	ρ_a	0.18 (0.02)	0.18 (0.16, 0.19)	0.10, 0.26	0
			ρ_k	0.18 (0.02)	0.17 (0.16, 0.19)	0.10, 0.26	0
			ρ_b	0.13 (0.02)	0.13 (0.12, 0.15)	0.07, 0.20	0
		0.53 (1.13)	ρ_a	0.38 (0.03)	0.38 (0.36, 0.40)	0.27, 0.50	0
			ρ_k	0.38 (0.03)	0.38 (0.36, 0.40)	0.26, 0.49	0
			ρ_b	0.32 (0.04)	0.32 (0.29, 0.34)	0.20, 0.45	0

Notes. The data was generated for a 3-level ordinal outcome with proportions in each category of 0.20, 0.70, and 0.1 for both treatment groups.

Table 5.4: Results showing the relationship between the ICC on the underlying continuous latent variable (ρ_l) and ICCs calculated on the 4-level ordinal outcome, ANOVA (ρ_a) and kappa (ρ_k), when the number of clusters is large. The dichotomised version of the ordinal outcome assumes an overall prevalence of 0.3 and its associated ICC is given by (ρ_b). It was assumed there was no effect of treatment.

Parameters (1000 simulations)				Descriptive statistics			
Total clusters	Cluster size	$\rho_l(\sigma_b^2)$	ICC	Mean (SD)	Median (IQR)	min, max	% negative
100	5	0.01 (0.01)	ρ_a	0.01 (0.03)	0.01 (-0.01, 0.03)	-0.10, 0.12	407(41%)
			ρ_k	0.01 (0.03)	0.01 (-0.01, 0.03)	-0.10, 0.12	409(41%)
			ρ_b	0.01 (0.03)	0.00 (-0.02, 0.03)	-0.10, 0.12	448(45%)
		0.08 (0.09)	ρ_a	0.07 (0.04)	0.07 (0.04, 0.09)	-0.05, 0.19	29(3%)
			ρ_k	0.07 (0.04)	0.07 (0.04, 0.09)	0.05, 0.19	29(3%)
			ρ_b	0.05 (0.04)	0.05 (0.02, 0.07)	-0.05, 0.18	90(9%)
		0.16 (0.19)	ρ_a	0.14 (0.04)	0.13 (0.11, 0.16)	0.01, 0.30	0(0%)
			ρ_k	0.13 (0.04)	0.13 (0.10, 0.16)	0.01, 0.30	0(0%)
			ρ_b	0.09 (0.04)	0.09 (0.07, 0.12)	-0.02, 0.27	9(1%)
		0.25 (0.33)	ρ_a	0.21 (0.05)	0.21 (0.18, 0.24)	0.06, 0.37	0(0%)
			ρ_k	0.21 (0.05)	0.21 (0.18, 0.24)	0.06, 0.36	0(0%)
			ρ_b	0.15 (0.05)	0.15 (0.12, 0.18)	-0.03, 0.33	1($<$ 1%)
		0.53 (1.13)	ρ_a	0.45 (0.05)	0.46 (0.42, 0.49)	0.26, 0.59	0(0%)
			ρ_k	0.45 (0.05)	0.45 (0.41, 0.48)	0.26, 0.59	0(0%)
			ρ_b	0.34 (0.06)	0.34 (0.30, 0.38)	0.15, 0.55	0(0%)
	50	0.01 (0.01)	ρ_a	0.01 (0.00)	0.01 (0.01, 0.01)	-0.00, 0.02	16(2%)
			ρ_k	0.01 (0.00)	0.01 (0.01, 0.01)	-0.00, 0.02	16(2%)
			ρ_b	0.01 (0.00)	0.01 (0.00, 0.01)	-0.01, 0.02	46(5%)
		0.08 (0.09)	ρ_a	0.07 (0.01)	0.07 (0.06, 0.08)	0.04, 0.10	0(0%)
			ρ_k	0.07 (0.01)	0.07 (0.06, 0.08)	0.03, 0.10	0(0%)
			ρ_b	0.05 (0.01)	0.05 (0.04, 0.05)	0.02, 0.07	0(0%)
		0.16 (0.19)	ρ_a	0.14 (0.02)	0.14 (0.12, 0.15)	0.08, 0.20	0(0%)
			ρ_k	0.13 (0.02)	0.13 (0.12, 0.15)	0.08, 0.20	0(0%)
			ρ_b	0.09 (0.01)	0.09 (0.08, 0.10)	0.05, 0.14	0(0%)
		0.25 (0.33)	ρ_a	0.21 (0.03)	0.21 (0.19, 0.23)	0.14, 0.30	0(0%)
			ρ_k	0.21 (0.02)	0.21 (0.19, 0.22)	0.14, 0.30	0(0%)
			ρ_b	0.15 (0.02)	0.15 (0.13, 0.16)	0.09, 0.23	0(0%)
		0.53 (1.13)	ρ_a	0.46 (0.04)	0.46 (0.43, 0.48)	0.33, 0.57	0(0%)
			ρ_k	0.45 (0.04)	0.45 (0.43, 0.47)	0.33, 0.56	0(0%)
			ρ_b	0.34 (0.04)	0.34 (0.32, 0.36)	0.23, 0.44	0(0%)

Notes. The data was generated for a 4-level ordinal outcome with proportions in each category of 0.20, 0.50, 0.20 and 0.1 for both treatment groups.

Table 5.5: Results showing the relationship between the ICC on the underlying continuous latent variable (ρ_l) and ICCs calculated on the 5-level ordinal outcome, ANOVA (ρ_a) and kappa (ρ_k), when the number of clusters is large. The dichotomised version of the ordinal outcome assumes an overall prevalence of 0.4 and its associated ICC is given by (ρ_b). It was assumed there was no effect of treatment.

Parameters (1000 simulations)				Descriptive statistics			
Total clusters	Cluster size	$\rho_l(\sigma_b^2)$	ICC	Mean (SD)	Median (IQR)	min, max	% negative
100	5	0.01 (0.01)	ρ_a	0.01 (0.03)	0.01 (-0.01, 0.03)	-0.09, 0.12	392(39%)
			ρ_k	0.01 (0.03)	0.01 (-0.01, 0.03)	-0.09, 0.12	392 (39%)
			ρ_b	0.01 (0.03)	0.01 (-0.02, 0.03)	-0.10, 0.14	417 (42%)
		0.08 (0.09)	ρ_a	0.07 (0.04)	0.07 (0.05, 0.10)	-0.05, 0.22	20 (2%)
			ρ_k	0.07 (0.04)	0.07 (0.05, 0.10)	-0.05, 0.22	20 (2%)
			ρ_b	0.05 (0.04)	0.05 (0.03, 0.08)	-0.05, 0.21	81 (8%)
		0.16 (0.19)	ρ_a	0.15 (0.04)	0.15 (0.12, 0.18)	0.01, 0.29	0
			ρ_k	0.15 (0.04)	0.15 (0.12, 0.18)	0.01, 0.29	0
			ρ_b	0.10 (0.04)	0.10 (0.07, 0.13)	-0.03, 0.28	3($<$ 1%)
		0.25 (0.33)	ρ_a	0.23 (0.05)	0.23 (0.19, 0.26)	0.09, 0.37	0
			ρ_k	0.22 (0.05)	0.23 (0.19, 0.25)	0.09, 0.37	0
			ρ_b	0.16 (0.04)	0.16 (0.13, 0.19)	0.02, 0.30	0
		0.53 (1.13)	ρ_a	0.49 (0.05)	0.49 (0.45, 0.52)	0.31, 0.62	0
			ρ_k	0.48 (0.05)	0.49 (0.45, 0.52)	0.31, 0.61	0
			ρ_b	0.35 (0.05)	0.36 (0.32, 0.39)	0.20, 0.53	0
	50	0.01 (0.01)	ρ_a	0.01 (0.00)	0.01 (0.01, 0.01)	-0.00, 0.02	9 (1%)
			ρ_k	0.01 (0.00)	0.01 (0.01, 0.01)	-0.00, 0.02	9(1%)
			ρ_b	0.01 (0.00)	0.01 (0.00, 0.01)	-0.00, 0.02	43 (4%)
		0.08 (0.09)	ρ_a	0.08 (0.01)	0.08 (0.07, 0.08)	0.04, 0.12	0
			ρ_k	0.07 (0.01)	0.07 (0.07, 0.08)	0.04, 0.11	0
			ρ_b	0.05 (0.01)	0.05 (0.05, 0.06)	0.02, 0.08	0
		0.16 (0.19)	ρ_a	0.14 (0.02)	0.14 (0.13, 0.16)	0.09, 0.21	0
			ρ_k	0.14 (0.02)	0.14 (0.13, 0.16)	0.08, 0.21	0
			ρ_b	0.10 (0.02)	0.10 (0.09, 0.11)	0.06,0.15	0
		0.25 (0.33)	ρ_a	0.23 (0.03)	0.23 (0.21, 0.24)	0.15, 0.30	0
			ρ_k	0.22 (0.03)	0.22 (0.21, 0.24)	0.15, 0.30	0
			ρ_b	0.16 (0.02)	0.16 (0.14, 0.17)	0.10, 0.21	0
		0.53 (1.13)	ρ_a	0.49 (0.04)	0.49 (0.46, 0.51)	0.39, 0.59	0
			ρ_k	0.48 (0.04)	0.48 (0.46, 0.51)	0.38, 0.59	0
			ρ_b	0.35 (0.03)	0.35 (0.33, 0.38)	0.25, 0.46	0

Notes. The data was generated for a 5-level ordinal outcome with proportions in each category of 0.20, 0.20, 0.30, 0.20 and 0.1 for both treatment groups.

Table 5.6: Results showing the relationship between the ICC on the underlying continuous latent variable (ρ_l) and ICCs calculated on the 4-level ordinal outcome, ANOVA (ρ_a) and kappa (ρ_k), when the number of clusters is small. The dichotomised version of the ordinal outcome assumes an overall prevalence of 0.3 and its associated ICC is given by (ρ_b). It was assumed there was no effect of treatment.

Parameters (1000 simulations)				Descriptive statistics			
Total clusters	Cluster size	$\rho_l(\sigma_b^2)$	ICC	Mean (SD)	Median (IQR)	min, max	% negative
10	5	0.01 (0.01)	ρ_a	0.01 (0.11)	0.01 (-0.06, 0.08)	-0.22, 0.45	486(49%)
			ρ_k	0.01 (0.10)	0.00 (-0.06, 0.07)	-0.20, 0.43	485(49%)
			ρ_b	0.01 (0.11)	-0.00 (-0.08, 0.08)	-0.21, 0.44	505(51%)
		0.08 (0.09)	ρ_a	0.06 (0.12)	0.05 (-0.02, 0.14)	-0.20, 0.46	320(32%)
			ρ_k	0.06 (0.11)	0.05 (-0.02, 0.13)	-0.18, 0.43	321(32%)
			ρ_b	0.04 (0.11)	0.03 (-0.04, 0.11)	-0.20, 0.59	381(38%)
		0.16 (0.19)	ρ_a	0.13 (0.14)	0.12 (0.02, 0.22)	-0.18, 0.66	189(19%)
			ρ_k	0.12 (0.13)	0.11 (0.02, 0.21)	-0.16, 0.63	191(19%)
			ρ_b	0.09 (0.13)	0.09 (-0.01, 0.18)	-0.19, 0.56	263(26%)
		0.25 (0.33)	ρ_a	0.20 (0.14)	0.19 (0.09, 0.29)	-0.17, 0.64	77(8%)
			ρ_k	0.18 (0.13)	0.18 (0.09, 0.27)	-0.15, 0.61	79(8%)
			ρ_b	0.14 (0.14)	0.12 (0.03, 0.23)	-0.18, 0.65	153(15%)
		0.53 (1.13)	ρ_a	0.43 (0.16)	0.44 (0.32, 0.55)	-0.11, 0.83	6(1%)
			ρ_k	0.40 (0.16)	0.41 (0.30, 0.52)	-0.10, 0.81	6(1%)
			ρ_b	0.33 (0.18)	0.33 (0.21, 0.45)	-0.14, 0.80	33(3%)
	50	0.01 (0.01)	ρ_a	0.01 (0.01)	0.01 (-0.00, 0.02)	-0.02, 0.07	292(29%)
			ρ_k	0.01 (0.01)	0.01 (-0.00, 0.01)	-0.01, 0.06	290(29%)
			ρ_b	0.01 (0.01)	0.00 (-0.00, 0.01)	-0.02, 0.05	370(37%)
		0.08 (0.09)	ρ_a	0.07 (0.04)	0.07 (0.04, 0.09)	-0.01, 0.26	9(1%)
			ρ_k	0.06 (0.03)	0.06 (0.04, 0.08)	-0.01, 0.24	9(1%)
			ρ_b	0.05 (0.03)	0.04 (0.02, 0.07)	-0.01, 0.22	22(2%)
		0.16 (0.19)	ρ_a	0.13 (0.06)	0.12 (0.09, 0.17)	-0.01, 0.44	2($<$ 1%)
			ρ_k	0.12 (0.06)	0.11 (0.08, 0.15)	-0.01, 0.41	2($<$ 1%)
			ρ_b	0.09 (0.05)	0.08 (0.06, 0.12)	-0.01, 0.34	4($<$ 1%)
		0.25 (0.33)	ρ_a	0.21 (0.08)	0.21 (0.14, 0.26)	0.03, 0.45	0(0%)
			ρ_k	0.19 (0.08)	0.19 (0.13, 0.24)	0.03, 0.42	0(0%)
			ρ_b	0.15 (0.07)	0.14 (0.10, 0.19)	0.01, 0.38	0(0%)
		0.53 (1.13)	ρ_a	0.44 (0.12)	0.44 (0.36, 0.52)	0.07, 0.75	0(0%)
			ρ_k	0.41 (0.12)	0.41 (0.33, 0.49)	0.06, 0.73	0(0%)
			ρ_b	0.33 (0.12)	0.33 (0.24, 0.41)	0.03, 0.66	0(0%)

Notes. The data was generated for a 4-level ordinal outcome with proportions in each category of 0.20, 0.50, 0.20 and 0.1 for both treatment groups.

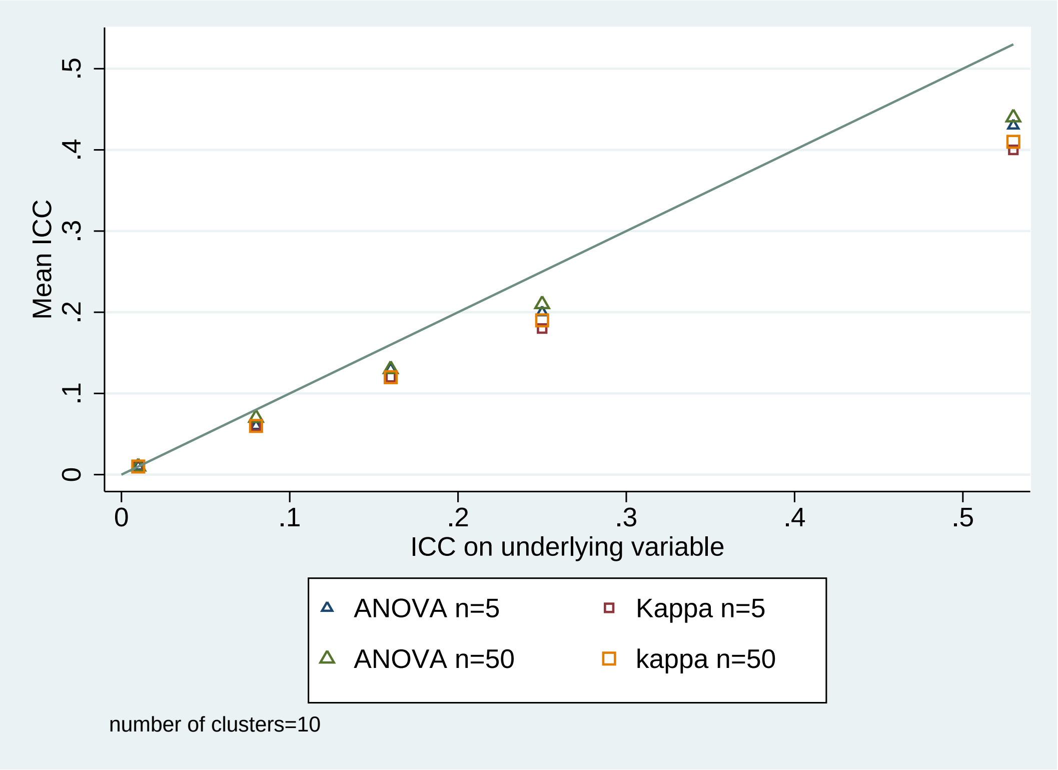


Figure 5.4: Relationship between ICC estimators for a 4-level ordinal variable with a total of 10 clusters, of size 5 or 50. It was assumed there was no effect of treatment.

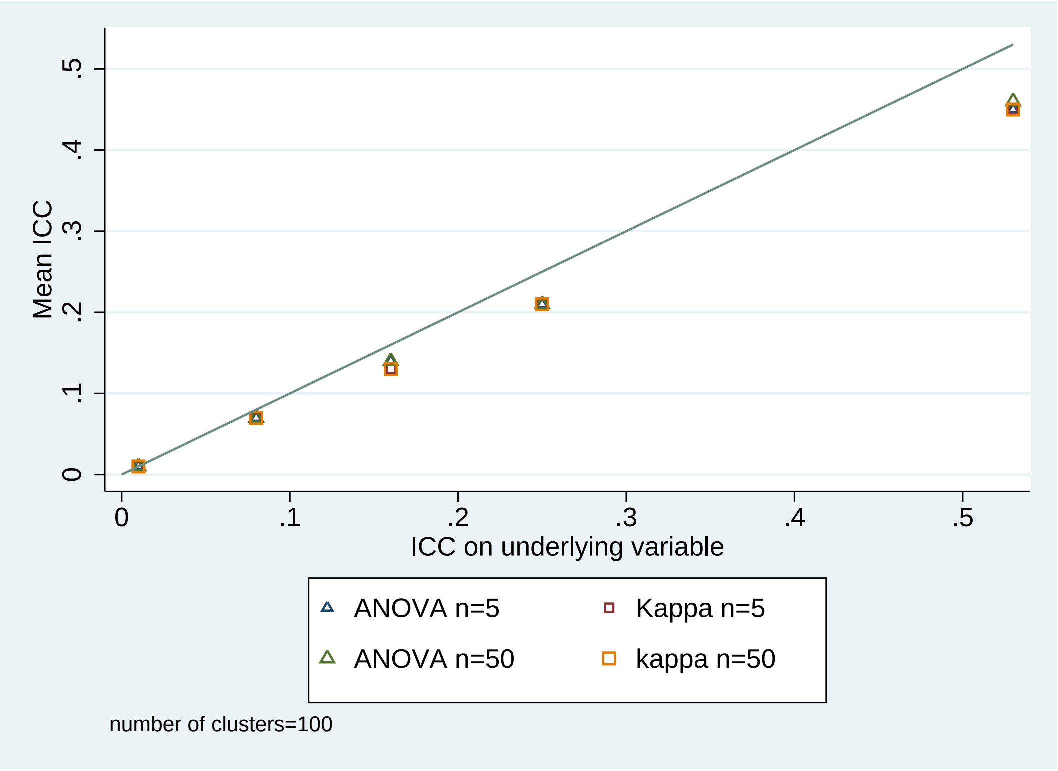


Figure 5.5: Relationship between ICC estimators for the 4-level ordinal variable with a total of 100 clusters, of size 5 or 50. It was assumed there was no effect of treatment.

Table 5.7: Results showing the relationship between the ICC on the underlying continuous latent variable (ρ_l) and ICCs calculated on the 4-level ordinal outcome (different proportions from those used in main simulations) ANOVA (ρ_a) and kappa (ρ_k), when the number of clusters is large. It was assumed there was no effect of treatment.

Parameters (1000 simulations)				Descriptive statistics			
Total clusters	Cluster size	$\rho_l(\sigma_b^2)$	ICC	Mean (SD)	Median (IQR)	min, max	% negative
100	5	0.01 (0.01)	ρ_a	0.01 (0.03)	0.01 (-0.01, 0.03)	-0.09, 0.16	395 (40%)
			ρ_k	0.01 (0.03)	0.01 (-0.01, 0.03)	-0.09, 0.16	395 (40%)
		0.08 (0.09)	ρ_a	0.07 (0.04)	0.07 (0.05, 0.10)	-0.03, 0.21	22 (2%)
			ρ_k	0.07 (0.04)	0.07 (0.05, 0.10)	-0.03, 0.21	21 (2%)
		0.16 (0.19)	ρ_a	0.14 (0.04)	0.14 (0.11, 0.17)	0.00, 0.29	0(0%)
			ρ_k	0.14 (0.04)	0.14 (0.11, 0.17)	0.00, 0.28	0 (0%)
		0.25 (0.33)	ρ_a	0.22 (0.05)	0.22 (0.19, 0.25)	0.08, 0.36	0(0%)
			ρ_k	0.21 (0.04)	0.22 (0.19, 0.24)	0.08, 0.35	0 (0%)
		0.53 (1.13)	ρ_a	0.47 (0.05)	0.47 (0.44, 0.50)	0.30, 0.63	0(0%)
			ρ_k	0.46 (0.05)	0.46 (0.43, 0.49)	0.30, 0.62	0(0%)
	50	0.01 (0.01)	ρ_a	0.01 (0.00)	0.01 (0.01, 0.01)	-0.00, 0.02	11 (1%)
			ρ_k	0.01 (0.00)	0.01 (0.01, 0.01)	-0.00, 0.02	11 (1%)
		0.08 (0.09)	ρ_a	0.07 (0.01)	0.07 (0.06, 0.08)	0.04, 0.12	0(0%)
			ρ_k	0.07 (0.01)	0.07 (0.06, 0.08)	0.04, 0.11	0(0%)
		0.16 (0.19)	ρ_a	0.14 (0.02)	0.14 (0.13, 0.15)	0.08, 0.20	0(0%)
			ρ_k	0.14 (0.02)	0.14 (0.12, 0.15)	0.08, 0.20	0(0%)
		0.25 (0.33)	ρ_a	0.22 (0.02)	0.22 (0.20, 0.23)	0.14, 0.29	0(0%)
			ρ_k	0.21 (0.02)	0.21 (0.20, 0.23)	0.14, 0.29	0(0%)
		0.53 (1.13)	ρ_a	0.47 (0.03)	0.47 (0.44, 0.49)	0.37, 0.59	0(0%)
			ρ_k	0.46 (0.03)	0.46 (0.44, 0.48)	0.36, 0.59	0(0%)

Notes. The data was generated for a 4-level ordinal outcome with proportions in each category of 0.10, 0.30, 0.40 and 0.20 for both treatment groups.

5.8.2 Power of a trial designed using the design effect with each ICC estimate

Tables 5.3, 5.4 and 5.5 show the ANOVA ICC and kappa-type ICCs to be asymptotically equivalent in samples with a large number of clusters. For this reason it was not necessary to assess the empirical power of both estimates in the design effect. I chose to only consider use of the ANOVA ICC as it was seen to be slightly larger than the kappa-type ICC in small samples and hence would provide a slightly larger sample size estimate in these situations.

For fixed values of the ICC on the latent continuous variable (ρ_l) 1000 datasets were generated for clusters of size 5, 10 and 50 under two estimates of treatment effect for the 3, 4, and 5 level ordinal outcomes. The size of each dataset was calculated by using Whitehead's formula for individually randomised trials multiplied by the design effect. The design effect used the estimate of the ANOVA ICC that corresponds to the ICC on the latent continuous variable used to generate the data for that scenario, identified in the previous section.

Tables 5.8, 5.9, and 5.10 summarise the empirical power when using the ANOVA ICC estimate in the design effect. The evaluation of power is only sensible if the Type I error rate for the test statistic is controlled at 5%. For the Wald test the Type I error is valid when there are at least 40 clusters per arm. For situations with less than 40 clusters the Type I error is inflated and therefore the expected power will also be inflated. This phenomenon can be seen in the Tables 5.8, 5.9, and 5.10. In chapter three the eleven trials identified with ordinal outcomes often included only a small number of clusters. The types of clusters in these trials were health care practices or schools which would suggest that the number of clusters available would not necessarily be restricted. Only one trial provided a sample size calculation and therefore the small numbers of clusters used in these trials may not be typical of future, appropriately designed, cluster randomised trials with ordinal outcomes. For these reasons I did not evaluate the use of small sample corrections in my simulations and I am discounting those situations where the number of clusters is less than 40. Future work is required to generalise my results to situations with a small number of clusters.

For the 3-level ordinal outcome all calculated empirical powers are larger than 90% , the additional power over 90% ranges from 0.9% to 4.20%. The standard error of the simulated power ranges from 0.73 to 0.91.

For the 4-level outcome the empirical powers are all above or very close to the expected 90%, the additional power over 90% ranges from 0% to 2.8% . The standard error of the simulated power ranges from 0.82 to 0.95. For the 5-level outcome we see some situations where the number of clusters is greater than 40 but 90% power is not achieved. However, the difference in power ranges from -1.4% to 2.7% which is still within the limit of 2% I specified for empirical power to be considered sufficiently close to nominal power. The sample size using the ANOVA estimate produces an adequately powered trial.

Table 5.8: Empirical power when using the ANOVA estimate of the ICC for the 3-level ordinal outcome in the design effect for sample size calculation. For each combination of cluster size and ICC: 1000 datasets generated with a 90% target level of power

Fixed Design Parameters			log odds=0.493		log odds=0.887	
			C	**Empirical power**	C	**Empirical power**
Cluster size	$\rho_a(\rho_l)$	Design effect		(SE, $\hat{\theta}-\theta$)		(SE, $\hat{\theta}-\theta$)
5	0.01 (0.01)	1.04	80	92.7 (0.82, 2.7)	24	92.8 (0.82, 2.8)
	0.06 (0.08)	1.24	95	90.2 (0.86, 0.2)	29	92.1 (0.85, 2.1)
	0.11 (0.16)	1.44	111	91.7 (0.87, 1.7)	33	92.0 (0.86, 2.0)
	0.18 (0.25)	1.72	132	92.4 (0.84, 2.4)	39	91.1 (0.90, 1.1)
	0.38 (0.53)	2.52	193	91.8 (0.87, 1.8)	57	90.9 (0.91, 0.9)
10	0.01 (0.01)	1.09	42	92.7 (0.82, 2.7)	13	92.4 (0.84, 2.4)
	0.06 (0.08)	1.54	59	94.2 (0.74, 4.2)	18	91.4 (0.89, 1.4)
	0.11 (0.16)	1.99	77	91.9 (0.86, 1.9)	23	90.4 (0.93, 0.4)
	0.18 (0.25)	2.62	101	92.2 (0.85, 2.2)	30	90.8 (0.91, 0.8)
	0.38 (0.53)	4.42	169	91.8 (0.87, 1.8)	50	91.7 (0.87, 1.7)
50	0.01 (0.01)	1.49	12	95.9 (0.63, 5.9)	4	97.4 (0.50, 7.4)
	0.06 (0.08)	3.94	31	94.2 (0.74, 4.2)	9	92.6 (0.83, 2.6)
	0.11 (0.16)	6.39	49	92.1 (0.85, 2.1)	15	92.5 (0.83, 2.5)
	0.18 (0.25)	9.82	76	92.8 (0.82, 2.8)	23	93.2 (0.80, 3.2)
	0.38 (0.53)	19.62	150	92.2 (0.85, 2.2)	45	92.3 (0.84, 2.3)

Notes. Assuming a 3-level ordinal outcome with proportions 0.20, 0.70, and 0.1 in the control group and C clusters per group. Empirical power, $\hat{\theta}$ is calculated as the proportion of fitted probit models with a treatment effect significant at the 5% level. $\hat{\theta}-\theta$ represents the absolute difference between empirical and nominal power

The ICC on the underlying latent continuous variable was shown to be larger than the ANOVA ICC and therefore will produce sample sizes that produce more power than required, as can be seen in Table 5.11 for a 4-level outcome. All calculated empirical powers were larger than 90%, the additional power over 90% ranges from 1.6% to 5.3%. The standard error of the simulated power ranges from 0.67 to 0.88.

Table 5.9: Empirical power when using the ANOVA estimate of the ICC for the 4-level ordinal outcome in the design effect for sample size calculation. For each combination of cluster size and ICC: 1000 datasets generated with a 90% target level of power

Fixed Design Parameters			log odds=0.493		log odds=0.887	
			C	**Empirical power**	C	**Empirical power**
Cluster size	$\rho_a(\rho_l)$	Design effect		(SE, $\hat{\theta} - \theta$)		(SE, $\hat{\theta} - \theta$)
5	0.01 (0.01)	1.04	64	90.2 (0.94, 0.2)	20	89.5 (0.97, -0.5)
	0.07 (0.08)	1.28	78	92.0 (0.86, 2.0)	25	91.2 (0.90, 1.2)
	0.14 (0.16)	1.56	95	90.8 (0.91, 0.8)	30	90.0 (0.95, 0.0)
	0.21 (0.25)	1.84	112	90.1 (0.94, 0.1)	35	89.5 (0.97, -0.5)
	0.46 (0.53)	2.84	173	90.8 (0.91, 0.8)	54	91.3 (0.89, 1.3)
10	0.01 (0.01)	1.09	34	89.9 (0.95, -0.1)	11	91.5 (0.88, 1.5)
	0.07 (0.08)	1.63	50	90.0 (0.95, 0.0)	16	92.4 (0.84, 2.4)
	0.14 (0.16)	2.26	69	91.5 (0.88, 1.5)	22	93.5 (0.78, 3.5)
	0.21 (0.25)	2.89	88	92.1 (0.85, 2.1)	28	91.2 (0.90, 1.2)
	0.46 (0.53)	5.14	157	91.5 (0.88, 1.5)	49	91.6 (0.88, 1.6)
50	0.01 (0.01)	1.49	10	95.5 (0.67, 5.5)	3	94.8 (0.70, 4.8)
	0.07 (0.08)	4.43	27	90.9 (0.91, 0.9)	9	94.0 (0.75, 4.0)
	0.14 (0.16)	7.86	48	90.2 (0.94, 0.2)	15	90.9 (0.91, 0.9)
	0.21 (0.25)	11.29	69	91.9 (0.86, 1.9)	22	91.4 (0.89, 1.4)
	0.46 (0.53)	23.54	144	92.0 (0.86, 2.0)	45	92.8 (0.82, 2.8)

Notes. Assuming a 4-level ordinal outcome with proportions 0.20, 0.50, 0.20 and 0.1 in the control group and C clusters per group. Empirical power, $\hat{\theta}$ is calculated as the proportion of fitted probit models with a treatment effect significant at the 5% level. $\hat{\theta} - \theta$ represents the absolute difference between empirical and nominal power

Table 5.10: Empirical power when using the ANOVA estimate of the ICC for the 5-level ordinal outcome in the design effect for sample size calculation. For each combination of cluster size and ICC: 1000 datasets generated with a 90% target level of power

Fixed Design Parameters			log odds=0.493		log odds=0.887	
			C	**Empirical power**	C	**Empirical power**
Cluster size	$\rho_a(\rho_l)$	Design effect		(SE, $\hat{\theta}-\theta$)		(SE, $\hat{\theta}-\theta$)
5	0.01 (0.01)	1.04	58	90.1 (0.94, 0.1)	18	89.2 (0.98, -0.8)
	0.08 (0.08)	1.32	73	91.8 (0.87, 1.8)	23	92.1 (0.85, 2.1)
	0.14 (0.16)	1.56	86	88.6 (1.00, -1.4)	27	90.1 (0.94, 0.1)
	0.23 (0.25)	1.92	106	91.1 (0.90, 1.1)	34	90.4 (0.93, 0.4)
	0.49 (0.53)	2.96	163	89.8 (0.96, -0.2)	51	88.8 (0.99, -1.2)
10	0.01 (0.01)	1.09	30	88.6 (1.00, -1.4)	10	91.6 (0.88, 1.6)
	0.08 (0.08)	1.72	48	92.3 (0.84, 2.3)	15	91.4 (0.89, 1.4)
	0.14 (0.16)	2.26	63	89.6 (0.97, -0.4)	20	91.5 (0.88, 1.5)
	0.23 (0.25)	3.07	85	90.7 (0.92, 0.7)	27	91.5 (0.88, 1.5)
	0.49 (0.53)	5.41	149	89.9 (0.95, -0.1)	47	91.6 (0.88, 1.6)
50	0.01 (0.01)	1.49	9	93.2 (0.80, 3.2)	3	96.5 (0.58, 6.5)
	0.08 (0.08)	4.92	28	93.2 (0.79, 3.2)	9	94.0 (0.75, 4.0)
	0.14 (0.16)	7.86	44	88.6 (1.00, -1.4)	14	91.7 (0.87, 1.7)
	0.23 (0.25)	12.27	68	91.1 (0.90, 1.1)	22	91.3 (0.89, 1.3)
	0.49 (0.53)	25.01	138	92.4 (0.84, 2.4)	44	92.7 (0.82, 2.7)

Notes. Assuming a 5-level ordinal outcome with proportions 0.20, 0.20, 0.30, 0.20 and 0.1 in the control group and C clusters per group. Empirical power, $\hat{\theta}$ is calculated as the proportion of fitted probit models with a treatment effect significant at the 5% level. $\hat{\theta}-\theta$ represents the absolute difference between empirical and nominal power

Table 5.11: Empirical power when using the ICC estimate for the underlying continuous variable for the 4-level ordinal outcome in the design effect for sample size calculation. For each combination of cluster size and ICC: 1000 datasets generated with a 90% target level of power

Fixed Design Parameters			log odds=0.493		log odds=0.887	
			C	**Empirical power**	C	**Empirical power**
Cluster size	ρ_l	Design effect		(SE,$\hat{\theta}-\theta$)		(SE,$\hat{\theta}-\theta$)
5	0.01	1.04	64	91.7 (0.87, 1.7)	20	89.4 (0.97, -0.6)
	0.08	1.32	81	91.6 (0.88, 1.6)	25	90.3 (0.94, 0.3)
	0.16	1.64	100	93.1 (0.80, 3.1)	31	93.2 (0.80, 3.2)
	0.25	2.00	122	92.0 (0.86, 2.0)	38	93.8 (0.76, 3.8)
	0.53	3.12	190	92.4 (0.84, 2.4)	59	94.1 (0.75, 4.1)
10	0.01	1.09	34	91.4 (0.89, 1.4)	11	91.6 (0.88, 1.6)
	0.08	1.72	53	92.6 (0.83, 2.6)	17	92.7 (0.82, 2.7)
	0.16	2.44	75	93.8 (0.76, 3.8)	23	94.2 (0.74, 4.2)
	0.25	3.25	99	94.3 (0.73, 4.3)	31	93.1 (0.80, 3.1)
	0.53	5.77	176	93.5 (0.78, 3.5)	55	93.9 (0.76, 3.9)
50	0.01	1.49	10	96.4 (0.59, 6.4)	3	95.7 (0.64, 5.7)
	0.08	4.92	30	94.5 (0.72, 4.5)	10	94.7 (0.71, 4.7)
	0.16	8.84	54	95.3 (0.67, 5.3)	17	93.8 (0.76, 3.8)
	0.25	13.25	81	93.8 (0.76, 3.8)	25	95.4 (0.66, 5.4)
	0.53	26.97	164	94.3 (0.73, 4.3)	51	93.4 (0.79, 3.4)

Notes. Assuming a 4-level ordinal outcome with proportions 0.20, 0.50, 0.20 and 0.1 in the control group and C clusters per group. Empirical power, $\hat{\theta}$ is calculated as the proportion of fitted probit models with a treatment effect significant at the 5% level. $\hat{\theta}-\theta$ represents the absolute difference between empirical and nominal power

5.8.3 Power of a trial designed using the design effect under minor deviations from the proportional odds assumption

For the four-level outcome Table 5.12 shows the effect on power when there is a minor deviation from the assumption of proportional odds, in this case the treatment effect for the first category was assumed to be different (smaller) than for the other categories. Power was calculated using Whitehead's formula multiplied by the design effect with the treatment effect based on either the smallest log-odds, 0.35 or the average log-odds of 0.45.

Use of the smallest log-odds in the power calculation resulted in substantially larger sample sizes than the average log-odds. The smallest log-odds produced overly conservative designs, the additional power over 90% ranged from 7.5% to 8.9% and the standard error of the power estimates ranged from 0.33 to 0.49. Use of the average log-odds resulted in trials which were slightly underpowered with the difference in power from the expected 90% ranging from -2.6% to 0.3%. The standard error of the power estimate ranges from 0.94 to 1.05.

Table 5.12: Empirical power when using the ANOVA estimate of the ICC for the ordinal outcome in the design effect for sample size calculation when there is a minor violation of the proportional odds assumption. For each combination of cluster size and ICC: 1000 datasets were generated with a 90% target level of power

Fixed Design Parameters			log odds=0.35		log odds=0.45	
				Empirical power		**Empirical power**
Cluster size	$\rho_a(\rho_l)$	Design effect	C	(SE, $\hat{\theta} - \theta$)	C	(SE, $\hat{\theta} - \theta$)
5	0.01 (0.01)	1.04	127	97.6 (0.48, 7.6)	77	88.8 (0.99, -1.2)
	0.07 (0.08)	1.28	157	98.7 (0.36, 8.7)	95	88.5 (1.00, -1.5)
	0.14 (0.16)	1.56	191	98.6 (0.37, 8.6)	116	89.4 (0.97, -0.6)
	0.21 (0.25)	1.84	225	98.9 (0.33, 8.9)	136	87.4 (1.05, -2.6)
	0.46 (0.53)	2.84	347	98.2 (0.42, 8.2)	210	88.2 (1.02, -1.8)
10	0.01 (0.01)	1.09	67	98.8 (0.34, 8.8)	41	87.8 (1.04, -2.2)
	0.07 (0.08)	1.63	100	98.7 (0.36, 8.7)	61	88.8 (0.99, -1.2)
	0.14 (0.16)	2.26	138	98.4 (0.40, 8.4)	84	89.8 (0.96, -0.2)
	0.21 (0.25)	2.89	177	98.4 (0.40, 8.4)	107	89.5 (0.97, -0.5)
	0.46 (0.53)	5.14	314	97.5 (0.49, 7.5)	190	89.4 (0.97, -0.6)
50	0.01 (0.01)	1.49	19	98.8 (0.34, 8.8)	11	91.7 (0.87, 1.7)
	0.07 (0.08)	4.43	55	98.2 (0.42, 8.2)	33	90.2 (0.94, 0.2)
	0.14 (0.16)	7.86	96	98.4 (0.40, 8.4)	59	89.7 (0.96, -0.3)
	0.21 (0.25)	11.29	138	98.2 (0.42, 8.2)	84	90.3 (0.94, 0.3)
	0.46 (0.53)	23.54	288	98.7 (0.36, 8.7)	174	87.9 (1.03, -2.1)

Notes. Assuming a 4-level ordinal outcome and C clusters per group. Empirical power, $\hat{\theta}$ is calculated as the proportion of fitted probit models with a treatment effect significant at the 5% level. $\hat{\theta} - \theta$ represents the absolute difference between empirical and nominal power

5.8.4 Power of a trial designed using the design effect under alternative analysis method

The data generation model assumed a probit regression model for simplicity. However, use of the logit link is a more popular analysis method. Using the ANOVA estimate of the ICC in the design effect the simulations for the four-level outcome variable were repeated with analysis via a logit link rather than probit. The results are summarised in Table 5.13. All empirical powers were above or very close to the calculated 90%. The difference in empirical to expected power ranged from -1.6% to 2.4% . The standard error of calculated power ranged from 0.84 to 1.01.

Table 5.13: Empirical power when using the ANOVA estimate of the ICC for the ordinal outcome in the design effect for sample size calculation followed by an analysis with a logit model. For each combination of cluster size and ICC: 1000 datasets generated with a 90% target level of power

Fixed Design Parameters			log odds=0.493		log odds=0.887	
			C	**Empirical power**	C	**Empirical power**
Cluster size	$\rho_a(\rho_l)$	Design effect		(SE, $\hat{\theta}-\theta$)		(SE, $\hat{\theta}-\theta$)
5	0.01 (0.01)	1.04	62	88.4 (1.01, -1.6)	20	89.7 (0.96, -0.3)
	0.07 (0.08)	1.28	76	90.8 (0.91, 0.8)	25	91.9 (0.86, 1.9)
	0.14 (0.16)	1.56	93	90.9 (0.91, 0.9)	30	91.0 (0.91, 1.0)
	0.21 (0.25)	1.84	109	90.6 (0.92, 0.6)	35	92.2 (0.85, 2.2)
	0.46 (0.53)	2.84	169	90.2 (0.94, 0.2)	54	89.8 (0.96, -0.2)
10	0.01 (0.01)	1.09	33	91.9 (0.86, 1.9)	11	93.9 (0.76, 3.9)
	0.07 (0.08)	1.63	49	91.6 (0.88, 1.6)	16	92.1 (0.85, 2.1)
	0.14 (0.16)	2.26	67	90.6 (0.92, 0.6)	22	90.9 (0.91, 0.9)
	0.21 (0.25)	2.89	86	92.1 (0.85, 2.1)	28	91.2 (0.90, 1.2)
	0.46 (0.53)	5.14	153	88.8 (0.99, -1.2)	49	90.6 (0.92, 0.6)
50	0.01 (0.01)	1.49	9	92.8 (0.82, 2.8)	3	92.7 (0.82, 2.7)
	0.07 (0.08)	4.43	27	91.5 (0.88, 1.5)	9	92.2 (0.85, 2.2)
	0.14 (0.16)	7.86	47	90.8 (0.91, 0.8)	15	92.1 (0.85, 2.1)
	0.21 (0.25)	11.29	67	89.8 (0.96, -0.2)	22	92.5 (0.83, 2.5)
	0.46 (0.53)	23.54	140	90.7 (0.92, 0.7)	45	92.4 (0.84, 2.4)

Notes. Assuming a 4-level ordinal outcome with proportions 0.20, 0.50, 0.20 and 0.1 in the control group and C clusters per group. Empirical power, $\hat{\theta}$ is calculated as the proportion of fitted probit models with a treatment effect significant at the 5% level. $\hat{\theta}-\theta$ represents the absolute difference between empirical and nominal power

5.9 Discussion

5.9.1 Main findings

To explore the relationship between ICC estimators

With a small number of clusters results showed that the ICC on the latent response tended to be largest followed by the ANOVA and kappa-type ICCs. As the level of clustering increased so too did the difference between the ANOVA and Kappa-type estimates with the ANOVA estimate being consistently larger. With a large number of clusters results also showed that the ICC on the latent response was largest, however, the ANOVA and kappa-type ICCs were almost identical. This was expected as it has been shown that these two estimators are asymptotically equivalent as the number of clusters increases.[125]

For each scenario investigated as the number of ordinal categories increased so too did the estimated ANOVA and kappa-type ICCs. This was expected because as the number of ordinal categories increases the variable more resembles a continuous variable and therefore we would expect the ICC to tend towards the ICC calculated on an assumed underlying continuous variable.

By comparing two possible patterns in the expected proportions across categories for a 4-level ordinal outcome the observed ANOVA ICC was shown to depend upon both the number of categories but also the proportions observed in each category. The two patterns of proportions explored both had a fairly even spread in the proportions expected in each category and the difference in observed ANOVA ICCs for the two categorisations was small. Larger deviations from an even spread of proportions across categories might have a more substantial impact on the ANOVA ICC.

To determine which ICC results in an adequately powered trial

The use of the ANOVA ICC estimate in the design effect resulted in adequately powered trials. The empirical power was within 2% for 3-, 4- and 5-level outcomes. The efficiency of Whiteheads method increases with the number of ordinal categories and is most efficient when the proportions in each ordinal category are evenly spread. However, once you go beyond five categories further efficiency gains are marginal. In my simulations as the number of categories increased the proportions in each ordinal category subsequently became more evenly spread. The ANOVA ICC estimate appeared to be conservative when the spread was less even. Hence, I saw a slight decrease in power as the number of categories increased but became more evenly spread.

The largest differences between nominal and empirical power were in situations with a small number of clusters, which was expected due to the inflated Type I error rates for these situations.

Use of the ICC of the underlying latent variable in the design effect resulted in overly conservative sample sizes, having an additional 1.6% to 5.3% power over the required 90%.

To determine the effect of non-proportional odds on power

I considered the situation where a minor deviation to proportional odds occurred. For the 4-level ordinal outcome, there are three possible ways to dichotomise the outcome to calculate the log-odds of being in category q or better. I assumed that two of these were the same and one was slightly lower, but all indicated a beneficial treatment effect. I deemed this a minor deviation to the proportional odds assumptions. For this situation a sample size based upon an average estimate of log-odds was shown to result in a marginally underpowered trial. Power calculations based on the smallest log-odds were overly conservative.

There are alternative situations which might also be classified as minor deviations from proportional odds for example the situation where one odds ratio is slightly larger than the other two. Use of the average log odds in the design effect for these situations was not explored and may not necessarily result in a marginally underpowered trial.

The use of the design effect sample size approach is not recommended for situations in which major deviations to proportional odds occur such as all the log odds being very different, or some log-odds showing inconsistency around the effect of treatment.

To determine whether similar conclusions are made with alternative analysis methods

Sample size using the ANOVA ICC and analysis via a logit link also resulted in adequately powered trials, analysis with the probit link was slightly more conservative. Due to the similarity in shape of the logistic and normal distributions the fit of a model using either of these links should be similar and therefore this result was expected.

Despite the difference in interpretation and magnitude between a random effects and GEE model the significance of the treatment effect is likely to be similar.[130] Therefore I expect the results of my simulation to be applicable if a GEE model was used to analyse the data. As Stata cannot be used to fit GEE models for ordinal outcomes I did not test this within my simulations. Given more time this could have been done using an alternative software package, such as SAS.

5.9.2 Strengths and limitations

There are several strengths to the research described within this chapter. The simulation study was planned and reported following the most commonly used guidance for reporting simulation studies and is therefore described in sufficient detail for this work to be fully reproduced by others. The scenarios chosen for the simulations are largely reflective of the characteristics of real life cluster randomised trials that have used ordinal outcomes, except that designs with a small number of clusters were excluded. Appropriate analysis methods when the number of clusters is small require some adjustment to account for the inflated Type I error and these methods have not been well established for ordinal outcomes. To explore different analysis methods for clustered ordinal outcomes when the number of clusters is small was beyond the scope of this thesis. Some approaches that might be appropriate are discussed in the final chapter of this thesis alongside some practical guidance with regard to sample size calculation in general for clustered ordinal outcomes.

I explored the relationship between ICC estimators under the simplifying assumption of no treatment effect. This assumption implies that the ICC is the same in both treatment arms and thus using these ICC estimates in sample size calculations would be equivalent to using an ICC estimate based on control data alone. However, like binary data I identified from the simulation study that the value of the ANOVA ICC for the ordinal outcome was dependent upon the proportions observed in each ordinal category. This would suggest that in the presence of a treatment effect the ANOVA ICC estimates would be different across treatment arms and in these situations it is more appropriate to use a pooled ANOVA ICC estimate in the sample size calculation.[147] To assess the implication that my assumption of no treatment effect may have had on my simulation results I re-ran one of the simulation scenarios from Table 5.4 this time assuming the same log-odds values of 0.493 and 0.887 that were used in the later sample size calculations (4-level outcome, 100 clusters of size 50 with an underlying latent ICC of 0.53). The results showed that for these treatment effects there was only a minor difference in the observed ANOVA ICCs across the treatment groups and the pooled ANOVA ICC estimate was the same as, or very close to, the ANOVA ICC estimate calculated when assuming no treatment effect. Therefore my initial assumption is unlikely to have affected my results significantly.

My simulations generated clustered ordinal data using the latent variable approach. However, as discussed in Section 5.3 this is not the only method that might be used for data generation and it is unclear as to whether a different method would have any impact upon my findings.

The ANOVA ICC was shown to depend upon both the number of ordinal categories and the proportions observed in each category. In this research I have considered scenarios where there is a fairly even spread in the proportions expected in each category, this is the situation for which Whitehead states his method is most efficient. In situations where this is not the case the estimate of the ANOVA ICC may be less similar to the latent variable ICC and therefore the performance of the ANOVA ICC in the design effect may be affected.

5.9.3 Comparison with other work

Gao[125] investigated analysis strategies for clustered ordinal data, with a focus on adjusted Cochran-Armitage tests. She considered the use of both the kappa-type ICC and ANOVA ICCs in the test statistic, and showed that the Cochran-Armitage test had greatest power with the kappa-type ICC estimate. Gao saw similar results to mine in that the ANOVA and kappa-type ICC estimates were asymptotically equivalent as the number of clusters increased. Her work used datasets generated using marginal models and hence the relationship with the ICC on the underlying continuous variable was not included.

There has been no other work which has considered the power of the design effect method for sample size calculations with ordinal outcomes analysed via a random effects model.

5.9.4 Implications

In this research I have identified that using the ANOVA estimate of the ICC, calculated by assigning equally spaced numerical values to the ordinal outcome, within the design effect for sample size calculations results in adequately powered trials. However, before implementing this method the researcher must ask themselves two questions: Is a good estimate of the ICC available? and is the assumption of proportional odds reasonable? Both of these elements will impact upon the power of the trial. With only minor deviations to the proportional odds assumptions the use of the design effect and an analysis which assumes proportional odds may result in only marginal over or under powering. Major deviations are likely to require alternative analysis methods and sample size would be best estimated through simulation. The sensitivity of the sample size calculation to the range of plausible ICCs should be examined. If no reasonable estimates are available researchers might consider a sample size calculation and analysis based upon the dichotomised version of the outcome, for which ICC estimates may be more readily available.

With binary outcomes the overall prevalence of the observed endpoint is a single proportion. Therefore in simulation studies of binary outcomes it is straightforward to examine several proportions to gain insight into emerging patterns as the prevalence increases for example 10%, 30% and 70%. For ordinal outcomes the situation is less straightforward. The number of ways that the proportions can occur across the ordinal categories is numerous and it is not easy to try and systematically explore all possible combinations or identify patterns. In Whitehead's work for individually randomised trials he considered two patterns. The first where there was an even spread across categories and the second where one category was dominant. He showed that an even spread across categories was a more efficient design. In this research I have focused on scenarios where there is generally an even spread of the proportions expected in each ordinal category. For situations in which there is a less even spread, perhaps one or more categories are dominating, the ordinal outcome looks less like a continuous outcome and hence the ANOVA estimate of the ICC may be less similar to the ICC on the latent variable. Therefore using the ANOVA estimate of the ICC in the design effect for these situations may not perform as well. Simulation studies should be conducted in these situations to confirm the sample size required.

This work has provided some guidance for sample size calculation for those designing trials with ordinal outcomes. However, in order to move forward estimates of the required ANOVA ICC are needed. I would therefore recommend that authors reporting results with ordinal outcomes report the ANOVA ICC and also provide the estimates of each log-odds in order that the reader may evaluate the assumption of proportional odds.

My work will impact those working in fields where ordinal outcomes are prevalent. However, there are still many other design aspects of cluster randomised trials for which the corresponding sample size development is still lacking. These areas are identified and discussed in the next chapter. In the final chapter I bring all my research together to discuss the future of sample size calculations for CRTs and formulate clear practical guidance for sample size calculations with ordinal outcomes.

Chapter 6

Remaining methodological gaps in sample size methods for CRTs

In Chapter Two I reviewed the published literature on sample size calculation for cluster randomised trials up to the year 2011, I identified 85 papers reporting sample size methods. However, of these only two were applicable to ordinal outcomes, and the development of one of these methods became the focus of my research. Therefore, there is much scope for developing some of the methods reported in the remaining papers or identifying new areas for development where there are currently no published methods available.

In this chapter I return my attention to my review of sample size methods. I provide an update to the review which additionally includes methods published between 2011 and 2015. The aim of this chapter is to describe where methods are lacking or need further development to encourage or enable routine use for a given design, outcome, or analysis.

6.1 Methods

The methods of the review have been described in detail within Chapter Two. In summary the review was conducted using electronic online databases, a personal collection of 41 articles on sample size in CRT's provided by SE, key text books on cluster randomised trials,[6–8, 32] and special issue journals on cluster randomised trials.[61–64] The electronic search was conducted using the online databases PubMed and Web of Science. The initial search was conducted on the 31st March 2011. Given the time that has elapsed since conducting the review the online databases were searched again in August 2015 and the additional methods identified are included within this chapter.

An article or method was included in the review if it provided a method of sample size calculation for cluster randomised trials, via a formula, simulation or other approach. The first paper to report a particular methodology was included in the review; subsequent papers describing the same approach were excluded. The two electronic databases searched, PubMed and Web of Science contained articles from 1946 and 1970 onwards respectively, no further date restrictions were applied.

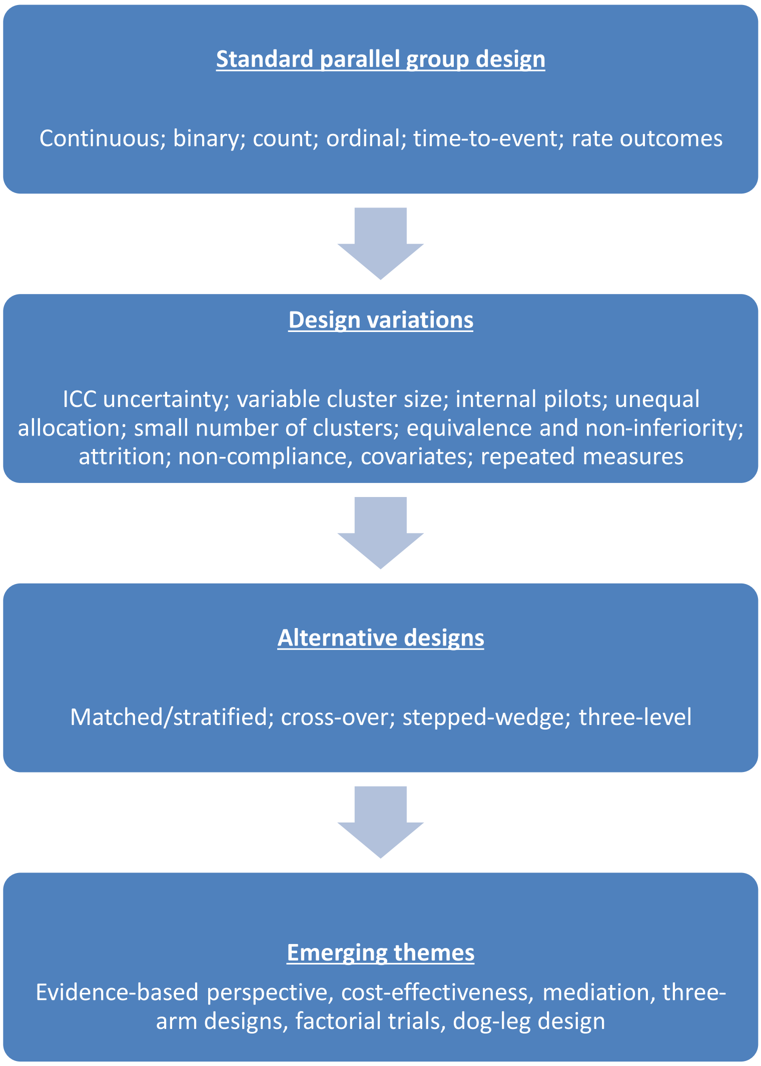
6.2 Results

The structure of this results section is summarised in Figure 6.1. This structure was chosen to mirror that of the published version of my initial review, which summarised sample size methods available for CRTs for a given design, outcome and analysis.[60] In contrast to the published review the focus of this chapter is to highlight where there are gaps in the methodology.

The results start with the simplest and most common design for a cluster randomised trial; the two-arm, completely randomised, parallel-group trial with fixed cluster sizes.

The methodology gaps for variations or adaptations to the simple design are discussed next. Variations to the design include: variability in cluster size or attrition; uncertainty around the ICC; unequal allocation; and inclusion of baseline measurements or repeated measures.

The third section discusses alternative design choices such as the cross-over, stepped-wedge, matched and three-level designs. The section concludes with a brief description of emerging topics that were identified in the 2015 update to the review.



Standard parallel group design

Continuous; binary; count; ordinal; time-to-event; rate outcomes

Design variations

ICC uncertainty; variable cluster size; internal pilots; unequal allocation; small number of clusters; equivalence and non-inferiority; attrition; non-compliance, covariates; repeated measures

Alternative designs

Matched/stratified; cross-over; stepped-wedge; three-level

Emerging themes

Evidence-based perspective, cost-effectiveness, mediation, three-arm designs, factorial trials, dog-leg design

Figure 6.1: Sample size methodology and gaps in the literature for CRTs: Flow diagram describing the order in which these are presented within the chapter

6.2.1 Standard parallel-group, two-arm design

Table 6.1 summarises the methodology available by outcome type and analysis method for the standard parallel-group, completely randomised design with fixed cluster sizes. I consider each of the outcome types in turn.

Continuous and binary outcomes

The surge in the development of sample size methods for cluster randomised trials started with methods for binary and continuous outcomes analysed at the cluster-level, described in the seminal paper by Cornfield in 1978[31] and the 1981 paper by Donner, Birkett and Buck.[52] Cluster-level summaries can often be considered continuous regardless of the nature of the variable at the individual level. Therefore these methods are often applicable to alternative types of individual-level outcomes. With the development of statistical software, individual-level analyses have now become a more popular analysis and Shih has demonstrated that the simple design effect described by Donner, Birkett and Buck for binary and continuous outcomes can be used to calculate the sample size when the planned analysis is by GEE.[83] The reason for this is that for continuous outcomes with equal cluster sizes, as assumed here, the cluster-level and individual-level analyses (population-averaged or cluster-specific approaches) are equivalent. However, in many situations the assumption of equal sized clusters is not realistic.

For continuous outcomes it is most common to assume the variable is normally distributed. Rosner and Glynn[161] have presented the only sample size approach for non-normally distributed outcomes analysed with a clustered version of the Wilcoxon test, but their method requires a large number of calculations and associated SAS macros for implementation. There is therefore, scope to develop simpler or alternative methods for non-normally distributed continuous outcomes.

In a cluster randomised design the sample size calculation requires estimates of both the number of clusters and the cluster size. In the majority of situations there will be constraints on the values of either of these parameters. For example if the cluster is a GP practice there will be a finite number available within a specified geographical location. However, in a minority of situations the values of these may be unconstrained and several different combinations may lead to equally

Outcome measure	Analysis	Reference
Continuous	Cluster-level	20, 52, 104, 159, 160
	Adjusted test	161 [162]
	Mixed model	99
	GEE	83
Binary	Cluster-level	20, 31, 52, 104, 159, 160 [163]
	Mixed model	105
	GEE	83
Count	GEE	73
Ordinal	GEE	65
	Mixed model	33
Time-to-event	Cluster-level	72, 110
	Mixed model	119 [164]
	Marginal model	71, 89 [165]
Rate	Cluster-level	20

Table 6.1: Results from my systematic review of sample size methods for CRTs: Sample size methods for the standard two-arm, parallel group, equal allocation, fixed cluster sizes completely randomised design. Those references identified in the update to the review are enclosed in square brackets.

powered trials. In these situations the cost of each design becomes important. If the total budget for sampling and measuring clusters and individuals is fixed the optimal design is one that maximises the precision of the treatment effect and power. If the required power and precision of the treatment effect are fixed the optimal design is one which minimises the total costs of measuring clusters and individuals. The total budget constraint is a combination of the cost per subject and cost per cluster. This optimization problem has been considered by Connelly,[104] Moerbeek[105] and Raudenbush.[99] A potential limitation of these methods is the simplicity of the cost functions used, which assume fixed costs over time and in each treatment group. In 2011 Tokola et al opted for a more general approach which allows costs in each treatment group to be different, both at the cluster and individual level.[162]

The majority of methods available for binary and continuous outcomes assume the ICC as a measure of between-cluster variance. The coefficient of variation in outcome, k, has only been proposed for cluster-level analysis by Hayes and Bennett.[20] For continuous outcomes the majority of sample size calculations make the assumption that the measure of correlation, be it the ICC or coefficient of variation in outcome, is the same in each treatment group. However, if the coefficient of variation is the same in each treatment group the ICC will not be, and vice versa.[7] Therefore the use of these different correlation measures will produce different sample size requirements. The assumption of a constant ICC is reasonable if the intervention effect is likely to be constant across clusters. The assumption of a constant k is reasonable if the intervention effect is likely to be proportional to the cluster mean.[21]

Similarly for binary outcomes different sample size requirements are calculated depending upon whether the ICC or coefficient of variation is used in the calculation. For binary outcomes there is an additional complication that the between-cluster variance also depends upon the value of the overall outcome proportion. The use of the ICC is recommended for sample size calculations of binary outcomes, unless the proportion is very small.[21]

Due to the high dependence on prevalence of the ICC for binary outcomes ICC values from one study may not be generalizable to seemingly similar studies that have different outcome prevalence. To address this problem in 2011 Crespi et al proposed a parameter which they call R, that helps to isolate the part of the ICC that measures dependence among responses within a cluster from the outcome prevalence $\rho = \frac{(R-1)\pi}{(1-\pi)}$ where ρ is the ICC and π the outcome proportion.[163] However, this is not necessarily a new concept as it can be easily shown that $R - 1 = k^2$ where k is the coefficient of variation used in the sample size methods by Hayes.[20]

To summarise, for the standard parallel-group trial with binary or continuous outcomes there are few areas that require future development. Sample size methods for these situations, using a variety of analysis methods are well established and have also been extended to incorporate budget constraints. Only one sample size approach considered non-normally distributed continuous outcomes and there is scope to consider whether this approach might be simplified or alternative methods developed.

Count outcomes

Use of the standard design effect has been shown to be appropriate for count outcomes when analysed with a GEE model.[73] However, the authors did not provide any indication of how this method might perform for alternative analyses such as the mixed model and so there is scope for some further research here. The only other sample size approach for count data has been via simulation for cross-over designs.[113]

Ordinal outcomes

Methods available for ordinal outcomes and how they might be developed further were described in detail in Chapter Two and updated in Chapter Five in the light of my research results. These summaries are not repeated here and no further methods were identified in the update to the review.

Time-to-event outcomes

For the individually randomised trial there are two common formulae used for sample size calculation for time-to-event outcomes, those by Schoenfeld[3] and Freedman.[4] Each of these methods has been adapted for the clustered case for cluster-level analyses, mixed, and marginal models by Gang, Jahn-Eimermacher, and Xie respectively[71, 72, 119] A further approach assuming an alternative marginal model was also considered by Manatunga however, it does not result in a simple explicit formula.[89] A simulation-based sample size calculation was proposed by Jung which assumed a weighted rank test with allowance for variable cluster size.[166]

Since the publication of my initial review two further sample size methods for time-to-event outcomes have been proposed by Zhong and Cook in 2014[165] and Moerbeek in 2012.[164] The method by Zhong assumes a semi-parametric proportional hazards model fitted under a working independence assumption with robust variance estimates. Kendall's τ is used to describe the association within clusters and the required number of clusters is calculated for both right-censored and interval-censored time-to-event outcomes. The method by Moerbeek considers discrete event times based on a generalised linear mixed model.

The definition of the ICC used within the papers for time-to-event outcomes was not always easy to determine. There is scope to look at the ICCs for time-to-event data in more detail, perhaps reviewing the alternative estimators and presenting some real life estimates. The use of time-to-event outcomes is less common than binary or continuous outcomes. In my published review of 166 CRTs that reported a sample size calculation no papers were identified with time-to-event outcomes, one was identified to have an ordinal outcome, in thirteen papers the outcome was unclear and in the remaining papers the outcome was either binary, continuous or a rate.[122] Some practical guidance would be useful to describe the issues common to time-to-event designs and to explain the differences and appropriateness of the various proposed sample size methods under different circumstances.

Rate outcomes

When it comes to rate outcomes an ICC cannot be defined and therefore the coefficient of variation in outcome must be used in sample size calculations. In my review there was only one calculation applicable to rate outcomes for the simple design assuming a cluster-level analysis.[20] There is scope to explore the performance of this method, or develop new methods, for individual-level analysis methods.

6.2.2 Adaptations to the standard parallel-group design

In this subsection I consider some of the additional aspects related to design or analysis of a parallel-group trial that one may wish to additionally account for in the sample size calculation. Some of these aspects may be under the control of the investigators, such as the inclusion of covariates or repeated measurements in the analysis and others may be less under their control such as variability

Table 6.2: Results from my systematic review of sample size methods for CRTs: Sample size methodology for design adaptations to the standard two-arm, parallel-group, completely randomised design. Those references identified in the update to the review are enclosed in square brackets

Adaptation	Outcome measure	Analysis	Reference
Design			
ICC uncertainty	Continuous	Cluster-level	118
		Adjusted test	121
		Mixed model	116–118 [167]
		GEE	117, 118
	Binary	Cluster-level	114
Variable cluster size	Continuous	Cluster-level	92, 168, 169
		Adjusted test	74
		Mixed model	95
		GEE	86
	Binary	Cluster-level	92, 112, 168
		Adjusted test	76
		Mixed model	93
		GEE	75, 86
	Time-to-event	Cluster-level	110
Internal pilot	Continuous	Mixed model	120 [170, 171]
		GEE	88
	Binary	Cluster-level	[172]
		GEE	88
Unequal allocation	Continuous	Cluster-level	169
		Mixed model	101
Small number of clusters	Continuous	Cluster-level	173, 174 [175]
	Binary	Cluster-level	173 [175]
Equivalence	Continuous	Adjusted test	33
	Binary	Adjusted test	78
Non-inferiority	Binary	Adjusted test	176
Attrition	Continuous	Adjusted test	85
		Mixed model	177
	Binary	Adjusted test	85
Non-compliance	Binary	Adjusted test	176, 178

Table 6.3: Results from my systematic review of sample size methods for CRTs: Sample size methodology for analysis adaptations to the standard two-arm, parallel-group, completely randomised design. Those references identified in the update to the review are enclosed in square brackets

Adaptation	Outcome measure	Analysis	Reference
Analysis			
Inclusion of covariates	Continuous	Cluster-level	84, 179
		Mixed model	96, 99, 102, 109, 180, 181 [182]
		GEE	86, 181, 183
	Binary	Mixed model	96, 107, 180
		GEE	86, 111, 181, 183, 184
Inclusion of repeated measures	Continuous	Mixed model	177, 185–188 [189, 190]
		GEE	87
	Binary	Marginal model	87 [191]

in cluster size and attrition.

Tables 6.2 and 6.3 summarise the methodology available for adaptations to the standard parallel-group design categorised by outcome measure and analysis method. What is immediately striking from these tables is that all the methods described for these variations are applicable only to binary and continuous outcomes, the exception being when cluster size is variable with a time-to-event outcome which has been considered by Jung via simulation methods.[110] Given that some of these variations from the standard design are quite common there is a considerable lack of methods available to those who wish to perform an individual-level analysis using an outcome that is not binary or continuous.

I now discuss each adaptation in turn.

Uncertainty around the estimate of the ICC

The estimate of the ICC used in the sample size calculation can have a substantial impact upon the resulting sample size, yet there is often a large amount of uncertainty surrounding the estimate.

Informal methods to address the problem have been to opt for a conservative estimate, but this may result in unnecessarily large trials. Turner and Spiegelhalter have described more formal methods for incorporating ICC uncertainty into the sample size calculation. For continuous outcomes these methods make distributional assumptions for one or many previously observed ICC values and from these calculate a distribution for the power.[116–118] These methods adopt a Bayesian perspective but assume that the analysis will follow a classical approach, mixed model or GEE. ICC uncertainty for binary outcomes has received less attention, Feng considers the effect of two possible ICC estimators on sample size via simulation with a cluster-level analysis assumed.[114]

The sample size formulae based around the coefficient of variation in outcome are not always as simple as the design effect and therefore the effect of miss-specification of the estimate of coefficient of variation in outcome is not as obvious. There has been no formal work that considers how potential uncertainty around the estimate of the coefficient of variation in outcome could be incorporated into the trial design, or whether the methods described for ICC uncertainty could translate to uncertainty in the coefficient of variation in outcome.

Where the cluster size and number of clusters are not fixed or constrained for the standard parallel group design methods have been derived to solve the optimality problem of finding the numbers of individuals and clusters that will 1) maximise power and precision of the treatment effect under a fixed budget or 2) minimise total costs given a fixed power and precision of the treatment effect. In addition to the possible oversimplified cost functions used in these methods they also assume that the number sampled from each cluster is the same and the ICC is known, assumptions which may not be realistic. To overcome the issue of unknown ICC van Breukelen derives a Maximin design based on relative efficiency. This provides a design which is robust against miss-specification of the ICC and cost-effective.[167] The Maximin design is found using a series of steps. First the parameter space for the ICC is defined and then the design space is defined for the number of clusters and cluster size. For each ICC value in the parameter space the locally optimal design is calculated using methods by Raudenbush or Moerbeek.[99, 105] The relative efficiency (RE) of each design in the design space is compared to the locally optimal design. For each design in the design space its minimum RE within the ICC parameter space is calculated. The maximin design is that which has the highest minimum RE among all designs in the design space.

Variable cluster sizes

The standard design effect commonly applied to binary and continuous outcomes in the standard parallel-group design makes the assumption that the number of observations from each cluster in the analysis is the same. This may be a reasonable assumption in some situations such as trials of ophthalmology where the cluster is a person and measurements are taken on eyes. However, in other studies where the cluster might be a GP surgery or hospital and we intend to take measurements on the entire cluster we are likely to experience drop out within clusters. At the design stage it is good practice to account for or at least consider the impact of variable clusters. A simple approach is to replace cluster size in the design effect with the mean cluster size but this will result in an underestimated sample size, more so as the variation in cluster size increases. Alternatively, using the maximum cluster size may be overly conservative. Replacement of cluster size with the harmonic mean of the sample size in each cluster provides an appropriate sample size when using the cluster-level approach with coefficient of variation described by Hayes.[7]

Of all the adaptations to the simple parallel-group trial variable cluster sizes has received the most attention. Methods for continuous and binary variables assuming both cluster-level and individual-level analyses have been derived. These methods can be divided into those that require each cluster size to be known in advance[75, 86, 92] and those which require only an estimate of the mean and standard deviation of cluster size to be known.[74, 76, 93, 168] A simple design effect has been derived for the latter approach.[168] Although this design effect has been derived for cluster-level analyses it provides a conservative estimate in the case of individual-level analyses. However, it has been argued by van Breukelen that the approximation of efficiency loss used in many of these approaches is inaccurate and he proposes two simple alternatives[95, 192]

The main development required for the case of variable cluster sizes is towards methods for outcomes that are not continuous or binary.

Internal and external pilot studies

A pilot study is a small study conducted before the main trial which aims to refine the design and procedures such as recruitment, randomisation, data collection and follow up to be used in the main study. The data from an external pilot is not intended to be included with the data from the main trial, hence an external pilot provides more flexibility to change and refine the design and procedures. The sample size for a pilot study should not be calculated as we would for the main trial, i.e. on the expected difference between treatment groups. Instead it should be powered to meet the aims of the pilot study which are usually based around the feasibility of the main study, for example whether the required recruitment rate can be met or the required rate of follow up met. For binary outcomes Ahn et al consider sample size calculations based upon a hypothesised proportion for a single arm clustered design with varying cluster size. The authors suggest that their method can be used to calculate sample sizes for pilot or early stage trials.[172]

If good estimates of the ICC and other required parameters are not available at the design stage it may be possible to re-estimate the sample size once the trial is running, using an internal pilot, as described by Lake et al[120] and Yin and Shen.[88] The data from an internal pilot will be included in the analysis of the data from the main trial. An internal pilot is most suited to trials that recruit a large number of clusters over a relatively long time period. The methods developed so far for sample size re-estimation using an internal pilot assume analysis by a mixed model with a continuous outcome or by a GEE model with binary or continuous outcomes. Internal pilots are less common in cluster randomised trials than they are in the individually randomised setting and there is much scope for further investigation to establish the best practice for their use, for example determining at what point an interim estimate of the ICC could be considered stable and hence used to re-estimate the sample size appropriately.

The method by Lake[120] focuses on situations in which there are a large number of small clusters. In 2012 van Schie and Moerbeek proposed sample size re-estimation methods for situations where there is a limited number of clusters but where the clusters are potentially large and one can continue to recruit participants in a participating cluster after sample size re-estimation in order to reach the required sample size.[170] In 2012 van Breukelen published simpler guidance for sample size calculations with unknown ICC and variable cluster sizes. His recommendations were to base initial sample size calculations upon the midpoint of an assumed ICC range and then re-estimate the ICC from the data once the trial is running. The sample size is then increased only if the ICC is larger than the midpoint. The authors state that the final analysis can be conducted without the need to account for this interim look.[171]

Further guidance on the design of pilot studies is provided by Eldridge and Kerry.[21]

Allocation ratio

Equal allocation provides the most efficient design and was often assumed for simplicity in the majority of methodology identified in the review. However, if the costs per sampling unit are different in each treatment group an equal allocation ratio may produce a smaller sample size for a fixed budget than an unequal allocation. Optimal unequal allocation of clusters to treatment arm can reduce the overall cost of the study and attain higher power than the equal allocation design.[101] Besides cost, unequal allocation may be desired if we wish to gain more information about one of the treatment arms. In situations where a design effect is applied directly to the sample size calculation under individual randomisation unequal allocation may be incorporated in the usual way to the formula for individual randomisation. There is very little published work which focused on the considerations of unequal allocation.

Small/fixed number of clusters

Many of the formulae in the review were based upon the normal distribution and therefore are appropriate when a large number of clusters are to be recruited. When the number of clusters is small these approximations will underestimate the required sample size.

A simple solution to this problem, suggested by Snedecor and Cochran is to add one cluster per arm when testing at the 5% level,[193] implemented in the formulae by Hayes.[20] Alternatively the normal distribution in the formula can be replaced by the t-distribution, or methods based upon the non-central t-distribution used.[169, 173]

Campbell first described methods that can aid in calculating the number of subjects per cluster when the number of clusters is fixed[194] Hemming et al has discussed this further for both binary and continuous outcomes. A trial with a limited number of clusters is feasible if the required power can be obtained by increasing the number of individuals sampled within the cluster. As a simple check the trial will be feasible if the number of clusters is greater than the product of the number of individuals needed under individual randomisation and the ICC. If the design is not feasible either the power must be reduced or the detectable difference must be increased.[175]

It is currently recommended that trials with a small number of clusters be avoided. In addition to the fact that sample size methods may be inappropriate many analysis methods do not perform well with a small number of clusters and imbalance in the cluster characteristics is also more likely to occur when the number of clusters is limited.[21] These two issues would indicate that, at present, further development of sample size calculations that perform well with a small number of clusters may not be useful. Sample size by simulation may be the best approach in these situations.

Equivalence and non-inferiority

Non-inferiority and equivalence designs are less common in the clustered setting. For equivalence designs with binary or continuous outcomes the standard design effect can be used to multiply the sample size calculated under individual randomisation.[33,78] There are no methods described for equivalence designs with alternative outcomes.

For non-inferiority designs Lui and Chang have derived an approach where the treatment effect is based upon the effect among compliers, along similar lines to the Complier Average Causal Effect (CACE) analyses sometimes seen for individually randomised trials. The calculation of the variance for this complier treatment effect is complex meaning this method is not straightforward to implement.[176] Given the role that non-compliance can play in the evaluation of non-inferiority it may not be sensible to try and simplify this method by ignoring non-compliance. For individually randomised trials it is well established that an intention to treat (ITT) analysis (i.e. include non-compliers) of a non inferiority trial may lead to a diluted treatment effect and therefore makes it easier to declare non-inferiority. A per protocol analysis (i.e. exclude non-compliers) of such a design may be biased as the benefits of randomisation are lost and the power is reduced.

Non-inferiority designs have also been considered under an evidence-based perspective to sample size estimation and will be described later in this chapter in the section on emerging themes in sample size methods.[195]

Attrition

In cluster randomised trials attrition among members of a cluster is a common problem. If the proportion that are lost to follow up is θ then the required sample size can be calculated by dividing CRT sample size formulas by $(1 - \theta)$. Alternatively, assuming equal drop out per cluster, n can be replaced by θn in the design effect. However, these methods overestimate and underestimate sample size respectively.

For continuous and binary outcomes, analysed at the individual level using either the individual-level t-test or chi-squared test suitably adjusted for clustering,[6] a simple design effect has been proposed by Taljaard to calculate the sample size.[85] In addition to the ICC this design effect requires an estimate of the probability that an outcome is observed and an intracluster correlation coefficient for the binary missingness indicator, defined as 0 if the outcome is missing and 1 otherwise.

When there is no attrition the individual-level t-test adjusted for clustering is identical to both the standard two-sample t-test based on cluster means and the test of the regression coefficient in a mixed-effects regression model with no other covariates. The drawback of this sample size method is that the analysis does not allow for inclusion of covariates and estimates of the ICC for the missing data mechanism are not routinely published, making the choice of an appropriate figure more difficult. Further work could be done to look at the performance of this method under alternative, and more commonly used, analysis methods and to summarise patterns in the ICC for the missing data mechanism from real life trials. Although Roy has considered an alternative approach for mixed models and drop-out for longitudinal clustered designs his method is iterative and therefore fairly complex.[177]

In 2014 Corrigan et al considered the impact of cluster composition changes, namely clusters merging post-randomisation, on the design and analysis of CRTs. The issue is related to variability in cluster size and loss-to-follow up of clusters; cluster sizes change when clusters merge and we in effect have drop out of an entire cluster. In their simulation studies merging clusters had a detrimental effect on study power. However, the merging of clusters in the same treatment group resulted in only a small loss in power because the ICC decreased with the merge. The authors recommendations are that allowance for cluster merges in the sample size calculation should depend on the perceived likelihood of merges given the costs involved in recruitment of additional clusters.[196]

Non-compliance

In a truly pragmatic trial the effect of the intervention is usually assessed in the presence of non-compliance using an ITT analysis and during the trial compliance may not be measured or actively sought. Therefore accounting for non-compliance at the design stage may not be necessary. An exception to this is the non-inferiority design where ignoring non-compliance may mean non-inferiority is declared more easily.

In trials where you wish to estimate the effect of treatment for those who comply a per-protocol analysis is often conducted. However by excluding participants such an analysis suffers from a reduction in power and the benefits of randomisation are lost. Complier Average Causal Effect models have been suggested as a method to estimate the treatment effect under compliance without the disadvantages of a per protocol analysis. For individually randomised designs CACE analyses are still not routinely used and hence it may be too early to expect routine use of appropriate analysis methods for non-compliance in the clustered case. However they have been considered for both non-inferiority and superiority clustered designs.[176, 178]

Inclusion of baseline measurements

The inclusion of covariates, which are correlated with the outcome, into the analysis will likely reduce the between-cluster variation, and a reduced ICC leads to a reduction in the required sample size.

Inclusion of the baseline value of the outcome has the biggest effect on reducing the ICC. The inclusion of a baseline measurement of the primary outcome introduces additional sources of correlation. In a repeated cross-sectional sample different individuals are measured at each time point. This leads to two sources of correlation: the correlation of outcomes from individuals within a cluster at the same time point (our familiar ICC) and the correlation between baseline and follow-up outcomes for individuals within a cluster (cluster autocorrelation). In a cohort sample the same individuals are measured at baseline and follow up and therefore an additional correlation is encountered across time points on the same individual conditional on the cluster (subject autocorrelation). The cohort sample is more efficient. However, cohort designs can suffer from larger loss-to-follow up than the cross-sectional design and therefore their sample size must be inflated to account for this. The relative efficiency of the cohort design to the cross sectional design has been quantified by Feldman.[96]

For continuous outcomes and a cluster-level analysis of covariance a simple design effect is derived by Teerenstra for the sample size calculation that allows either a cohort sample, cross-sectional sample or a mixture of the two.[84, 179] Cohort and cross-sectional designs for binary and continuous outcomes have also been considered by Preisser et al[183, 184] their sample size method is based upon cluster-level summary statistics but uses an estimating equations approach to estimate the ICCs.

Like so many other methods the practical use of these methods is dependent upon finding good estimates of the additional correlation parameters and further work is required to do this.

Inclusion of other covariates

To account for the inclusion of baseline covariates Neuhaus and Segal[180] have suggested that in general multiplying the ICC by the ICC for the individual-level covariate results in an estimate of the adjusted ICC . This adjusted ICC can then be used in the standard design effect.

When considering the inclusion of covariates other than the baseline measurement of the outcome much of the work has focused around the cost-benefit of their inclusion. The inclusion of the covariate being most cost efficient when the cost of measurement is small and the correlation between the covariate and outcome is large. Raudenbush derived the optimal sample sizes at each level to minimise the standard error of the treatment effect for a continuous outcome under a fixed budget for both individual and cluster-level covariates.[99] His work has been extended for the three-level design by Konstantopoulus.[182] Moerbeek has been very active in this area defining optimal sample sizes at each level to minimise the variance of the treatment effect under a given cost constraint for continuous outcomes[108, 109] and binary outcomes with a binary covariate.[107]

Given there is little guidance about how to choose covariates in a cluster randomised trial and correlation estimates may be difficult to find a conservative approach would be to ignore the covariates in the sample size calculation.

Inclusion of repeated measurements

In a longitudinal cluster randomised design there is a three-level structure with outcomes measured at specific time points within subjects, within clusters. The final analysis model now contains fixed effects for time and the treatment-by-time interaction. The identified sample size methods for these trials were considerably more complex than the methods seen for other scenarios, and the calculations more substantial. The added complexity came from: allowing for different hypothesised paths of the intervention effect over time by Koepsell;[185] incorporation of differential drop-out by Roy;[177] the introduction of random coefficients to the model by Murray;[187]marginal models by Liu[87] and Reboussin,[191] testing of the treatment-by-time interaction by Heo[186] and testing the treatment effect at the final time point with incorporation of information from the entire study period by Heo.[188]

Since the initial review Heo has continued to develop new methods in this area. In 2013 Heo looked at including a subject-specific random slope and in 2014 adapted a previously derived method to allow for anticipated attrition.[186, 189]

This area of sample size methodology would benefit from a detailed evaluation and comparison of each of these methods in order to provide some explicit guidance about when each approach might be suitable and the information that is required to implement each method.

6.2.3 Alternative designs

Sample size methodology for alternative designs is summarised in Table 6.4. Like the adaptations presented in the previous section this methodology has largely focused on continuous and binary outcomes.

Stratification and matching

Generally speaking cluster randomised trials recruit a smaller number of units than an individually randomised trial. This means there is potential for baseline imbalances in cluster characteristics across the treatment groups. Matching or stratification is used is help balance the treatment groups for characteristics that are thought to affect the outcome. In a matched-pair design similar clusters,

Table 6.4: Results from my systematic review of sample size methods for CRTs: Sample size methodology for alternative designs to the standard two-arm, parallel-group, completely randomised design. Those references identified in the update to the review are enclosed in square brackets.

Trial design	Outcome measure	Analysis	Reference
Matched/stratified	Continuous	Cluster-level	20, 160, 197
		Mixed model	198
		Bayesian	115
	Binary	Cluster-level	20, 97, 160, 197, 199
		Mixed model	198
		Adjusted test	200
	Rate	Cluster-level	20, 201
Cross-over	Continuous	Cluster-level	77, 113, 174
		Mixed model	98
	Binary	Cluster-level	113 [202]
	Count	Cluster-level	113
Stepped-wedge	Continuous	Mixed model	82, 203 [204, 205]
	Binary	Mixed model	[205]
	Time-to-event	Equivalent to Cox proportional hazards	[206]
Three-level	continuous	Mixed model	79, 81, 103, 106 [182, 207]
		GEE	80
	Binary	GEE	80

similarities being defined on aspects such as size or geographical location, are matched. One cluster is allocated to control and the other to the treatment group. The reduction of the between-cluster variance induced by matching can provide efficiency in sample size. It should be noted that any potential gain in efficiency may be lost if clusters drop out, which renders the matched pair unusable in the analysis. However, for trials that recruit a small number of relatively large clusters ignoring matching in the analysis of a matched design has been shown to be valid and efficient and can avoid the problem of lost clusters from a matched analysis.[208]

Analysis of a matched-pair design is conducted at the cluster level and sample size calculations have been proposed for continuous, binary and rate outcomes using either the coefficient of variation in outcome within matched pairs,[20] direct measures of the between and within-cluster variances[160, 198, 199, 201] or the correlation in the outcome between matched pairs.[97] To account for the small number of clusters randomised two, rather than one, cluster should be added per group to account for the use of the normal approximation in sample size formulae. This is incorporated into the formula by Hayes.[20]

Stratification is related to the matched-pair design in that there are several clusters within a stratum, rather than two. A simple sample size method has been developed for binary outcomes[200] and a less common Bayesian approach developed for continuous outcomes.[115]

As the impact of stratification is difficult to ascertain in advance, recommendations are to ignore it in the sample size calculation for a more conservative estimate.[21] Therefore, at present, there is probably less need for further development of these methods.

Cross-over designs

Cross-over designs are useful when the availability of clusters is limited as these designs require a smaller number of clusters than a standard parallel-group trial. When different subjects from each cluster are included in separate periods of the trial the design has a cross-sectional sample. Alternatively each subject could be included in both periods within the cluster, a cohort sample. With the cohort design the treatment effect is calculated within subjects, within clusters so both between-cluster and between-subject variations are eliminated making this this most efficient of the cluster randomised cross-over designs. The relative efficiency of the cross-over design with either cohort or cross-sectional sample over the parallel group cluster randomised design have been quantified for continuous outcomes with an assumed analysis by mixed model.[98]

For the cohort sample a simple design effect has been proposed by Giraudeau for continuous outcomes analysed at the cluster level.[77] In 2015 this methodology was extended to accommodate imbalances in cluster sizes with a focus on binary outcomes.[202]

Methods for continuous, binary and count outcomes for the cross-over design have also been explored via simulation.[113]

Future extensions for the cross-over design could be to include allowance for covariates.

Stepped-wedge design

In the stepped-wedge design all clusters receive the control intervention at baseline. At points in the trial one or more clusters will cross-over to receive the treatment intervention, with all clusters receiving the treatment intervention by the end of the trial. This design is similar to the cross-over design except that cross-over is in one direction and staggered over time. The point at which a cluster crosses over is randomly determined at the beginning of the trial. See Figure 6.2.

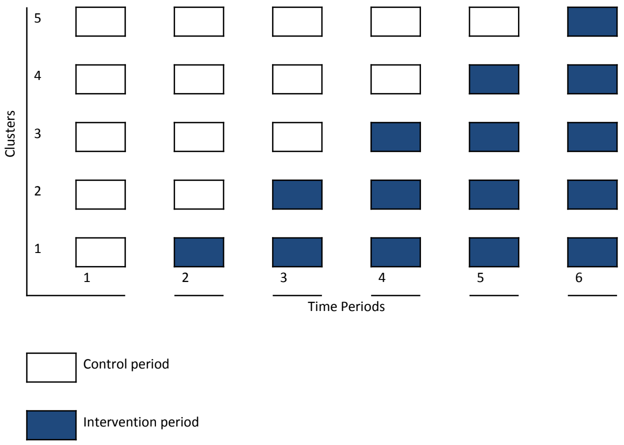


Figure 6.2: Graphical representation of the cluster randomised stepped-wedge design

A stepped wedge design might be used when the implementation of the intervention can only be performed sequentially across clusters or when it is believed the intervention will do more good than harm and so it is thought to be unethical to deny clusters from receiving the intervention during the trial.

These designs are increasing in popularity but guidance regarding best practice around their design and analysis is limited. The first published guidance by Hussey in 2005[203] has since been extended to develop a design effect for the stepped wedge design with continuous outcomes analysed by mixed model by Woertman et al.[82] The appropriate use of this design effect has been further clarified by Hemming and Girling.[209] These calculations have now been implemented in the statistical software Stata.[210]

For time-to-event outcomes Moulton et al propose an adaptation of Hayes formula for rates to account for the stepped-wedge design.[206]

Power calculations for stepped-wedge designs have assumed a cross-sectional design with equally spaced steps, an equal number of clusters randomised at each step and data collected at every step. Methods to account for some variations to these assumptions have been proposed by Hemming, Lilford and Girling in 2014.[204] The variations dealt with include multiple layers of clustering or incomplete data i.e. at some steps for some clusters data are not collected or do not contribute to the analysis. At present only a simulation-based approach has been proposed which can address some of the more complex variations in design, including a cohort design.[205] However, calculations based on a cross-sectional design are likely to be conservative for the cohort design.

Methodological developments in the design, conduct and analysis of stepped wedge designs appear to be occurring at a rapid pace and so there is a lot of scope for development of sample size methodology in parallel.

Three-level cluster randomised trial

Three-level cluster randomised trials commonly occur in educational research where pupils (level 1) are sampled within classrooms (level 2) and randomisation occurs at the school level (level 3). The total variance now includes the variance between schools, the variance between classrooms within schools and the variance of students within classrooms and schools (see Figure 6.3). Aggregating the data at the classroom or school levels reduces this design to the standard parallel group trial where sample size methodology is widely available.

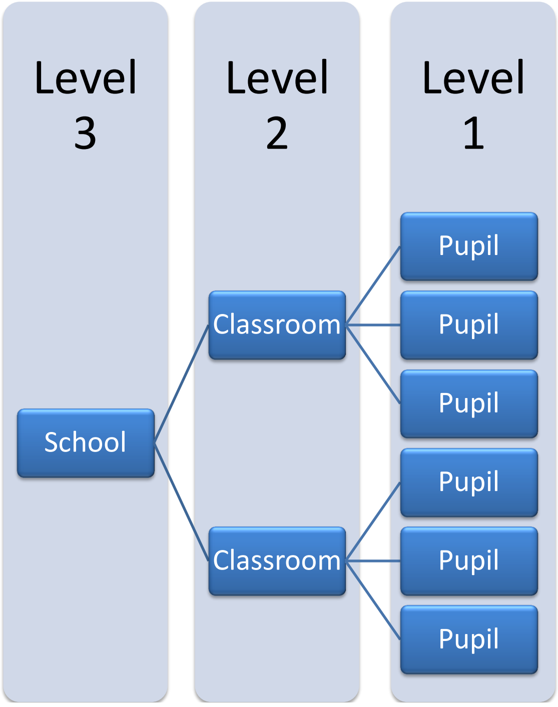


Figure 6.3: Graphical representation of the cluster randomised three-level design

The three-level design lends itself to an individual-level analysis by mixed model or GEE and simple design effects have been proposed for these methods for binary or continuous outcomes by Teerenstra[79,80] and Heo.[81] For the three-level design two ICCs are required, in this example one for students within schools and the second for students within classrooms. The calculation of the optimal sample sizes at each level under a cost constraint has also been considered by Moerbeek[106] and Konstantopoulos.[103,182]

Most methods discussed so far for three-level designs have assumed that randomisation will take place at the highest level e.g. school. If randomisation were to occur at the second-level these designs can be thought of as multi-center cluster randomised trials. Cunningham and Johnson have proposed a simple design effect for randomisation at the lower levels.[207] The same is not true for longitudinal CRTs where outcomes are measured at specific time points within subjects, within clusters. If randomisation were to occur at the second-level the design would be equivalent to a longitudinal multi-center trial.

The most obvious development of the design effect for three-level trials would be to incorporate variable cluster sizes and other adaptations to the standard design as described in Table 6.2.

6.2.4 Emerging themes

The majority of methods identified in the review derived a sample size calculation to reach a pre-specified level of power for a superiority analysis of the primary outcome. Exceptions to this were those methods based on a non-inferiority design or those which optimised the cluster size and number of clusters to provide the maximum precision under a fixed budget constraint. In the update to the review four additional motivations to sample size calculations emerged and are briefly described here.

The evidence-based perspective: Rotondi and Donner took an evidence-based approach to sample size determination. The appropriate sample size is derived based upon its potential impact on the literature i.e. the trial should be large enough to establish whether there is a treatment effect on its own but to also provide a definitive answer when used in subsequent meta-analysis.[195]

Powering for tests of mediation: Mediation analysis is undertaken in order to explain the process by which the intervention affects response. Using simulation methods Hox et al derive the lowest number of clusters required to accurately test and estimate mediation in cluster randomised trials both when maximum likelihood and Bayesian methods are used in estimation.[211]

Powering for cost-effectiveness: In 2014 Manju et al considered optimal sample sizes at the individual and cluster levels under a cost constraint where the outcome is the cost-effectiveness of treatments on a continuous scale.[212] Their approach uses a maximin design and therefore is robust to miss-specification of the parameters such as the ICC.

Powering for a pre-specified confidence interval width : The final approach taken by Pornprasertmanit and Schneider is described as the accuracy in parameter estimation (AIPE) approach.[213] This method helps researchers to find the smallest sample size that will ensure that the confidence interval around the treatment effect will be sufficiently narrow to be informative. Their methods are also extended to include a covariate and deal with unequal cluster sizes.

The remaining emerging themes centred on aspects of the design: three-arm trials; factorial trials; and the dog-leg design.

Three-arm trial: For a three-arm cluster randomised trial the simplest approach to sample size estimation would be to assume that three independent comparisons between the groups are to be made and the maximum sample size is then used for each treatment group.[33] Methods of calculation based upon an overall test of treatment effect have recently been proposed.[214]

Factorial trial: Two methods for factorial designs have been proposed since the original review. The first by Dziak in 2012 for continuous outcomes can accommodate a pre-test measure of the outcome.[215] The second approach in 2015, also for continuous outcomes, by Lemme et al calculates the optimal numbers at each unit in order to minimise the variance of the treatment effect estimator under a total budget constraint and heterogeneous variances across treatment groups. The authors conclude that the 2x2 factorial design is quite robust against heterogeneity of variance and any loss in efficiency can be compensated by the addition of one of two clusters per treatment group.

The dog-leg design: The stepped-wedge design is a form of cross forward design where all clusters cross-over to the intervention arm at some point during the trial. These designs often require a large number of individuals, as repeated cross sectional samples are taken from each cluster. An incomplete cross forward design can reduce the number of individuals required as it leaves gaps in the assessment schedule in some of the arms. The dog-leg design has been proposed as the simplest incomplete cross forward design by Hooper and Bourke which can potentially reduce the number of individuals required and aid researchers to meet ethical and financial requirements for limiting the number of research participants. The name dog-leg comes from the pictorial representation made by the assessment schedule.[216]

6.3 Discussion

In the original review of sample size methods 85 papers were identified published over the 33 years spanning 1978 and 2011. When this review was updated in August 2015 an additional 28 papers were identified published over the 4 years between 2012 and 2015 (see appendix vi for details). This shows the methodology is still increasing.

Papers which made reference to a particular trial as the motivation behind the proposed sample size method were in the minority. Therefore, this rapid increase in methodology may not necessarily reflect a trend towards more varied and/or complex designs but instead may indicate that methods are being developed which are yet to have practical applications.

The focus of my thesis has been on sample size and analysis methods for ordinal outcomes. In this chapter I have taken a side step from this and presented a very broad overview of all the methodology available. I now describe some of the gaps in the methodology that I consider most striking or that I can see would be of interest to the applied statistician. Statisticians with a more detailed knowledge of specific areas for example time-to-event data or longitudinal designs may identify more specific issues that I have not raised here.

For the standard parallel group trial methods are available across a range of outcome measures: continuous, binary, count, ordinal, time-to-event and rates. In this thesis I have further developed the design effect method for ordinal outcomes to provide guidance around its use. However, for variations to this design and alternative design choices the methodology almost solely centres on binary and continuous outcomes.

In the vast majority of methods homogeneity of the between-cluster correlation across treatment groups is assumed and has rarely been challenged. It would be interesting to look further into this to see how reasonable this assumption is in different situations.

Cluster randomised trials with longitudinal designs produced some of the most complex sample size methods. This complexity makes it difficult to identify how they should be implemented and understand the differences between them and the situations to which they can be applied. Further work to consider a comparison of these methods and provide simple advice on their use is needed.

For outcomes which are not binary or continuous and designs other than the standard parallel-group trial many of the sample size methods require estimates of additional parameters. For example for time-to-event outcomes an ICC must be defined and estimated, when cluster size is variable a coefficient of variation in cluster size is required, or when attrition is expected an estimate for the probability that the outcome is missing and an ICC of the missing data mechanism are needed.

These parameters are not yet well established or routinely reported. Finding appropriate estimates to use is therefore one of the biggest barriers to the practical application of these methods. There is scope for much work to be done in this area. Summaries of these parameters from real life data across a range of health areas are needed. More awareness amongst researchers and journal reviewers about the need to report these estimates for the design of future trials would also be helpful.

This is the first comprehensive review of sample size methodology for cluster randomised trials. The full details, including formulae for the methods identified in this review have been described in the associated publication.[60] Given that sample size methods appear to still be expanding. I plan to publish future updates to this review. A strength of the results presented here are that the areas where there is the biggest need for sample size development are immediately highlighted. A thorough critique and comparison of the methods within each section was beyond the scope of this thesis but may reveal some further areas that warrant development.

In this chapter I have highlighted several avenues for future exploration in sample size calculations for cluster randomised trials. One of the most useful aspects of my research on ordinal outcomes was my review of published clustered randomised trials (Chapter Three). This provided great insight into whether there was a need to develop methods for ordinal outcomes and the characteristics of these trials then guided the development of the sample size method, tailoring it to the needs of researchers to make it more practical. Before embarking on the development of any of the methodological gaps highlighted in this chapter I would strongly advise researchers to conduct similar reviews of CRTs to inform their research so that we see more pragmatic methodology, which is actually needed, being developed. My research into ordinal outcomes also raised several questions about situations which are not uncommon such as the appropriate analysis method if the number of clusters is small, what reasonable estimates of the ICC may be and what to do if non-proportional odds are suspected. These issues will be discussed further in the final chapter but I think they highlight the fact that there are still many questions of a very practical nature worth exploring before we move on to develop more complex methods to deal with design variations.

Chapter 7

Discussion and conclusions

The focus of this thesis was to provide a unique contribution towards the review and development of sample size methods for cluster randomised trials, with particular emphasis on ordinal outcomes. The specific objectives are provided in Figure 7.1

Prior to my research none of the objectives of this thesis had been fully addressed in the literature. Several reviews of sample size methodology for CRTs had been published but many of these were not designed to provide a comprehensive overview of all the methods available and the most recent of these was published in 2007. At the time of starting my research there was no proposed simple method for sample size calculations for ordinal outcomes. During the course of the research use of the design effect was proposed but there was little guidance around best practice for its use, including which estimate of the ICC should be used and under which circumstances the method should be used.

Although not in the context of sample size some research had proposed the same ICC estimators as used in this thesis for ordinal outcomes and explored the relationship between them. However, this work assumed a marginal model for both data generation and analysis and hence these results were not directly comparable to the mixed effects context used in my research.

The focus of many of the published reviews of CRTs was the reporting or methodology quality of the included trials. Mine was the first, which I am aware of, to explore the prevalence of ordinal

1. Provide a comprehensive summary of sample size methods available for CRTs (Chapter Two)

2. Identify the distinct processes that have been used to derive sample size methods for CRTs (Chapter Two)

3. Develop sample size methodology with a particular focus on ordinal outcomes, achieved via the following steps:

→ Identify the prevalence of sample size methods for ordinal outcomes (Chapter Two)
→ Determine the prevalence and design characteristics of CRTs with ordinal primary outcomes (Chapter Three)
→ Summarise proposed estimators of the ICC for ordinal outcomes (Chapter Four)
→ Explore the relationship between ICC estimators via simulation (Chapter Five)
→ Determine which ICC provides empirical power closest to the nominal value when used in the design effect for sample size calculation (Chapter Five)
→ Explore the effect on power under variations of the assumptions underlying use of the design effect (Chapter Five)

4. Provide a comprehensive summary of the remaining gaps in sample size methodology for CRTs (Chapter Six)

Figure 7.1: Research objectives of the thesis

outcomes in cluster designs and summarise the design features common to these trials.

In this chapter I summarise the new knowledge around sample size calculation for cluster randomised trials generated as a direct result of my research and describe its significance, strengths, and weaknesses. The main results of my research are translated into explicit practical guidance for those wishing to conduct sample size calculations using ordinal outcomes. The chapter concludes with a description of future research opportunities generated from this work, those specific to ordinal outcomes and those of a broader nature.

7.1 Main findings

In this section I describe the main findings in relation to the thesis objectives described in Figure 7.1

7.1.1 Sample size methods available for CRTs

My systematic review of sample size methods for cluster randomised trials identified a large body of literature, 85 papers. In the 2015 update to the review a further 28 papers were identified, which shows that sample size methodology is still rapidly developing. The literature is dominated by methods applicable to continuous or binary outcomes. Under the standard, parallel group design some methods have been developed for alternative outcomes such as rates, time-to-event and ordinal outcomes but methods available for these types of outcomes under variations or adaptations to the standard design are considerably lacking.

The intracluster correlation coefficient is more frequently used as a measure of between-cluster correlation than the coefficient of variation in outcome. This is most likely due to the wide availability of ICC estimates and literature on the patterns in ICCs for different outcomes and cluster types. More recently an alternative measure of correlation, R, has been proposed specifically for binary outcomes that is not dependent upon the overall prevalence in the way the ICC is.[163]

As the complexity in the design increases the sample size methods often require estimates of further parameters over and above the ICC and cluster size, for example additional measures of correlation are needed for the pre- post- design. The opportunity to use any of the sample size methods identified depends upon the availability and quality of estimates for the required parameters. Many of the sample size methods papers requested that authors report the required estimates with the results of their trial. However, for this to happen more awareness by researchers of these sample size methods and how to calculate the required parameters is needed. It is not clear how this should be done. Recommendations in the CONSORT guidelines might at first seem the obvious place to raise awareness. However, the aim of the CONSORT statement extension for CRTs is to provide a minimum set of reporting requirements, these sample size parameters are additional to the minimum requirements and often trial specific. To recommend reporting via CONSORT would be inappropriate.

In 2015 a summary of methodology for sample size calculations in cluster randomised trials was published by Gao et al.[217] In comparison to my review this review focuses on a smaller range of design situations but additionally includes worked examples of sample size calculations for each type of outcome.

7.1.2 Approaches to sample size derivation

In the majority of papers the sample size method was derived from first principles (n=44, 52%). Using this approach the treatment effect, with its corresponding variance, is defined alongside a definition of the analysis method and an appropriate test statistic, from which a statement about power or sample size can be generated. Other approaches to derivation included: calculating a ratio of the variances of alternative treatment effect estimates; defining the optimal allocation of clusters and cluster size, subject to optimality criteria such as a cost constraint; sample size by simulation; Bayesian methods; or the adaptation of a pre-existing sample size method.

7.1.3 Sample size methods for ordinal outcomes

In my systematic review of sample size methods two methods were identified relevant to ordinal outcomes.[33, 65] The method by Campbell and Walters proposed multiplication of the sample size assuming individual randomisation by the design effect and was the focus of my research. My aim was to determine which ICC estimator should be used and explore how this method performed under design characteristics common to ordinal outcome trials. These design characteristics were identified from my review of 300 cluster randomised trials, of which 11 (4%) were found to have ordinal primary outcomes. Nine (82%) of these outcomes had between 3 and 5 ordinal categories and these outcomes most often measured an aspect of behaviour 9/11 (82%). The design of these trials were most often parallel group 10 (91%), two-arm 6 (55%), completely randomised 8 (73%), and used a cohort sample 9 (82%). Both the total number of clusters randomised and the number of enrolled participants per arm tended to be small.

Three estimators for the ICC in the design effect were considered. The ICC of an underlying latent continuous variable, the ANOVA ICC calculated by assigning numerical equally spaced values to the ordinal categories and a weighted kappa-type ICC measuring chance-corrected agreement.

The results showed that when the number of clusters was large the ICC on the latent response was largest, the ANOVA and kappa-type ICCs were smaller and almost identical. These observed patterns were consistent across the 3-, 4- and 5-level outcome variables. For each scenario investigated as the number of ordinal categories increased the estimated ANOVA and Kappa-type ICCs became closer to the ICC on the latent response.

The use of the ANOVA ICC estimate in the design effect resulted in adequately powered trials, the empirical power was never more than 2% below nominal power. Sample size was more conservative when the proportions in each ordinal category were less evenly spread (see section 5.9.1). Power was only marginally decreased under a minor deviation from the proportional odds assumption.

7.1.4 The future of sample size methods

In chapter Six I presented a summary of the 85 sample size methods for CRTs that were identified in my original review and an additional 28 methods identified in an update to the review conducted in 2015.

For variations to the standard parallel group design such as variability in cluster size, attrition, and inclusion of baseline covariates or repeated measures there is much scope for further development for outcomes which are not binary or continuous.

The practical use of sample size methods rely on the ability to find suitable estimates of the required parameters. Rather than developing new methods of sample size calculation provision of some real life estimates required by these existing methods would go a long way to allowing researchers to implement these methods.

In some papers a particular trial example provided the motivation behind the development of a sample size method and hence a practical application was clear. Where no motivating example was given it was difficult to see how or why these methods would be used, particularly when the method or assumed analysis was computationally complex. Methodologists involved in the development of these methods and statisticians involved in their implementation may wish to consider whether there is a future in developing complex formulae that may be inaccessible to many researchers. Perhaps sample size by simulation may be a better approach to cope with increasing complexity in design, explore ranges of parameter estimates and most importantly reflect analysis methods for which software is available.

7.2 Strengths

Limited guidance has been published around the methodological aspects of conducting and reporting reviews of methodology. My review of sample size methods was implemented using robust methods that aimed to mirror the approaches used in more conventional systematic reviews. These approaches included developing and agreeing a protocol and validation of electronic search terms, data collection and review coverage. This has provided a reliable comprehensive review of sample size methods that future researchers can use to further develop sample size methods.

In this thesis I chose to extend the design effect approach to sample size calculation for ordinal outcomes. The design effect is a familiar and simple approach to sample size calculation and the underlying assumption of analysis via a proportional odds model is the most commonly used method of analysing ordinal data. Contribution to enabling widespread use of the design effect approach will likely have a larger impact in improving trial design than developing a new, but potentially more complicated method.

One of the greatest strengths of my work is its strong generalisability to real-life ordinal outcome trials. The design of the simulation studies were informed by the characteristics of trials that have used ordinal outcomes, identified from a review of 300 cluster randomised trials representative of all areas of health research. This means the results and recommendations for sample size calculation can immediately be applied to similar trials with a reasonable number of clusters. As there was no data available to inform the estimates of ICC investigated in the simulations a wide range of estimates were used to cover the majority of situations.

7.3 Limitations

Since I conducted the review of sample size methods for CRTs there has been 28 new methods published. These new methods were described in Chapter Six. However, with such an increase in sample size methodology over the last five years my review of sample size methods, published in 2015, will quickly be out of date and an update to the published paper will be required.

The prevalence of trials with ordinal primary outcomes was identified from a review of 300 cluster randomised trial results published between 2000 and 2008. Eleven trials with ordinal primary outcomes were identified. It would have been interesting to gain an understanding of whether investigators were actively avoiding using ordinal outcomes due to the lack of sample size methodology available or whether alternative outcomes or methods of analysis were just clinically more relevant. Some qualitative work interviewing or surveying investigators may have provided some insight here.

The characteristics of the eleven trials with ordinal primary outcomes informed the parameters under which I evaluated the design effect methodology. A more up-to-date review, or one which was more focused on areas commonly implementing ordinal outcomes, such as trials in stroke, may have provided more data on which to base the simulation studies. However, this would have been a substantial amount of work which was beyond the time frame of this research and the results would have been less generalizable to all areas of health research.

Despite investigator willingness it was not possible to gain access to the data for any of the 11 trials with ordinal outcomes. This lack of available data or summaries of ICCs for ordinal outcomes meant that a very broad range of possible ICCs were used in the simulation studies and they may not all be applicable to real life trials.

No single method for generating clustered ordinal outcomes has been recommended. I investigated several methods and considered alternative statistical software. This process took considerable time and meant that there was less time for investigating a wider range of scenarios, such as variable cluster sizes and methods for dealing with a small number of clusters, in the simulation study. It is not known whether an alternative data generation method would affect the results.

7.4 Comparison to other work

In this research I extended the design effect approach for sample size calculations for ordinal outcomes to include recommendations on how to calculate an appropriate estimate of the ICC. The calculations involved in this approach are simple and easy to implement and the assumed analysis is a straight forward extension of the proportional odds regression model commonly used for ordinal outcomes in independent data. Kim et al have proposed a sample size method for correlated ordinal outcomes in the longitudinal setting. Their approach assumes a GEE analysis and involves substantially more calculations than the design effect approach, as the cluster size increases the calculation burden also increases. Further work is needed to evaluate this method in the CRT context[65]

In 2012 Ruochu Gao published her thesis on the statistical analysis of correlated ordinal data in cluster randomised trials. Her work was restricted to 3-level ordinal outcomes and community intervention trials, where a small number of large clusters are expected.[125] Although her work did not propose methods for sample size determination she did consider the validity and power of different analysis methods, the focus being on the non-parametric adjusted Cochran-Armitage test and small sample adjustments to the Wald test from a GEE model with clustered ordinal data generated using a marginal model.

7.5 Implications

In this section I translate the results of my research into a series of recommendations for the design of cluster randomised trials with ordinal outcomes. These recommendations are summarised at the end in Figure 7.2.

7.5.1 Choosing the number of categories

When the investigator is in control of the number of categories to use in the ordinal outcome the method by Whitehead has shown that sample size can be most dramatically reduced when increasing the number of categories beyond two, but there is little to be gained in increasing the number of categories beyond five.[2] Depending upon the sample size an increased number of categories may also lead to some categories containing very few observations which may impact the analysis. Whitehead's method is most efficient when the proportions are equally spread across all outcome categories and least efficient when one category is dominant. These results are equally applicable to the clustered design.

7.5.2 Estimating the ICC

For an ordinal outcome the design effect can be calculated using the ANOVA estimate of the ICC, assuming equally spaced numerical scores assigned to the ordinal categories. The value of the ANOVA ICC depends upon the number of categories and the proportions expected within each category. If a weighted kappa-type estimate is used similar levels of power to the ANOVA will be achieved when the number of clusters is large or the level of clustering is small. In the remaining situations the kappa-type estimate is smaller than the ANOVA and hence will lead to reduced statistical power if used in the design effect. Use of the ICC on the latent scale will always lead to a conservative estimate of sample size compared to the ANOVA or kappa-type ICC, but less so as the number of ordinal categories increases. The latent variable ICC is provided in the output from random effects models fitted in Stata and hence estimates may be more available for this ICC than the ANOVA and kappa-type ICC. Other approaches to assigning scores to the ordinal outcome should be used with caution, the performance of the ANOVA ICC estimate in the design effect for these situations is not known.

If no estimates of the ICC are available researchers may consider getting an estimate by utilising estimates or patterns seen for continuous or binary versions of similar outcomes and exploring the sensitivity of the sample size estimate to a range of values. However, a conservative estimate may result in an unnecessarily large trial. There have been no methods that have looked at formally incorporating ICC uncertainty for ordinal outcomes into sample size calculations. Methods developed for continuous outcomes may provide a starting point for some development in this area.[116–118]

Researchers could consider re-estimating the ICC value and sample size part way through the trial, in an internal pilot. However, these methods are only really appropriate when a large number of clusters are to be recruited over a long period of time.

7.5.3 Dealing with a small numbers of clusters

When the analysis is performed at the cluster level the addition of one cluster per arm has been suggested to account for a small number of clusters. When an individual-level analysis is planned it is sensible to also add one cluster per arm which will go some way to reduce the impact of any clusters dropping out. However, when the number of clusters is small the Wald test of the treatment effect from an individual-level analysis by random effects regression does not perform well. In these situations the Type I error is inflated and therefore the possibility of finding a statistically significant treatment effect when no such effect exists is increased. In Chapter Five I looked at whether comparison to the t-distribution, rather than the Normal distribution, may improve the type I error rate of the Wald test for small number of clusters. However, in many situations this reduced the Type I error below 5%. I removed this issue from my simulations by only considering scenarios with at least 40 clusters per arm where the Type I error rate was maintained at 5%.

Another approach to correct for the inflated Type I error rate with a small number of clusters is to use an approximated Wald F test to test the treatment effect in a random effects model. With this test there are several possibilities for estimating the denominator degrees of freedom (DDF). Overestimation will lead to an increased Type I error rate and underestimation will lead to a conservative test and loss of power. Numerous methods have been proposed to approximate the DDF, two of the most well-known are the Kenward-Roger and Satterthwaite methods.[218, 219] The Kenward and Roger method generates a more conservative test than the Satterthwaite method. These corrections tend to perform best under a balanced design, alternative methods may be required when cluster sizes vary and have been evaluated for binary and continuous outcomes.[220, 221] These correction methods are now part of the Stata 14 update to the *mixed* command. I am not aware of any work evaluating these methods for ordinal outcomes or considering the sample size implications of their use. I would recommend using simulation methods to explore the sample size requirements for ordinal outcomes employing these analysis corrections.

Although GEE models are not the focus of the thesis I consider their small sample adjustments briefly here. The Wald test can be similarly applied in the GEE case i.e. the treatment effect divided by its standard deviation and the result compared to a Normal distribution. There are two possible estimates of the variance from the GEE model: the model based variance or the sandwich (robust) estimator. The sandwich estimator is considered robust to miss-specification of the correlation structure but is biased downwards when the number of clusters is small. Some bias-corrected sandwich estimators have been proposed when there are a small number of clusters, such as those by Kauermann and Carroll, Fay and Graubard, Mancl and DeRouen and Morel and Bokossa.[222–225] These approaches have been evaluated via simulation for cluster randomised trials with binary outcomes.[226] The Kauermann and Carroll method works well with moderate variation in cluster size. The authors also proposed a sample size formula for the minimum number of clusters required when using the Kauermann and Carroll correction.

In her research on analysis methods for ordinal outcomes Gao explored the power and Type I error via simulation of the adjusted Cochran-Armitage test and the GEE model with correction and modification strategies applied to the Wald test and score test. Her recommendations were to use the bias-corrected sandwich variance estimator of Mancl and DeRouen.[125]

Brennan Kahan and colleagues from the Pragmatic Clinical Trials Unit (PCTU) at Queen Mary University of London have been doing some research into how big the issue of increased Type I error is among cluster randomised trials. Of 100 randomly selected cluster randomised trials 65% were identified as being at risk from increased Type I error. The majority of these performed an analysis at the individual level without an appropriate small-sample correction (work not yet published).

7.5.4 Choosing the analysis method

Due to the similarity between the logistic and normal distributions the design effect approach for ordinal outcomes can be used when the analysis is via a proportional odds cumulative logit model or ordered probit regression model. The choice between a random-effects model and marginal model should not be crucial to inferential conclusions if the variance components are similar in the two groups.[130] Therefore I would expect the design effect to be applicable to a GEE analysis. This was not tested in my simulations as it is currently not possible to fit a GEE model to ordinal outcomes in Stata. It should also be remembered that the GEE analysis is only valid under the strong assumption of Missing Completely At Random (MCAR). When choosing the analysis method one should also consider the number of clusters to be randomised. If the number of clusters is small small-sample correction methods will need to be applied and for ordinal outcomes there is no consensus over which to use, as described in the previous section.

7.5.5 Dealing with non-proportional odds

In order to use Whitehead's method of sample size calculation the proportions expected in each ordinal category must be known. From these we can calculate the log odds ratios for dichotomisation at each ordinal category and make an assessment of whether the assumption of proportional odds are reasonable.

With the minor deviation from the proportional odds assumption explored in this research use of the design effect to calculate sample size (with treatment effect estimated as the average log-odds) followed by a random effects ordinal regression analysis with probit link resulted in a marginally underpowered trial. Given the difficulty in systematically exploring all possible deviations from proportionality here I would recommend that sample size requirements be confirmed via simulation where non-proportional odds are present. Future reporting of trial results should contain the proportions observed in each category across treatment groups and the associated log-odds ratios so that these estimates may be used in future sample size calculations and the appropriateness of the proportional odds assumption observed.

When larger deviations from proportional odds occur the assumptions of the design effect approach are not met and a simulation approach to sample size is required. However, the alternative analysis methods required for non-proportional odds, and the software to implement them, have received little attention for the clustered case. There are three alternative analyses to consider when the proportional odds assumption is violated i) dichotomise the outcome and carry out a random effects logistic regression ii) assume the outcome is nominal and conduct a multinomial regression or iii) fit a model that allows for non-proportional or partial proportional odds. The first of these approaches is appealing as sample size methods and analysis are well established, but the required sample size is likely to be substantially larger than if the ordered nature of the data had been accounted for. The second approach appeals due to the fact that multinomial random effects regression methods are available in statistical software, but it is not clear whether this method would result in different conclusions compared to a method that takes the ordinal nature of the data into account, as is the case for unclustered data. The final approach is appealing as it maintains the ordinal nature of the data while allowing for non-proportional odds or partial proportional odds and the effects of each category can be seen within the same model. The partial proportional odds method described by Hedeker and Mermelstein[143] has been implemented in an extension to the MIXOR package available in R for mixed effects ordinal regression.[144]

Some simulation methods for sample size are likely to be required when non-proportional odds is expected. This requires some further research to recommend methods and associated software for generating clustered ordinal data with non-proportional odds.

7.5.6 Incorporating baseline measurements

Many of the trials with ordinal primary outcomes identified used a cohort design with an analysis that incorporated baseline measurements. The incorporation of baseline measurements would require an estimate of an appropriately adjusted ICC value to use in the design effect. As ICCs are not routinely available and there is likely to be considerable uncertainty around any ICC estimate I suggest that, at present, baseline measurements are ignored for the purpose of sample size calculations using the design effect for a conservative estimate.

7.5.7 Dealing with cluster size variability

As with design effects used in other settings if there is a large amount of variability in cluster size , i.e. a coefficient of variation greater than 0.23 use of the mean cluster size in the design effect will likely underestimate the required sample size. A design effect which incorporates cluster size variability has been proposed for binary and continuous outcomes but its use for ordinal outcomes has not yet been evaluated.[168] If variability in cluster size is expected I would suggest that a simulation approach to sample size should be taken. If a small number of clusters are to be randomised then small-sample adjustments to the analysis behave differently under variable cluster sizes and should also be explored in the simulations.

7.6 Future work

In this section I describe the future work which has been generated from the work conducted in this thesis.

7.6.1 Guidance on parameter estimation

A common thread throughout this thesis is that the proposed sample size formulae are only useful if we have good estimates to use with them, finding reasonable estimates is the biggest practical barrier to their use. The design effect for ordinal outcomes is no exception. Future work is needed to explore patterns in ICCs for different types of outcomes and clusters to mirror the literature that is available for binary and continuous outcomes.

When reporting the results of the trial it is imperative to the design of future trials that researchers report the ICC value for the ordinal outcome alongside an explicit description of the method of calculation and the proportions observed in each ordinal category for each treatment group.

7.6.2 Research impact

In Chapter Three I identified eleven trials that had an ordinal primary outcome. In only two of these trials the analysis was performed on the ordinal version of the outcome, in others it was either

1. Sample size method
Use Whitehead's formula for individually randomised trials multiplied by the design effect, $1 + (n - 1)\rho$

2. Number of categories
Whitehead's method is most efficient when the proportions are equally spread across all outcome categories. There is little gain in using more than 5 categories.

3. Estimating the ICC
ICC should be calculated using an ANOVA estimate, assuming equally spaced numerical scores assigned to the ordinal categories

→ When the number of clusters is large a weighted kappa-type ICC estimate may alternatively be used
→ Use of the ICC on the latent scale will lead to a conservative estimate of sample size, but less so as the number of categories increases
→ Recommend at present that incorporation of baseline measurements are ignored for a conservative sample size calculation

4. Clusters
Cluster size is assumed fixed

→ If cluster size variability, measured by the coefficient of variation, is greater than 0.23 I recommend sample size by simulation instead of the design effect approach

5. Analysis
Assumed analysis is a random-effects ordered regression model assuming proportional odds and missing data is missing at random. Performs well when the number of clusters is large, around 40 per group

→ With minor deviations to proportional odds the design effect approach may still be appropriate
→ With larger deviations to proportional odds alternative analyses are required and sample size should be calculated by appropriate simulation
→ If the number of clusters is small there is a risk of an inflated Type I error. Small sample correction methods are required and sample size via simulation is recommended

Figure 7.2: Recommendations and considerations for sample size calculation for cluster randomised trials with ordinal outcomes

dichotomised or treated as continuous. It is unclear whether ordinal outcomes are less frequently used because the sample size and analysis methods are not as well understood as those for binary and continuous outcomes or that treating the outcome as binary or continuous rather than ordinal is actually more clinically relevant. The answer to this question was not addressed in this research. However, if the reason is due to the former then I expect to see an increase in ordinal outcome trials after the dissemination of the guidance contained within this chapter. Measuring the impact of my research in this way is something I wish to consider in the future.

7.6.3 Comparison of sample size methods

With such a large number of sample size methods identified for cluster randomised trials in my review it was often difficult to identify the specific methodological differences between similar methods and the potential impact these may have on the resulting sample size. It was not easy to determine how much was gained in sample size efficiency by using a complex approach over a simpler method, and under which circumstances the gain is maximised. In particular, methods for time-to-event outcomes or longitudinal designs were very complex. More work is needed to evaluate the performance of competing approaches under various conditions to aid researchers in sample size decision making. The need for this work was also recognised by Campbell.[55]

7.6.4 Reporting guidelines for methodology papers

In conducting this research I read many methodological papers describing proposed sample size calculations for cluster randomised designs. Despite there being 113 papers describing sample size methods only a handful of these are highly cited. The reason why some methods are not being used could be down to the fact that no practical application has yet arisen, or the new methodology may be considered to provide minimal benefit in light of the added complexity, or no reasonable estimates can be found for the required parameters. I hypothesise that the way in which many methodologies are currently reported has a large influence on their accessibility to a wide audience. I found the reporting quality to be highly variable among the papers and the key assumptions and parameter definitions underpinning each method were often difficult to locate or comprehend. An improvement in the reporting quality may lead to increased identification and uptake of new methods.

Within health research improvements in research reporting are being driven by the creation of reporting guidelines such as CONSORT.[49] Currently there are no guidelines that I am aware of for the reporting of methodology papers in health research. In future research I intend to examine more formally the quality of reporting of statistical methodology using the sample from this research and identify the need for a reporting guideline to be developed. I wish to highlight potential items for inclusion on any subsequent reporting checklist and provide examples of good reporting practices.

7.6.5 Extending sample size methods for ordinal outcomes

The development of statistical methodology in general commonly starts with continuous or binary outcomes, as these are the most frequently used and simplest type of data. Extensions for alternative outcomes and then clustered data follow later, if at all. This is certainly true for ordinal outcomes with both sample size and analysis methods limited for clustered data. The use of the design effect in the context of cluster randomised trials with ordinal outcomes assumes: proportional odds; a reasonable number of clusters; fixed cluster size; and an ANOVA ICC assuming equally spaced numerical values are assigned to each ordinal category. It is reasonable to say that more often than not one of these assumptions may not be satisfied. In the future it would be useful to look at the behaviour of the design effect under different violations of these assumptions and be able to extend or develop new methods of sample size for these circumstances. In order to do this further work is required to provide guidance around such aspects as: the analysis method to use under non-proportional odds with or without variable clusters; which small-sample correction should be applied in the analysis when the number of clusters is small; and which method should be used for generating clustered ordinal data under which circumstances.

In addition to undertaking some of the above future research I intend to publish my work from this thesis on ICC selection when using the design effect for sample size calculation with clustered ordinal outcomes.

7.7 Conclusions

A large amount of literature has been dedicated to sample size calculations for cluster randomised trials. The early work concentrated on the standard parallel group trial with fixed cluster sizes and binary or continuous outcomes. In recent years the literature has expanded to include alternative designs and allowance for variations from the standard design such as variable cluster sizes. However, binary and continuous outcomes still remain the assumption for the vast majority of sample size methods. Trials with alternative types of outcomes, such as ordinal, are conducted and guidance is needed in how to calculate their sample size.

In this thesis I focused on ordinal outcomes for which the proposed sample size approach is to calculate the sample size under individual randomisation and multiply by the standard design effect, which is a function of the cluster size and intracluster correlation coefficient.

At the start of this research the biggest practical barriers to implementation of this approach were that there were no recommendations as to how the ICC should be calculated or under which circumstances the method performed adequately.

In this thesis I evaluated the empirical power of using three alternative ICC estimators in the design effect under scenarios common to ordinal outcome trials. I demonstrated that the ANOVA ICC, by assigning numerical equally spaced values to the ordinal categories, performs sufficiently.

In conclusion, I have provided practical guidance for the strategy to be adopted by researchers calculating sample sizes for cluster randomised trials with ordinal outcomes which was not previously available.

References

[1] S. Pocock. *Clinical Trials A practical Approach.* John Wiley & Sons Inc, Chichester, 1983.

[2] J. Whitehead. Sample size calculations for ordered categorical data. *Statistics in Medicine*, 12:2257–2271, 1993.

[3] D. Schoenfeld. Sample size formula for the proportional-hazards regression model. *Biometrics*, 39:499–503, 1983.

[4] L. Freedman. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine*, 1:121–129, 1982.

[5] J. Sterne and G. Davey Smith. Sifting the evidence - what's wrong with significance tests? *BMJ*, 322:226–231, 2001.

[6] A. Donner and N. Klar. *Design and Analysis of Cluster Randomisation Trials in Health Research.* Arnold, London, 2000.

[7] R. Hayes and L. Moulton. *Cluster randomised trials.* Chapman & Hall/CRC, Boca Raton, 2009.

[8] O. Ukoumunne, M. Gulliford, S. Chinn, J. Sterne, and P. Burney. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assessment*, 3 (5), 1999.

[9] A. Gregory, J. Ramsay, R. Agnew-Davies, K. Baird, A. Devine, D. Dunne, S. Eldridge, A. howell, M. Johnson, C. Rutterford, D. Sharp, and G. Feder. Primary care identification and referral to improve safety of women experiencing domestic violence (iris): protocol for a pragmatic cluster randomised controlled trial. *BMC Public Health*, 10:54, 2010.

[10] B. Husebo, C. Ballard, R. Sandvik, O. Nilsen, and N. Aarsland. Efficacy of treating pain to reduce behavioural disturbances in residents of nursing homes with dementia: cluster randomised clinical trial. *BMJ*, page 343:d4065, 2011.

[11] R. Hayes, N. Alexander, S. Bennet, and S. Cousens. Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Statistical Methods in Medical Research*, 9:95–116, 2000.

[12] PJ. Hannan, DM. Murray, and DR. et al Jacobs. Parameters to aid in the design and analysis of community trials: intraclass correlations from the minnesota heart health program. *Epidemiology*, 5:88–95, 1994.

[13] DM. Murray, DJ. Catellier, and PJ. et al Hannan. School-level intraclass correlation for physical activity in adolescent girls. *Medicine and Science in Sports and Exercise*, 36:876–882, 2004.

[14] MK. Campbell, PM. Fayers, and JM. Grimshaw. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clinical Trials*, 2:99–107, 2005.

[15] M. Taljaard, A. Donner, and J. et al Villar. Intracluster correlation coefficients from the 2005 who global survey on maternal and perinatal health: implications for implementation research. *Paediatric and Perinatal Epidemiology*, 22:117–125, 2008.

[16] G. Adams, M. Gulliford, O. Ukoumunne, S. Chinn, and M. Campbell. Geographical and organisational variation in the structure of primary care services: Implications for study design. *Journal of Health Services Research and Policy*, 8:87–93, 2003.

[17] G. Adams, MC. Gulliford, OC. Ukoumunne, S. Eldridge, S. Chinn, and MJ. Campbell. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, 57 (8):785–794, 2004.

[18] MC. Gulliford, G. Adams, OC. Ukoumunne, R. Latinovic, S. Chinn, and MJ. Campbell. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *Journal of Clinical Epidemiology*, 58:246–251, 2005.

[19] A. Donner. An empirical study of cluster randomization. *International Journal of epidemiology*, 111:283–286, 1982.

[20] R. Hayes and S. Bennett. Simple sample size calculation for cluster- randomized trials. *International Journal of Epidemiology*, 28:319–326, 1999.

[21] S. Eldridge and S. Kerry. *A practical guide to cluster randomised trials in health services research*. Wiley, Chichester, United Kingdom, 2012.

[22] SP. Varnell, DM. Murray, JB. Janega, and JL. Blitstein. Design and analysis of group-randomised trials: a review of recent practices. *American Journal of Public Health*, 94:393–399, 2004.

[23] P. Isaakidid and JP. Ioannidis. Evaluation of cluster randomized controlled trials in sub saharan africa. *American Journal of Epidemiology*, 158:921–6, 2003.

[24] N. Mdege, M. Man, C. Taylor, and D. Torgerson. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology*, 64:936–948, 2011.

[25] S. Eldridge, D. Ashby, G. Feder, A. Rudnicka, and O. Ukoumunne. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical Trials*, 1:80–90, 2004.

[26] M. Bland. Cluster randomised trials in the medical literature: two bibliometric surveys. *BMC Medical Research Methodology*, 4:21, 2004.

[27] LN. Handlos, H. Chakraborty, and PK Sen. Evaluation of cluster-randomized trials on maternal and child health research in developing countries. *Tropical Medicine and International Health*, 8:947–956, 2009.

[28] 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*, 13:127, 2013.

[29] K. Diaz-Ordaz, A. Slowther, R. Potter, and S. Eldridge. Consent processes in cluster-randomised trials in residential facilities for older adults: a systematic review of reporting practices and proposed guidelines. *BMJ Open*, 3:e003057, 2013.

[30] N. Ivers, M. Taljaard, S. Dixon, C. Bennett, A. McRae, J. Taleban, Z. Skea, J. Brehaut, R. Boruch, M. Eccles, J. Grimshaw, C. Weijer, M. Zwarenstein, and A. Donner. Impact of consort extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. *BMJ*, page 343:d5886, 2011.

[31] J. Cornfield. Randomization by group: a formal analysis. *American Journal of Epidemiology*, 108:100–102, 1978.

[32] D. Murray. *Design and analysis of group-randomized trials.* Oxford University Press, New York, 1998.

[33] M. Campbell and S. Walters. *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research.* Wiley, Chichester, 2014.

[34] A. Donner, K. Stephen Brown, and P. Brasher. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979-1989. *International Journal of Epidemiology*, 19:795–800, 1990.

[35] J. Simpson, N. Klar, and A. Donner. Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Public Health*, 85:1378–1383, 1995.

[36] JH. Chuang, G. Hripcsak, and RA. Jenders. Considering clustering: a methodological review of clinical decision support system studies. *Proceedings of the AMIA Symposium*, pages 146–50, 2000.

[37] S. Puffer, D. Torgerson, and J. Watson. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *BMJ*, 327:785, 2003.

[38] DM. Murray, SL. Pals, JL. Blitstein, CM. Alfano, and J. Lehman. Design and analysis of group-randomized trials in cancer: a review of current practices. *Journal of the National Cancer Institute*, 100:483–91, 2008.

[39] S. Eldridge, D. Ashby, C. Bennett, M. Wakelin, and G. Feder. Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ*, 336:876, 2008.

[40] R. Bowater, S. Abdelmalik, and R. Liford. The methodological quality of cluster randomised controlled trials for managing tropical parasitic disease: a review of trials published from 1998 to 2007. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 103:429–436, 2009.

[41] S. Walleser, S. Hill, and L. Bero. Characteristics and quality of reporting of cluster randomized trials in children: reporting needs improvement. *Journal of Clinical Epidemiology*, 64:1331–1340, 2011.

[42] C. Crespi, A. Maxwell, and S. Wu. Cluster randomized trials of cancer screening interventions: Are appropriate statistical methods being used? *Contemporary Clinical Trials*, 32:477–484, 2011.

[43] G. Brierley, S. Brabyn, D. Torgerson, and J. Watson. Bias in recruitment to cluster randomized trials: a review of recent publications. *Journal of Evaluation in Clinical Practice*, 18:878–886, 2012.

[44] B. Giraudeau, A. Caille, A. Le Gouge, and P. Ravaud. Participant informed consent in cluster randomized trials: Review. *PloS ONE*, 5:e40436, 2012.

[45] R. Froud, K. Diaz Ordaz, VCC. Marinho, and A. Donner. Quality of cluster randomized controlled trials in oral health: a systematic review of reports published between 2005 and 2009. *Community Dentistry and Oral Epidemiology*, 40:3–14, 2012.

[46] C. Sutton, C. Watkins, and P. Dey. Illustrating problems faced by stroke researchers: a review of cluster-randomized controlled trials. *International Journal of Stroke*, 8:566–574, 2013.

[47] C. Begg, M. Cho, S. Eastwood, R. Horton, D. Moher, I. Olkin, R. Pitkin, D. Rennie, KF. Schulz, D. Simel, and DF. Stroup. Improving the quality of reporting of randomized controlled trials. the consort statement. *Journal of the American Medical Association*, 276:637–9, 1996.

[48] D. Moher, KF. Schulz, and DG. Altman. The consort statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*, 357:1191–4, 2001.

[49] K. Schulz, D. Altman, and D. for the CONSORT Group Moher. Consort 2010: Updated guidelines for reporting parallel group randomised trials. *British Medical Journal*, 340:698–702, 2010.

[50] MK. Campbell, DR. Elbourne, and DG. Altman. Consort statement: extension to cluster randomised trials. *British Medical Journal*, 328:702–8, 2004.

[51] MK. Campbell, G. Piaggio, DR. Elbourne, and DG. Altman. Consort 2010 statement: extension to cluster randomised trials. *British Medical Journal*, 345:e5661, 2012.

[52] A. Donner, N. Birkett, and C. Buck. Randomization by cluster. sample size requirements and analysis. *American Journal of Epidemiology*, 114:906–914, 1981.

[53] Klar.N. and A. Donner. Current and future challenges in the design and analysis of cluster randomization trials. *Statistics in Medicine*, 20:3729–3740, 2001.

[54] D. Murray, S. Varnell, and L. Blitstein. Design and analysis of group-randomized trials: A review of recent developments. *American Journal of Public Health*, 94:423–432, 2004.

[55] MJ. Campbell, A. Donner, and N. Klar. Developments in cluster randomized trials and statistics in medicine. *Statistics in Medicine*, 26:2–19, 2007.

[56] A. Steptoe, S. Kerry, E. Rink, and S. Hilton. The impact of behavioral counseling on stage of change in fat intake, physical activity, and cigarette smoking in adults at increased risk of coronary heart disease. *American Journal of Public Health*, 91:265–269, 2001.

[57] The Optimising Analysis of Stroke Trials (OAST) collaboration. Calculation of sample size for stroke trials assessing functional outcome: comparison of binary and ordinal approaches. *International Journal of Stroke*, 3:78–84, 2008.

[58] PR. Williamson, D. Altman, J. Blazeby, M. Clarke, and E. Gargon. Driving up the quality and relevance of research through the use of agreed core outcomes. *Journal of Health Services Research and Policy*, 17:1–2, 2012.

[59] R. Hooper. Versatile sample size calculation using simulation. *Stata Journal*, 13:21–38, 2013.

[60] C. Rutterford, A. Copas, and S. Eldridge. Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, pages 1051–1067, 2015.

[61] Statistics in Medicine. *Special Issue: Design and Analysis of Cluster Randomized trials*, volume 20(3). Wiley, February 2001. `http://onlinelibrary.wiley.com/doi/10.1002/1097-0258(20010215)20:3%3C%3E1.0.CO;2-B/issuetoc`.

[62] Clinical Trials. *Special Issue for cluster randomized trials*, volume 2(2). Sage, April 2005. `http://ctj.sagepub.com/content/2/2.toc`.

[63] Statistical Methods in Medical Research. *Special Issue for cluster randomized trials*, volume 9(2). Sage, April 2000. `http://smm.sagepub.com/content/9/2.toc`.

[64] American Journal of Epidemiology. *Special Issue for community intervention trials*, volume 142(6). Oxford Journals, September 1995. `http://aje.oxfordjournals.org/content/142/6.toc`.

[65] H. Kim, J. Williamson, and C. Lyles. Sample size calculations for studies with correlated ordinal outcomes. *Statistics in Medicine*, 24:2977–2987, 2005.

[66] M. Campbell. *Statistics at Square Two: Understanding Modern Statistical Applications in Medicine.* British Medical Journal, London, 2001.

[67] J. Rochon. Application of gee procedures for sample size calculations in repeated measures experiments. *Statistics in Medicine*, 17:1643–1658, 1998.

[68] S. Lipsitz, K. Kim, and L. Zhao. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, 13:1149–1163, 1994.

[69] KY. Liang and SL. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

[70] C. Ahn, M. Heo, and S. Zhang. *Sample size calculations for clustered and longitudinal outcomes in clinical research.* Chapman & Hall/CRC Biostatistics Series, Boca Raton, 2014.

[71] T. Xie and J. Waksman. Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. *Statistics in Medicine*, 22:2835–2846, 2003.

[72] R. Gangnon and M. Kosorok. Sample-size formula for clustered survival data using weighted log-rank statistics. *Biometrika*, 91:263–275, 2004.

[73] A. Amatya, D. Bhaumikb, and R. Gibbons. Sample size determination for clustered count data. *Statistics in Medicine*, 32:4162–4179, 2013.

[74] A. Manatunga and M. Hudgens. Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal*, 43:75–86, 2001.

[75] W. Pan. Sample size and power calculations with correlated binary data. *Controlled Clinical Trials*, 22:211–227, 2001.

[76] S. Kang, C. Ahn, and S. Jung. Sample size calculation for dichotomous outcomes in cluster randomization trials with varying cluster size. *Drug information journal*, 37:109–114, 2003.

[77] B. Giraudeau, P. Ravaud, and A. Donner. Sample size calculation for cluster randomized crossover trials. *Statistics in Medicine*, 27:5578–5585, 2008.

[78] A. Donner. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics*, 47:95–113, 1998.

[79] S. Teerenstra, M. Moerbeek, T. Achterberg, B. Pelzer, and G. Borm. Sample size calculations for 3-level cluster randomised trials. *Clinical Trials*, 5:486–495, 2008.

[80] S. Teerenstra, B. Lu, J. Preisser, T. Achterberg, and G. Borm. Sample size considerations for gee analyses of three-level cluster randomized trials. *Biometrics*, 66:1230–1237, 2010.

[81] M. Heo and A. Leon. Statistical power and sample size requirements for three level hierarchical cluster randomised trials. *Biometrics*, 64:1256–1262, 2008.

[82] W. Woertman, E. de Hoop, M. Moerbeek, SU. Zuidema, DL. Gerritsen, and S. Teerenstra. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *Journal of Clinical Epidemiology*, 66:752–758, 2013.

[83] W. Shih. Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalised estimating equations. *Biometrical journal*, 39:899–908, 1997.

[84] T. Teerenstra, S. Eldridge, M. Graff, E. de Hoop, and G. Borm. A simple sample size formula for analysis of covariance in cluster randomized trials. *Statistics in Medicine*, 2012.

[85] M. Taljaard, A. Donner, and N. Klar. Accounting for expected attrition in the planning of community intervention trials. *Statistics in Medicine*, 26:2615–2628, 2007.

[86] G. Liu and K. Liang. Sample size calculations for studies with correlated observations. *Biometrics*, 53:937–947, 1997.

[87] A. Liu, W. Shih, and E. Gehan. Sample size and power determination for clustered repeated measurements. *Statistics in Medicine*, 21:1787–1801, 2002.

[88] G. Yin and Y. Shen. Adaptive design and estimation in randomized clinical trials with correlated observations. *Biometrics*, 61:362–369, 2005.

[89] A. Manatunga and S. Chen. Sample size estimation for survival outcomes in cluster-randomization studies with small cluster sizes. *Biometrics*, 56:616–621, 2000.

[90] A. Agresti and R. Natarajan. Modeling clustered ordered categorical data: A survey. *International Statistical Review*, 69:345–371, 2001.

[91] BS. Everitt and A Skrondal. *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge, 2010.

[92] S. Kerry and M. Bland. Unequal cluster sizes for trials in english and welsh general practice: implications for sample size calculations. *Statistics in Medicine*, 20:377–390, 2001.

[93] M. Candel and G. Van Breukelen. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order pql mixed logistic regression. *Statistics in Medicine*, 29:1488–1501, 2010.

[94] Z. You, O. Williams, I. Aban, E. Kabagambe, H. Tiwari, and G. Cutter. Relative efficiency and sample size for cluster randomized trials with variable cluster sizes. *Clinical Trials*, 8:27–36, 2011.

[95] G. Van Breukelen, M. Candel, and M. Berger. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*, 26:2589–2603, 2007.

[96] H. Feldman. Cohort versus cross-sectional design in large field trials: Precision, sample size, and a unifying model. *Statistics in Medicine*, 13:61–78, 1994.

[97] L. Freedman, S. Green, and D. Byar. Assessing the gain in efficiency due to matching in a community intervention study. *Statistics in Medicine*, 9:943–952, 1990.

[98] C. Rietbergen and M. Moerbeek. The design of cluster randomized crossover trials. *Journal of Educational and Behavioral Statistics*, 36:472–490, 2011.

[99] S. Raudenbush. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2:173–185, 1997.

[100] T. Snijders and R. Bosker. Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, 18:237–259, 1993.

[101] X. Liu. Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomisation. *Journal of Educational and Behavioural Statistics*, 28:231–248, 2003.

[102] S. McKinlay. Cost-efficient designs of cluster unit trials. *Preventive medicine*, 23:606–611, 1994.

[103] S. Konstantopoulos. Incorporating cost in power analysis for three-level cluster randomized designs. *Evaluation review*, 33:335–357, 2009.

[104] L. Connelly. Balancing the number and size of sites: an economic approach to the optimal design of cluster samples. *Controlled Clinical Trials*, 24:544–559, 2003.

[105] M. Moerbeek, G. Van Breukelen, and M. Berger. Optimal experimental design for multilevel logistic models. *Journal of the Royal Statistical Society series D The Statistician*, 50:17–30, 2001.

[106] M. Moerbeek, G. Van Breukelen, and M. Berger. Design issues for experiments in multilevel populations. *Journal of Educational and Behavioural Statistics*, 25:271–284, 2000.

[107] M. Moerbeek and C. Maas. Optimal experimental design for multilevel logistic models with two binary predictors. *Communications in statistics-Theory and Methods*, 34:1151–1167, 2005.

[108] M. Moerbeek, G. Van Breukelen, and M. Berger. Optimal experimental design for multilevel models with covariates. *Communications in statistics-Theory and Methods*, 30:2683–2697, 2001.

[109] M. Moerbeek. Power and money in cluster randomized trials: when is it worth measuring a covariate. *Statistics in Medicine*, 25:2607–2617, 2006.

[110] S. Jung. Sample size calculation for weighted rank tests comparing survival distributions under cluster randomisation: A simulation method. *Journal of Biopharmaceutical Statistics*, 17:839–849, 2007.

[111] S. Hendricks, J. Wassell, J. Collins, and S. Sedlak. Power determination for geographically clustered data using generalised estimating equations. *Statistics in Medicine*, 15:1951–1960, 1996.

[112] T. Braun. A mixed model formulation for designing cluster randomized trials with binary outcomes. *Statistical modelling*, 3:233–249, 2003.

[113] NG. Reich, JA. Myers, D. Obeng, AM. Milstone, and TM. Perl. Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. *PLoS One*, 7:e35564, 2012.

[114] Z. Feng and J. Grizzle. Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculation. *Statistics in Medicine*, 11:1607–1614, 1992.

[115] T. Kikuchi and J. Gittins. A behavioural bayes approach for sample size determination in cluster randomized clinical trials. *Journal of the Royal Statistical Society*, 59:875–888, 2010.

[116] D. Spiegelhalter. Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine*, 20:435–452, 2001.

[117] R. Turner, T. Prevost, and S. Thompson. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Statistics in Medicine*, 23:1195–1214, 2004.

[118] R. Turner, S. Thompson, and D. Spiegelhalter. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials*, 2:108–118, 2005.

[119] A. Jahn-Eimermacher, l K. Inge, and A. Schneider. Sample size in cluster-randomized trials with time to event as the primary endpoint. *Statistics in Medicine*, 32:639–512, 2013.

[120] S. Lake, E. Kammann, N. Klar, and R. Betensky. Sample size re-estimation in cluster randomization trials. *Statistics in Medicine*, 21:1337–1350, 2002.

[121] S. Mukhopadhyay and S. Looney. Quantile dispersion graphs to compare the efficiencies of cluster randomized designs. *Journal of Applied Statistics*, 36:1293–1305, 2009.

[122] C. Rutterford, M. Taljaard, S. Dixon, A. Copas, and S. Eldridge. Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review. *Journal of Clinical Epidemiology*, 68:716–723, 2014.

[123] P. Howlin, K. Gordon, G. Pasco, A. Wade, and T. Charman. The effectiveness of picture exchange communication system (pecs) training for teachers of children with autism: a pragmatic, group randomised controlled trial. *Journal of Child Psychology and Psychiatry*, 48:473–481, 2007.

[124] G. Brody, V. Murry, S. Kogan, M. Gerrard, F. Gibbons, V. Molgaard, A. Brown, T. Anderson, Y. Chen, Z. Luo, and T. Wills. The strong african american families program: A cluster-randomized prevention trial of long-term effects and a mediational model. *Journal of Consulting and Clinical Psychology*, 74:356–366, 2006.

[125] R Gao. Statistical analysis of correlated ordinal data: Application to cluster randomization trials. *University of Western Ontario - Electronic Thesis and Dissertation Repository. Paper 1696*, page http://ir.lib.uwo.ca/etd/1696, 2013.

[126] A. Agresti. A survey of models for repeated ordered categorical response data. *Statistics in Medicine*, 8:1209–24, 1989.

[127] A. Agresti. Modelling ordered categorical data: recent advances and future challenges. *Statistics in Medicine*, 18:2191–2207, 1999.

[128] I. Liu and A. Agresti. The analysis of ordered categorical data: An overview and survey of recent developments. *Sociedad de Estadistica e Investigacion Operativa*, 14:1–73, 2005.

[129] R. Lall, M. Campbell, S. Walters, K. Morgan, and MRC CFAS Co-operative. A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research*, 11:49–67, 2002.

[130] A Agresti. *Analysis of ordinal categorical data.* Wiley, New Jersey, 2010.

[131] B. Roozenbeek, H. Lingsma, P. Perel, P. Edwards, I. Roberts, G. Murray, A. Maas, and E. Steyerberg. The added value of ordinal analysis in clinical trials: an example in traumatic brain injury. *Critical Care*, 15:R127, 2011.

[132] P. Bath, C. Geeganage, L. Gray, T. Collier, and S. Pocock. Use of ordinal outcomes in vascular prevention trials: comparison with binary outcomes in published trials. *Stroke*, 39:2817–2823, 2008.

[133] L. Sullivan and R. D'Agostino. Robustness and power of analysis of covariance applied to ordinal scaled data as arising in randomized controlled trials. *Statistics in Medicine*, 22:1317–1334, 2003.

[134] T. Heeren and R. D'Agostino. Robustness of the two independent samples t-test when applied to ordinal scaled data. *statistics in Medicine*, 6:79–90, 1987.

[135] S. Walters, M. Campbell, and R. Lall. Design and analysis of trials with quality of life as an outcome: a practical guide. *Journal of Biopharmaceutical Statistics*, 11:155–176, 2001.

[136] S. Walters and M. Campbell. The use of bootstrap methods for estimating sample size and analysing health-related quality of life outcomes. *Statistics in Medicine*, 24:1075–1102, 2005.

[137] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B*, 42:109–142, 1980.

[138] R. Brant. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46:1171–1178, 1990.

[139] B. Peterson and F.E. Harrell. Partial proportional odds models for ordinal response variables. *Applied Statistics*, 39:205–217, 1990.

[140] B. Rosner and D. Grove. Use of the mann-whitney u-test for clustered data. *Statistics in Medicine*, 18:1387–1400, 1999.

[141] D. Hedeker and R. Gibbons. A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 50:933–944, 1994.

[142] R. Little and D. Rubin. *Statistical Analysis with Missing Data.* Wiley, New York, 1987.

[143] D. Hedeker and R. Mermelstein. A multilevel thresholds of change model for analysis of states of change data. *Multivariate Behavioral Research*, 33:427–455, 1998.

[144] D. Hedeker and R. Gibbons. A computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine*, 49:157–176, 1996.

[145] M. Ridout, C. Demetrio, and D. Firth. Estimating intraclass correlation for binary data. *Biometrics*, 55:137–148, 1999.

[146] A. Donner. A review of inference procedures for the intraclass correlation-coefficient in the one-way random effects model. *International statistical review*, 54:67–82, 1986.

[147] S. Eldridge, O. Ukoumunne, and J. Carlin. The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions. *International Statistical Review*, 77:378–394, 2009.

[148] H. Goldstein, W. Browne, and J. Rasbash. Partitioning variation in multilevel models. *Understanding Statistics*, 1:223–232, 2002.

[149] P. Rothery. A nonparametric measure of intraclass correlation. *Biometrika*, 66:629–639, 1979.

[150] R. Muller and P. Buttner. A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, 13:2465–2476, 1994.

[151] S Shirahata. Intraclass rank tests for independence. *Biometrika*, 68:451–456, 1981.

[152] S. Priebe, L. Kelley, S. Omer, E. Golden, S. Walsh, H. Khanom, D. Kingdon, C. Rutterford, P. McCrone, and R. McCabe. The effectiveness of a patient-centred assessment with a solution-focused approach (dialog+) for patients with psychosis: A pragmatic cluster-randomised controlled trial in community care. *Psychotherapy and Psychosomatics*, 84:304–313, 2015.

[153] J. Williamson, K. Kim, and S. Lipsitz. Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association*, 90:1432–1437, 1995.

[154] A. Burton, D. Altman, P. Royston, and R. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25:4279–4292, 2006.

[155] SJ. Gange. Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician*, 49:134–138, 1995.

[156] A. Biswas. Generating correlated ordinal categorical random samples. *Statistics and probability letters*, 70:25–35, 2004.

[157] H. Demirtas. A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation*, 76:1017–1025, 2006.

[158] S. Jung and S. Kang. Tests for 2 x k contingency tables with clustered ordered categorical data. *Statistics in Medicine*, 20:785–794, 2001.

[159] S. Kerry and M. Bland. Trials which randomize practices ii: sample size. *Family Practice*, 15:84–87, 1998.

[160] F. Hsieh. Sample size formulae for intervention studies with the cluster as unit of randomisation. *Statistics in Medicine*, 8:1195–1201, 1988.

[161] B. Rosner and R. Glynn. Power and sample size estimation for the clustered wilcoxon test. *Biometrics*, pages 1–8, 2010.

[162] K. Tokola, D. Larocque, J. Nevalainen, and H. Oja. Power, sample size and sampling costs for clustered data. *Statistics and probability letters*, 81:852–860, 2011.

[163] C. Crespi, W. Wong, and S. Wu. A new dependence parameter approach to improve the design of cluster randomized trials with binary outcomes. *Clinical Trials*, 8:687–698, 2011.

[164] M. Moerbeek. Sample size issues for cluster randomized trials with discrete-time survival endpoints. *Methodology*, 8:146–158, 2012.

[165] Y. Zhong and R. Cook. Sample size and robust marginal methods for cluster-randomized trials with censored event times. *Statistics in Medicine*, 34:901–923, 2015.

[166] S. Jung and J. Jeong. Rank tests for clustered survival data. *Lifetime Data Analysis*, 9:21–33, 2003.

[167] G. Van Breukelen and M. Candel. Efficient design of cluster randomized and multicentre trials with unknown intraclass correlation. *Statistics Methods in Medical Research*, pages 1–18, 2011.

[168] S. 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*, 35:1292–1300, 2006.

[169] D. Hoover. Power for t-test comparisons of unbalanced cluster exposure studies. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 79:278–294, 2002.

[170] van Schie. S. and M. Moerbeek. Re-estimating sample size in cluster randomised trials with active recruitment within clusters. *Statistics in Medicine*, 33:3253–3268, 2014.

[171] G. Van Breukelen and M. Candel. Calculating sample sizes for cluster randomized trials: We can keep it simple and efficient! *Journal of Clinical Epidemiology*, 65:1212–1218, 2012.

[172] C. Ahn, F. Hu, and S-C. Lee. Relative efficiency of unequal versus equal cluster sizes for the nonparametric weighted sign test estimators in clustered binary data. *Drug information Journal*, 46:428–433, 2012.

[173] A. Donner and N. Klar. Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology*, 49:435–439, 1996.

[174] D. Harrison and A. Brady. Sample size and power calculations using the noncentral t-distribution. *Stata Journal*, 4:142–153, 2004.

[175] K. Hemming, A. Girling, A. Sitch, J. Marsh, and R. Lilford. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Medical Research Methodology*, 11:102, 2011.

[176] K. Lui and K. Chang. Test non-inferiority and sample size determination based on the odds ratio under a cluster randomized trial with noncompliance. *Journal of Biopharmaceutical Statistics*, 21:94–110, 2011.

[177] A. Roy, D. Bhaumik, S. Aryal, and R. Gibbons. Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics*, 63:699–707, 2007.

[178] K. Lui and K. Chang. Sample size determination for testing equality in a cluster randomized trial with noncompliance. *Journal of biopharmaceutical Statistics*, 21:1–17, 2011.

[179] D. Murray and P. Hannan. Planning for the appropriate analysis in school-based drug-use prevention studies. *Journal of Consulting and Clinical Psychology*, 58:458–468, 1990.

[180] J. Neuhaus and M. segal. Design effects for binary regression models fitted to dependent data. *Statistics in Medicine*, 12:1259–1268, 1993.

[181] X. Tu, J. Kowalski, J. Zhang, K. Lynch, and P. Crits-Christoph. Power analyses for longitudinal trials and other clustered designs. *Statistics in Medicine*, 23:2799–2815, 2004.

[182] S. Konstantopoulos. Optimal sampling of units in three-level cluster randomized designs: An ancova framework. *Educational and Psychological Measurement*, 71:798–813, 2011.

[183] J. Preisser, M. Young, D. Zaccaro, and M. Wolfson. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Statistics in Medicine*, 22:1235–1254, 2003.

[184] J. Preisser, B. Reboussin, E. Song, and M. Wolfson. The importance and role of intracluster correlations in planning cluster trials. *Epidemiology*, 18:552–560, 2007.

[185] T. Koepsell, D. Martin, P. Diehr, B. Psaty, E. Wagner, E. 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*, 44:701–713, 1991.

[186] M. Heo and A. Leon. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials. *Statistics in Medicine*, 28:1017–1027, 2009.

[187] D. Murray, J. Blitstein, P. Hannan, W. Baker, and L. Lytle. Sizing a trial to alter the trajectory of health behaviours: Methods, parameter estimates, and their application. *Statistics in Medicine*, 26:2297–2316, 2007.

[188] M. Heo, Y. Kim, X. Xue, and M. Kim. Sample size requirement to detect an intervention effect at the end of follow-up in a longitudinal cluster randomized trial. *Statistics in Medicine*, 29:382–390, 2009.

[189] M. Heo. Impact of subject attrition on sample size determinations for longitudinal cluster randomized clinical trials. *Journal of Biopharmaceutical Statistics*, 24:507–522, 2014.

[190] M. Heo, X. Xue, and M. Kim. Sample size requirement to detect an intervention by time interaction in longitudinal cluster randomized trials with random slopes. *Computational Statistics and Data Analysis*, 60:169–178, 2013.

[191] B. Reboussin, J. Preisser, E-Y. Song, and M. Wolfson. Sample size estimation for alternating logistic regressions analysis of multilevel randomized community trials of under-age drinking. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175:691–712, 2012.

[192] G. Van Breukelen and M. Candel. Comments on 'efficiency loss because of varying cluster size in cluster randomized trials is smaller than literature suggests'. *Statistics in Medicine*, 31:397–400, 2012.

[193] WG. Snedecor, GW. Cochran. *Statistical Methods.* Iowa State University Press, Ames, IA, 1980.

[194] MJ. Campbell. Cluster randomized trials in general (family) practice research. *Statistical Methods in Medical Research*, 9:81–94, 2000.

[195] M. Rotondi and A. Donner. Sample size estimation in cluster randomized trials: An evidence-based perspective. *Computational Statistics and Data Analysis*, 56:1174–1187, 2012.

[196] N. Corrigan, M. Bankart, L. Gray, and K. Smith. Changing cluster composition in cluster randomised controlled trials: design and analysis considerations. *Trials*, 15:184, 2014.

[197] Z. Feng and B. Thompson. Some design issues in a community intervention trial. *Controlled Clinical Trials*, 23:431–449, 2002.

[198] S. Thompson, S. Pyke, and R. Hardy. The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. *Statistics in Medicine*, 16:2063–2079, 1997.

[199] D. Byar. The design of cancer prevention trials. *Recent Results in Cancer Research*, 111:34–48, 1988.

[200] A. Donner. Sample size requirements for stratified cluster randomization designs. *Statistics in Medicine*, 11:743–750, 1992.

[201] M. Shipley, P. Smith, and M. Dramaix. Calculation of power for matched pair studies when randomization is by group. *International Journal of Epidemiology*, 18:457–461, 1989.

[202] A. Forbes, M. Akram, D. Pilcher, J. Cooper, and R. Bellomo. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: Application to studies of near-universal interventions in intensive care. *Clinical Trials*, 12:34–44, 2015.

[203] M. Hussey and J. Hughes. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, 28:182–191, 2007.

[204] K. Hemming, R. Lilford, and A. Girling. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *statistics in Medicine*, 34:181–196, 2015.

[205] G. Baio, A. Copas, G. Ambler, J. Hargreaves, E. Beard, and R. Omar. Sample size calculation for a stepped wedge trial. *Trials*, 16:354, 2015.

[206] L. Moulton, J. Golub, B. Burovni, S. Cavalcante, A. Pacheco, and V. et al Saraceni. Statistical design of thrio: a phased implementation clinic-randomized study of tuberculosis preventive therapy intervention. *Clinical Trials*, 4:190–9, 2007.

[207] T. Cunningham and R. Johnson. Design effects for sample size computation in three-level designs. *Statistical Methods in Medical Research*, page 0962280212460443, 2012.

[208] A. Donner, M. Taljaard, and N. Klar. The merits of breaking the matches: a cautionary tale. *Statistics in Medicine*, 26:2036–51, 2007.

[209] K. Hemming and A.. Girling. The efficiency of stepped wedge vs. cluster randomized trials: Stepped wedge studies do not always require a smaller sample size. *Journal of Clinical Epidemiology*, 66:1427–1428, 2013.

[210] K. Hemming and A. Girling. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata Journal*, 14:363–380, 2014.

[211] J. Hox, M. Moerbeek, A. Kluytmans, and R. van de Schoot. Analyzing indirect effects in cluster randomized trials. the effect of estimation method, number of groups and group sizes on accuracy and power. *Frontiers in Psychology*, 5:78, 2014.

[212] A. Manju, M. Candel, and M. Berger. Sample size calculation in cost-effectiveness cluster randomized trials: optimal and maximin approaches. *Statistics in Medicine*, 33:2538–2553, 2014.

[213] S. Pornprasertmanit and W. Schneider. Accuracy in parameter estimation in cluster randomized designs. *Psychological Methods*, 19:356–379, 2014.

[214] X.S. Liu. Statistical power in three-arm cluster randomized trials. *Evaluation and the Health Professions*, 37:470–487, 2014.

[215] J. Dziak, I. Nahum-Shani, and L. Collins. Multilevel factorial experiments for developing behavioral interventions: Power, sample size and resource considerations. *Psychological Methods*, 17:153–175, 2012.

[216] R. Hooper and L Bourke. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ*, 350:h2925, 2015.

[217] F. Gao, A. Earnest, D. Matchar, M. Campbell, and D. Machin. Sample size calculations for the design of cluster randomized trials: A summary of methodology. *Contemporary Clinical Trials*, 42:41–50, 2015.

[218] M. Kenward and J. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53:983–97, 1997.

[219] F. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2:10–4, 1946.

[220] P. Li and D. Redden. Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology*, 15:38, 2015.

[221] J. Johnson, S. Kreidler, D. Catellier, D. Murray, K. Muller, and D. Glueck. Recommendations for choosing an analysis method that controls type i error for unbalanced cluster sample designs with gaussian outcomes. *Statistics in Medicine*, 2015.

[222] G. kauermann and R Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96:1387–1396, 2001.

[223] M. Fay and B. Graubard. Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics*, 57:1198–1206, 2001.

[224] L. Mancl and T. DeRouen. A covariance estimator for gee with improved small-sample properties. *Biometrics*, 57:126–134, 2001.

[225] J. Morel, M. Bokossa, and N. Neerchal. Small sample correction for the variance of gee estimators. *Biometrical Journal*, 45:395–409, 2003.

[226] P. Li and D. Redden. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine*, 34:281–296, 2015.

Appendices

i (Chapter two) Protocol for the systematic review of sample size methodology

Systematic review of sample size methodology for cluster randomised trials

Protocol

Version 1.0 31/03/2011

Clare Rutterford
School of Medicine and Dentistry,
Queen Mary University of London,
c.m.rutterford@qmul.ac.uk

Contents

1.1 Background

1.1.1 An introduction to cluster randomised trials

Well designed and conducted Randomised Controlled Trials (RCT's) are seen as the gold standard in research design for the evaluation of new interventions for improving participant outcomes. In this research design we compare the outcomes of those participants who recieve the new intervention with a control group who receive either standard treatment or a placebo. In order to be able to reliably conclude the effect of the treatment the characteristics of these two groups of individuals should be similar so that the only systematic difference between them is the treatment received. This is acheieved by randomly allocating each individual to one of the treatment arms under study, this helps to avoid potential selection bias where participants with certain characteristics, such as severe disease, may end up more likley in one of the treatment arms [7].

There are some circumstances when it becomes necessary to randomise groups of individuals together, as a unit, rather than individually [4]. This process is known as cluster or group randomisation. The cluster may be large, such as a hospital or General Practice surgery, or may be smaller like a family. When we conduct a cluster randomised trial we may sample an entire cluster such as all participants registered at a General Practice, or take a subsample for inclusion into the trial. Throughout this report, when we refer to cluster size, we are specifically refering to the sample of the cluster that is to be included in the analysis, which may or may not be the entire cluster.

1.1.2 Sample size calculations in cluster randomised trials

In designing a randomised controlled trial we must determine the number of participants we need to recruit in order to be able to make meaningful and precise conclusions about our treatment effect. This is done through the use of sample size or power calculations. Power calculations require knowledge or decisions about the primary outcome measure, how the data will be analysed, the Type I error to be tolerated , the amount of statistical power required, and an estimate of the expected treatment difference [7]. Randomisation by cluster presents additional complexity in both the design and analysis, compared to individually randomised trials and the sample size calculation requires estimates of two more parameters, the cluster size and the intracluster correlation coefficient.

The response of individuals within a cluster are likely to be more similar than the responses of those from different clusters i.e. the data within a cluster is correlated and the data between clusters is assumed to be independent. The magnitude of this clustering is quantified by a parameter known as the intracluster correlation coefficient (ICC), usually deoted by ρ. When $\rho = 0$ we have, in effect, statistical independence between members of a cluster. When $\rho = 1$ the opposite is true and we have total dependence among members of a cluster. An estimate of the ICC is required for the sample size calculation. For a researcher this presents difficulties as historically ICCs have not always been explicitly published with the reults of CRT's. Where ICC's are provided they are often acompanied by wide confidence intervals, indicating a lack of precision. Further difficulties arise as to whether an adjusted or unadjusted ICC has been reported, what if any the adjustments are, and whether these adjustments are relevent to the reader.

The expected value of the ICC will vary with the choice of outcome measure and the type of unit under study. The outcomes from small units such as a family are likely to be more correlated than outcomes from larger units. It is important to note that each unit has an underlying ICC for a particular outcome. If you increase the number you sample from each unit you will increase the precision with which you can estimate the ICC but will not impact upon its underlying value.

The effect of clustering is that the information gained is less than an individually randomised trial of the same size, making it a less efficient design [2]. It was proposed by Donner, Birkett and Buck [3] that a sample size calculated assuming individual randomisation can be inflated by a Variance Inflation Factor (VIF), or Design Effect (DE)to provide the sample size required under cluster randomisation to reach the required level of statistical power. This design effect is given by

$$DE = 1 + (m-1)\rho \tag{1.1}$$

where m is the number of individuals per cluster and ρ is the intracluster correlation coefficient.

1.1.3 Motivation for this review

Donner and Klar [4] describe several methodological reviews of cluster randomised trials. The consensus of these reviews were that between-cluster variation was accounted for in the design of only around 19% of the CRTs studied and was accounted for at the analysis stage by only 57%.

The authors speculate that the reason why many CRT designs do not include formal sample size calculations may be due to

1. Inaccesability of formulas
2. Difficulties with estimation of the intracluster correlation, required for the calculation
3. CRTs enrolling large numbers of participants may give the misleading impression of extensive statistical power

With the recent extension of the CONSORT statement to provide guidelines for the reporting of cluster randomised trials [1] researchers are becoming more aware of cluster randomised trials and the need to account for the clustering at both the design and analysis stages. The CONSORT extension also reccomends that researchers report the ICC for their outcomes. This development will go some way towards enabling researchers to find appropriate ICC estimates for their sample size calculations, and ensuring that the appropriate sample size methodology is implemented.

However we still need to tackle the issue of inaccesability of formulae. The widely used methodology for sample size calculation in cluster randomised trials is to calculate the sample size under the assumption of an individually randomised trial, with equivalent design features, and multiply this by the design effect, equation 1.1. This standard methodology is described for trials where the analysis may be a comparison of means, proportions, or incidence rates and the design may be completely randomised, matched pair or stratified [4].

This methodology assumes that the size of each cluster is the same, a condition which does not tend to hold true for the majority of trial designs.

1.2 Objectives of the review

- To produce a thorough review of the exisiting state of knowledge of sample size calculations for cluster randomised trials
- To identify gaps in the knowledge, particularly for ordinal, count and time-to-event data.

1.3 Search methodology

1.3.1 Inclusion/exclusion criteria

When searching for papers on sample size calculations in cluster randomised trials, from an examination of the literature, we are likely to encounter the following main types of paper:

1. Those which derive some or all methodological aspects of sample size formulae. This may be done using either a Bayesian of frequentist approach.
2. Those which provide suggested adaptations and extensions to previously derived formulae.
3. Papers evaluating or comparing current methodology.
4. Papers which focus on the calculation/use of the intracluster correlation or coefficient of variation.
5. Papers with a broader focus that discuss design features in general for cluster randomised trials.
6. Overviews or summaries of the current methodology in a particular area.
7. Papers with a general discussion on the effects of clustering upon sample size and power.
8. Papers describing application tools for implentation of the methodology.

9. Protocols or design papers for specific cluster randomised trials, describing the sample size calculation, as applied in the trial design.

10. Reports of results from cluster randomised trials.

To fulfil the objectives of this review we aim to collate articles of type 1-6, papers reporting specific trials are likely to be numerous and an examination of current practices is beyond the scope of this review. The above list is not exhaustive, however it does show that the types of paper can vary greatly making it necessary to take a broad approach to defining inclusion and exclusion criteria.

we have defined inclusion and exclusion criteria based on the following categories:
Data Type: Articles will be eligible for the review if the methods described can be applied to clustered or correlated data, where the correlation of data is present in at least one of the comparison groups, and hence could be applied to cluster randomised trials.

Content: We shall exclude papers reporting design or results of **specific** randomised trials, or those which provide a very general discussion on the effects of clustering or correlated data.

Methodology: Articles will be eligible for the review if they discuss any aspect relating to the methodology of sample size or power. This includes those papers which may discuss or simulate only components of the sample size.

Methodology may be described under a frequentist or Bayesian approach, both shall be included in this review.

Language: Papers included in the review will be restricted to those written in English. To conduct this search in other languages would not be feasible within our resources due to time and cost involved in translation of manuscripts.

All articles retrived during the serach will be examined for eligibility using these criteria, as summarised in table 1.1. The inclusion/exclusion criteria has been piloted on 20 search results in order

	Inclusion	**Exclusion**
Data	Clustered data	Non-clustered
Content	Methodological piece or simulation study. Bayesian or frequentist approach.	Non-methodological. Description of a specific trial.
Methodology	Aspect of sample size	No sample size
Language	English	Non-English

Table 1.1: Inclusion/exclusion critera

to identify any need for revision.

1.3.2 Review sources

This review will be conducted using electronic searches, personal collections of articles on sample size in CRT's, key text books on cluster randomised trials, and discussions with experts in the field of cluster randomised trials.

The electronic databases used in the search will be PubMed and Web of Science.

PubMed database:

The MEDLINE database is the largest component of the PubMed database. The subject scope of searches conducted in the MEDLINE database is biomedicine and health articles dating from 1946 onwards. Articles in the MEDLINE database are indexed using Medical Subject Heading (MeSH) terms. MeSH terms are used by indexers to provide a consistent way to index articles which may use different terminology to describe the same concepts.

In addition to the MEDLINE database PubMed also contains additional references such as those which are yet to be indexed with MeSH and citations that preceed the date that a journal was selected for MEDLINE indexing.

The search terms for the electronic databases was guided by the personal collection of articles from

Professor Sandra Eldridge. I examined the list of 41 papers available in the personal collection of articles, appendix A.1. Thirty five of these were available via PubMed, and so, had associated MeSH terms. One hundred and ninety nine MeSH terms were associated with these 35 papers. The most frequent being “randomized controlled trials as Topic” (n=28), “cluster analysis” (n=20), “research design” (n=16), and “sample size”(n=20).

Due to the wide variety of terminology we can expect to see in our target papers the use of MeSH terms would be ideal. However it is clear that the use of MeSH terms alone would not be sufficient to provide us with a comprehensive review of the literature. Therefore we will suppliment our search with further search terms based on freetext.

The definition of the MeSH term “Cluster analysis” does not include reference to cluster randomised trials [1]. Instead it describes cluster analysis which is an unrelated method of analysis. However from an examination of the search results retrieved with only this term it is clear that cluster randomised trials have been indexed with this term. Therefore a search using the term “cluster analysis” alone will retrieve some results irrelevant to this review. There is no alternative MeSH term for cluster randomised trials.

Similarly a search on “randomized controlled trials as topic” will retrieve clinical trial reports, there may also be instances where papers describe sample size calculations for clustered data which may not be in the context of clinical trials, but could be applied to them. With these in mind I shall not restrict the search using the MeSH term for randomised controlled trials or research design.

The MeSH terms I have therefore chosen to search on is the combination of “cluster analysis” and “sample size”.

The freetext search will be conducted on the article title only. This is to improve the efficency of

[1] MeSH description of cluster analysis is “A set of statistical methods used to group variables or observations into strongly inter-related subgroups. In epidemiology it may be used to analyse a closely grouped series of events or cases of disease or other health related phenonomen with well-defined distribution patterns in relation to time or place or both”

our search by avoiding extraction of articles reporting results from clinical trials which are likely to be identified when using a broader search strategy on title and abstract.

It is expected that the term “sample size” will be the term most frequently used in the title of relevent articles. However the term “Design effect” is common in cluster randomised trials. The term “power” is commonly used in discussion of sample size, however as this has a wide number of meanings, we will retrict its search to those trials which additionaly mention clustering in some form.

The list below is a summary of the search terms for the PubMed database.

Search terms:

1. cluster analysis[MeSH] AND sample size[MeSH]
2. “sample size” [Title]
3. “design effect”[Title] OR “design effects”[Title] OR “variance inflation factor”[Title]
4. (design*[Title] OR plan*[Title] OR siz*[Title]) AND cluster*[Title] [2]
5. power [title] AND cluster*[Title]
6. “intraclass correlation*”[Title] OR “interclass correlation*”[Title] OR “intracluster correlation*”[Title] OR “coefficient of variation”[Title] OR “between cluster”[Title]
7. coefficient[Title] AND variation[Title]
8. (design[Title] OR matching[Title]) AND community[Title]
9. power[Title] AND correlated[Title]
10. number[Title] AND clusters[Title]
11. 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10

Monthly email notifications based on these search items will be set up within the database for the duration of the project in order to keep informed of new developments.

Within PubMed the word “of” is a stopword and is ignored when searching for a phrase. Therefore rather than search on the phrase “coefficient of variation” I have chosen to search on the seperate components, as described in 7 above. Similarly with the phrase “number of clusters”.

Web of Science:

The Web of Science database contains seven citation databases. We will conduct this search using the “Science citation index expanded database” only. This database contains articles accross 150 scientific disciplines from the year 1970.

The Web of Science does not index articles with MeSH so this search will not be performed. We will conduct the search based upon a search of the artice titles using the free text search items as described for the PubMed search.

Web of Science treats the word “of” as a placeholder when contained in the search for a phrase. Therefore a search with the phrase “coefficient of variation” will retrieve all results where the two terms are seperated by one word, such as “coefficient and variation” and “coefficient of variation”. We shall use this search rather than the search on the two seperate components which was required when searching Pubmed. The word “between” is similarly treated as a placeholder in Web of Science. This means that a search on ”“between cluster” will retrieve any article with the word cluster in the title. This retrieves a vast number of articles, with many irrelevant articles relating to disease clusters and laborartory data. Therefore the search term “between cluster” will be excluded from the Web of Science search.

The searches within Web of Science will additionally be refined by limiting the search to articles

[2] The “*” is used as a wildcard to search on all words that are formed from the truncated word entered. For example siz* will find all words formed from siz such as “size” and “sizing”.

only, thereby excluding conference proceedings and letters etc.

Monthly notifications based on these search items will be set up within the database for the duration of the project in order to keep informed of new developments.

Text books:

There are currently four books on cluster randomised trials, from Donner and Klar [4], Murray [6], Ukoumunne [8], and Hayes [5]. These will be reviewed, after the electronic search and data extraction is complete, for any methodology not already present in the review.

Expert opinion:

The following experts in the field will be approached to provide opinion on the final list of included papers to determine their thoughts upon its coverage. This list of experts may evolve as we conduct the review and identify other individuals making significant contributions to the field.

Obi Ukoumunne
Mike Campbell
Steven Tereenstra
G.J.P. van Breukelen
Allan Donner

1.3.3 Search strategy

The search strategy is summarised in figure 1.1.

The search terms will be implemented in the electronic databases.The title of each paper will be examined for inclusion into the review following the inclusion/exclusion criteria described in table 1.1. If eligibility is unclear from the title, the abstract and then full text will be examined until eligibility is clear. Any papers where eligibility is still unclear after examination of the full text will be discussed with Professor Sandra Eldridge and Dr Andrew Copas.

The full text will be extracted for all papers eligible for inclusion in the review.The references provided in each article will be examined for eligibility into the review following the process described above.

We will conduct a search in both PubMed and Web of Science on the authors of each paper included in the review and examine the results against the inclusion and exclusion criteria for potential inclusion into the review.

Any new eligible articles will be added to the review and the process of reference and author searching will continue until it is demmed that we have reached sufficient saturation of new methodology or concepts.

1.3.4 Validation of search methodology

We have attempted to validate the appropriateness of the proposed search strategy in two ways.

First the titles of the papers included in the personal collection were compared to our search terms to see if the papers would have been identified from the proposed electronic search. In the majority of cases it was found that the papers would have been identified. For those which were not identified we examined the citations for these papers, in all cases these articles were cited by a paper that would have been included in our review based upon the search terms. As we are searching references of all included papers these papers would have been identified.

For our second validation technique we wanted to compare our search strategy with a “gold standard” search. We conducted a search of the Statistics in Medicine journal as a “gold standard” search as it was considered most likely to contain articles on methodology of sample size calculations. Using

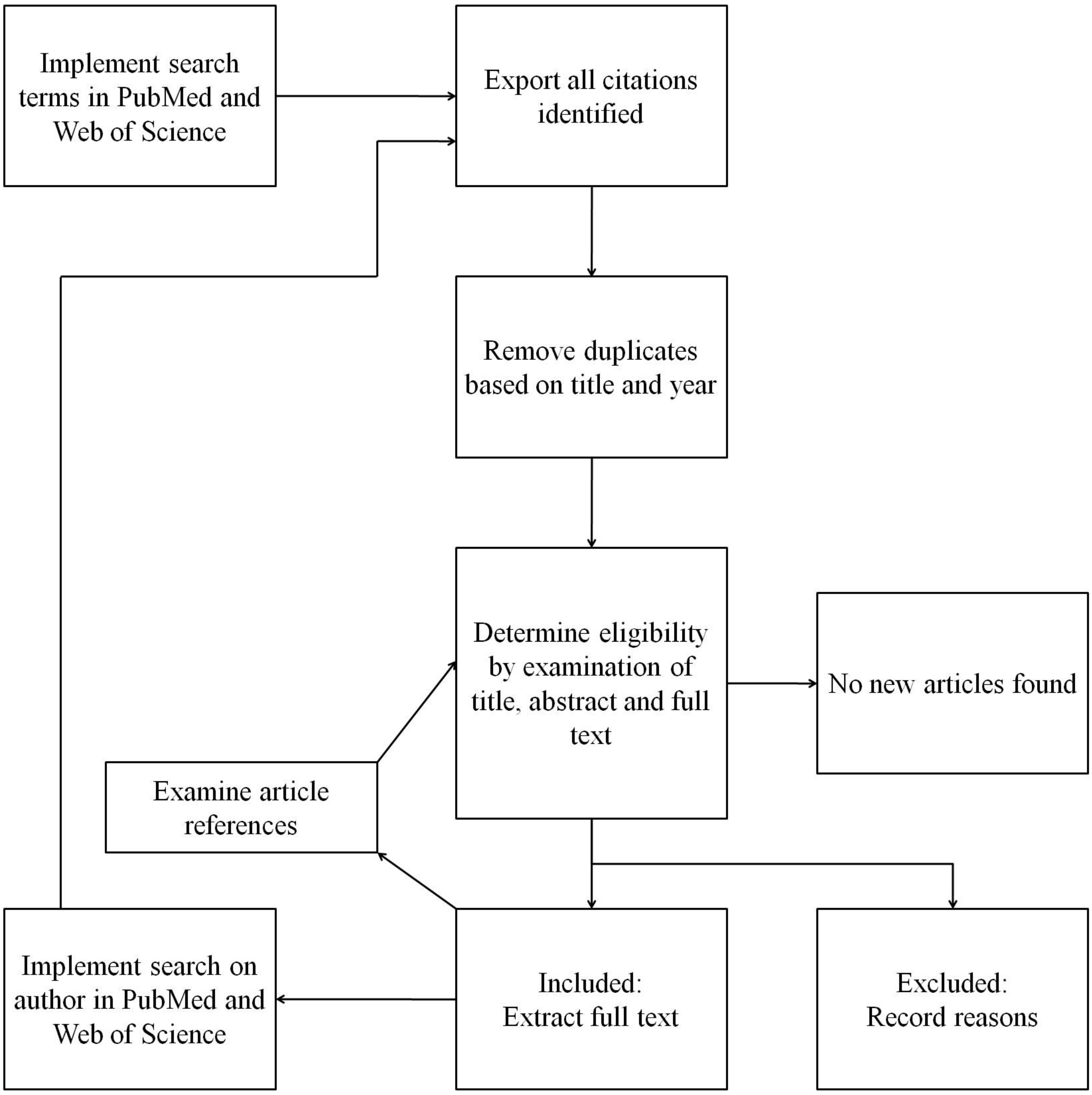

Implement search terms in PubMed and Web of Science
Export all citations identified
Remove duplicates based on title and year
Determine eligibility by examination of title, abstract and full text
No new articles found
Examine article references
Implement search on author in PubMed and Web of Science
Included: Extract full text
Excluded: Record reasons

Figure 1.1: Flow diagram of literature search process

PubMed all articles published in Statistics and Medicine from 1982 to March 2011 were retrieved. Each article title was examined; any titles with reference to clustered data or cluster randomised trials were exported. This search produced just less than 200 results. These results were then further examined, on title alone, to reduce the list to those which would be eligible into the review or would warrant further examination. These results were then compared to our proposed search terms to see if they would have been identified. This highlighted the fact that there were a number of articles about the intracluster correlation coefficient which were currently not included in our search terms. This was added to the search terms for the review along with its alternative descriptions.

Many of the early articles made reference to community intervention trials rather than cluster randomized trials, for which they are now more commonly known. The term community was added to the search terms in combination with design or matching to pick up these articles. “Number of clusters” was also a common phrase used to describe sample size and so was added to the search terms.

1.4 Data management

A database will be created in Microsoft Access 2010 to store and organise the details of all of the references located as part of the systematic review.

The database contains a data entry form to allow the user to enter additional information, navigate through the database and categorise the papers for inclusion into the review or for further follow up or discussion. The variable names, description and entry values can be found in appendix B.1 alongside a screen print of the data entry form.

Naming conventions when exporting the results from the two electronic databases, PubMed and Web of Science need to be standardised to ensure compatibility with each other and the Access database. To do this the located references will first be extracted to Excel, where variable names can be changed to those used in the Access database. The Excel files can then be directly imported in a batch to the Access database, where each imported result will be given a unique numeric ID auto generated in the database.

Within web of science the following variables for each paper can be extracted in separate fields in a format suitable for importing into Excel: title, authors, journal, publication year, volume, issue, start page, end page, publication year, and abstract.

PubMed does not provide an easy method with which to export the search results in a format compatible with Excel. However they can be imported to Excel using a tool called FLink (Frequency-weighted Links) provided by the National Center for Biotechnology information.

This tool exports the following variables for each paper into separate columns: PubMed ID; Authors, Year of publication; month of publication; title; and a summary of all this information in one column.

Duplicates, based upon year and title of publication within the Access database will be located and removed using the “find duplicates” query in Access. Preference for inclusion will be given to the more detailed Web of Science entries.

For each record the user will navigate through the database, categorising the papers for inclusion or exclusion into the review. The reason for exclusion at each stage will be recorded by the user from a list of options which have been proposed based upon a test sample of 20 search results. Where the list does not provide an adequate description the user may add a reason using free text.

The information missing from PubMed produced results, in particular the abstract, will only be added to the database if the paper is eligible for the review or where abstract information is required to aid determination of eligibility.

Once all original papers have been categorised the full text will be located for all eligible papers. These will be stored in both hardcopy and electronic format, this may involve scanning of papers currently not available in electronic format. paper copies will be stored in alphabetical order based upon the surname of the first author.

Further searches on the reference lists of each paper will be conducted by hand. As these reference lists are not always available electronically only those deemed eligible will be entered into the database.

Further searches on the first author surname with first initial, will be conducted in both PubMed and Web of Science and these results will be entered into the database following the methodology of the original search.

This process will continue for new papers found until either no more additional papers are located or it is deemed that further searching will not significantly improve the methodological coverage of the review.

The process of data extraction to the Access database and data entry has been tested prior to the start of the review.

1.5 Data extraction

A data extraction template will be devloped and piloted on 5 papers eligible for inclusion into the review. The test papers were selected in order to test the extraction template on different types and styles of article. This highlighted the need for some variation in the extraction template for different styles of articles. All extraction templates can be found in Appendix C.

The data extraction will be mainly qualitative. The completed data extraction form will be stored both electronically and in hard copy format with each article and will be linked to the datbase entry through the inclusion of the database generated identification number.

The data extraction process will be validated by a double data extraction for 10 randomly selected articles. Five will be extracted by Professor Eldridge and five by Dr Copas. A discussion will take place to agree any differences in interpretation.

During the data extraction process any articles with ambiguity will be discussed with SA and AC.

1.6 Synthesis of results

The results of the search strategy will be summarised in a flow diagram as in figure 1.2.

The results of the search will be provided in a narrative form. The discussion will be similarly structured to the systematic review of organisation-based interventions in health care conducted by OC Ukoumunne [8]. Each trial design will be discussed in turn (with equal and unequal allocation ratio handled in turn), and within each design the methodological aspects for different outcome types will be summarised. Proposed summary tables of the information are provided in appendix E.

- completly randomised design
  - Continuous data

    Analysis methods, measures of correlation, measures of cluster size
  - Binary data

    Analysis methods, measures of correlation, measures of cluster size
  - Ordinal data

    Analysis methods, measures of correlation, measures of cluster size
  - Time-to-event data

    Analysis methods, measures of correlation, measures of cluster size
  - Count data

    Analysis methods, measures of correlation, measures of cluster size
- matched-pairs design . . .
- stratfied design . . .
- Other designs (e.g. crossover, factorial) . . .

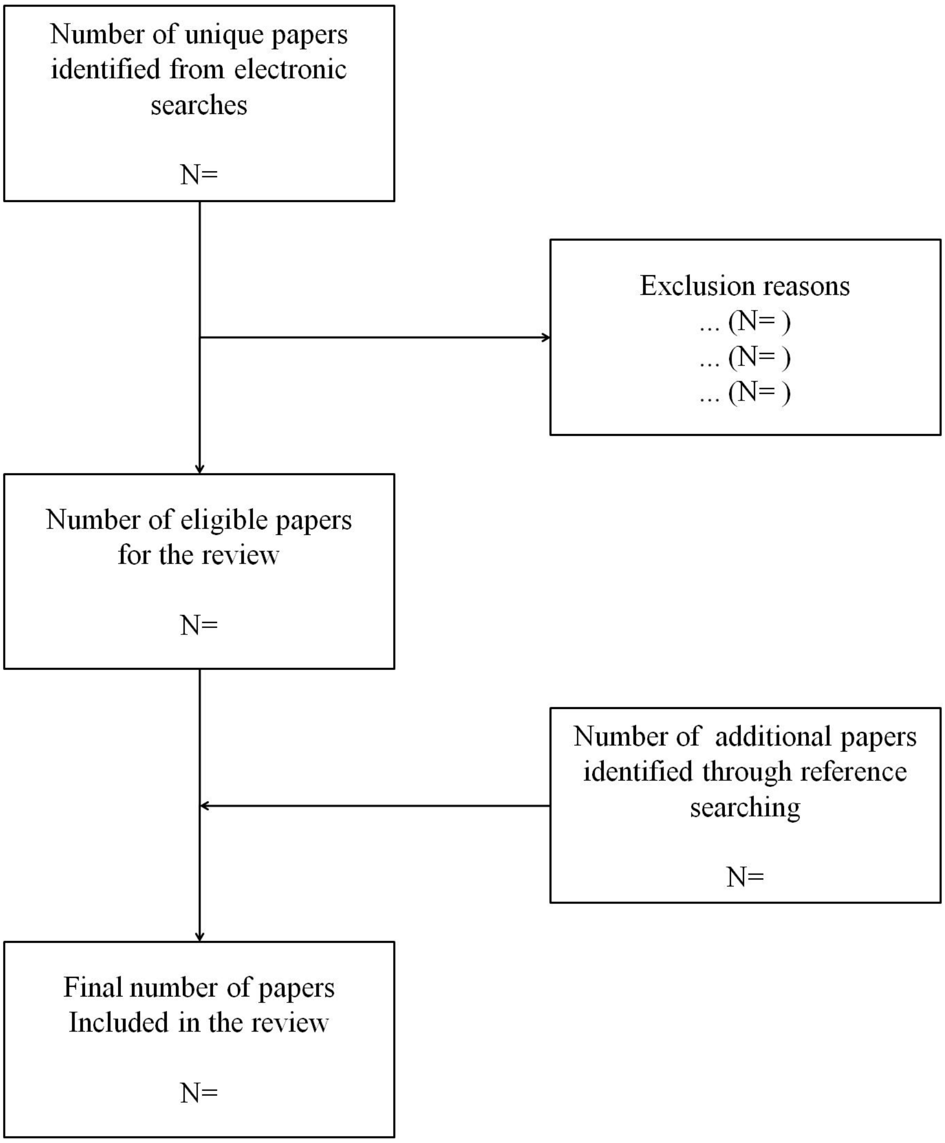

Number of unique papers identified from electronic searches
N=
Exclusion reasons
... (N=)
... (N=)
... (N=)
Number of eligible papers for the review
N=
Number of additional papers identified through reference searching
N=
Final number of papers Included in the review
N=

Figure 1.2: Flow diagram of literature search results

Bibliography

[1] Marion K Campbell, Diana R Elbourne, and Douglas Altman. Consort statement: extension to cluster randomised trials. *British Medical Journal*, 328:702–708, 2004.

[2] J. Cornfield. Randomization by group: a formal analysis. *American Journal of Epidemiology*, 108:100–102, 1978.

[3] A Donner, Birkett N, and Buck C. Randomization by cluster. sample size requirements and analysis. *American Journal of Epidemiology*, 114:906–914, 1981.

[4] Allan Donner and Neil Klar. *Design and Analysis of Cluster Randomisation Trials in Health Research*. Arnold, London, 2000.

[5] R. Hayes and L Moulton. *Cluster randomised trials*. Chapman & Hall/CRC, Boca Raton, 2009.

[6] David M. Murray. *Design and analysis of group-randomized trials*. Oxford University Press, New York, 1998.

[7] Stuart J Pocock. *Clinical Trials A practical Approach*. John Wiley & Sons Inc, Chichester, 1983.

[8] O. C. Ukoumunne, M. C. Gulliford, S. Chinn, J. A. Sterne, and P. G Burney. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assessment*, 3 (5), 1999.

Appendix A: Personal collection

[1] P. M. E. Altham. Discrete variable analysis for individuals grouped into families. *Biometrika*, 63(2):263–269, 1976.

[2] N. B. Baskerville, W. Hogg, and J. Lemelin. The effect of cluster randomization on sample size in prevention research. *J Fam Pract*, 50(3):W241–6, 2001.

[3] C. Butler and M. Bachmann. Design and analysis of studies evaluating smoking cessation interventions where effects vary between practices or practitioners. *Fam Pract*, 13(4):402–7, 1996.

[4] M. K. Campbell, S. Thomson, C. R. Ramsay, G. S. MacLennan, and J. M. Grimshaw. Sample size calculator for cluster randomized trials. *Comput Biol Med*, 34(2):113–25, 2004.

[5] J. E. Cohen. Distribution of chi-squared statistic under clustered sampling from contingency-tables. *Journal of the American Statistical Association*, 71(355):665–670, 1976.

[6] J. Cornfield. Randomization by group - formal analysis. *American Journal of Epidemiology*, 108(2):100–102, 1978.

[7] A. Donner. Approaches to sample size estimation in the design of clinical trials–a review. *Stat Med*, 3(3):199–214, 1984.

[8] A. Donner. Sample-size requirements for stratified cluster randomization designs. *Statistics in Medicine*, 11(6):743–750, 1992.

[9] A. Donner, N. Birkett, and C. Buck. Randomization by cluster - sample-size requirements and analysis. *American Journal of Epidemiology*, 114(6):906–914, 1981.

[10] 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.

[11] T. N. Flynn, E. Whitley, and T. J. Peters. Recruitment strategies in a cluster randomized trial - cost implications. *Statistics in Medicine*, 21(3):397–405, 2002.

[12] M. H. Gail, D. P. Byar, T. F. Pechacek, and D. K. Corle. Aspects of statistical design for the community intervention trial for smoking cessation (commit). *Control Clin Trials*, 13(1):6–21, 1992.

[13] L. Guittet, B. Giraudeau, and P. Ravaud. A priori postulated and real power in cluster randomized trials: mind the gap. *BMC Med Res Methodol*, 5:25, 2005.

[14] L. Guittet, P. Ravaud, and B. Giraudeau. Planning a cluster randomized trial with unequal cluster sizes: practical issues involving continuous outcomes. *BMC Med Res Methodol*, 6:17, 2006.

[15] R. J. Hayes and S. Bennett. Simple sample size calculation for cluster-randomized trials. *Int J Epidemiol*, 28(2):319–26, 1999.

[16] S. A. Hendricks, J. T. Wassell, J. W. Collins, and S. L. Sedlak. Power determination for geographically clustered data using generalized estimating equations. *Stat Med*, 15(17-18):1951–60, 1996.

[17] F. Y. Hsieh. Sample size formulae for intervention studies with the cluster as unit of randomization. *Stat Med*, 7(11):1195–201, 1988.

[18] F. Y. Hsieh, P. W. Lavori, H. J. Cohen, and J. R. Feussner. An overview of variance inflation factors for sample-size calculation. *Eval Health Prof*, 26(3):239–57, 2003.

[19] S. H. Jung, S. H. Kang, and C. Ahn. Sample size calculations for clustered binary data. *Stat Med*, 20(13):1971–82, 2001.

[20] D. Kang, J. B. Schwartz, and D. Verotta. A sample size computation method for non-linear mixed effects models with applications to pharmacokinetics models. 2004.

[21] S. H. Kang, C. W. Ahn, and S. H. Jung. Sample size calculation for dichotomous outcomes in cluster randomization trials with varying cluster size. *Drug Information Journal*, 37(1):109–114, 2003.

[22] S. M. Kerry and J. M. Bland. Sample size in cluster randomisation. *BMJ*, 316(7130):549, 1998.

[23] S. M. Kerry and J. M. Bland. Trials which randomize practices ii: sample size. *Family Practice*, 15(1):84–87, 1998.

[24] S. Lake, E. Kammann, N. Klar, and R. Betensky. Sample size re-estimation in cluster randomization trials. *Statistics in Medicine*, 21(10):1337–1350, 2002.

[25] E. W. Lee and N. Dubin. Estimation and sample size considerations for clustered binary responses. *Stat Med*, 13(12):1241–52, 1994.

[26] A. K. Manatunga and S. Chen. Sample size estimation for survival outcomes in cluster-randomized studies with small cluster sizes. *Biometrics*, 56(2):616–621, 2000.

[27] A. K. Manatunga, M. G. Hudgens, and S. D. Chen. Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal*, 43(1):75–86, 2001.

[28] S. M. McKinlay. Cost-efficient designs of cluster unit trials. *Prev Med*, 23(5):606–11, 1994.

[29] R. M. Mickey, G. D. Goodwin, and M. C. Costanza. Estimation of the design effect in community intervention studies. *Stat Med*, 10(1):53–64, 1991.

[30] L. A. Moye. Sizing clinical trials with variable endpoint event rates. *Stat Med*, 16(20):2267–82, 1997.

[31] A. Munoz and B. Rosner. Power and sample-size for a collection of 2x2 tables. *Biometrics*, 40(4):995–1004, 1984.

[32] J. M. Neuhaus and M. R. Segal. Design effects for binary regression-models fitted to dependent data. *Statistics in Medicine*, 12(13):1259–1268, 1993.

[33] W. Pan. Sample size and power calculations with correlated binary data. *Controlled Clinical Trials*, 22(3):211–227, 2001.

[34] J. S. Preisser, M. L. Young, D. J. Zaccaro, and M. Wolfson. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med*, 22(8):1235–54, 2003.

[35] M. M. Shoukri and S. W. Martin. Estimating the number of clusters for the analysis of correlated binary response variables from unbalanced data. *Stat Med*, 11(6):751–60, 1992.

[36] D. J. Torgerson and M. K. Campbell. Economic notes - use of unequal randomisation to aid the economic efficiency of clinical trials. *British Medical Journal*, 321(7263):759–759, 2000.

[37] G. J. P. van Breukelen, M. J. J. M. Candel, and M. P. F. Berger. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*, 26(13):2589–2603, 2007.

[38] J. Whitehead. Sample size calculations for ordered categorical data. *Stat Med*, 12(24):2257–71, 1993.

[39] R. F. Woolson, J. A. Bean, and P. B. Rojas. Sample size for case-control studies using cochran's statistic. *Biometrics*, 42(4):927–32, 1986.

[40] T. L. Xie and J. Waksman. Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. *Statistics in Medicine*, 22(18):2835–2846, 2003.

[41] K. H. Zou and S. L. Normand. On determination of sample size in hierarchical binomial models. *Stat Med*, 20(14):2163–82, 2001.

Appendix B: Validation set

[1] J. M. Albert. Estimating efficacy in clinical trials with clustered binary responses. *Stat Med*, 21(5):649–61, 2002.

[2] P. C. Austin. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Stat Med*, 26(19):3550–65, 2007.

[3] L. A. Beckett. Community-based studies of alzheimer's disease: statistical challenges in design and analysis. *Stat Med*, 19(11-12):1469–80, 2000.

[4] D. G. Bonett. Sample size requirements for estimating intraclass correlations with desired precision. *Stat Med*, 21(9):1331–5, 2002.

[5] M. J. Campbell, A. Donner, and N. Klar. Developments in cluster randomized trials and statistics in medicine. *Stat Med*, 26(1):2–19, 2007.

[6] M. K. Campbell, J. Mollison, and J. M. Grimshaw. Cluster trials in implementation research: estimation of intracluster correlation coefficients and sample size. *Stat Med*, 20(3):391–9, 2001.

[7] M. J. Candel and G. J. Van Breukelen. Varying cluster sizes in trials with clusters in one treatment arm: sample size adjustments when testing treatment effects with linear mixed models. *Stat Med*, 28(18):2307–24, 2009.

[8] M. J. Candel and G. J. Van Breukelen. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order pql mixed logistic regression. *Stat Med*, 29(14):1488–501, 2010.

[9] J. C. Cappelleri and N. Ting. A modified large-sample approach to approximate interval estimation for a particular intraclass correlation coefficient. *Stat Med*, 22(11):1861–77, 2003.

[10] B. Carter. Cluster size variability and imbalance in cluster randomized controlled trials. *Stat Med*, 29(29):2984–93, 2010.

[11] A. Donner. Sample size requirements for stratified cluster randomization designs. *Stat Med*, 11(6):743–50, 1992.

[12] G. Duran Pacheco, J. Hattendorf, Jr. Colford, J. M., D. Mausezahl, and T. Smith. Performance of analytical methods for overdispersed counts in cluster randomized trials: sample size, degree of clustering and imbalance. *Stat Med*, 28(24):2989–3011, 2009.

[13] S. Eldridge, C. Cryer, G. Feder, and M. Underwood. Sample size calculations for intervention trials in primary care randomizing by primary care group: an empirical illustration from one proposed intervention trial. *Stat Med*, 20(3):367–76, 2001.

[14] Z. Feng and J. E. Grizzle. Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculation. *Stat Med*, 11(12):1607–14, 1992.

[15] T. N. Flynn, E. Whitley, and T. J. Peters. Recruitment strategies in a cluster randomized trial–cost implications. *Stat Med*, 21(3):397–405, 2002.

[16] B. Giraudeau. Model mis-specification and overestimation of the intraclass correlation coefficient in cluster randomized trials. *Stat Med*, 25(6):957–64, 2006.

[17] B. Giraudeau and J. Y. Mary. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Stat Med*, 20(21):3205–14, 2001.

[18] B. Giraudeau, P. Ravaud, and A. Donner. Sample size calculation for cluster randomized cross-over trials. *Stat Med*, 27(27):5578–85, 2008.

[19] M. Gonen. Sample size and power for mcnemar's test with clustered data. *Stat Med*, 23(14):2283–94, 2004.

[20] S. A. Hendricks, J. T. Wassell, J. W. Collins, and S. L. Sedlak. Power determination for geographically clustered data using generalized estimating equations. *Stat Med*, 15(17-18):1951–60, 1996.

[21] M. Heo, Y. Kim, X. Xue, and M. Y. Kim. Sample size requirement to detect an intervention effect at the end of follow-up in a longitudinal cluster randomized trial. *Stat Med*, 29(3):382–90, 2010.

[22] M. Heo and A. C. Leon. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials. *Stat Med*, 28(6):1017–27, 2009.

[23] F. Y. Hsieh. Sample size formulae for intervention studies with the cluster as unit of randomization. *Stat Med*, 7(11):1195–201, 1988.

[24] I. Irigoien and C. Arenas. Inca: new statistic for estimating the number of clusters and identifying atypical units. *Stat Med*, 27(15):2948–73, 2008.

[25] S. H. Jung and C. Ahn. Sample size estimation for gee method for comparing slopes in repeated measurements data. *Stat Med*, 22(8):1305–15, 2003.

[26] S. H. Jung, S. H. Kang, and C. Ahn. Sample size calculations for clustered binary data. *Stat Med*, 20(13):1971–82, 2001.

[27] S. M. Kerry and J. M. Bland. Unequal cluster sizes for trials in english and welsh general practice: implications for sample size calculations. *Stat Med*, 20(3):377–90, 2001.

[28] H. Y. Kim, J. M. Williamson, and C. M. Lyles. Sample-size calculations for studies with correlated ordinal outcomes. *Stat Med*, 24(19):2977–87, 2005.

[29] N. Klar and A. Donner. The merits of matching in community intervention trials: a cautionary tale. *Stat Med*, 16(15):1753–64, 1997.

[30] N. Klar and A. Donner. Current and future challenges in the design and analysis of cluster randomization trials. *Stat Med*, 20(24):3729–40, 2001.

[31] S. Lake, E. Kammann, N. Klar, and R. Betensky. Sample size re-estimation in cluster randomization trials. *Stat Med*, 21(10):1337–50, 2002.

[32] E. W. Lee and N. Dubin. Estimation and sample size considerations for clustered binary responses. *Stat Med*, 13(12):1241–52, 1994.

[33] A. C. Leon, M. Heo, J. J. Teres, and T. Morikawa. Statistical power of multiplicity adjustment strategies for correlated binary endpoints. *Stat Med*, 26(8):1712–23, 2007.

[34] J. D. Lewsey. Comparing completely and stratified randomized designs in cluster randomized trials when the stratifying factor is cluster size: a simulation study. *Stat Med*, 23(6):897–905, 2004.

[35] A. Liu, W. J. Shih, and E. Gehan. Sample size and power determination for clustered repeated measurements. *Stat Med*, 21(12):1787–801, 2002.

[36] T. Loeys, S. Vansteelandt, and E. Goetghebeur. Accounting for correlation and compliance in cluster randomized trials. *Stat Med*, 20(24):3753–67, 2001.

[37] K. J. Lui. A note on the effect of the intraclass correlation in the multiple reading procedure with a unanimity rule. *Stat Med*, 11(2):209–18, 1992.

[38] D. C. Martin, P. Diehr, E. B. Perrin, and T. D. Koepsell. The effect of matching on the power of randomized community intervention studies. *Stat Med*, 12(3-4):329–38, 1993.

[39] W. L. May and W. D. Johnson. The validity and power of tests for equality of two correlated proportions. *Stat Med*, 16(10):1081–96, 1997.

[40] I. U. Mian and M. M. Shoukri. Statistical analysis of intraclass correlations from multiple samples with applications to arterial blood pressure data. *Stat Med*, 16(13):1497–514, 1997.

[41] R. M. Mickey, G. D. Goodwin, and M. C. Costanza. Estimation of the design effect in community intervention studies. *Stat Med*, 10(1):53–64, 1991.

[42] M. Moerbeek. Power and money in cluster randomized trials: when is it worth measuring a covariate? *Stat Med*, 25(15):2607–17, 2006.

[43] H. Moore, C. Summerbell, A. Vail, D. C. Greenwood, and A. J. Adamson. The design features and practicalities of conducting a pragmatic cluster randomized trial of obesity management in primary care. *Stat Med*, 20(3):331–40, 2001.

[44] R. Muller and P. Buttner. A critical discussion of intraclass correlation coefficients. *Stat Med*, 13(23-24):2465–76, 1994.

[45] J. M. Neuhaus and M. R. Segal. Design effects for binary regression models fitted to dependent data. *Stat Med*, 12(13):1259–68, 1993.

[46] I. Perisic and B. Rosner. Comparisons of measures of interclass correlations: the general case of unequal group size. *Stat Med*, 18(12):1451–66, 1999.

[47] G. Piaggio, G. Carroli, J. Villar, A. Pinol, L. Bakketeig, P. Lumbiganon, P. Bergsjo, Y. Al-Mazrou, H. Ba'aqeel, J. M. Belizan, U. Farnot, and H. Berendes. Methodological considerations on the design and analysis of an equivalence stratified cluster randomization trial. *Stat Med*, 20(3):401–16, 2001.

[48] J. S. Preisser, B. Lu, and B. F. Qaqish. Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations. *Stat Med*, 27(27):5764–85, 2008.

[49] J. S. Preisser, M. L. Young, D. J. Zaccaro, and M. Wolfson. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med*, 22(8):1235–54, 2003.

[50] G. M. Raab and I. Butcher. Balance in cluster randomized trials. *Stat Med*, 20(3):351–65, 2001.

[51] S. Ren, S. Yang, and S. Lai. Intraclass correlation coefficients and bootstrap methods of hierarchical binary outcomes. *Stat Med*, 25(20):3576–88, 2006.

[52] D. E. Schaubel. Variance estimation for clustered recurrent event data with a small number of clusters. *Stat Med*, 24(19):3037–51, 2005.

[53] M. M. Shoukri and S. W. Martin. Estimating the number of clusters for the analysis of correlated binary response variables from unbalanced data. *Stat Med*, 11(6):751–60, 1992.

[54] D. J. Spiegelhalter. Bayesian methods for cluster randomized trials with continuous responses. *Stat Med*, 20(3):435–52, 2001.

[55] S. G. Thompson. The merits of matching in community intervention trials: a cautionary tale. *Stat Med*, 17(18):2149–52, 1998.

[56] S. G. Thompson, S. D. Pyke, and R. J. Hardy. The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. *Stat Med*, 16(18):2063–79, 1997.

[57] A. Thomson, R. Hayes, and S. Cousens. Measures of between-cluster variability in cluster randomized trials with binary outcomes. *Stat Med*, 28(12):1739–51, 2009.

[58] L. Tian. Inferences on the common coefficient of variation. *Stat Med*, 24(14):2213–20, 2005.

[59] L. Tian. On confidence intervals of a common intraclass correlation coefficient. *Stat Med*, 24(21):3311–8, 2005.

[60] J. M. Tielsch and Jr. West, K. P. Cost and efficiency considerations in community-based trials of vitamin a in developing countries. *Stat Med*, 9(1-2):35–41; discussion 41–3, 1990.

[61] X. M. Tu, J. Kowalski, P. Crits-Christoph, and R. Gallop. Power analyses for correlations from clustered study designs. *Stat Med*, 25(15):2587–606, 2006.

[62] X. M. Tu, J. Kowalski, J. Zhang, K. G. Lynch, and P. Crits-Christoph. Power analyses for longitudinal trials and other clustered designs. *Stat Med*, 23(18):2799–815, 2004.

[63] R. M. Turner, R. Z. Omar, and S. G. Thompson. Constructing intervals for the intracluster correlation coefficient using bayesian modelling, and application in cluster randomized trials. *Stat Med*, 25(9):1443–56, 2006.

[64] R. M. Turner, A. T. Prevost, and S. G. Thompson. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med*, 23(8):1195–214, 2004.

[65] O. C. Ukoumunne. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Stat Med*, 21(24):3757–74, 2002.

[66] O. C. Ukoumunne, A. C. Davison, M. C. Gulliford, and S. Chinn. Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Stat Med*, 22(24):3805–21, 2003.

[67] G. J. van Breukelen, M. J. Candel, and M. P. Berger. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Stat Med*, 26(13):2589–603, 2007.

[68] P. Vargha. A critical discussion of intraclass correlation coefficients. *Stat Med*, 16(7):821–3, 1997.

[69] L. A. Waller, E. G. Hill, and R. A. Rudd. The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. *Stat Med*, 25(5):853–65, 2006.

[70] P. M. Westgate and T. M. Braun. Improving small-sample inference in group randomized trials with binary outcomes. *Stat Med*, 30(3):201–10, 2011.

[71] T. Xie and J. Waksman. Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. *Stat Med*, 22(18):2835–46, 2003.

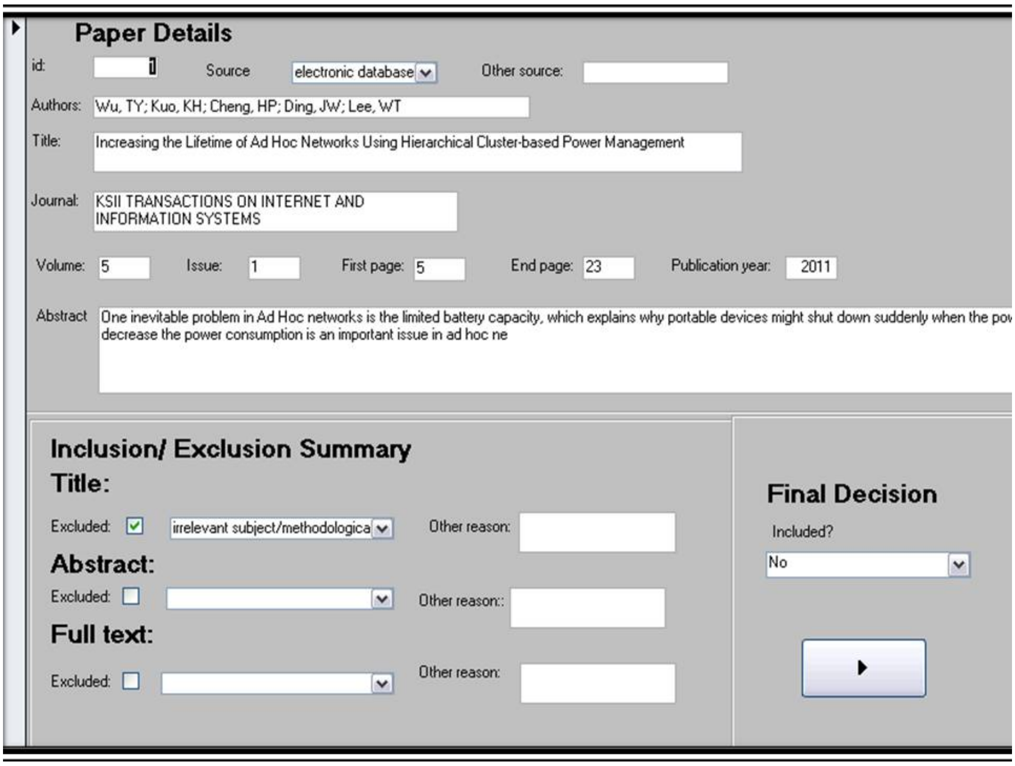
[72] P. L. Yudkin and M. Moher. Putting theory into practice: a cluster randomized trial with a small number of clusters. *Stat Med*, 20(3):341–9, 2001.

[73] K. H. Zou and M. P. McDermott. Higher-moment approaches to approximate interval estimation for a certain intraclass correlation coefficient. *Stat Med*, 18(15):2051–61, 1999.

Appendix C: Database

Database field	**Data type**	**Description**	**Field values**
id	Text	Unique numeric identifier for paper	
authors	Text	List of authors	
year	Number	Year of publication	
title	Text	Title of paper	
journal	Text	Name of journal	
volume	Number	Volume	
issue	Number	Issue	
bp	Number	Beginning page	
ep	Number	End page	
abstract	Text	Abstract	
exc_title	Text	Exclusion from review based on title	Yes/No
exc_tit_reason	Text	Reason for exclusion based on title	irrelevent subject/methodological area, not clustered data, not related to sample size, analysis method, too general, other
exc_tit_reason_other	Text	Reason for exclusion based on title, other	
exc_abs	Text	Exclusion from review based on abstract	Yes/No
exc_abs_reason	Text	Reason for exclusion based on abstract	irrelevent subject/methodological area, not clustered data, not related to sample size, analysis method, too general, other
exc_abs_reason_other	Text	Reason for exclusion based on abstract, other	
exc_ft	Text	Exclusion from review based on full text	Yes/No
exc_ft_reason	Text	Reason for exclusion based on full text	irrelevent subject/methodological area, not clustered data, not related to sample size, analysis method, too general, other
exc_ft_reason_other	Text	Reason for exclusion based on full text, other	
source	Text	Search source of paper	electronic database, reference search, author search, book, other
source_other	Text	other source	
review	Text	meets inclusion into the review	
comment	Text	Comments field	

Table 1.2: Data base description



Paper Details

id: 1 Source: electronic database Other source:

Authors: Wu, TY; Kuo, KH; Cheng, HP; Ding, JW; Lee, WT

Title: Increasing the Lifetime of Ad Hoc Networks Using Hierarchical Cluster-based Power Management

Journal: KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS

Volume: 5 Issue: 1 First page: 5 End page: 23 Publication year: 2011

Abstract One inevitable problem in Ad Hoc networks is the limited battery capacity, which explains why portable devices might shut down suddenly when the pov
decrease the power consumption is an important issue in ad hoc ne

Inclusion/ Exclusion Summary

Title:

Excluded: ☑ irrelevant subject/methodologica Other reason:

Abstract:

Excluded: □ Other reason::

Full text:

Excluded: □ Other reason:

Final Decision

Included?

No

Figure 1.3: Screen print of database

Appendix D: Data extraction forms

1.1 Paper details:

database ID:

Paper type: Derivation or adaptation to formulae

First Author Surname:

Year of Publication:

Journal name:

Full title:

1.2 Trial design components:

Id number for trial design:		
Outcome type:		
Number of groups:		
Trial design:		
Matching:		
Allocation ratio:		
Cluster size:		

1.3 Formulae described:

1. **Id number for trial design: 1**

   (a) **Reference number of the relevent formula in paper or link to previously extracted formula by database id: id number for method:**

   (b) **Re-produce formula below:**

   (c) **List any relevent references provided for this formula:**

2. **Id for trial design: 2**

   (a) **Reference number of the relevent formula in paper or link to previously extracted formula by database id: id number for method:**

   (b) **Re-produce formula below:**

   (c) **Define components of formula:**

   (d) **List any relevent references provided for this formula:**

1.4 Correlation measure:

1. **Id for trial design:** 1

   (a) **Name of the term used to estimate correlation:**

   (b) **Reference number of the relevent formula in paper:**

   (c) **Re-produce formula below or name the method used:**

   (d) **Define components of formula:**

   (e) **List any relevent references provided for this formula:**

2. **Id for trial design:** 2

(a) **Name of the term used to estimate correlation:**

(b) **Reference number of the relevent formula in paper:**

(c) **Re-produce formula below or name the method used:**

(d) **Define components of formula:**

(e) **List any relevent references provided for this formula:**

1.5 Analysis method:

1. **Id for trial design:** 1

   (a) **Hypothesis to be tested:**

   (b) **Name of the method assumed for analysis:**

   (c) **List any relevent references provided for this method:**

2. **Id for trial design:** 2

   (a) **Hypothesis to be tested:**

   (b) **Name of the method assumed for analysis:**

(c) **List any relevent references provided for this method:**

1.6 Assumptions:

1. **Id for trial design:** 1

   (a) **State assumptions of the sample size methodology:**

2. **Id for trial design:** 2

   (a) **State assumptions of the sample size methodology:**

1.7 Strengths/weaknesses:

Describe any simulation study of the formulae provided:

Is an example of formula use provided:

1. **Id for trial design:** 1

   (a) **Provide the strengths as stated by the author:** none provided

   (b) **Provide the weaknesses as stated by the author:** none provided

2. **Id for trial design:** 2

(a) **Provide the strengths as stated by the author:** none provided

(b) **Provide the weaknesses as stated by the author:** none provided

1.8 Extensions:

1. **Id for extension:** 1

   (a) **Id for trial design:**

   (b) **Describe the extension provided:**

   (c) **Reference number of its relevent formula in paper:**

   (d) **Re-produce formula below:**

   (e) **Define components of formula:**

   (f) **List any relevent references provided for this extension:**

   (g) **Provide the strengths of the extension as stated by the author:**

   (h) **Provide the weaknesses of the extension as stated by the author:**

2. **Id for extension:** 2

(a) **Id for trial design:**

(b) **Describe the extension provided:**

(c) **Reference number of its relevent formula in paper:**

(d) **Re-produce formula below:**

(e) **Define components of formula:**

(f) **List any relevent references provided for this extension:**

(g) **Provide the strengths of the extension as stated by the author:**

(h) **Provide the weaknesses of the extension as stated by the author:**

1.1 Paper details:

Database ID:

Type of paper: Assessment of current methodology

First Author Surname:

Year of Publication:

Journal name:

Full title:

1.2 Methods to be assessed:

1. **Id number for method:**

   (a) **Reference number of the relevent formula in paper:**

   (b) **Re-produce or describe formula below:**

   (c) **State assumptions of the methodology:**

   (d) **Define components of formula:**

   (e) **List any relevent references provided for this formula:**

2. **Id number for method:**

   (a) **Reference number of the relevent formula in paper:**

(b) **Re-produce or describe formula below:**

(c) **State assumptions of the methodology:**

(d) **Define components of formula:**

(e) **List any relevent references provided for this formula:**

1.3 Simulated data:

1. **Id for simulation:**

   (a) **Objective of the simulation:**

   (b) **Model used to simulate the data:**

   (c) **simulation method:** Monte Carlo simulations

   (d) **Range of variables used in the simulations:**

   (e) **Procedure:**

   (f) **Results:**

2. **Id for simulation:**

(a) **Objective of the simulation:**

(b) **Model used to simulate the data:**

(c) **simulation method:**

(d) **Range of variables adjusted in the simulations:**

(e) **Procedure:**

(f) **Results:**

1.4 Conclusions and Recommendations:

1. **Provide the recommendations as stated by the author:**

1.5 Strengths/weaknesses:

1. **Provide the strengths as stated by the author:**

2. **Provide the weaknesses as stated by the author:**

1.6 Extensions:

1. **Id for extension:**

(a) **Describe the extension provided:**

(b) **Reference number of its relevent formula in paper:**

(c) **Re-produce formula below:**

(d) **Define components of formula:**

(e) **List any relevent references provided for this extension:**

(f) **Provide the strengths of the extension as stated by the author:**

(g) **Provide the weaknesses of the extension as stated by the author:**

Appendix E: Proposed tables

Data Type:	**Analysis Methods:**	**Corresponding database ID:**
Continuous		
Binary		
Ordinal		
Time-to-event		
Count		

Table 1.3: **Completely randomised design, equal allocation:** Description of analysis methods assumed in sample size calculations

Data Type:	**Measure of correlation:**	**Corresponding database ID:**
Continuous		
Binary		
Ordinal		
Time-to-event		
Count		

Table 1.4: **Completely randomised design, equal allocation:** Description of measures of correlation used in sample size calculations

Data Type:	**Measure of cluster size:**	**Corresponding database ID:**
Continuous		
Binary		
Ordinal		
Time-to-event		
Count		

Table 1.5: **Completely randomised design, equal allocation:** Description of measures of cluster size used in sample size calculations

ii (Chapter two) Literature review, list of included papers

1. Amatya A, Bhaumik D and Gibbons RD. Sample size determination for clustered count data. *Stat Med*. 2013.

2. Braun T. A mixed model formulation for designing cluster randomized trials with binary outcomes. *Statistical Modelling*. 2003; 3: 233-49.

3. Byar DP. The design of cancer prevention trials. *Recent Results Cancer Res*. 1988; 111: 34-48.

4. Campbell CJWSJ. *How to design, analyse and report cluster randomised trials in medicine and health related research*. Chichester,UK: Wiley, 2014.

5. Candel MJ and Van Breukelen GJ. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Stat Med*. 2010; 29: 1488-501.

6. Connelly LB. Balancing the number and size of sites: an economic approach to the optimal design of cluster samples. *Control Clin Trials*. 2003; 24: 544-59.

7. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol*. 1978; 108: 100-2.

8. Donner A. An empirical study of cluster randomization. *Int J Epidemiol*. 1982; 11: 283-6.

9. Donner A. Sample size requirements for stratified cluster randomization designs. *Stat Med*. 1992; 11: 743-50.

10. Donner A. Some aspects of the design and analysis of cluster randomization trials. *Journal of the Royal Statistical Society Series C-Applied Statistics*. 1998; 47: 95-113.

11. Donner A, Birkett N and Buck C. Randomization by cluster- sample size requirements and analysis. *American Journal of Epidemiology*. 1981; 114: 906-14.

12. Donner A and Klar N. Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology*. 1996; 49: 435-9.

13. Eldridge SM, Ashby D and Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol*. 2006; 35: 1292-300.

14. Feldman HA and McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. *Stat Med*. 1994; 13: 61-78.

15. Feng Z and Grizzle JE. Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculation. *Stat Med*. 1992; 11: 1607-14.

16. Feng Z and Thompson B. Some design issues in a community intervention trial. *Control Clin Trials*. 2002; 23: 431-49.

17. Freedman LS, Green SB and Byar DP. Assessing the gain in efficiency due to matching in a community intervention study. *Stat Med*. 1990; 9: 943-52.

18. Gangnon R and Kosorok M. Sample-size formula for clustered survival data using weighted log-rank statistics. *Biometrika*. 2004; 91: 263-75.

19. Giraudeau B, Ravaud P and Donner A. Sample size calculation for cluster randomized cross-over trials. *Statistics in Medicine*. 2008; 27: 5578-85.

20. Harrison DA and Brady AR. Sample size and power calculations using the noncentral t-distribution. *Stata Journal*. 2004; 4: 142-53.

21. Hayes R and Bennett S. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*. 1999; 28: 319-26.

22. Hedges L and Hedberg E. Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*. 2007; 29: 60-87.

23. Hendricks S, Wassell J, Collins J and Sedlak S. Power determination for geographically clustered data using generalized estimating equations. *Statistics in Medicine*. 1996; 15: 1951-60.

24. Heo M, Kim Y, Xue X and Kim M. Sample size requirement to detect an intervention effect at the end of follow-up in a longitudinal cluster randomized trial. *Statistics in Medicine*. 2010; 29: 382-90.

25. Heo M and Leon A. Statistical Power and Sample Size Requirements for Three Level Hierarchical Cluster Randomized Trials. *Biometrics*. 2008; 64: 1256-62.

26. Heo M and Leon A. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials. *Statistics in Medicine*. 2009; 28: 1017-27.

27. Hoover D. Power for t-test comparisons of unbalanced cluster exposure studies. *Journal of Urban Health-Bulletin of the New York Academy of Medicine*. 2002; 79: 278-94.

28. Hsieh F. Sample-size formulas for intervention studies with the cluster as unit of randomisation. *Statistics in Medicine*. 1988; 7: 1195-201.

29. Hussey M and Hughes J. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*. 2007; 28: 182-91.

30. Jahn-Eimermacher A, Ingel K and Schneider A. Sample size in cluster-randomized trials with time to event as the primary endpoint. *Stat Med*. 2013; 32: 739-51.

31. Jung S. Sample size calculation for weighted rank tests comparing survival distributions under cluster randomization: a simulation method. *Journal of Biopharmaceutical Statistics*. 2007; 17: 839-49.

32. Kang S, Ahn C and Jung S. Sample size calculation for dichotomous outcomes in cluster randomization trials with varying cluster size. *Drug Information Journal*. 2003; 37: 109-14.

33. Kerry S and Bland J. Trials which randomize practices II: sample size. *Family Practice*. 1998; 15: 84-7.

34. Kerry S and Bland J. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Statistics in Medicine*. 2001; 20: 377-90.

35. Kikuchi T and Gittins J. A behavioural Bayes approach for sample size determination in cluster randomized clinical trials. *Journal of the Royal Statistical Society Series C-Applied Statistics*. 2010; 59: 875-88.

36. Kim HY, Williamson JM and Lyles CM. Sample-size calculations for studies with correlated ordinal outcomes. *Stat Med*. 2005; 24: 2977-87.

37. Koepsell T, Martin D, Diehr P, et al. 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*. 1991; 44: 701-13.

38. Konstantopoulos S. Incorporating Cost in Power Analysis for Three-Level Cluster-Randomized Designs. *Evaluation Review*. 2009; 33: 335-57.

39. Lake S, Kammann E, Klar N and Betensky R. Sample size re-estimation in cluster randomization trials. *Statistics in Medicine*. 2002; 21: 1337-50.

40. Liu A, Shih W and Gehan E. Sample size and power determination for clustered repeated measurements. *Statistics in Medicine*. 2002; 21: 1787-801.

41. Liu G and Liang K. Sample size calculations for studies with correlated observations. *Biometrics*. 1997; 53: 937-47.

42. Liu X. Statistical Power and Optimum Sample Allocation Ratio for Treatment and Control Having Unequal Costs per Unit of Randomization. *Journal of Educational and Behavioral Statistics*. 2003; 28: 231-48.

43. Lui K and Chang K. Sample Size Determination for Testing Equality in a Cluster Randomized Trial with Noncompliance. *Journal of Biopharmaceutical Statistics*. 2011; 21: 1-17.

44. Lui K and Chang K. Test Non-Inferiority and Sample Size Determination Based on the Odds Ratio Under a Cluster Randomized Trial with Noncompliance. *Journal of Biopharmaceutical Statistics*. 2011; 21: 94-110.

45. Manatunga A and Chen S. Sample size estimation for survival outcomes in cluster-randomized studies with small cluster sizes. *Biometrics*. 2000; 56: 616-21.

46. Manatunga A, Hudgens M and Chen S. Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal*. 2001; 43: 75-86.

47. McKinlay S. Cost-efficient designs of cluster unit trials. *Preventive Medicine*. 1994; 23: 606-11.

48. Mickey R, Goodwin G and Costanza M. Estimation of the design effect in community intervention studies. *Statistics in Medicine*. 1991; 10: 53-64.

49. Moerbeek G, Van Breukelen G and Berger M. Design issues for experiments in multilevel populations. *Journal of educational and Behavioural Statistics*. 2000; 25: 271-84.

50. Moerbeek M. Power and money in cluster randomized trials: When is it worth measuring a covariate? *Statistics in Medicine*. 2006; 25: 2607-17.

51. Moerbeek M and Maas C. Optimal experimental designs for multilevel logistic models with two binary predictors. *Communications in Statistics-Theory and Methods*. 2005; 34: 1151-67.

52. Moerbeek M, Van Breukelen G and Berger M. Optimal experimental designs for multilevel logistic models. *Journal of the Royal Statistical Society Series D-the Statistician*. 2001; 50: 17-30.

53. Moerbeek M, Van Breukelen G and Berger M. Optimal experimental designs for multilevel models with covariates. *Communications in Statistics-Theory and Methods*. 2001; 30: 2683-97.

54. Mukhopadhyay S and Looney S. Quantile dispersion graphs to compare the efficiencies of cluster randomized designs. *Journal of Applied Statistics*. 2009; 36: 1293-305.

55. Murray D. *Design and Analysis of Group-Randomized Trials*. Oxford University Press, 1998.

56. Murray DM, Blitstein JL, Hannan PJ, Baker WL and Lytle LA. Sizing a trial to alter the trajectory of health behaviours: methods, parameter estimates, and their application. *Stat Med*. 2007; 26: 2297-316.

57. Murray DM and Hannan PJ. Planning for the appropriate analysis in school-based drug-use prevention studies. *J Consult Clin Psychol*. 1990; 58: 458-68.

58. Neuhaus JM and Segal MR. Design effects for binary regression models fitted to dependent data. *Stat Med*. 1993; 12: 1259-68.

59. Pan W. Sample size and power calculations with correlated binary data. *Control Clin Trials*. 2001; 22: 211-27.

60. Preisser JS, Reboussin BA, Song EY and Wolfson M. The importance and role of intracluster correlations in planning cluster trials. *Epidemiology*. 2007; 18: 552-60.

61. Preisser JS, Young ML, Zaccaro DJ and Wolfson M. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med*. 2003; 22: 1235-54.

62. Raudenbush S. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*. 1997; 2: 173-85.

63. Reich NG, Myers JA, Obeng D, Milstone AM and Perl TM. Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. *PLoS One*. 2012; 7: e35564.

64. Rietbergen C and Moerbeek M. The Design of Cluster Randomized Crossover Trials. *Journal of Educational and Behavioral Statistics*. 2011; 36: 472-90.

65. Rosner B and Glynn R. Power and Sample Size Estimation for the Clustered Wilcoxon Test. *Biometrics*. 2011; 67: 646-53.

66. Roy A, Bhaumik D, Aryal S and Gibbons R. Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics*. 2007; 63: 699-707.

67. Shih W. Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimating equations. *Biometrical Journal*. 1997; 39: 899-908.

68. Shipley M, Smith P and Dramaix M. Calculation of power for matched pair studies when randomization is by group. *International Journal of Epidemiology*. 1989; 18: 457-61.

69. Snijders T and Bosker R. Standard errors and sample sizes for 2-level research. *Journal of Educational Statistics*. 1993; 18: 237-59.

70. Spiegelhalter D. Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine*. 2001; 20: 435-52.

71. Taljaard M, Donner A and Klar N. Accounting for expected attrition in the planning of community intervention trials. *Statistics in Medicine*. 2007; 26: 2615-28.

72. Teerenstra S, Eldridge S, Graff M, de Hoop E and Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med*. 2012.

73. Teerenstra S, Lu B, Preisser J, van Achterberg T and Borm G. Sample Size Considerations for GEE Analyses of Three-Level Cluster Randomized Trials. *Biometrics*. 2010; 66: 1230-7.

74. Teerenstra S, Moerbeek M, van Achterberg T, Pelzer B and Borm G. Sample size calculations for 3-level cluster randomized trials. *Clinical Trials*. 2008; 5: 486-95.

75. Thompson S, Pyke S and Hardy R. The design and analysis of paired cluster randomized trials: An application of meta-analysis techniques. *Statistics in Medicine*. 1997; 16: 2063-79.

76. Tu X, Kowalski J, Zhang J, Lynch K and Crits-Christoph P. Power analyses for longitudinal trials and other clustered designs. *Statistics in Medicine*. 2004; 23: 2799-815.

77. Turner R, Prevost A and Thompson S. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Statistics in Medicine*. 2004; 23: 1195-214.

78. Turner R, Thompson S and Spiegelhalter D. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials*. 2005; 2: 108-18.

79. van Breukelen GJ, Candel MJ and Berger MP. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Stat Med*. 2007; 26: 2589-603.

80. van Breukelen GJ, Candel MJ and Berger MP. Relative efficiency of unequal cluster sizes for variance component estimation in cluster randomized and multicentre trials. *Stat Methods Med Res*. 2008; 17: 439-58.

81. Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL and Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol*. 2013; 66: 752-8.

82. Xie T and Waksman J. Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. *Stat Med*. 2003; 22: 2835-46.

83. Yin G and Shen Y. Adaptive design and estimation in randomized clinical trials with correlated observations. *Biometrics*. 2005; 61: 362-9.

84. You Z, Williams OD, Aban I, Kabagambe EK, Tiwari HK and Cutter G. Relative efficiency and sample size for cluster randomized trials with variable cluster sizes. *Clin Trials*. 2011; 8: 27-36.

85. Zucker DM, Lakatos E, Webber LS, et al. Statistical design of the Child and Adolescent Trial for Cardiovascular Health (CATCH): implications of cluster randomization. *Control Clin Trials*. 1995; 16: 96-118.

iii (Chapter three) Data abstraction form for review of 300 cluster randomised trials

Reporting of sample size calculations in cluster randomised trials.
Data Abstraction form

Note: Items previously abstracted will be automatically uploaded to the abstraction Access database.

1. Study ID (this is the Medline UID number – also the pdf file name) □
2. Reviewer name □
3. Publication year □
4. Journal name □

Study Design

Note: For items colored in black the data have been collected previously and will not be re-abstracted, although in some cases described below data may need to be updated. These items will be automatically uploaded to the abstraction Access database. They are included here to ensure consistency in abstraction of other data, for example when talking about the primary outcome, and also included for completeness of information regarding sample size reporting.

5. Were primary outcome measure(s) identified by authors? (Authors clearly distinguished between main (or primary) and secondary outcomes measures?) (Note: Not acceptable if authors merely stated primary **objectives** without operationalizing in terms of specific variables.)

   1 Yes (specify number) □
   2 No

6. For quality control purposes, state the single **primary outcome** identified during the initial review. NOTE: when multiple outcomes are specified by authors, the main outcome will be the first outcome stated in the abstract or the analysis.

   □

7. If a sample size calculation is not presented for the outcome described in Q6, state the single outcome that the sample size calculation has been performed for and data abstraction will be based on. (Note: if more than one outcome possible, choose the first outcome described.)

   □ □ Same as above

8. Trial design at cluster-level:

1	Parallel trial (clusters independently randomized to different treatments with or without pre-test)	
2	Factorial trial (specify factors and levels) (e.g., 2x2)	□
3	Cross-over trial	
4	Other (specify) (e.g., latin squares, split-plot, stepped wedge)	□

9. Method of random allocation:

1	Completely randomized design (unrestricted randomization)	
2	Stratified design	
3	Pair-matched design	
4	Other (specify) (e.g., minimization algorithm)	□

10. Trial design at patient-level (primary outcome):

1	Nested cross-sectional design (each patient measured only once or different patients measured each time point)
2	Nested cohort design (same patients measured at different time points or in continuous surveillance) (NOTE: Patient-level attrition is possible in a cohort design but NOT in a cross-sectional design)
3	Primary outcome evaluated on both cross-sectional and cohort components

11. Number of study arms (we need to know if there were multiple intervention or control arms so that patient numbers can be divided up appropriately when assessing imbalances):

a) Intervention arms	□
b) Control arms	□

Sample size

Note: The data collected in this section should be based upon the sample size outcome as defined in Q7. Where this is different from Q6 please update previously abstracted data i.e. Q12 for this section.

12. Sample size / power calculations presented?

1	Not presented or presented for substudy or outcome regarded as secondary only
2	**Patient**-level accounting for ICC (“Sample size was based on a significant effect size of 0.5, incorporated an ICC of 0.05 and was based on enrollment of 4 patients per physician”; “Based on a mean (SD) number of admission days per resident enrolled, within cluster variance of 2 days and between-cluster variance of 3 days and 10 residents per nursing home”. Usually will involve stating at least the average cluster size and the ICC/ design effect/ overdispersion factor/within-and between-cluster variance or stated that accounting for clustering without reporting value of ICC.)
3	**Cluster**-level (Should be clear that cluster-level summary data are used for calculation e.g., “sample size was based on the hospital as the unit of analysis…assuming a rate of episiotomy of 42% at baseline, with a standard deviation of 15%, we need 18 hospitals to identify a decrease in episiotomy rate.” Use of standard deviation in the case of proportions indicates that binary data was summarized at cluster-level and treated as continuous data for the purpose of sample size calculation.)
4	**Patient**-level without accounting for ICC (usually difficult to tell whether at patient- or cluster-level unless specifically stated)
5	Unclear whether at patient- or cluster-level or whether accounted for clustering (e.g., “sample size was calculated to give a power of 80% of detecting a difference of 1 SD at 5% significance in mean diagnosis concordance score”; “sample size of 500 participants would result in 80% power to detect a difference of 10 points between groups”)
6	Other (specify) (e.g., based on intermediate level of clustering) □

Note: Questions 13-29 are not applicable for papers which do not report a sample size calculation. Not applicable will be automatically generated in the database for all the relevant papers, based upon the answer to question 12. Those where question 12 is updated will be updated manually

13. Describe the method or citation of sample size calculation that has been used (for example: Hsieh, Donner and Klar, Hayes and Bennett, standard two-sample t-test adjusted for clustering):

14. What type of data has been assumed for the sample size calculation? (Note: this question aims to address any occasions where ordinal or count data may have analyzed as a dichotomous or continuous variable for simplicity.)

1	Binary (two categories)
2	Categorical (more than two categories, but no natural ordering to the categories)
3	Ordinal (more than two ordered categories)

4	Continuous (normal or non-normal)
5	Count data
6	Time to event data
7	Unclear
8	Other
9	Not applicable

15. What is the type I error rate that has been assumed for the sample size calculation?

1	5%
2	1%
3	Unclear or not stated
4	Other
5	Not applicable

16. What is the power (1 minus the Type II error rate) that has been assumed for the sample size calculation? (Note: record "at least 80%" as 80%).

1	80%
2	90%
3	Unclear or not stated
4	Other
5	Not applicable

17. Does the sample size calculation assume a one or two-sided test?

1	1-sided
2	2-sided
3	Unclear or not stated
4	Other
5	Not applicable

18. How is correlation within clusters described in the sample size calculation?

1	Intracluster correlation
2	Coefficient of variation
3	Unclear or not stated
4	Other (for example between and within cluster variance) □
5	Not applicable (i.e. sample size does not account for clustering or sample size calculation is at cluster level)

19. State the value of the assumed correlation

□ Not provided/Not Applicable

20. Does the sample size account for attrition (either of clusters or individuals)? NOTE: if not stated then assume "no".

1 Yes
2 No
3 Unclear
4 Other
5 Not applicable

21. Does the sample size account for cluster size imbalance? NOTE: if not stated (even if the calculation was at cluster level) then assume "no".

1 Yes
2 No
3 Unclear
4 Other
5 Not applicable

22. Justification for the ICC or other correlation measure assumed in the sample size calculation:

1 No justification
2 Results from a previous trial
3 A preliminary/pilot study
4 Observational data
5 Results from a systematic review
6 Baseline data
7 Other (please explain)
8 Not Applicable

23. Justification for the control group effect (e.g., proportion, mean or standard deviation) assumed in the sample size calculation:

1 No justification
2 Results from a previous trial
3 A preliminary/pilot study
4 Observational data

5 Results from a systematic review
6 Baseline data
7 Other (please explain) □
8 Not applicable

24. Justification for the treatment group effect (e.g., proportion, mean or standard deviation) assumed in the sample size calculation: Where multiple treatment groups, quote the values for the treatment group which produced/drove the sample size requirement.

1 No justification
2 Analogy to another trial or treatment
3 Clinical relevance
4 Observational data
5 Results from a systematic review
6 Other (please explain) □
7 Not applicable

25. State the target total number of clusters required for the analysis. (Note: this is before adjustments for attrition.)

□ □ Not provided/unclear

26. State the target total number of individuals. (Note: this may need to be calculated: e.g., number of clusters multiplied by average cluster size.)

□ □ Not provided /unclear □ NA (e.g., sample size calculation is at cluster level)

27. Assumed effect in the control arm: (Note: for dichotomous and time-to-event outcomes, provide the rate of events; for continuous outcomes, provide the mean and standard deviation.)

Proportion

□ □ Not provided/unclear

Mean	Standard deviation

28. Assumed effect in the treatment arm: (Note: for dichotomous and time-to-event outcomes, provide the rate of events; for continuous outcomes, provide the mean and standard deviation. Where multiple treatment groups, quote the values for the treatment group which produced/drove the sample size requirement. With papers that do not provide a SD for the treatment arm, we will assume a common standard deviation)

Proportion

□ Not provided/unclear

Mean Standard deviation (if SD not stated, enter the same value as for the control arm)

Note: Q29 is only to be completed if Q27 and Q28 are incomplete and only a standardized effect size has been provided.

29. State the standardized effect size if provided (e.g., the difference in means divided by the standard deviation, Relative Risk, or Odds Ratio):

Effect size

□ Not provided/Not applicable

Specify the type of effect (e.g., difference in means, Relative Risk, Odds Ratio)

Analysis

Note: The data collected in this section should be based upon the sample size outcome as defined in Q7. Where this is different from Q6 please update previously extracted data i.e. Q32 and Q40 for this section.

30. State the total number of clusters used in the analysis of the sample size outcome. (Note: try to limit the use of the "unclear" option.)

□ Not provided/unclear

31. State the total number of individuals used in the analysis of the sample size outcome. (Note: try to limit the use of the “unclear” option.)

□ □ Not provided/unclear

32. Analysis for **sample size** outcome: (NOTE: primary and sample size outcome will be the same if 6 and 7 are the same)

1	At **patient-level** accounting for ICC (e.g., using mixed-effects logistic regression, GEE taking account of clustering by physician, Chi-square statistic adjusted for clustering, random effect for physician, hierarchical modeling, multi-level modeling, alternating logistic regression)	
2	At **cluster-level** (clearly stated that analysis at cluster-level, e.g., “analyses performed using patient-level variables aggregated at the provider-level”, analysis was based on hospital rates, t-test weighted by inverse variance etc.)	
3	At patient-level **not** accounting for ICC (more difficult to distinguish, e.g., multivariable regression analysis of patient-level data with no mention of clustering, or standard 2-sample test on patient-level data without mention clustering or stated that since ICCs were low, clustering was ignored in presentation of results)	
4	Unclear whether at patient-level or cluster-level or whether accounted for clustering	
5	Other (specify) (e.g., based on intermediate level of clustering, both individual-level and cluster-level analyses used for primary outcome analysis)	□

33. Describe the analysis method or citation (e.g. adjusted two-sample test, permutation test, GEE, hierarchical model):

□

34. What type of data is the raw sample size outcome at the level (cluster or individual) corresponding to the analysis? (Note: this is a description of the data as it was measured, regardless of how it may have been treated at sample size or analysis stages.)

1	Binary (two categories)
2	Categorical (more than two categories, but no natural ordering to the categories)
3	Ordinal (more than two and <7 ordered categories)
4	Continuous, normal or non-normal
5	Count data

6 Time to event data
7 Unclear
8 Other (specify) □

35. How has the data for the sample size outcome been used for the analysis?

1 Binary (two categories)
2 Categorical
(more than two categories, but no natural ordering to the categories)
3 Ordinal (more than two ordered categories)
4 Continuous, normal or non-normal
5 Count data
6 Time to event data
7 Unclear
8 Other (specify) □

36. State the value of the observed intracluster correlation for the sample size outcome (or tick if not provided)

□ □ Not provided

37. Observed effect in the control arm: (Note: for dichotomous and time-to-event outcomes, provide the rate of events; for continuous outcomes, provide the mean and standard deviation. Where multiple time-points are given use the time-point corresponding to the sample size calculation; if not provided use the final time-point.)

Proportion

□ □ Not provided

Mean	Standard deviation

38. Observed effect in the treatment arm: (Note: for dichotomous and time-to-event outcomes, provide the rate of events; for continuous outcomes, provide the mean and standard deviation. Where multiple treatment groups, please quote the values for the treatment group which produced the sample size requirement, or the treatment group producing the largest difference).

Proportion

□ □ Not provided

Mean	Standard deviation

Note: Q39 is only to be completed if Q37 and Q38 are incomplete and only a standardized effect size has been provided.

39. State the unadjusted effect size if provided (e.g., the standardized difference in means, Relative Risk, or Odds Ratio):

Effect size

□ Not provided/Not applicable

Specify the type of effect (e.g., standardized difference in means, Relative Risk, Odds Ratio)

40. Is the sample size outcome reported as statistically significant?

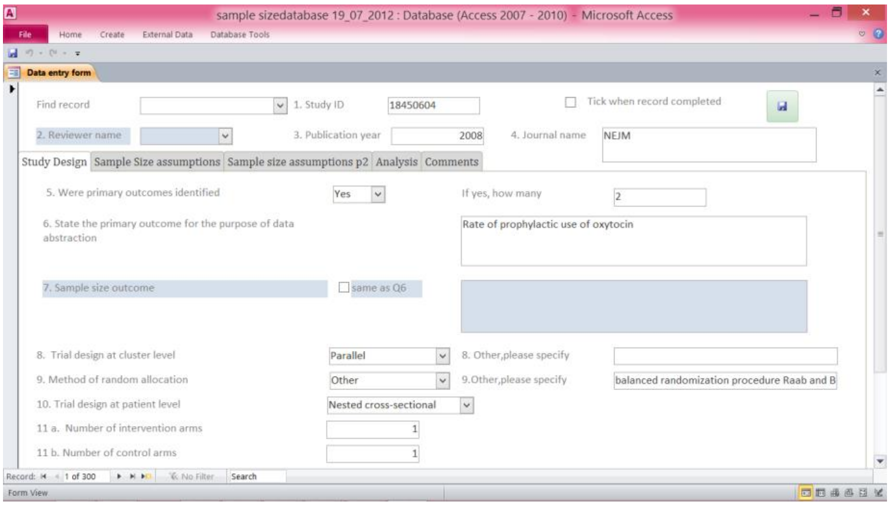
1 Yes
2 No

41. Do authors reference a separate publication which may provide further details on items required in this data abstraction form? (Note: If yes, please refer to the separate publication to complete this abstraction form.)

1 Yes
2 No

Comments:

iv (Chapter three) Database screen shot for review of 300 cluster randomised trials



Find record

1. Study ID 18450604

□ Tick when record completed

2. Reviewer name

3. Publication year 2008

4. Journal name NEJM

Study Design | Sample Size assumptions | Sample size assumptions p2 | Analysis | Comments

5. Were primary outcomes identified: Yes

If yes, how many: 2

6. State the primary outcome for the purpose of data abstraction: Rate of prophylactic use of oxytocin

7. Sample size outcome □ same as Q6

8. Trial design at cluster level: Parallel

8. Other,please specify

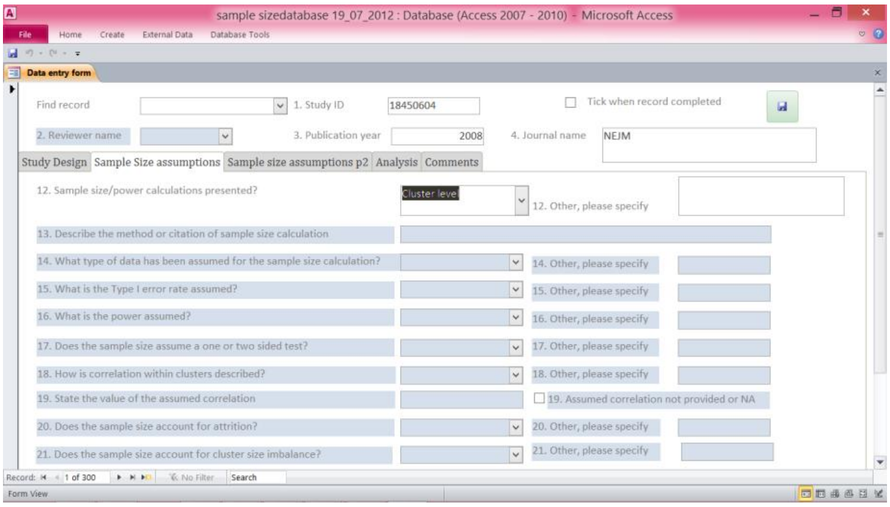
9. Method of random allocation: Other

9.Other,please specify: balanced randomization procedure Raab and B

10. Trial design at patient level: Nested cross-sectional

11 a. Number of intervention arms: 1

11 b. Number of control arms: 1



Find record

1. Study ID 18450604

□ Tick when record completed

2. Reviewer name

3. Publication year 2008

4. Journal name NEJM

Study Design | Sample Size assumptions | Sample size assumptions p2 | Analysis | Comments

12. Sample size/power calculations presented? Cluster leve

12. Other, please specify

13. Describe the method or citation of sample size calculation

14. What type of data has been assumed for the sample size calculation?

14. Other, please specify

15. What is the Type I error rate assumed?

15. Other, please specify

16. What is the power assumed?

16. Other, please specify

17. Does the sample size assume a one or two sided test?

17. Other, please specify

18. How is correlation within clusters described?

18. Other, please specify

19. State the value of the assumed correlation

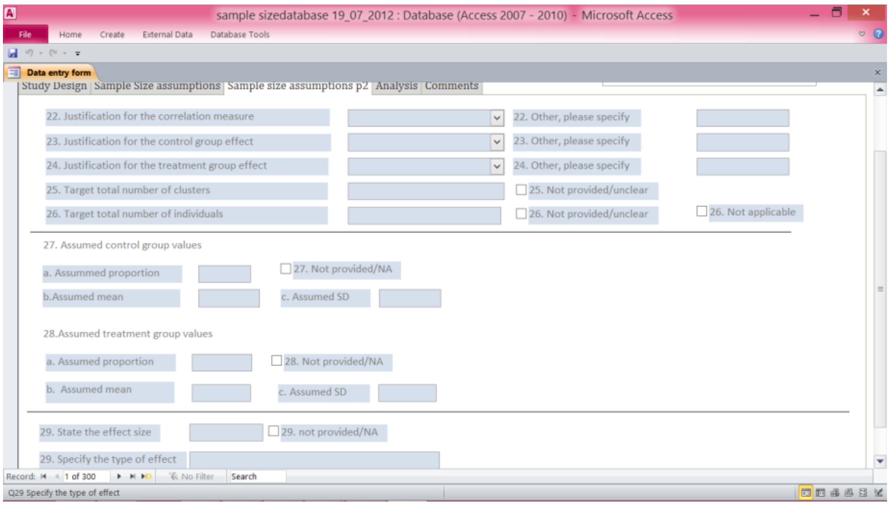
□ 19. Assumed correlation not provided or NA

20. Does the sample size account for attrition?

20. Other, please specify

21. Does the sample size account for cluster size imbalance?

21. Other, please specify



22. Justification for the correlation measure | 22. Other, please specify

23. Justification for the control group effect | 23. Other, please specify

24. Justification for the treatment group effect | 24. Other, please specify

25. Target total number of clusters | □ 25. Not provided/unclear

26. Target total number of individuals | □ 26. Not provided/unclear | □ 26. Not applicable

27. Assumed control group values

a. Assummed proportion | □ 27. Not provided/NA

b.Assumed mean | c. Assumed SD

28.Assumed treatment group values

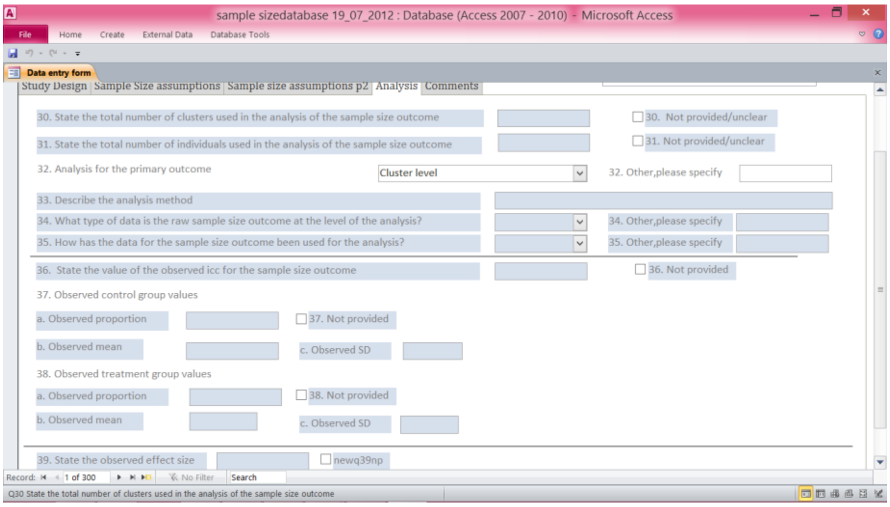
a. Assumed proportion | □ 28. Not provided/NA

b. Assumed mean | c. Assumed SD

29. State the effect size | □ 29. not provided/NA

29. Specify the type of effect

Q29 Specify the type of effect



30. State the total number of clusters used in the analysis of the sample size outcome | □ 30. Not provided/unclear

31. State the total number of individuals used in the analysis of the sample size outcome | □ 31. Not provided/unclear

32. Analysis for the primary outcome: Cluster level | 32. Other,please specify

33. Describe the analysis method

34. What type of data is the raw sample size outcome at the level of the analysis? | 34. Other,please specify

35. How has the data for the sample size outcome been used for the analysis? | 35. Other,please specify

36. State the value of the observed icc for the sample size outcome | □ 36. Not provided

37. Observed control group values

a. Observed proportion | □ 37. Not provided

b. Observed mean | c. Observed SD

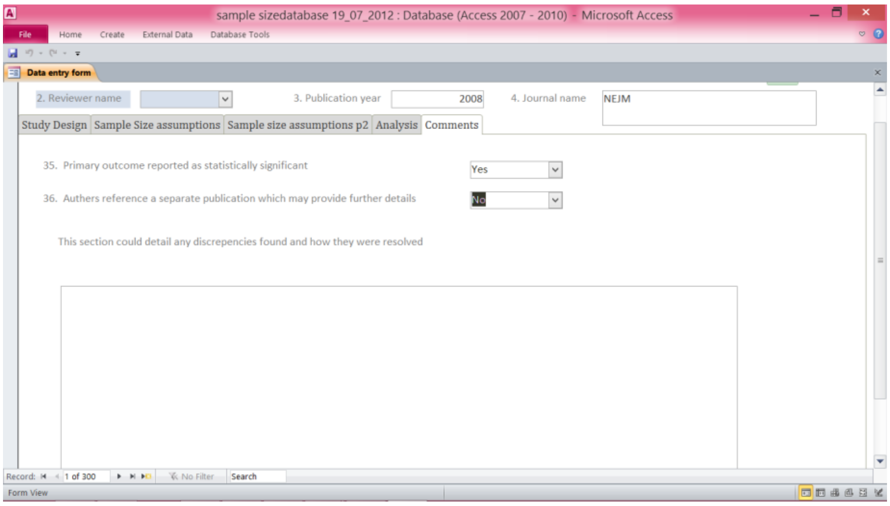
38. Observed treatment group values

a. Observed proportion | □ 38. Not provided

b. Observed mean | c. Observed SD

39. State the observed effect size | □ newq39np

Q30 State the total number of clusters used in the analysis of the sample size outcome



2. Reviewer name

3. Publication year 2008

4. Journal name NEJM

35. Primary outcome reported as statistically significant Yes

36. Authers reference a separate publication which may provide further details No

This section could detail any discrepencies found and how they were resolved

Record: 1 of 300

v (Chapter five) Example of simulation code

```
log using "de_simulations.log",replace

*program drop s_cluster_ologit

program define s_cluster_ologit,rclass

syntax, CLUSTERSPERGRP(integer) SIGMAV(real) CSIZE(integer)

drop _all

clear


set obs `=2*`clusterspergrp''

gen group=mod(_n,2)


gen b2=-0.52*sqrt(1+`sigmav')



*cluster level errors are normally distributed N(0,sigma^2 v)

gen ybar=rnormal(b2*group,sqrt(`sigmav'))


gen clusterid=_n

expand `csize'

gen outcome=rnormal(ybar,1)

gen outcome_normal=(outcome)/(sqrt(1+`sigmav'))


gen ordinal=1 if outcome_normal<=-.84162123
```

```
replace ordinal=2 if outcome_normal<=.52440051 & ordinal==.

replace ordinal=3 if outcome_normal<=1.2815516 & ordinal==.

replace ordinal=4 if outcome_normal>1.2815516 & ordinal==.


tab ordinal group,col matcell(freqs)


return scalar cat1_group0=(freqs[1,1]/(`clusterspergrp'*`csize'))*100

return scalar cat2_group0=(freqs[2,1]/(`clusterspergrp'*`csize'))*100

return scalar cat3_group0=(freqs[3,1]/(`clusterspergrp'*`csize'))*100

return scalar cat4_group0=(freqs[4,1]/(`clusterspergrp'*`csize'))*100


return scalar cat1_group1=(freqs[1,2]/(`clusterspergrp'*`csize'))*100

return scalar cat2_group1=(freqs[2,2]/(`clusterspergrp'*`csize'))*100

return scalar cat3_group1=(freqs[3,2]/(`clusterspergrp'*`csize'))*100

return scalar cat4_group1=(freqs[4,2]/(`clusterspergrp'*`csize'))*100



xtset clusterid

capture noisily {

    xtoprobit ordinal group

     return scalar p=2*normal(-abs(_b[group]/_se[group]))

            return scalar p_t=2*t(2*`clusterspergrp'-2,-abs(_b[group]/_se[group]))
```

```
        matrix b=e(b)

            return scalar coef=round(b[1,1],0.001)

            return scalar cor_icc_ord=e(sigma_u)^2/(e(sigma_u)^2+(1))


**Alternative ICC estimates


*anova-assuming scale assigned to ordinal outcome with equal distance

loneway ordinal clusterid

oneway ordinal clusterid

*calculate the ICC myself as Stata automatically truncates at 0

return scalar icc_ordinal_anova=((r(mss)/r(df_m))-
(r(rss)/r(df_r)))/((r(mss)/r(df_m))+(`csize'-1)*(r(rss)/r(df_r)))


}


end


clear


simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
, saving(pow1) reps(1001) seed(421892): s_cluster_ologit,clusterspergrp(20)
sigmav(0.01) csize(5)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
,saving(pow2) reps(1001) seed(136787): s_cluster_ologit,clusterspergrp(25)
sigmav(0.09) csize(5)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
,saving(pow3) reps(1000) seed(0034525): s_cluster_ologit,clusterspergrp(30)
sigmav(0.19) csize(5)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
, saving(pow4) reps(1000) seed(422101): s_cluster_ologit,clusterspergrp(35)
sigmav(0.33) csize(5)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
, saving(pow5) reps(1000) seed(5359235): s_cluster_ologit,clusterspergrp(54)
sigmav(1.13) csize(5)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
, saving(pow6) reps(1002) seed(32124): s_cluster_ologit,clusterspergrp(11) sigmav(0.01)
csize(10)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
,saving(pow7) reps(1000) seed(689987): s_cluster_ologit,clusterspergrp(16)
sigmav(0.09) csize(10)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
,saving(pow8) reps(1000) seed(123434): s_cluster_ologit,clusterspergrp(22)
sigmav(0.19) csize(10)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
, saving(pow9) reps(1000) seed(494329): s_cluster_ologit,clusterspergrp(28)
sigmav(0.33) csize(10)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
, saving(pow10) reps(1000) seed(56024): s_cluster_ologit,clusterspergrp(49)
sigmav(1.13) csize(10)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
, saving(pow11) reps(1003) seed(7458): s_cluster_ologit,clusterspergrp(3) sigmav(0.01)
csize(50)
```

```
simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
,saving(pow12) reps(1000) seed(96456): s_cluster_ologit,clusterspergrp(9) sigmav(0.09)
csize(50)

simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
,saving(pow13) reps(1000) seed(79002): s_cluster_ologit,clusterspergrp(15)
sigmav(0.19) csize(50)

simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
, saving(pow14) reps(1000) seed(57477): s_cluster_ologit,clusterspergrp(22)
sigmav(0.33) csize(50)

simulate  anova_ordinalicc=r(icc_ordinal_anova) pvalue=r(p)  pvaluet=r(p_t)  trt=r(coef)
icc_ordmodel=r(cor_icc_ord) cat1_group0=r(cat1_group0) cat2_group0=r(cat2_group0)
cat3_group0=r(cat3_group0) cat4_group0=r(cat4_group0) cat1_group1=r(cat1_group1)
cat2_group1=r(cat2_group1) cat3_group1=r(cat3_group1) cat4_group1=r(cat4_group1)
, saving(pow15) reps(1000) seed(222234): s_cluster_ologit,clusterspergrp(45)
sigmav(1.13) csize(50)



log close
```

vi (Chapter six) Literature review update, list of included papers

1. Ahn C, Hu F and Lee SC. Relative Efficiency of Unequal Versus Equal Cluster Sizes for the Nonparametric Weighted Sign Test Estimators in Clustered Binary Data. *Drug Inf J*. 2012; 46: 428-33.

2. Baio G, Copas A, Ambler G, Hargreaves J, Beard E and Omar RZ. Sample size calculation for a stepped wedge trial. *Trials*. 2015; 16: 354.

3. Corrigan N, Bankart MJ, Gray LJ and Smith KL. Changing cluster composition in cluster randomised controlled trials: design and analysis considerations. *Trials*. 2014; 15: 184.

4. Crespi CM, Wong WK and Wu S. A new dependence parameter approach to improve the design of cluster randomized trials with binary outcomes. *Clin Trials*. 2011; 8: 687-98.

5. Cunningham TD and Johnson RE. Design effects for sample size computation in three-level designs. *Stat Methods Med Res*. 2012.

6. Dziak JJ, Nahum-Shani I and Collins LM. Multilevel factorial experiments for developing behavioral interventions: power, sample size, and resource considerations. *Psychol Methods*. 2012; 17: 153-75.

7. Forbes AB, Akram M, Pilcher D, Cooper J and Bellomo R. Cluster randomised crossover trials with binary data and unbalanced cluster sizes: application to studies of near-universal interventions in intensive care. *Clin Trials*. 2015; 12: 34-44.

8. Hemming K, Girling AJ, Sitch AJ, Marsh J and Lilford RJ. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Med Res Methodol*. 2011; 11: 102.

9. Hemming K, Lilford R and Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple-level designs. *Stat Med*. 2015; 34: 181-96.

10. Heo M. Impact of subject attrition on sample size determinations for longitudinal cluster randomized clinical trials. *J Biopharm Stat*. 2014; 24: 507-22.

11. Heo M, Xue X and Kim MY. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials with random slopes. *Computational statistics & data analysis*. 2013; 60: 169-78.

12. Hooper R and Bourke L. Cluster randomised trials with repeated cross sections: alternatives to parallel group designs. *BMJ*. 2015; 350: h2925.

13. Hox JJ, Moerbeek M, Kluytmans A and van de Schoot R. Analyzing indirect effects in cluster randomized trials. The effect of estimation method, number of groups and group sizes on accuracy and power. *Frontiers in psychology*. 2014; 5: 78.

14. Konstantopoulos S. Optimal Sampling of Units in Three-Level Cluster Randomized Designs: An ANCOVA Framework. *Educational and Psychological Measurement*. 2011; 71: 798-813.

15. Lemme F, van Breukelen GJ, Candel MJ and Berger MP. The effect of heterogeneous variance on efficiency and power of cluster randomized trials with a balanced 2 x 2 factorial design. *Stat Methods Med Res*. 2015.

16. Liu XS. Statistical power in three-arm cluster randomized trials. *Evaluation & the health professions*. 2014; 37: 470-87.

17. Manju MA, Candel MJ and Berger MP. Sample size calculation in cost-effectiveness cluster randomized trials: optimal and maximin approaches. *Stat Med*. 2014; 33: 2538-53.
18. Moerbeek M. Sample Size Issues for Cluster Randomized Trials With Discrete-Time Survival Endpoints. *Methodology*. 2012; 8: 146-58.
19. Moerbeek M. Sample size issues for cluster randomized trials with discrete-time survival endpoints. *Methodology*. 2012; 8: 146-58.
20. Moulton LH, Golub JE, Durovni B, et al. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clin Trials*. 2007; 4: 190-9.
21. Pornprasertmanit S and Schneider WJ. Accuracy in parameter estimation in cluster randomized designs. *Psychol Methods*. 2014; 19: 356-79.
22. Reboussin BA, Preisser JS, Song EY and Wolfson M. Sample size estimation for alternating logistic regressions analysis of multilevel randomized community trials of under-age drinking. *Journal of the Royal Statistical Society Series A*. 2012; 175.
23. Rotondi M and Donner A. Sample size estimation in cluster randomized trials: An evidence-based perspective. *Computational Statistics and Data Analysis*. 2012; 56: 1174-87.
24. Tokola K, Laracque D, Nevalainen J and Oja H. Power, sample size and sampling costs for clustered data. *Statistics and Probability Letters*. 2011; 81: 852-60.
25. van Breukelen GJ and Candel MJ. Efficient design of cluster randomized and multicentre trials with unknown intraclass correlation. *Stat Methods Med Res*. 2011.
26. van Breukelen GJ and Candel MJ. Calculating sample sizes for cluster randomized trials: we can keep it simple and efficient! *J Clin Epidemiol*. 2012; 65: 1212-8.
27. van Schie S and Moerbeek M. Re-estimating sample size in cluster randomised trials with active recruitment within clusters. *Stat Med*. 2014; 33: 3253-68.
28. Zhong Y and Cook RJ. Sample size and robust marginal methods for cluster-randomized trials with censored event times. *Stat Med*. 2015; 34: 901-23.

vii Poster presentations

Rutterford *et al. Trials* 2013, **14**(Suppl 1):P122
http://www.trialsjournal.com/content/14/S1/P122

TRIALS

POSTER PRESENTATION **Open Access**

Inadequate reporting of sample size calculations in cluster randomised trials: a review

Clare Rutterford[1*], Monica Taljaard[2], Stephanie Dixon[3], Andrew Copas[4], Sandra Eldridge[1]

From 2nd Clinical Trials Methodology Conference: Methodology Matters
Edinburgh, UK. 18-19 November 2013

Objectives

To assess the adequacy of reporting sample size calculations in published cluster randomised trials (CRTs) and to evaluate the accuracy and justifications behind the a priori estimates used.

Methods

A review was conducted of 166 CRTs reporting sample size calculations published between 2000 and 2008. Each trial was reviewed independently by two statisticians. The adequacy of the reporting of key elements in the CONSORT recommendations for CRTs was evaluated. Comparisons were made between the authors' a priori assumptions and values then observed in the trial.

Results

Of 166 trials, only 56 (34%) reported all key elements of sample size calculations in line with CONSORT recommendations. Elements specific to CRTs were the worst reported: the number of clusters or average cluster size was specified in only 94 (57%) and a measure of intracluster correlation coefficient (ICC) in only 86 (52%). Only 20 papers (12%) reported a priori and observed ICC values. In the majority of these reports, the a priori estimate for the ICC was conservative compared to the observed value. Few authors provided justifications for their choice of a priori estimates. Not unexpectedly, trials which reported no statistically significant difference were more likely to observe effect sizes smaller than the assumed clinically important difference.

Conclusions

Even with the CONSORT extension to CRTs, the reporting of sample size calculations in CRTs remains below that necessary for transparent reporting. Further awareness is needed to encourage the reporting of observed ICCs in order to evaluate the choice of a priori estimates and interpret the trial results.

Authors' details
[1]Queen Mary University of London, London, UK. [2]Ottawa Hospital Research Institute, Ottawa Hospital, Ottawa, Canada. [3]Schulich School of Medicine and Dentistry, University of Western Ontario, London, Canada. [4]MRC London Hub for Trials Methodology Research, London, UK.

Published: 29 November 2013

doi:10.1186/1745-6215-14-S1-P122
Cite this article as: Rutterford *et al.*: Inadequate reporting of sample size calculations in cluster randomised trials: a review. *Trials* 2013 14(Suppl 1):P122.

[1]Queen Mary University of London, London, UK
Full list of author information is available at the end of the article

Barts and The London School of Medicine and Dentistry

Blizard Institute, Centre for Primary Care and Public Health

Reporting of Sample Size Calculations in Cluster Randomised trials is Inadequate: a Review

Clare Rutterford[1], Monica Taljaard[2], Stephanie Dixon[3], Andrew Copas[4], Sandra Eldridge[1]

[1]Centre for Primary Care and Public Health, Barts and The London School of Medicine and Dentistry, Queen Mary University of London ,United Kingdom [2]Ottawa Hospital Research Institute, Ottawa Hospital, Ottawa, Ontario, Canada, and Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada, [3]Department of Epidemiology and Biostatistics, Schulich School of Medicine and Dentistry, University of Western Ontario, London, Ontario, Canada, [4]Hub for Trials Methodology Research, MRC Clinical Trials Unit at University College London, London, United Kingdom

INTRODUCTION

- Transparent reporting of the sample size calculation is important to show that a trial is designed to adequately address the research question
- Without transparent reporting the results from trials that may be well conducted risk having little impact
- The CONSORT statement extension for cluster randomised trials (CRTs)[1,2] describes five required elements for the transparent reporting of sample size calculations. These include whether/how the sample size calculation accounted for within cluster correlation. Failure to account for clustering in the calculation risks an increased Type II error rate
- Using a large sample of published cluster randomised trials we review the adequacy of sample size reporting in cluster randomised trials

METHODS

Data source: A previously published[3] review of 300 published reports of CRTs in primary care and public health from 2000-2008

Data abstraction for each trial included:

- The estimates reported within the sample size calculation for target difference, power, type I error, number of clusters, and within cluster correlation
- The sample size methodology used
- The observed values of the estimates at the end of the trial

Data validation: Data abstraction was conducted independently in pairs (by CR, MT and SD) and discrepancies were resolved by consensus

Data Analysis:

- Proportions reporting each CONSORT required sample size element
- Summary of sample size methodology used
- Discrepancies between *a-priori* estimates and observed data summarised

RESULTS

Figure 1: CONSORT required elements specific to cluster randomisation are worst reported

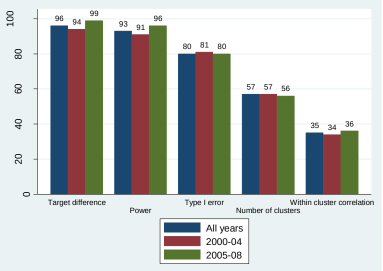
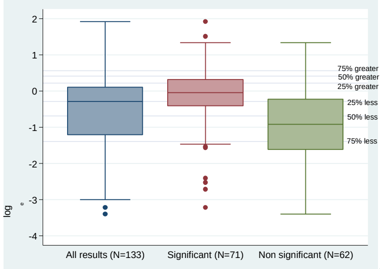


Table 1: Sample size methodology practices for 55% of trials that reported a calculation

Sample size method	N 166	% 55
Patient level accounting for correlation	91	55
Cluster level	9	5
Patient level without accounting for correlation	48	29
Unclear /Other	18	11
Accounted for attrition	38	23
Accounted for variable cluster sizes	1	1

Figure 2: Trials with non-significant results have larger target difference a-priori



CONCLUSIONS

- Sample size elements specific to CRTs were the worst reported. There was no real improvement seen in the years post CONSORT (2005-08)
- 55% of trials reported a sample size calculation. Of these, only 55% clearly accounted for clustering
- Additional adjustments for attrition and variable cluster sizes were made in only 38 (23%) and 1(1%) of trials. Failure to consider these may lead to an underestimate of the sample size required
- The observed treatment difference was often smaller than the *a-priori* target difference. This could be because interventions are very often ineffective or because in some cases investigators choose inappropriately large target differences. Similar discrepancies have been seen in individually randomised trials[4]

RECOMMENDATIONS

- Journals and peer reviewers should implement stricter requirements for authors to follow the CONSORT statement and its extensions
- More consideration should be given to the choice of the target difference in the sample size calculation

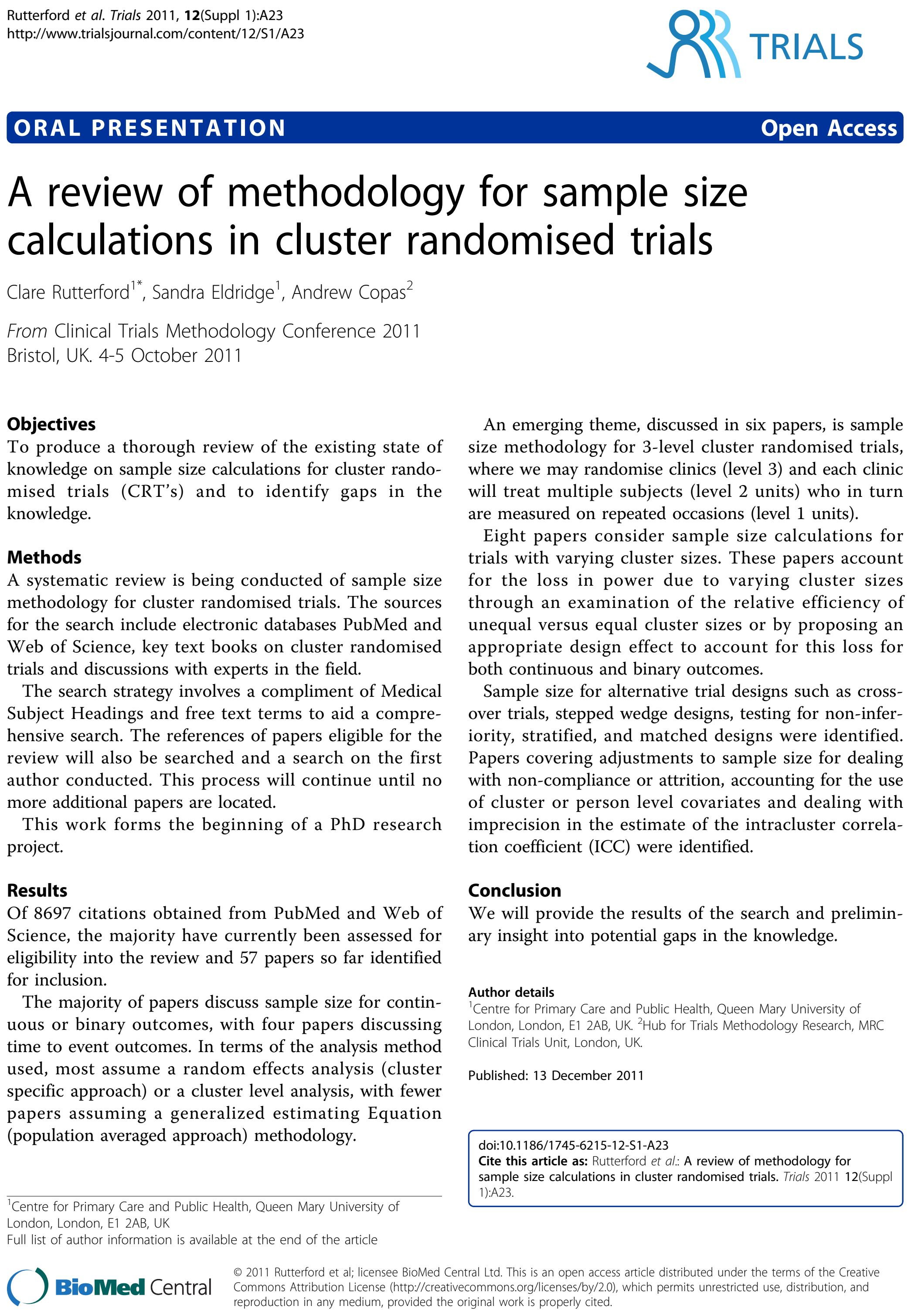
REFERENCES

1. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. BMJ2004;328:702-8
2. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345:e5661.
3. Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. BMJ. 2011;343:d5886.
4. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. BMJ. 2009;338:b1732.

PCTU Ottawa Hospital Research Institute Institut de recherche de l'Hôpital d'Ottawa MRC Clinical Trials Unit

www.blizard.qmul.ac.uk

viii Oral presentations



Rutterford *et al. Trials* 2011, **12**(Suppl 1):A23
http://www.trialsjournal.com/content/12/S1/A23

TRIALS

ORAL PRESENTATION **Open Access**

A review of methodology for sample size calculations in cluster randomised trials

Clare Rutterford[1*], Sandra Eldridge[1], Andrew Copas[2]

From Clinical Trials Methodology Conference 2011
Bristol, UK. 4-5 October 2011

Objectives

To produce a thorough review of the existing state of knowledge on sample size calculations for cluster randomised trials (CRT's) and to identify gaps in the knowledge.

Methods

A systematic review is being conducted of sample size methodology for cluster randomised trials. The sources for the search include electronic databases PubMed and Web of Science, key text books on cluster randomised trials and discussions with experts in the field.

The search strategy involves a compliment of Medical Subject Headings and free text terms to aid a comprehensive search. The references of papers eligible for the review will also be searched and a search on the first author conducted. This process will continue until no more additional papers are located.

This work forms the beginning of a PhD research project.

Results

Of 8697 citations obtained from PubMed and Web of Science, the majority have currently been assessed for eligibility into the review and 57 papers so far identified for inclusion.

The majority of papers discuss sample size for continuous or binary outcomes, with four papers discussing time to event outcomes. In terms of the analysis method used, most assume a random effects analysis (cluster specific approach) or a cluster level analysis, with fewer papers assuming a generalized estimating Equation (population averaged approach) methodology.

An emerging theme, discussed in six papers, is sample size methodology for 3-level cluster randomised trials, where we may randomise clinics (level 3) and each clinic will treat multiple subjects (level 2 units) who in turn are measured on repeated occasions (level 1 units).

Eight papers consider sample size calculations for trials with varying cluster sizes. These papers account for the loss in power due to varying cluster sizes through an examination of the relative efficiency of unequal versus equal cluster sizes or by proposing an appropriate design effect to account for this loss for both continuous and binary outcomes.

Sample size for alternative trial designs such as cross-over trials, stepped wedge designs, testing for non-inferiority, stratified, and matched designs were identified. Papers covering adjustments to sample size for dealing with non-compliance or attrition, accounting for the use of cluster or person level covariates and dealing with imprecision in the estimate of the intracluster correlation coefficient (ICC) were identified.

Conclusion

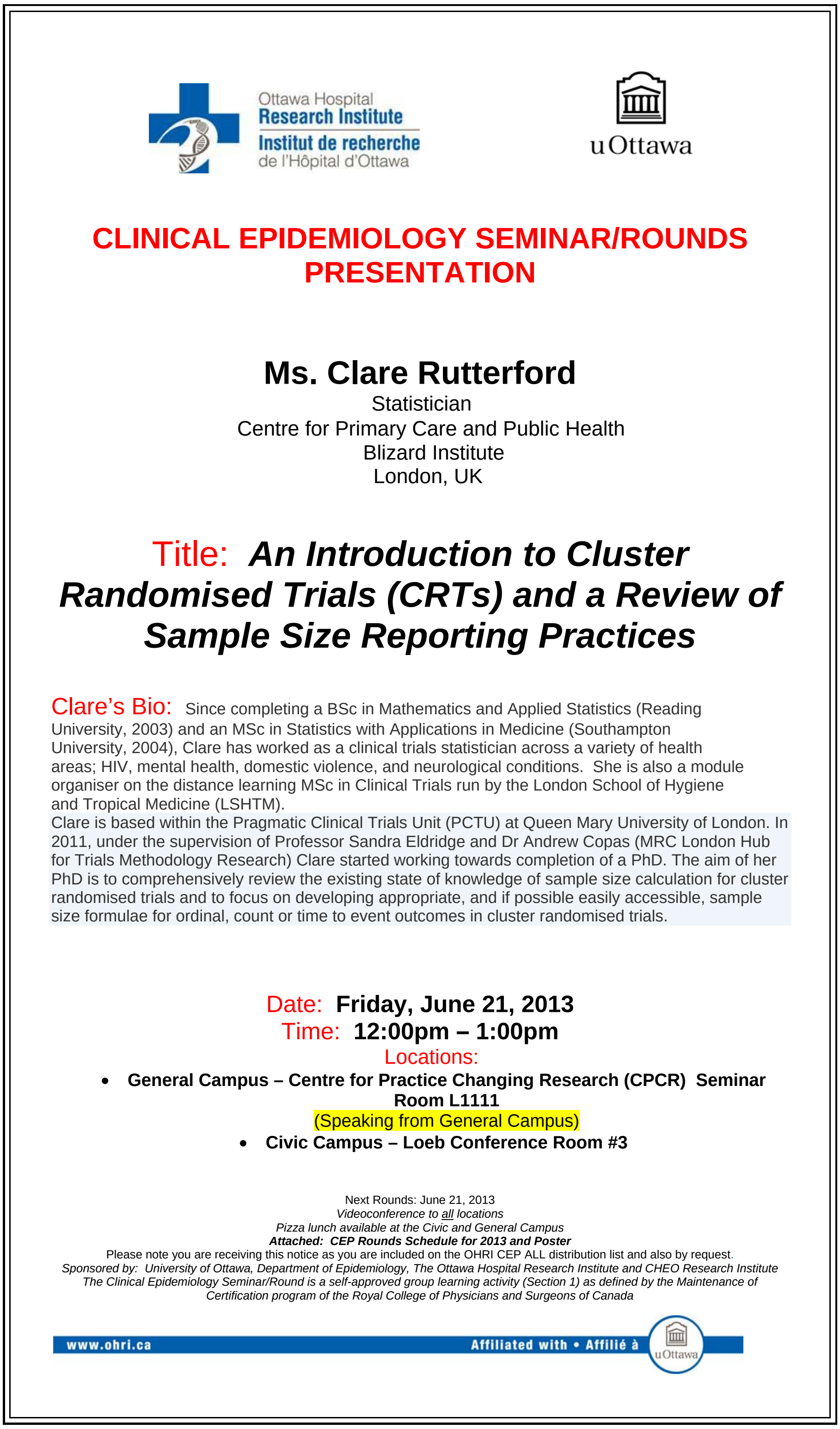
We will provide the results of the search and preliminary insight into potential gaps in the knowledge.

Author details
[1]Centre for Primary Care and Public Health, Queen Mary University of London, London, E1 2AB, UK. [2]Hub for Trials Methodology Research, MRC Clinical Trials Unit, London, UK.

Published: 13 December 2011

doi:10.1186/1745-6215-12-S1-A23
Cite this article as: Rutterford *et al.*: A review of methodology for sample size calculations in cluster randomised trials. *Trials* 2011 **12**(Suppl 1):A23.

[1]Centre for Primary Care and Public Health, Queen Mary University of London, London, E1 2AB, UK
Full list of author information is available at the end of the article



Ottawa Hospital Research Institute
Institut de recherche de l'Hôpital d'Ottawa

uOttawa

CLINICAL EPIDEMIOLOGY SEMINAR/ROUNDS PRESENTATION

Ms. Clare Rutterford
Statistician
Centre for Primary Care and Public Health
Blizard Institute
London, UK

Title: ***An Introduction to Cluster Randomised Trials (CRTs) and a Review of Sample Size Reporting Practices***

Clare's Bio: Since completing a BSc in Mathematics and Applied Statistics (Reading University, 2003) and an MSc in Statistics with Applications in Medicine (Southampton University, 2004), Clare has worked as a clinical trials statistician across a variety of health areas; HIV, mental health, domestic violence, and neurological conditions. She is also a module organiser on the distance learning MSc in Clinical Trials run by the London School of Hygiene and Tropical Medicine (LSHTM).

Clare is based within the Pragmatic Clinical Trials Unit (PCTU) at Queen Mary University of London. In 2011, under the supervision of Professor Sandra Eldridge and Dr Andrew Copas (MRC London Hub for Trials Methodology Research) Clare started working towards completion of a PhD. The aim of her PhD is to comprehensively review the existing state of knowledge of sample size calculation for cluster randomised trials and to focus on developing appropriate, and if possible easily accessible, sample size formulae for ordinal, count or time to event outcomes in cluster randomised trials.

Date: **Friday, June 21, 2013**
Time: **12:00pm – 1:00pm**
Locations:

- **General Campus – Centre for Practice Changing Research (CPCR) Seminar Room L1111**
(Speaking from General Campus)
- **Civic Campus – Loeb Conference Room #3**

Next Rounds: June 21, 2013
Videoconference to all locations
Pizza lunch available at the Civic and General Campus
Attached: CEP Rounds Schedule for 2013 and Poster
Please note you are receiving this notice as you are included on the OHRI CEP ALL distribution list and also by request.
Sponsored by: University of Ottawa, Department of Epidemiology, The Ottawa Hospital Research Institute and CHEO Research Institute
The Clinical Epidemiology Seminar/Round is a self-approved group learning activity (Section 1) as defined by the Maintenance of Certification program of the Royal College of Physicians and Surgeons of Canada

www.ohri.ca
Affiliated with • Affilié à uOttawa

Sample size calculation for cluster randomised trials with ordinal outcomes

Clare Rutterford[1], Andrew Copas[2], Sandra Eldridge[1]

[1] Centre for Primary Care and Public Health, Blizard Institute, Barts and The London School of Medicine and Dentistry, Yvonne Carter Building, 58 Turner Street, London, E1 2AB.

[2] Hub for Trials Methodology Research, MRC Clinical Trials Unit at University College London, United Kingdom

Contact details:

Clare: Telephone: 020 7882 2518, c.m.rutterford@qmul.ac.uk

Sandra: Telephone: 020 7882 2519, s.eldridge@qmul.ac.uk

Andrew: anc@ctu.mrc.ac.uk

Conference stream: Medical/biometrics/clinical trials

Presenter biography

After completing a BSc in Mathematics and Applied Statistics (Reading University, 2003) and an MSc in Statistics with Applications in Medicine (Southampton University, 2004), Clare spent two years as a statistician at the Medical Research Council Clinical Trials Unit, London, working on a large phase III HIV prevention trial, part of the Microbicides Development Programme. In 2007 Clare joined the Clinical Trials Unit at the Institute of Psychiatry, London, where she was the trial statistician for several randomised controlled trials in the areas of dementia, forensic mental health, Attention Deficit Hyperactivity Disorder, and neurological conditions.

She joined the Centre for Primary Care and Public Health in March 2009. Since October 2009 Clare has also been a module organiser on the distance learning MSc in Clinical Trials run by the London School of Hygiene and Tropical Medicine.

In 2011, under the supervision of Professor Sandra Eldridge (Centre for Primary Care and Public Health) and Dr Andrew Copas (MRC London Hub for Trials Methodology Research) Clare started working towards completion of a PhD. The aim of her PhD is to comprehensively review the existing state of knowledge of sample size calculation for cluster randomised trials and to focus on developing appropriate, and if possible easily accessible, sample size formulae for ordinal, count or time to event outcomes in cluster randomised trials.

Intended audience

This talk is intended for statisticians who are involved in trials or have an interest in sample size calculations and/or the use of ordinal outcomes.

Abstract (max 250 words)

Background

A common approach to sample size calculation for cluster randomised trials is to calculate the sample size assuming individual randomisation and multiply this by the design effect. Calculation of the design effect requires knowledge of the cluster size and intracluster correlation coefficient (ICC), a measure of the extent of clustering. It is not yet clear how well the design effect method works for ordinal outcomes.

Objectives

To evaluate the performance of the design effect approach in sample size calculation for cluster randomised trials with ordinal outcomes and to provide recommendations for the calculation of the ICC.

Methods

The performance of the design effect method was evaluated across a range of scenarios, chosen to reflect the characteristics of trials using ordinal outcomes.

For each scenario sample size was calculated using the design effect approach with three alternative estimators for the ICC: an ANOVA based estimate, a Kappa type estimate and the ICC on the assumed underlying continuous outcome. A thousand datasets of appropriate size were generated and each analysed via a random-effects model. Empirical power was calculated as the proportion of datasets with significant result.

The simulation studies were extended to explore performance under alternative analysis methods and variable cluster sizes.

Conclusions

The use of the design effect works well with ordinal data, except in the case of a small number of clusters. Power calculations using the ANOVA estimate of the ICC performed adequately, and that ICC is simple to calculate.

ix Publications

ELSEVIER

CrossMark

Journal of Clinical Epidemiology 68 (2015) 716−723

Journal of Clinical Epidemiology

Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: a review

Clare Rutterford[a,*], Monica Taljaard[b,c], Stephanie Dixon[d], Andrew Copas[e], Sandra Eldridge[a]

[a]*Centre for Primary Care and Public Health, Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Yvonne Carter Building, 58 Turner Street, London E1 2AB, UK*

[b]*Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa Hospital, Civic Campus, 1053 Carling Avenue, Civic Box 693, Ottawa, Ontario K1Y 4E9, Canada*

[c]*Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada*

[d]*Department of Epidemiology and Biostatistics, Schulich School of Medicine and Density, University of Western Ontario, 1151 Richmond St, London, Ontario N6A 3K7, Canada*

[e]*Hub for Trials Methodology Research, MRC Clinical Trials Unit at University College London, Aviation House, 125 Kingsway, London, WC2B 6NH, UK*

Accepted 17 October 2014; Published online 15 December 2014

Abstract

Objectives: To assess the quality of reporting and accuracy of a priori estimates used in sample size calculations for cluster randomized trials (CRTs).

Study Design and Setting: We reviewed 300 CRTs published between 2000 and 2008. The prevalence of reporting sample size elements from the 2004 CONSORT recommendations was evaluated and a priori estimates compared with those observed in the trial.

Results: Of the 300 trials, 166 (55%) reported a sample size calculation. Only 36 of 166 (22%) reported all recommended descriptive elements. Elements specific to CRTs were the worst reported: a measure of within-cluster correlation was specified in only 58 of 166 (35%). Only 18 of 166 articles (11%) reported both a priori and observed within-cluster correlation values. Except in two cases, observed within-cluster correlation values were either close to or less than a priori values.

Conclusion: Even with the CONSORT extension for cluster randomization, the reporting of sample size elements specific to these trials remains below that necessary for transparent reporting. Journal editors and peer reviewers should implement stricter requirements for authors to follow CONSORT recommendations. Authors should report observed and a priori within-cluster correlation values to enable comparisons between these over a wider range of trials. © 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Keywords: CONSORT statement; Reporting; Cluster randomized trial; Sample size; Intracluster correlation coefficient; Statistical methods

1. Introduction

In a cluster randomized trial (CRT), groups or "clusters," rather than the constituent individuals themselves, are randomly allocated to interventions [1,2]. A cluster could be, for example a medical practice, hospital, or community. Cluster randomization may be deemed necessary when randomization at individual level is impractical, for example, when the intervention is necessarily administered at the cluster level. There can also be scientific reasons to adopt cluster randomization, for example to avoid contamination between treatment groups or for reasons of administrative convenience or cost. In a CRT, the responses from different individuals within the same cluster are usually more similar than those from different clusters. The degree of this correlation is commonly quantified by the intracluster correlation coefficient (ICC) but can also be quantified using the coefficient of variation of the outcome, often referred to as k [3]. Sample size calculations for CRTs must take correlation into account to avoid potentially underestimating the required sample size. Donner et al. [4] proposed that a sample size calculated assuming individual randomization can be inflated by a design effect (DE) to reach the required level of statistical power under cluster randomization. This DE is.

Conflict of interest: None.

Funding: C.R. is funded by the Medical Research Council (ref:1031857). The funding agency had no role in the study design; collection, analysis, or interpretation of data; writing of the manuscript; or in the decision to submit the manuscript for publication.

* Corresponding author. Tel.: +44-(0)20-7882-2518; fax: +44 (0)20 7882 2552.

E-mail address: c.m.rutterford@qmul.ac.uk (C. Rutterford).

http://dx.doi.org/10.1016/j.jclinepi.2014.10.006

What is new?

- This is the first article to evaluate sample size reporting prevalence for each of the CONSORT 2004 recommended sample size descriptive elements for cluster randomized trials (CRTs).
- There is much room for improvement in sample size reporting. Sample size elements specific to CRTs were the worst reported.
- Comparisons between a small sample of a priori estimates and observed values of the within-cluster correlation show the a priori estimates to be reasonably accurate.
- There seems to be a discrepancy between endorsement of CONSORT by journals and implementation by authors.
- We recommend journals to consider making adherence to CONSORT guidelines and extension statements a condition of publication.

$$DE = 1 + (n - 1)\rho$$

where n is the number of individuals per cluster and ρ the ICC. For example, using this formula, a trial with 35 individuals per cluster and an anticipated ICC of 0.01 would require 34% more participants than the equivalent individually randomized design. When cluster sizes are unequal, n is usually replaced by the average cluster size, although this risks potentially underestimating the required sample size [5]. Since 2001, various methods for accounting for variable cluster size have been published [5–10].

Reporting how the sample size calculation was performed is important from both a scientific and ethical perspective to show that the trial was designed to adequately address the research question without wasting resources or exposing too many participants to potentially harmful interventions. The CONSORT statement recommends the reporting of 25 items related to the design, conduct, and analysis of randomized controlled trials. The statement was first published in 1996 [11] and revised in 2001 [12] and 2010 [13]. This latest revision is referred to as CONSORT 2010. The CONSORT statement was developed with the dual aims of standardizing reporting and facilitating transparency. The transparent reporting of sample size methods and assumptions provides some reassurance to the reader of the quality with which the trial has been conducted. Ideally, adequate reporting ensures that a reader can appraise the methodology; identify whether an appropriate and a priori calculation was performed for the study design; and assess whether the assumptions made in the sample size calculation were reasonable.

The item that relates to describing the sample size calculation in CONSORT 2010 recommends the following descriptive elements to be reported (1) the estimated outcomes in each group (which implies the minimum important treatment effect), (2) the level of significance [or the α (type I) error level], (3) the statistical power [or the β (type II) error level], and (4) for continuous outcomes, the assumed standard deviation of the measurements. The CONSORT statement was extended in 2004 [14] for the reporting of CRTs, and this was revised in 2012 [15] to be in line with CONSORT 2010. This extension includes adaptations to items relevant to the reporting of CRTs. The item that relates to describing the sample size calculation additionally recommends the reporting of two further descriptive elements (5) the number of clusters or the cluster size and (6) the ICC or coefficient of variation (k), along with a measure of its uncertainty. The 2012 revision additionally recommends specification of whether equal or unequal cluster sizes are assumed.

Adherence to all the reporting items provided in the 2004 CONSORT extension for CRTs has been reviewed in 23 trials in oral health [16], 300 randomly sampled trials [17], 106 trials in children [18], and 73 trials in residential facilities [19]. The presence of a sample size calculation in a trial report is considered an initial indication of reporting quality in the area of sample size. Across these reviews, the proportion presenting a calculation was 21 of 23 (91%), 164 of 300 (55%), 87 of 106 (82%), and 43 of 73 (59%), respectively. The third review also reports that 63 of 87 (72%) of trials reported all 2004 CONSORT recommended sample size descriptive elements. Unlike our review, none of these reviews identify which of the six sample size descriptive elements are reported and which are not.

To ensure methodological quality, a sample size calculation should be appropriate to the trial design. For a CRT, this implies a sample size calculation accounting for clustering, either by correctly accounting for the clustered nature of the data using a suitable estimate of within-cluster correlation or calculating the sample size from cluster-level measures. A cluster-level sample size provides the number of clusters required and may be undertaken when information on cluster size is unavailable at the design stage and/or the primary outcome itself takes some account of cluster size [20]. Many reviews of the methodological conduct of CRTs show that this methodological aspect is suboptimal: for example, in two reviews mentioned above, 100 of 164 (61%) and 15 of 21 (71%) trials that reported a sample size calculation also accounted for the clustered nature of the design [16,17].

In addition to using an appropriate methodology, it is vital that a sample size calculation is realistic and that, in principle, the required numbers can be recruited. Ideally, sample size calculations should use a within-cluster correlation based on the best available data and specify a minimum treatment effect that is both clinically important and is likely to be achievable based on evidence from previous trials of similar interventions. However, it is recognized that this is not always achievable as appropriate estimates may not be available or circumstances outside the investigators' control may make the same size unobtainable. A comparison between the estimates used in the sample size calculations and those observed at the end of the trial in a sample of individually randomized trials concluded that sample size calculations are often based on inaccurate assumptions [21]. We are unaware of any previous studies considering similar discrepancies for CRTs.

The review by Ivers et al. [17] found that the prevalence of reporting sample size calculations was low and that calculations were not always appropriate for the clustered nature of the trial. In this review, we use the same sample of trials to look in more detail at sample size calculations in these trials, assessing adherence to reporting descriptive sample size elements in the 2004 CONSORT extension. In particular, we aim to identify which elements are underreported, whether reporting has improved since the introduction of the 2004 CONSORT extension, and the accuracy of the a priori estimates used in the sample size calculation by making comparisons with their observed values at the end of the trial.

2. Materials and methods

2.1. Search strategy

We used a previously published review of 300 reports of CRTs randomly sampled from MEDLINE. The search strategy and characteristics of the included studies have been described in detail elsewhere [17]. Briefly, a publication was included if it was published in an English language journal between the years 2000 and 2008, and it was the main report of a CRT; trial protocols, pilot studies, secondary analyses of CRTs, and trials with households or families as clusters, with multistage designs, or presenting only baseline findings were excluded.

2.2. Data abstraction

We reviewed the sample to identify those that reported prospective sample size calculations. Data abstraction was conducted only on these trials. All abstraction relates to the outcome used in the sample size calculation. This was the primary outcome in all but 30 trials.

Descriptive information for each journal included year of publication, impact factor, and whether the journal endorsed the CONSORT statement (taken from the CONSORT Web site [22]). As timing and strength of endorsement are difficult to define, this variable was provided as a rough indicator of journal quality only and no formal comparisons of its impact were made. Descriptive information for each trial included trial design, method of randomization, health area, data type of the sample size outcome, and whether a statistically significant result was seen for the sample size outcome (as reported by authors).

For the primary objective of reporting quality, abstracted information for each trial included the values of the 2004 CONSORT required sample size elements and additionally whether adjustments had been made for variable cluster sizes or attrition. For the secondary objective of methodological quality and accuracy, abstracted information included whether the sample size accounted for clustering, including any methodology cited, any justifications provided for the estimates of the within-cluster variation and estimates of outcomes in the control and treatment groups, and the corresponding observed values of treatment effect and within-cluster correlation at the end of the trial. No assessment was made of whether the correct formulae had been referenced or implemented given the particular trial design or whether the sample size calculation could be reproduced.

The research team developed and piloted a data abstraction instrument and corresponding Access database for electronic data storage. Three experienced statisticians (C.R., M.T., and S.D.) abstracted the data for all the articles in rotating pairs. After each set of 10 trials had been abstracted, discrepancies were reviewed within the pair and resolved by discussion.

2.3. Data analysis

2.3.1. Description of sample

Characteristics of the journals and trials for articles containing sample size calculations are summarized using frequencies and percentages or medians and interquartile ranges.

2.3.2. Reporting according to CONSORT guidelines

For each trial, we describe the 2004 CONSORT required elements that were provided for the sample size calculation, with a maximum of six elements. The required elements were (1) the type I error rate, (2) power, (3) estimates of outcomes in each group or minimum important target effect, (4) the standard deviation for continuous outcomes, (5) the number of clusters or average cluster size, and (6) the assumed measure of intracluster correlation, design effect, or coefficient of variation. Trials that do not have a continuous sample size outcome would not be expected to provide a standard deviation, and similarly, trials for which a cluster-level analysis has been assumed in the sample size calculation are not required to provide a value of within-cluster correlation. For each trial, we calculate the maximum number of possible 2004 CONSORT elements that could have been reported (given the type of outcome) and the proportion that were reported. For completeness, we summarize the number of articles that make adjustments for attrition and variable cluster sizes, although we expect the latter to be low given that most publications on the subject were written after the sample sizes would have been decided for most of our trials, and the specific 2012 CONSORT recommendation for reporting this was introduced after our sample was collected.

Adherence to the 2004 CONSORT elements is compared for those articles published before (2000–2004) and after (2005–2008) its publication.

2.3.3. Assessment of methodological approach

The sample size methodology and justifications for the values of a priori estimates are summarized using frequencies and percentages. A scatter plot of a priori estimates and observed values of within-cluster correlations is presented.

Where possible, discrepancies are calculated between the minimum important target effect and the observed effect as (observed effect/minimum important target effect). The calculation is performed on the scale the authors used to report the effect, that is, absolute or relative differences, odds ratio, or hazard ratio. When the observed effect is smaller than the a priori minimum important target effect, this value is <1 regardless of measurement scale.

3. Results

Of the 300 articles, 166 trials (55%) reported a priori sample size calculations and are thus summarized. (This number differs from the 164 reported in the original review, which focused on sample size calculation for a particular variable identified by reviewers as primary.) The vast majority [155 of 166 (93%)] were parallel group trials, implementing simple or stratified randomization [125 of 166 (75%)], and used binary or continuous primary outcomes [136 of 166 (82%)] (Table 1).

The median impact factor of included journals was fairly low (3.6, interquartile range: 2.3–12.1); however, 99 of 166 (60%) of these were journals whose guidance recommends use of CONSORT. Nearly half (46%) were published in the years preceding the first publication of the CONSORT extension for CRTs.

3.1. Reporting according to CONSORT guidelines

Of the CONSORT required elements for sample size calculation, the most commonly reported were outcome levels in each group or minimum important target effect (160 of 166, 96%), power (155 of 166, 93%), and type I error rate (133 of 166, 80%), Table 2. The elements specific to CRTs were reported less frequently: number of clusters or cluster size [94 of 166 (57%)] and a measure of within-cluster correlation [58 of 166 (35%)].

No articles reported corresponding measures of uncertainty alongside the correlation estimates, although some [18 of 102 (18%)] assessed sample size sensitivity based on a range of within-cluster correlation values (Table 3).

The assumed standard deviation was reported in only 18 (32%) of the 56 articles with a continuous outcome. Only 38 (23%) of 166 trials reported explicitly accounting for attrition, and only one article accounted for variable cluster sizes (Table 2).

Table 1. Journal and trial characteristics of 166 cluster randomized trials reporting sample size calculations

Characteristic	*N* (%)
Journal: year of publication	
2000–2004	77 (46)
2005–2008	89 (54)
Journal: impact factor, median (IQR)	3.6 (2.3–12.1)
Journal: endorsement of CONSORT	99 (60)
Trial: design	
Parallel group	155 (93)
Factorial	5 (3)
Crossover	4 (2)
Stepped wedge	1 (1)
Balanced incomplete block design	1 (1)
Cross sectional[a]	56 (34)
Cohort[a]	110 (66)
Trial: method of random allocation:	
Completely randomized	61 (37)
Stratified	64 (39)
Within matched sets	31 (19)
Minimization	9 (5)
Other	1 (1)
Trial: health area	
Primary or hospital care	116 (70)
Public health	50 (30)
Trial: type of primary outcome	
Dichotomous	80 (48)
Continuous	56 (34)
Rate	16 (10)
Ordinal	1 (1)
Categorical	0 (0)
Count	0 (0)
Time to event	0 (0)
Unclear	13 (8)

Abbreviation: IQR, interquartile range.

[a] In a cohort design, repeated measurements are taken on the same individuals at each time point, and in a cross-sectional design, repeated measurements take place on different individuals.

Only 36 of 166 articles (22%) reported all the 2004 CONSORT required elements. There was some improvement in reporting over time with 23 of 89 trials (26%) reporting all the CONSORT required elements after the introduction of the extension, compared with only 13 of 77 (8%) before the 2004 CONSORT extension (Table 2). In particular, improvements were observed with respect to the standard deviation for continuous outcomes and the reported value of the within-cluster correlation. Categorization of journals as above or below the median impact factor indicated that articles in higher impact journals tended to report more CONSORT elements: in lower impact journals, 13 of 83 (16%) reported all required elements compared with 23 of 83 (28%) in higher impact journals.

3.2. Assessment of methodological approach

Of the 166 articles reporting a sample size calculation, 102 of 166 (61%) clearly accounted for the within-cluster

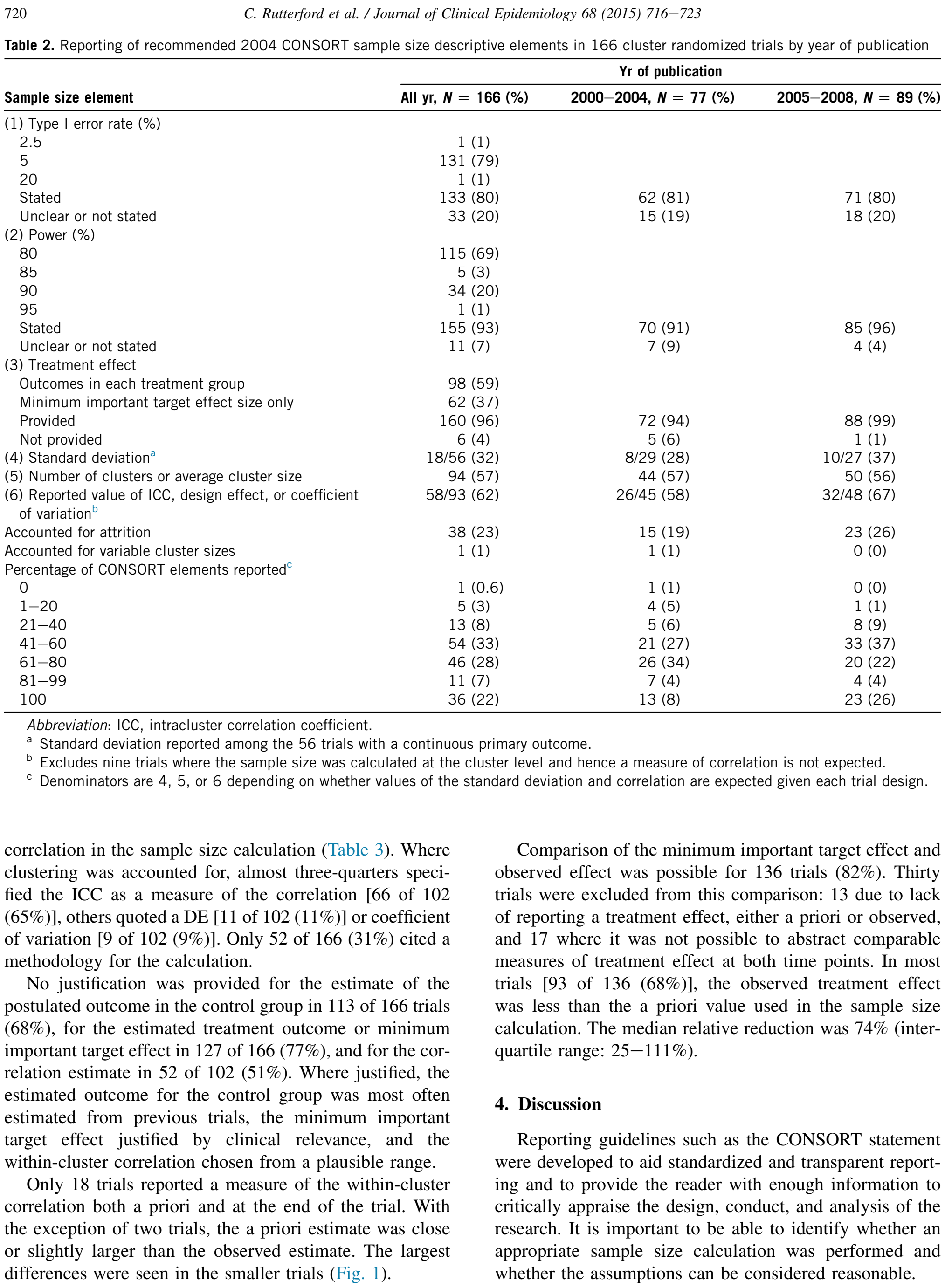


Table 2. Reporting of recommended 2004 CONSORT sample size descriptive elements in 166 cluster randomized trials by year of publication

	Yr of publication		
Sample size element	**All yr, N = 166 (%)**	**2000–2004, N = 77 (%)**	**2005–2008, N = 89 (%)**
(1) Type I error rate (%)			
2.5	1 (1)		
5	131 (79)		
20	1 (1)		
Stated	133 (80)	62 (81)	71 (80)
Unclear or not stated	33 (20)	15 (19)	18 (20)
(2) Power (%)			
80	115 (69)		
85	5 (3)		
90	34 (20)		
95	1 (1)		
Stated	155 (93)	70 (91)	85 (96)
Unclear or not stated	11 (7)	7 (9)	4 (4)
(3) Treatment effect			
Outcomes in each treatment group	98 (59)		
Minimum important target effect size only	62 (37)		
Provided	160 (96)	72 (94)	88 (99)
Not provided	6 (4)	5 (6)	1 (1)
(4) Standard deviation[a]	18/56 (32)	8/29 (28)	10/27 (37)
(5) Number of clusters or average cluster size	94 (57)	44 (57)	50 (56)
(6) Reported value of ICC, design effect, or coefficient of variation[b]	58/93 (62)	26/45 (58)	32/48 (67)
Accounted for attrition	38 (23)	15 (19)	23 (26)
Accounted for variable cluster sizes	1 (1)	1 (1)	0 (0)
Percentage of CONSORT elements reported[c]			
0	1 (0.6)	1 (1)	0 (0)
1–20	5 (3)	4 (5)	1 (1)
21–40	13 (8)	5 (6)	8 (9)
41–60	54 (33)	21 (27)	33 (37)
61–80	46 (28)	26 (34)	20 (22)
81–99	11 (7)	7 (4)	4 (4)
100	36 (22)	13 (8)	23 (26)

Abbreviation: ICC, intracluster correlation coefficient.

[a] Standard deviation reported among the 56 trials with a continuous primary outcome.

[b] Excludes nine trials where the sample size was calculated at the cluster level and hence a measure of correlation is not expected.

[c] Denominators are 4, 5, or 6 depending on whether values of the standard deviation and correlation are expected given each trial design.

correlation in the sample size calculation (Table 3). Where clustering was accounted for, almost three-quarters specified the ICC as a measure of the correlation [66 of 102 (65%)], others quoted a DE [11 of 102 (11%)] or coefficient of variation [9 of 102 (9%)]. Only 52 of 166 (31%) cited a methodology for the calculation.

No justification was provided for the estimate of the postulated outcome in the control group in 113 of 166 trials (68%), for the estimated treatment outcome or minimum important target effect in 127 of 166 (77%), and for the correlation estimate in 52 of 102 (51%). Where justified, the estimated outcome for the control group was most often estimated from previous trials, the minimum important target effect justified by clinical relevance, and the within-cluster correlation chosen from a plausible range.

Only 18 trials reported a measure of the within-cluster correlation both a priori and at the end of the trial. With the exception of two trials, the a priori estimate was close or slightly larger than the observed estimate. The largest differences were seen in the smaller trials (Fig. 1).

Comparison of the minimum important target effect and observed effect was possible for 136 trials (82%). Thirty trials were excluded from this comparison: 13 due to lack of reporting a treatment effect, either a priori or observed, and 17 where it was not possible to abstract comparable measures of treatment effect at both time points. In most trials [93 of 136 (68%)], the observed treatment effect was less than the a priori value used in the sample size calculation. The median relative reduction was 74% (interquartile range: 25−111%).

4. Discussion

Reporting guidelines such as the CONSORT statement were developed to aid standardized and transparent reporting and to provide the reader with enough information to critically appraise the design, conduct, and analysis of the research. It is important to be able to identify whether an appropriate sample size calculation was performed and whether the assumptions can be considered reasonable.

Table 3. Sample size approach and justification for values of a priori estimates for 166 cluster randomized trials

Sample size method	*N* = 166 (%)
(1) Sample size accounted for clustering	102/166 (61)
Intracluster correlation	66/102 (65)
Design effect	11/102 (11)
Coefficient of variation	9/102 (9)
Unclear or not stated	7/102 (7)
Cluster-level calculation	9/102 (9)
Did not account for clustering	48/166 (29)
Unclear	16/166 (10)
(2) Methodology cited	52/166 (31)
(3) Justification of a priori estimates	
(3a) The control group expected outcomes, N = 166	
No justification	113/166 (68)
Results from published data	39/166 (23)
A preliminary/pilot study	6/166 (4)
Conservative estimate	8/166 (5)
(3b) The treatment group expected outcome	*N* = 166
No justification	127/166 (77)
Results from published data	16/166 (10)
A preliminary/pilot study	3/166 (2)
Clinical relevance (no data referenced)	20/166 (12)
(3c) The within-cluster correlation estimate	*N* = 102
No justification	52/102 (51)
Results from published data	17/102 (17)
A preliminary/pilot study	6/102 (6)
Plausible range	18/102 (18)
NA due to cluster-level calculation	9/102 (9)

Abbreviation: NA, not applicable.

Among our included trials, adherence to the sample size recommendations of the 2004 CONSORT extension was poor, with elements specific to cluster randomized designs being the worst reported—a phenomenon seen in many reviews assessing the wider aspects of the 2004 CONSORT extension recommendations, for example [15]. Twenty-two percent of trials were compliant with all the 2004 CONSORT extension sample size element recommendations. This is comparable to results seen for individually randomized trials: In a review of 215 individually randomized trials published in 2005 and 2006 in six high-impact medical journals, 34% of trials included enough information for full replication of the sample size [21]. However, we are encouraged to see in our review some improvements over time in the number of trials reporting all the required CONSORT sample size elements with increases occurring in the reporting of the within-cluster correlation. This is particularly important given the large influence that this parameter has on the required sample size.

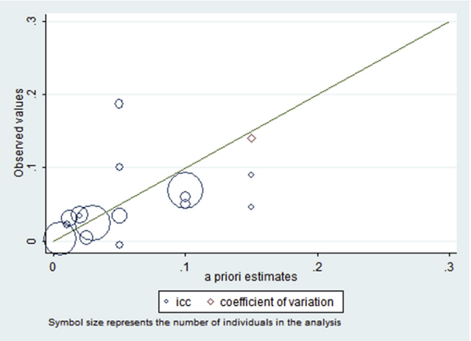


Fig. 1. A comparison of a priori and observed estimates of the within-cluster correlation.

Our sample contained only trials published until 2008. This allowed us to assess the immediate impact of the CONSORT extension statement in the 4 years after its publication. This review provides an important baseline with which to compare the impact of future CONSORT statements. We plan to repeat this review, including more recent articles, to assess the immediate effects of the 2012 revision once there has been sufficient time for it to take full effect and a large number of trials published. A secondary aim of a future review will be to look at the medium- to long-term impact of the 2004 CONSORT extension.

We acknowledge the limitation that the current review does not allow us to assess the medium- to long-term impact of the 2004 CONSORT extension. However, this sample is unique in both its size and its coverage of all medical areas and journals; hence, it provides us with an important insight into the immediate uptake of the CONSORT extension among the medical researcher field in general. Most of the other reviews of CRTs have contained less than 40 trials, with the largest containing 173 [16,19,23–33]. This sample of 300 is the largest to date, with the largest sample of identified sample size calculations. Previous reviews have also focused on particular areas of health such as stroke, oral health, or primary care or targeted particular high ranking journals during the search. Our sample was designed to be representative of CRTs across the health research field.

Additional adjustments for attrition and variable cluster sizes were made in only 38 (23%) and 1 (1%) of trials, respectively. We expected the latter figure to be low as relevant publications describing how to do this are relatively recent and the recommendation to report information on cluster size variability has only recently been included in CONSORT. Failure to take into account these additional considerations may lead to an underestimate of the sample size required, the effect of variable cluster size only being negligible when the coefficient of variation in cluster size is small (less than 0.23) [5]. Simple DE approaches to deal with variable cluster sizes are available [5–8].

Sixty percent of the articles included in our review were in journals whose recommendations mention CONSORT. Although we recognize that it is not known at what time each journal first endorsed CONSORT or whether the extension statements are similarly endorsed, there does seem to be a discrepancy between endorsement and use of CONSORT. The way in which medical journals incorporate CONSORT recommendations into their editorial process has been surveyed [34]. That survey found that 38% (62 of 165) and 3% (5 of 165) of journals recommend that authors comply with CONSORT and the cluster extension, respectively, within their instructions to authors, but only 37% (23 of 62) and 60% (3 of 5) make it a requirement. These figures are higher when surveying journal editors directly, with 69% (31 of 45) recommending that authors comply with the extension for CRTs. However, we recognize that endorsement of CONSORT by journals may not be the sole driver behind improving reporting quality. A recent review of reporting quality in CRTs concluded that quality of reporting and conduct was influenced more by the presence of a statistician (or quantitative researcher) among the authorship than a journal's endorsement of CONSORT [19].

From the original sample of 300 trials, only 166 (55%) presented an a priori sample size calculation. Where a sample size calculation was reported, clustering was accounted for in 102 of 166 (61%) of these trials. This figure is quite poor given the standard trial design adopted by many of the included studies, for which there are simple sample size methodologies available. A limitation of our study is that we did not assess whether each trial was appropriately powered. We did not reproduce sample size calculations or assess the assumptions made about cluster size, as a measure of observed cluster size variability was not collected and recommendations for reporting this have only been made recently.

In our sample, only a small number of trials reported both the a priori and observed correlation estimates. This was similarly seen in a review of CRTs in cancer screening where only 7 of 50 (14%) reported both a priori estimated and observed ICCs, and there was no evidence that its reporting improved after the CONSORT statement extension [35]. Given its influence, it is important for authors to provide the observed value—not only to aid the design of future trials, but also to allow one to assess the accuracy of the a priori estimates and hence the sample size. Albeit on a small and possibly unrepresentative sample, comparisons between a priori estimates and observed values of the within-cluster correlation showed that authors tended to assume conservative estimates. Part of the explanation for this may be that values of ICCs used in sample size calculations are usually not adjusted for covariates, whereas observed ICCs may be adjusted for covariates (and hence smaller). Furthermore, our results are consistent with part of any discrepancy between observed and assumed ICCs being due to sampling error; the discrepancy was generally larger in small trials. Whether our findings remain in a larger, more representative, sample remains to be seen. The a priori estimate of the within-cluster correlation may be taken from a previous trial or alternatively, if using the ICC, there are published summaries available for ICCs of specific outcomes [36−39]. In our sample, few authors provided an explanation for their assumed estimate.

The observed treatment effect was often smaller than the minimum important target effect used in the sample size calculation. This could be because interventions are often ineffective or because in some cases investigators power trials using minimum important target effects that could never be achieved with the type of intervention under evaluation. Careful consideration is required in determining the minimally important effect in trials of complex interventions, both in individually randomized and CRTs [1].

The results from our review show that there is much room for improvement in the conduct and reporting of sample size calculations in CRTs. We recommend that journals consider making adherence to CONSORT guidelines a condition of publication to aid improvement in the quality of reporting and methodological conduct of CRTs.

References

[1] Eldridge S, Kerry S. A practical guide to cluster randomised trials in health services research. Chichester: John Wiley & Sons; 2012.

[2] Donner A, Klar N. Design and analysis of cluster randomization trials in health research. Chichester: John Wiley & Sons Ltd; 2000.

[3] Hayes R, Moulton L. Cluster randomised trials. Boca Raton: Chapman & Hall; 2009.

[4] Donner A, Birkett N, Buck C. Randomization by cluster-sample size requirements and analysis. Am J Epidemiol 1981;114:906−14.

[5] Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. Int J Epidemiol 2006;35:1292−300.

[6] Kang S, Ahn C, Jung S. Sample size calculation for dichotomous outcomes in cluster randomization trials with varying cluster size. Drug Inf J 2003;37(1):109−14.

[7] Kerry S, Bland J. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. Stat Med 2001;20:377−90.

[8] Manatunga A, Hudgens M, Chen S. Sample size estimation in cluster randomized studies with varying cluster size. Biometrical J 2001; 43(1):75−86.

[9] Pan W. Sample size and power calculations with correlated binary data. Control Clin Trials 2001;22:211−27.

[10] van Breukelen GJ, Candel MJ. Calculating sample sizes for cluster randomized trials: we can keep it simple and efficient!. J Clin Epidemiol 2012;65:1212−8.

[11] Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637−9.

[12] Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001;357:1191−4.

[13] Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMC Med 2010;8:18.

[14] Campbell MK, Elbourne DR, Altman DG, group C. CONSORT statement: extension to cluster randomised trials. BMJ 2004;328:702−8.

[15] Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012;345:e5661.

[16] Froud R, Eldridge S, Diaz Ordaz K, Marinho VC, Donner A. Quality of cluster randomized controlled trials in oral health: a systematic review of reports published between 2005 and 2009. Community Dent Oral Epidemiol 2012;40(Suppl 1):3−14.

[17] Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials

on quality of reporting and study methodology: review of random sample of 300 trials, 2000-8. BMJ 2011;343:d5886.

[18] Walleser S, Hill SR, Bero LA. Characteristics and quality of reporting of cluster randomized trials in children: reporting needs improvement. J Clin Epidemiol 2011;64:1331−40.

[19] Diaz-Ordaz K, Froud R, Sheehan B, Eldridge S. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. BMC Med Res Methodol 2013; 13:127.

[20] Eccles M, Steen N, Grimshaw J, Thomas L, McNamee P, Soutter J, et al. Effect of audit and feedback, and reminder messages on primary-care radiology referrals: a randomised trial. Lancet 2001; 357:1406−9.

[21] Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. BMJ 2009;338:b1732.

[22] The CONSORT Group. CONSORT Endorsers-Journals. 2011. Available at http://www.consort-statement.org/about-consort/consort-endorsement/consort-endorsers—journals/. Accessed May 24, 2013.

[23] Bowater RJ, Abdelmalik SM, Lilford RJ. The methodological quality of cluster randomised controlled trials for managing tropical parasitic disease: a review of trials published from 1998 to 2007. Trans R Soc Trop Med Hyg 2009;103(5):429−36.

[24] Brierley G, Brabyn S, Torgerson D, Watson J. Bias in recruitment to cluster randomized trials: a review of recent publications. J Eval Clin Pract 2012;18:878−86.

[25] Diazordaz K, Slowther AM, Potter R, Eldridge S. Consent processes in cluster-randomised trials in residential facilities for older adults: a systematic review of reporting practices and proposed guidelines. BMJ 2013;3.

[26] Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979-1989. Int J Epidemiol 1990;19:795−800.

[27] Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomised trials: systematic review of recent trials. BMJ 2008;336:876−80.

[28] Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clin Trials 2004;1: 80−90.

[29] Giraudeau B, Caille A, Le Gouge A, Ravaud P. Participant informed consent in cluster randomized trials: review. PLoS One 2012;7(7): e40436.

[30] Mdege ND, Man MS, Taylor Nee Brown CA, Torgerson DJ. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. J Clin Epidemiol 2011;64:936−48.

[31] Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. BMJ 2003;327:785−9.

[32] Simpson JM, Klar N, Donnor A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. Am J Public Health 1995;85:1378−83.

[33] Sutton CJ, Watkins CL, Dey P. Illustrating problems faced by stroke researchers: a review of cluster-randomized controlled trials. Int J Stroke 2013;8(7):566−74.

[34] Hopewell S, Altman D, Moher D, Schulz K. Endorsement of the CONSORT statement by high impact factor medical journals: a survey of journal editors and journal ‘Instructions to Authors’. Trials 2008;9:20.

[35] Crespi CM, Maxwell AE, Wu S. Cluster randomized trials of cancer screening interventions: are appropriate statistical methods being used? Contemp Clin Trials 2011;32:477−84.

[36] Hannan PJ, Murray DM, Jacobs DR Jr, McGovern PG. Parameters to aid in the design and analysis of community trials: intraclass correlations from the Minnesota Heart Health Program. Epidemiology 1994; 5:88−95.

[37] Kelder SH, Jacobs DR Jr, Jeffery RW, McGovern PG, Forster JL. The worksite component of variance: design effects and the Healthy Worker Project. Health Educ Res 1993;8(4):555−66.

[38] Murray DM, Short B. Intraclass correlation among measures related to alcohol use by young adults: estimates, correlates and applications in intervention studies. J Stud alcohol 1995;56:681−94.

[39] Murray DM, Catellier DJ, Hannan PJ, Treuth MS, Stevens J, Schmitz KH, et al. School-level intraclass correlation for physical activity in adolescent girls. Med Sci Sports Exerc 2004;36:876−82.

Int. J. Epidemiol. Advance Access published July 13, 2015

International Journal of Epidemiology, 2015, 1–17
doi: 10.1093/ije/dyv113
Original article

Original article

Methods for sample size determination in cluster randomized trials

Clare Rutterford,[1]* Andrew Copas[2] and Sandra Eldridge[1]

[1]Centre for Primary Care and Public Health, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK and [2]Hub for Trials Methodology Research, MRC Clinical Trials Unit at University College London, London, UK

*Corresponding author. Centre for Primary Care and Public Health, Blizard Institute, Yvonne Carter Building, 58 Turner Street, London E1 2AB, UK. E-mail: c.m.rutterford@qmul.ac.uk

Accepted 2 June 2015

Abstract

Background: The use of cluster randomized trials (CRTs) is increasing, along with the variety in their design and analysis. The simplest approach for their sample size calculation is to calculate the sample size assuming individual randomization and inflate this by a design effect to account for randomization by cluster. The assumptions of a simple design effect may not always be met; alternative or more complicated approaches are required.
Methods: We summarise a wide range of sample size methods available for cluster randomized trials. For those familiar with sample size calculations for individually randomized trials but with less experience in the clustered case, this manuscript provides formulae for a wide range of scenarios with associated explanation and recommendations. For those with more experience, comprehensive summaries are provided that allow quick identification of methods for a given design, outcome and analysis method.
Results: We present first those methods applicable to the simplest two-arm, parallel group, completely randomized design followed by methods that incorporate deviations from this design such as: variability in cluster sizes; attrition; non-compliance; or the inclusion of baseline covariates or repeated measures. The paper concludes with methods for alternative designs.
Conclusions: There is a large amount of methodology available for sample size calculations in CRTs. This paper gives the most comprehensive description of published methodology for sample size calculation and provides an important resource for those designing these trials.

Key words: Sample size, cluster randomization, design effect

Key Messages

- There is a large body of literature on sample size calculations for cluster randomized trials.
- There are relatively simple and accessible methods to allow for design complexities such as variable cluster sizes; time-to-event outcomes; incorporation of baseline values and cross-over, stepped-wedge and matched designs.
- This is the most comprehensive resource to date for sample size methods for cluster randomized trials.
- There is scope for further methodological development.

Introduction

Cluster randomized trials

In a cluster randomized trial, groups or clusters, rather than individuals, are randomly allocated to intervention groups. This approach may be deemed necessary; if randomization at individual level is impractical, to avoid contamination between treatment groups, i.e. individuals in the control arm being exposed to the intervention; or for administrative or cost advantages. The rationale for cluster randomized trials has been described in detail elsewhere.[1–10]

The responses from individuals within a cluster are likely to be more similar than those from different clusters. This is because individuals within a cluster may share similar characteristics or be exposed to the same external factors associated with membership to a particular cluster. This lack of independence introduces complexity to the design and analysis. The degree of similarity, or clustering, is commonly quantified by the intracluster correlation coefficient (ICC) denoted in this article as ρ.

Obtaining a good sample size estimate is particularly important in cluster randomized trials due to the large cost that can be associated with recruiting an additional cluster as compared with recruiting an additional subject in an individually randomized trial. Equally important are the ethical implications of over- or under-recruitment where the addition or loss of one cluster may equate to a large number of individuals potentially being exposed to the risk of treatment, or lost.

A simple approach to sample size calculation

A consequence of clustering is that the information gained is less than that in an individually randomized trial of the same size, making randomization by cluster less efficient. This inefficiency was identified in the seminal paper by Cornfield that sparked the development of methodology for the design and analysis of cluster randomized trials.[11] It has been proposed by Donner, Birkett and Buck that a sample size calculated assuming individual randomization can be inflated by a Design Effect (DE) to reach the required level of statistical power under cluster randomization:[12]

$$DE = 1 + (n-1)\rho \qquad (1)$$

where n is the number of individuals per cluster and ρ the ICC.

Therefore for a comparison of means, in a two-arm trial with equal allocation the required the number of individuals per group, m, is calculated as:

$$\mathrm{m} = \frac{(\mathrm{Z}_{1-\alpha/2} + \mathrm{Z}_{1-\beta})^2 \; 2\sigma^2}{\Delta^2}\Big(1 + (n-1)\rho\Big) \qquad (2)$$

where Z_x is the x'th percentage point of the standard normal distribution, Δ the clinically important difference in treatment means and σ^2 the variance in the outcome.

Analyses may be conducted at either the cluster or individual level (see Eldridge and Kerry for a full discussion of analysis methods[1])

In cluster-level analyses, a cluster-level summary is calculated for each cluster, effectively reducing the data to one observation per cluster. The observations can then be treated as independent, and standard statistical analysis methods applied. The main advantages of cluster-level analyses are their simplicity and applicability to different types of outcomes. Disadvantages of this approach are that individual-level covariates cannot be included and the number of observations per group may be small. However, the two-sample t-test has been shown to be quite robust to deviations from normality and a small number of clusters per treatment group.[13]

Methods that use individual-level data but adjust for clustering can be used for analysis, such as the adjusted chi-square method for binary data, the adjusted two-sample t-test[2] or the non-parametric clustered Wilcoxon test for continuous data.[14] In this article, these are referred to as adjusted tests. The main drawback to these methods is that they do not allow for the inclusion of covariates.

Commonly individual-level analyses are conducted using a regression model that accounts for the clustered nature of the data and may include either cluster or individual level covariates. Mixed effects regression models are a cluster-specific method (henceforth referred to as mixed models) and Generalised Estimating Equations (GEE), a type of population-averaged or marginal method. Both approaches require a sufficient number of clusters for optimal performance; when the number of clusters is small, the mixed model is less biased than the GEE. The difference between these two approaches lies in the interpretation of the estimated treatment effect.[1]

In general, sample size requirements depend upon the proposed analysis method. In this paper we describe each sample size method alongside the analysis method for which it was designed. However, alternative analysis approaches may also be suitable. For example, with continuous outcomes a cluster-level analysis is equivalent to an individual-level analysis if all the clusters are the same size. When cluster size is variable, the assumptions underlying the cluster-level t-test are not met and a weighted t-test must be used to achieve adequate power and precision. Individual-level analyses naturally incorporate this weighting and so are more efficient than cluster-level analyses weighted by cluster size.[4] For continuous outcomes and equal-sized clusters, the cluster-specific and population-averaged methods for individual-level analyses are mathematically equivalent.

For binary outcomes, due to the transformation of the data onto the logistic scale, the treatment effects calculated under the cluster-specific and population-averaged methods are different. For binary outcomes, Austin *et al.*[15] compared the performance of three cluster-level methods: the t-test, the Wilcoxon rank-sum test and the permutation test, and three individual-level methods: the adjusted chi-square test, the mixed effects model and the GEE model. In the scenarios investigated, which included variable cluster sizes, the difference in power between these methods was negligible.

Measuring variability between clusters

A key parameter common to all sample size calculations for cluster randomized trials is the extent of similarity between units within a cluster. The measure used in the majority of sample size methodology is the ICC, usually denoted by the Greek letter ρ. The ICC can be interpreted as the proportion of variance due to between-cluster variation. When $\rho = 0$ there is statistical independence between members of a cluster, whereas when $\rho = 1$, all observations within a cluster are identical. A review of estimators for calculating the ICC for continuous and dichotomous outcomes can be found in the papers by Donner[16] and Ridout,[17] respectively. Properties of the ICC have been widely investigated and patterns in ICCs[18–22] and sources of ICC estimates[5,23–26] are available in the literature and have been summarized by Eldridge and Kerry.[1] An alternative measure to the ICC is the coefficient of variation in the outcome, denoted by k. This is calculated as the between-cluster standard deviation divided by the parameter of interest, i.e. the proportion, rate or mean, within each cluster.[27] This measure is particularly useful when the primary outcome variable is a rate, as an ICC cannot be calculated.[27]

When choosing an estimate of the ICC, in addition to the method of calculation, it is also important to identify whether the estimate has been adjusted for covariates. This can impact on its value and hence on the calculated sample size. Inclusion of the baseline value of an outcome as a covariate is arguably the strongest factor to reduce the ICC. However, this level of detail is not always explicitly reported alongside the ICC estimate.

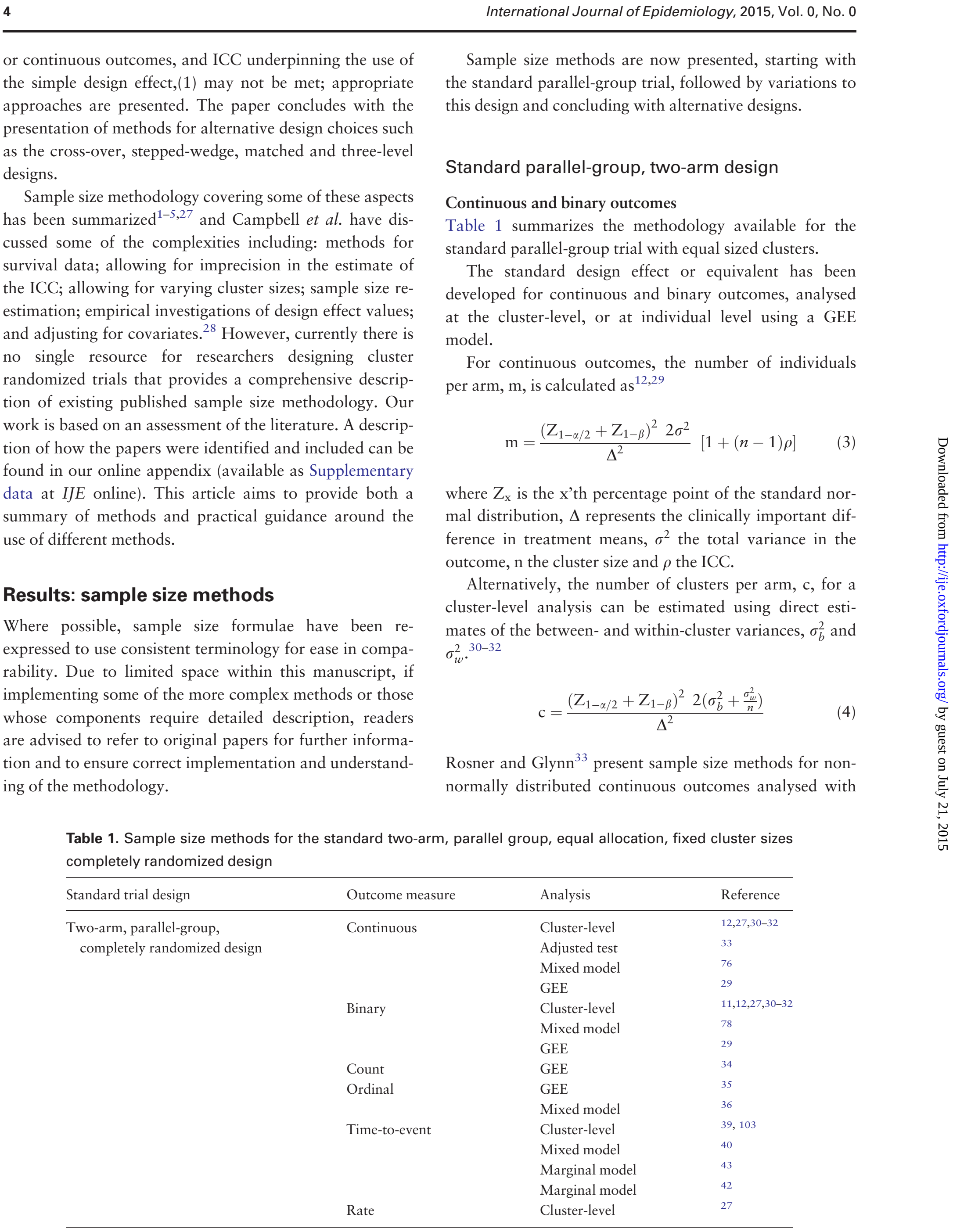
Comparison of ICC and coefficient of variation

Sample size calculations often make the assumption that the measure of correlation, be it the ICC or k, is the same in each treatment group. However, if the coefficient of variation is the same in each treatment group the ICC will not be, and vice versa.[4] Therefore the use of these different measures will produce different sample size requirements. The assumption of a constant ICC is reasonable if the intervention effect is likely to be constant across clusters. The assumption of a constant k is reasonable if the intervention effect is likely to be proportional to the cluster mean.[1]

Similarly for binary outcomes, different sample size requirements are calculated depending upon whether the ICC or coefficient of variation is used in the calculation. For binary outcomes there is an additional complication that the between-cluster variance also depends upon the value of the overall outcome proportion. The use of the ICC is recommended for sample size calculations of binary outcomes, unless the proportion is very small.[1]

Trial design features that impact on sample size

The most common and simplest design choice for a cluster randomized trial is the completely randomized, two-arm parallel-group design with fixed cluster sizes. In this paper, the methods appropriate for this design are discussed first. Variations to this design may be somewhat outside the investigator's control, such as variability in cluster size or attrition, or more within the investigator's control, such as choice of outcome measure or analysis method. With these variations, the assumptions of constant cluster size, binary



or continuous outcomes, and ICC underpinning the use of the simple design effect,(1) may not be met; appropriate approaches are presented. The paper concludes with the presentation of methods for alternative design choices such as the cross-over, stepped-wedge, matched and three-level designs.

Sample size methodology covering some of these aspects has been summarized[1–5,27] and Campbell *et al.* have discussed some of the complexities including: methods for survival data; allowing for imprecision in the estimate of the ICC; allowing for varying cluster sizes; sample size re-estimation; empirical investigations of design effect values; and adjusting for covariates.[28] However, currently there is no single resource for researchers designing cluster randomized trials that provides a comprehensive description of existing published sample size methodology. Our work is based on an assessment of the literature. A description of how the papers were identified and included can be found in our online appendix (available as Supplementary data at *IJE* online). This article aims to provide both a summary of methods and practical guidance around the use of different methods.

Results: sample size methods

Where possible, sample size formulae have been re-expressed to use consistent terminology for ease in comparability. Due to limited space within this manuscript, if implementing some of the more complex methods or those whose components require detailed description, readers are advised to refer to original papers for further information and to ensure correct implementation and understanding of the methodology.

Sample size methods are now presented, starting with the standard parallel-group trial, followed by variations to this design and concluding with alternative designs.

Standard parallel-group, two-arm design

Continuous and binary outcomes

Table 1 summarizes the methodology available for the standard parallel-group trial with equal sized clusters.

The standard design effect or equivalent has been developed for continuous and binary outcomes, analysed at the cluster-level, or at individual level using a GEE model.

For continuous outcomes, the number of individuals per arm, m, is calculated as[12,29]

$$\mathrm{m} = \frac{(\mathrm{Z}_{1-\alpha/2} + \mathrm{Z}_{1-\beta})^2 \; 2\sigma^2}{\Delta^2} \; [1 + (n-1)\rho] \qquad (3)$$

where Z_x is the x'th percentage point of the standard normal distribution, Δ represents the clinically important difference in treatment means, σ^2 the total variance in the outcome, n the cluster size and ρ the ICC.

Alternatively, the number of clusters per arm, c, for a cluster-level analysis can be estimated using direct estimates of the between- and within-cluster variances, σ_b^2 and σ_w^2.[30–32]

$$\mathrm{c} = \frac{(\mathrm{Z}_{1-\alpha/2} + \mathrm{Z}_{1-\beta})^2 \; 2(\sigma_b^2 + \frac{\sigma_w^2}{n})}{\Delta^2} \qquad (4)$$

Rosner and Glynn[33] present sample size methods for non-normally distributed continuous outcomes analysed with

Table 1. Sample size methods for the standard two-arm, parallel group, equal allocation, fixed cluster sizes completely randomized design

Standard trial design	Outcome measure	Analysis	Reference
Two-arm, parallel-group, completely randomized design	Continuous	Cluster-level	12,27,30–32
		Adjusted test	33
		Mixed model	76
		GEE	29
	Binary	Cluster-level	11,12,27,30–32
		Mixed model	78
		GEE	29
	Count	GEE	34
	Ordinal	GEE	35
		Mixed model	36
	Time-to-event	Cluster-level	39, 103
		Mixed model	40
		Marginal model	43
		Marginal model	42
	Rate	Cluster-level	27

an adjusted test, the clustered Wilcoxon test. This method requires a large number of calculations but can be implemented using SAS macros provided by the authors.

For binary outcomes, the number of individuals per arm, assuming a cluster-level analysis, is calculated as[12]

$$m = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \quad [P_1(1-P_1) + P_2(1-P_2)]}{\Delta^2} \times [1 + (n-1)\rho] \tag{5}$$

where P_1 is the probability of an event in the control group, and P_2 the probability of an event in the treatment group, and Δ represents the clinically important difference in treatment proportions, $P_1 - P_2$. The design effect can also be used to inflate the variance for the treatment effect described by a log odds ratio and assuming a GEE analysis.[29]

Alternatively, the number of clusters per group, assuming a cluster-level analysis can be calculated as[30,31]

$$c = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \ \ [2\sigma_b^2 + \frac{P_1(1-P_1)+P_2(1-P_2)}{n}]}{\Delta^2} \tag{6}$$

Simple methods are available for continuous and binary outcomes that use the coefficient of variation in outcome as a measure of correlation and assume a cluster-level analysis.[27] For continuous outcomes where μ_1 and μ_2 are the means in the control and intervention group, respectively, and σ_1 and σ_2 the associated within-cluster standard deviations, the number of clusters per group is shown as

$$c = 1 + \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \left[\frac{(\sigma_1^2+\sigma_2^2)}{n} + \ \ k^2(\mu_1^2 + \mu_2^2)\right]}{(\mu_1 - \mu_2)^2} \tag{7}$$

Similarly for binary outcomes where P_1 and P_2 are the proportions in the control and intervention group, respectively,

$$c = 1 + \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \left[\frac{P_1(1-P_1)}{n} + \frac{P_2(1-P_2)}{n} + k^2(P_1^2 + P_2^2)\right]}{(P_1 - P_2)^2} \tag{8}$$

One cluster per group has been added to account for the use of the normal approximation in the sample size calculation.

Count outcomes

For count outcomes, multiplication of the sample size calculation for ordinary Poisson regression by the standard design effect can be used to calculate the number of individuals per group, m, assuming fixed cluster size, and an analysis by GEE[34]

$$m = \frac{[Z_{\alpha/2}\sqrt{2} + Z_\beta\sqrt{[1 + e^{-\tilde{b}}}]]^2}{e^{\beta_0}\tilde{b}^2}[1 + (n-1)\rho] \tag{9}$$

where β_0 represents the event rate in the control group and $\tilde{b}$ is the treatment effect.

Ordinal outcomes

A method for correlated ordinal outcomes assuming a GEE analysis has been proposed.[35] This method has been described in the context of longitudinal data where the number of repeated measurements (or cluster size) is small and the number of clusters large. Its performance for smaller numbers of larger clusters is unknown and its implementation is best done via computer. More recently, Campbell and Walters[36] suggest multiplication of Whitehead's sample size calculation for ordinal outcomes in individually randomized trials by the design effect[37]

$$m = \frac{6[z_{1-\alpha/2} + z_{1-\beta}]^2/(\log OR)^2}{\left[1 - \sum_{i=1}^{I} \overline{\pi}_i{}^3\right]} \ [1 + (n-1)\rho] \tag{10}$$

$\overline{\pi}_i$ is the mean proportion expected in ordinal category i calculated as $\overline{\pi}_i = (\pi_{1i} + \pi_{2i})/2$ where π_{1i} and π_{2i} are the proportions in category i for the control and intervention groups. The treatment effect is given by the log odds ratio and a mixed model analysis is assumed.

Time-to-event outcomes

Methods have been suggested for time-to-event outcomes that adapt the formulae for individual randomization provided by Schoenfeld.[38]

The required number of individuals per group given by Schoenfeld's formula for individually randomized trials assuming equal allocation is

$$m_0 = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\log_e^2\theta \ (1 - P(C))} \tag{11}$$

where $P(C)$ is the probability of being censored and θ denotes the hazard ratio.

The standard design effect can be used to inflate the formula of Schoenfeld assuming the cluster-level weighted log-rank test.[39]

Jahn-Eimermacher *et al.*[40] present a simple formula for time-to-event outcomes adjusting Schoenfeld's formula and using the coefficient of variation in outcome as a measure of clustering and assuming a mixed model analysis using a shared frailty model, a popular method for the analysis of clustered time-to-event data. The number of clusters per group is given by

$$C \approx m_0 + (Z_{\alpha/2} + Z_\beta)^2 k^2 \frac{1+\theta^2}{(1-\theta)^2} \qquad (12)$$

where m_0 is the required number of clusters per group assuming uncorrelated data according to Schoenfeld (11) and k is the coefficient of variation in outcome.

Alternatively, Freedman's formula[41] for the number of events required under individual randomization can be multiplied by the design effect[42]

$$E = (Z_{1-\alpha/2} + Z_{1-\beta})^2 \frac{(1+\theta)^2}{(1-\theta)^2}[1 + (\overline{n} - 1)\rho] \qquad (13)$$

where $\overline{n}$ is the average cluster size, and analysis by marginal model is assumed.

Manatunga[43] considers time-to-event outcomes also assuming a marginal model, although the method does not provide a simple explicit formula.

Rate outcomes

The number of clusters per group, c, for rate outcomes in an unmatched design with cluster-level analysis is[27]

$$c = 1 + \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \left[\frac{r_1+r_2}{y} + \ k^2(r_1^2 + r_2^2)\right]}{(r_1 - r_2)^2} \qquad (14)$$

where y is the number of person-years in each cluster (assumed equal), k the coefficient of variation in the outcome and r_1 and r_2 the rates in the control and intervention group, respectively.

Variations to the standard parallel-group design

Table 2 provides a summary of all sample size methodology for variations to the standard parallel group trial. The key methods in each area are presented and discussed here.

Uncertainty around the estimate of the ICC

There is often large uncertainty around the estimate of the ICC, leading to wide confidence intervals. As the value of the ICC has a large impact upon the required sample size, it is sensible to consider the impact of its uncertainty. An informal method to address this problem has been to use a conservative estimate of the ICC in the sample size calculation; this provides a quick gauge of the impact of the ICC but could lead to unnecessarily large trials. Several authors have proposed formal methods of incorporating ICC uncertainty into the sample size calculation by making distributional assumptions for one or many previously observed ICC values and then calculating the corresponding distribution for the power.[44–47] Several of these methods adopt a Bayesian perspective but assume the analysis will follow a frequentist approach. Incorporating uncertainty about the ICC into the sample size calculation produces larger sample sizes than using a single estimate.

There may be situations where there are no good estimates of the ICC available for sample size calculations. This occurred in a trial of mental illness because the outcome measure was a newly adaptive questionnaire with unknown properties.[48] In these situations, several approaches might be considered: an educated estimate could be gained from assessment of published ICCs and known patterns in their behaviour for different outcome types and clusters; graphical methods that compare competing designs without requiring knowledge of the ICC[49]; or an internal pilot could be considered (see later section).

Variable cluster sizes

The use of the standard design effect assumes that the number of observations from each cluster to be included in the analysis is the same. In some situations such as ophthalmology studies where the cluster is a person and measurements are taken on eyes, this may be a reasonable assumption. However, in trials of primary care where the cluster may be a general practice or drop out may occur within clusters, it is more likely that clusters of variable size will be present in the analysis, and it is good practice to consider the potential impact of this at the design stage. If cluster sizes are variable, the use of the mean cluster size in the simple design effect will underestimate the required sample size, more so as the variation in cluster sizes increases. Use of the maximum cluster size as an alternative may be overly conservative. Methods to account for variable cluster size are recommended when cluster size variability is large, i.e. the coefficient of variation of cluster size, defined as the ratio of the standard deviation of cluster size S_n to mean cluster size $\overline{n}$, is greater than 0.23.[50]

The available methods to account for variable cluster size can be divided into two groups: I, those that require the size of each cluster to be known and II, those that require the mean and standard deviation of the distribution of cluster size.

Methods that require the size of each cluster to be known: Here the design effect is given by

$$DE = \frac{\overline{n}c}{\sum_{i=1}^{c} \frac{n_i}{1+(n_i-1)\rho}} \qquad (15)$$

where c represents the number of clusters per group, n_i the size of cluster i and $\overline{n}$ mean cluster size.

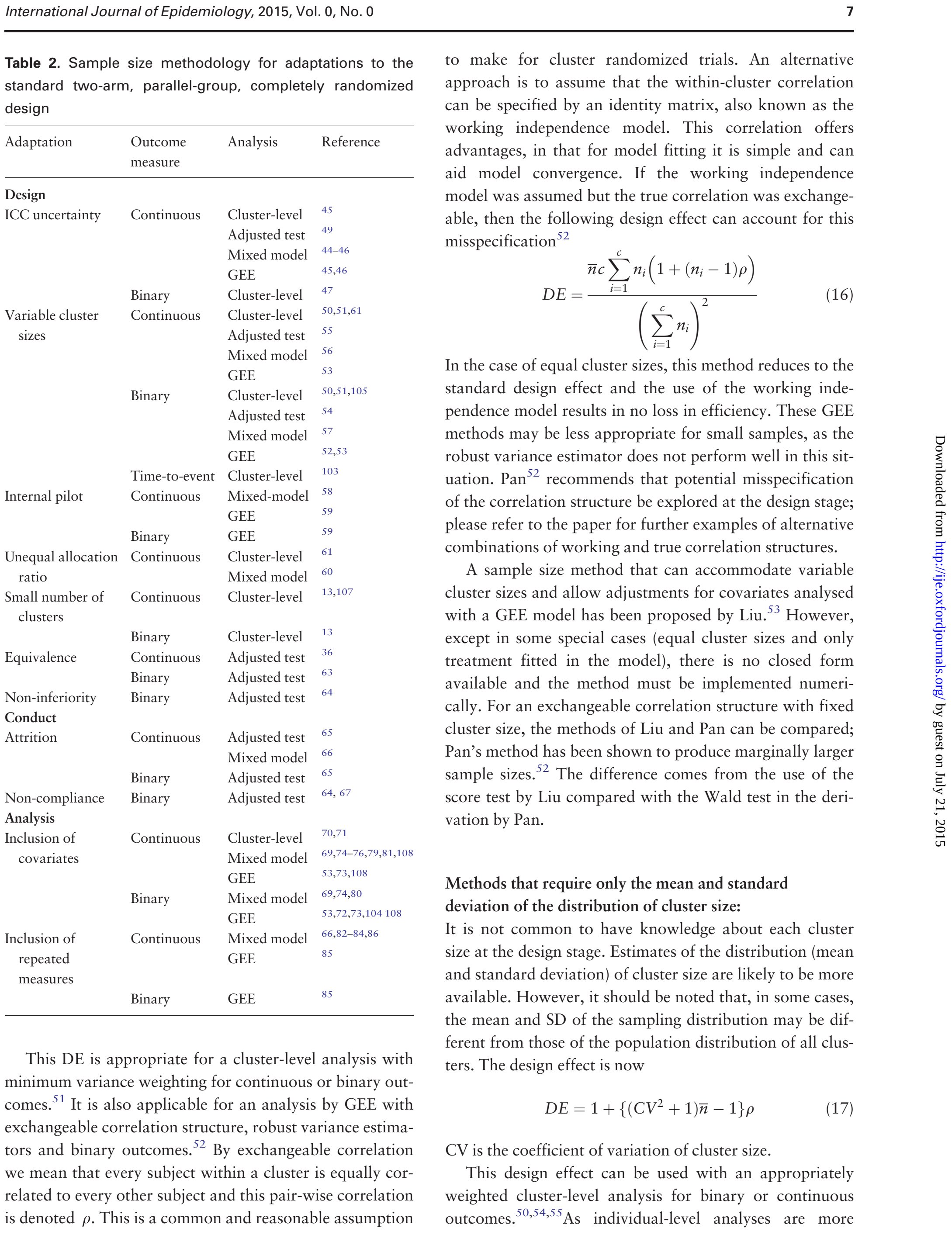


Table 2. Sample size methodology for adaptations to the standard two-arm, parallel-group, completely randomized design

Adaptation	Outcome measure	Analysis	Reference
Design			
ICC uncertainty	Continuous	Cluster-level	45
		Adjusted test	49
		Mixed model	44–46
		GEE	45,46
	Binary	Cluster-level	47
Variable cluster sizes	Continuous	Cluster-level	50,51,61
		Adjusted test	55
		Mixed model	56
		GEE	53
	Binary	Cluster-level	50,51,105
		Adjusted test	54
		Mixed model	57
		GEE	52,53
	Time-to-event	Cluster-level	103
Internal pilot	Continuous	Mixed-model	58
		GEE	59
	Binary	GEE	59
Unequal allocation ratio	Continuous	Cluster-level	61
		Mixed model	60
Small number of clusters	Continuous	Cluster-level	13,107
	Binary	Cluster-level	13
Equivalence	Continuous	Adjusted test	36
	Binary	Adjusted test	63
Non-inferiority	Binary	Adjusted test	64
Conduct			
Attrition	Continuous	Adjusted test	65
		Mixed model	66
	Binary	Adjusted test	65
Non-compliance	Binary	Adjusted test	64, 67
Analysis			
Inclusion of covariates	Continuous	Cluster-level	70,71
		Mixed model	69,74–76,79,81,108
		GEE	53,73,108
	Binary	Mixed model	69,74,80
		GEE	53,72,73,104 108
Inclusion of repeated measures	Continuous	Mixed model	66,82–84,86
		GEE	85
	Binary	GEE	85

This DE is appropriate for a cluster-level analysis with minimum variance weighting for continuous or binary outcomes.[51] It is also applicable for an analysis by GEE with exchangeable correlation structure, robust variance estimators and binary outcomes.[52] By exchangeable correlation we mean that every subject within a cluster is equally correlated to every other subject and this pair-wise correlation is denoted ρ. This is a common and reasonable assumption to make for cluster randomized trials. An alternative approach is to assume that the within-cluster correlation can be specified by an identity matrix, also known as the working independence model. This correlation offers advantages, in that for model fitting it is simple and can aid model convergence. If the working independence model was assumed but the true correlation was exchangeable, then the following design effect can account for this misspecification[52]

$$DE = \frac{\bar{n}c\sum_{i=1}^{c} n_i\left(1+(n_i-1)\rho\right)}{\left(\sum_{i=1}^{c} n_i\right)^2} \qquad (16)$$

In the case of equal cluster sizes, this method reduces to the standard design effect and the use of the working independence model results in no loss in efficiency. These GEE methods may be less appropriate for small samples, as the robust variance estimator does not perform well in this situation. Pan[52] recommends that potential misspecification of the correlation structure be explored at the design stage; please refer to the paper for further examples of alternative combinations of working and true correlation structures.

A sample size method that can accommodate variable cluster sizes and allow adjustments for covariates analysed with a GEE model has been proposed by Liu.[53] However, except in some special cases (equal cluster sizes and only treatment fitted in the model), there is no closed form available and the method must be implemented numerically. For an exchangeable correlation structure with fixed cluster size, the methods of Liu and Pan can be compared; Pan's method has been shown to produce marginally larger sample sizes.[52] The difference comes from the use of the score test by Liu compared with the Wald test in the derivation by Pan.

Methods that require only the mean and standard deviation of the distribution of cluster size:

It is not common to have knowledge about each cluster size at the design stage. Estimates of the distribution (mean and standard deviation) of cluster size are likely to be more available. However, it should be noted that, in some cases, the mean and SD of the sampling distribution may be different from those of the population distribution of all clusters. The design effect is now

$$DE = 1 + \{(CV^2+1)\bar{n} - 1\}\rho \qquad (17)$$

CV is the coefficient of variation of cluster size.

This design effect can be used with an appropriately weighted cluster-level analysis for binary or continuous outcomes.[50,54,55]As individual-level analyses are more efficient, it provides an overestimate of sample size required for most individual level analyses.

Van Breukelen[56] and Candel[57] propose the total number of clusters, as computed assuming equal cluster size and mixed model analysis, multiplied by the following design effect to account for variability in cluster size. It potentially has wide applicability as the authors suggest its use for correction of sample sizes calculated using any current formulae where equal-sized clusters are assumed.

$$DE \approx \frac{1}{1 - CV^2 \frac{\bar{n}}{\bar{n}+\frac{1-\rho}{\rho}}\left[1 - \frac{\bar{n}}{\bar{n}+\frac{1-\rho}{\rho}}\right]} \qquad (18)$$

The above DE is calculated via Taylor approximation but is considered to provide a good approximation for all reasonable distributions of cluster size. Heterogeneous variances across treatment groups can also be accommodated.[57]

Internal pilots

For trials that recruit a relatively large number of clusters over a fairly long period of time, it may be appropriate to re-estimate the sample size during the trial once information has been gained on the ICC and other nuisance parameters.[58,59] These methods assume a mixed model analysis for continuous outcomes and GEE for binary or continuous outcomes. The use of these internal pilots is less common in clustered trials and further investigation is required to determine best practice for their use, for example it is not known at which stage an interim estimate of the ICC can be considered stable and used to adequately re-estimate the sample size.

Allocation ratio

Design efficiency is maximized with equal allocation to treatment groups, and this has been assumed in the majority of the methodology presented here. However, there is an argument that unequal allocation may occasionally be desirable, particularly in cases where the costs associated with the intervention are high. Liu studies the optimal allocation of units to treatment group when the cost per cluster varies across the treatment groups, assuming a mixed model analysis.[60] The optimal cluster allocation ratio depends upon the cost ratio between the treatment and control.

Small number of clusters

The majority of the methods assume that a relatively large number of clusters is to be recruited, making the approximation to the normal distribution in the formulae appropriate. When the number of clusters is small, calculations based upon these approximations will likely underestimate the required sample size. In this case the normal distribution can be replaced by the t-distribution or methods based on the non-central t used. Donner[13] presents a power calculation based upon the non-central t-distribution with a simple non-centrality parameter for cluster-level analyses. Extensions to this non-centrality parameter can additionally allow for unbalanced designs.[61] As the percentage points of the non-central t-distribution are not routinely available in statistical texts, these methods are best implemented with a statistical package using the code provided by the authors.

Alternatively, Snedecor and Cochran[62] suggest adding one cluster per arm when testing at the 5% level and the number of clusters is small, which is incorporated into the formulae described by Hayes (equations 7, 8 and 14)[27] or could be added to the other formulae presented.

In general however, trials with a small number of clusters should be avoided. As well as the difficulties in sample size estimation, many analysis methods do not perform as well with a small number of clusters and imbalance in cluster characteristics across treatment groups is more likely to occur.[1]

Equivalence and non-inferiority

Non-inferiority and equivalence designs are less commonly used in cluster randomized trials. The methods presented here assume an analysis using an adjusted test. For equivalence designs, the standard design effect can be applied to the sample size calculated under individual randomization for binary outcomes[63]

$$m = \frac{2P(1-P)(Z_{1-\alpha} + Z_{1-\beta})^2}{d^2}[1 + (n-1)\rho] \qquad (19)$$

where P is the true event proportion in both groups and d represents the equivalence limit for the upper limit of the confidence interval of the difference in intervention proportion, and for continuous outcomes[36]

$$m = \frac{2(Z_{1-\alpha} + Z_{1-\beta})^2}{(d/\sigma)^2}[1 + (n-1)\rho] \qquad (20)$$

Here we have specified one-sided tests. To be conservative, two-sided tests could be used.

The calculation for the number of clusters per treatment group, c, in a non-inferiority trial with binary outcome, is[64]

$$c = \frac{(z_\alpha + z_\beta)^2 Var(\log(OR))}{(\log(d) - \log(OR))^2} \qquad (21)$$

where the relative treatment effect is measured by the odds ratio (OR) of a positive response among compliers and d represents the non-inferiority margin of the OR. This method additionally incorporates non-compliance and, due to this, the variance of this odds ratio is complex to calculate (see original paper).

Attrition

In a cluster randomized trial, individuals within a cluster may withdraw from the trial or an entire cluster may withdraw or not recruit any participants. Drop-out of entire clusters is relatively uncommon but could be incorporated into the sample size calculation by the addition of 1 or 2 extra clusters per treatment group.

Attrition among members of a cluster is a more common problem, particularly for cohort samples. Conventional approaches to account for such attrition are to divide the sample size by the anticipated follow-up rate or use the anticipated average cluster size in the calculation. However, these methods overestimate and underestimate, respectively, when cluster follow-up rates are highly variable or the cluster size or ICC is large. A design effect has been proposed for binary or continuous outcomes assuming adjusted tests, i.e. the individual-level t-test or chi-square test suitably adjusted for clustering[65]

$$DE = [1 + (n\pi - 1)\rho + (1 - \pi)[1 + (n - 1)\tau]\rho]/\pi \qquad (22)$$

π represents the probability of the outcome being observed. A binary missingness indicator variable is 0 if the outcome is missing and 1 otherwise. τ is the intracluster correlation coefficient for the missingness data mechanism, i.e. at its minimum $\tau = -\frac{1}{n-1}$ implies that all clusters have identical follow up rates and $\tau = 1$ implies all the missingness indicators are the same within a cluster (entire clusters are completely observed or completely missing). Currently estimates for τ are not routinely published with the results of trials and the authors recommend a sensitivity analysis using a range of plausible values.

Roy has also considered attrition for the longitudinal clustered design, assuming analysis with a mixed effects regression model.[66] The calculation uses an iterative method and allows for a differential drop-out across treatment groups and over time.

Non-compliance

Sample size requirements increase as the level of non-compliance increases. Methods which allow for non-compliance, where analysis is by an adjusted test, have been proposed for both non-inferiority and superiority designs.[64,67] However, the allowance for non-compliance makes the variance of the treatment effect more complex to calculate. These methods may be less applicable in pragmatic cluster randomized trials where the effect of the intervention is usually assessed in the presence of non-compliance. In a truly pragmatic trial, compliance may not be measured or actively encouraged.[68]

Inclusion of baseline measurements

Sample size calculations can be adapted to allow covariates in the analysis, as this may increase power by explaining variability and reducing the between-cluster variation, which is particularly important when the number of available clusters is limited or the cost of recruiting each additional cluster is high. Covariates may be collected at the level of the individual or the cluster and they may be demographic variables, such as age, or baseline measures of the primary outcome. Neuhaus and Segal[69] suggest, in general, that multiplication of the ICC by the ICC of any individual-level covariate provides an estimate of an adjusted ICC that can be used in the standard design effect, assuming a mixed model analysis.

Pre-post design

Inclusion of the baseline measurement of the primary outcome into the analysis is referred to as a pre-post design.

The nature of the correlation in a pre-post design will depend upon the population being sampled, for which there are two types: cross-sectional or cohort sample. With a cross-sectional sample, different individuals are measured at each time point. Here there are two sources of correlation to be accounted for: the correlation of outcomes from individuals within a cluster at the same time point (which can be thought of as the familiar ICC, ρ) and the correlation between baseline and follow-up outcomes for individuals within a cluster (referred to as the cluster auto correlation, ρ_c). With a cohort sample, the same individuals are measured at baseline and follow-up and the additional correlation across time points on the same individual conditional on the cluster is referred to as the subject autocorrelation, ρ_s.

Assuming a cluster-level ANCOVA, a relatively straightforward design effect can be used for the pre-post design.[70,71] The design effect can accommodate either the cross-sectional sample ($\rho_s = 0$), cohort sample or a mixture of the two[70]

$$DE = [1 + (n - 1)\rho] \times \left(1 - \left(\frac{n\rho}{1 + (n - 1)\rho}\rho_c + \frac{1 - \rho}{1 + (n - 1)\rho}\rho_s\right)^2\right) \qquad (23)$$

When the analysis is performed on change from baseline scores the design effect is

$$DE = [1 + (n-1)\rho] \times 2\left(1 - \left(\frac{n\rho}{1+(n-1)\rho}\rho_c + \frac{1-\rho}{1+(n-1)\rho}\rho_s\right)\right) \tag{24}$$

Preisser[72,73] focuses on binary outcomes with a GEE analysis. The number of clusters for the cross-sectional pre-post design is given as

$$\mathrm{c} = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \ (\sigma_1^2 + \sigma_2^2)}{n((\pi_{11} - \pi_{10}) - (\pi_{21} - \pi_{20}))^2} \tag{25}$$

where

$$\sigma_h^2 = [\pi_{h1}(1-\pi_{h1}) + \pi_{h0}(1-\pi_{h0})][1-(n-1)\rho] \\ -2n\rho_c\sqrt{\pi_{h1}(1-\pi_{h1}) + \pi_{h0}(1-\pi_{h0})}$$

and π_{ht} is the probability of the outcome for an individual at time t (0 = pre-test, 1 = post-test) from treatment group h (1 = control, 2 = intervention).

In terms of sample size, a cohort sample is more efficient, although it suffers from several drawbacks. To gain noticeable precision, the correlation across time points on the same individual must be fairly substantial. Cohort designs can also suffer from loss to follow-up and therefore require oversampling at baseline and attentive follow-up of individuals.

The sample size efficiency of the cohort design relative to the repeated cross-sectional design with 1 measurement on each individual at each time point, assuming a mixed model, has been quantified as[74,75]

$$RE = \frac{n(1-\rho_c)\sigma_b^2 + (1-\rho_s)\sigma_w^2}{n(1-\rho_c)\sigma_b^2 + \sigma_w^2} \tag{26}$$

Inclusion of other covariates

Although the inclusion of covariates can reduce the sample size requirements, there are costs associated with taking additional measurements. In a trial without covariates, suppose the total budget for the trial is summarized via the cost function $T = n\mathrm{C}c_1 + \mathrm{C}c_2$, where C is the total number of clusters, n the cluster size, c_1 the costs per individual and c_2 the costs per cluster. The number of clusters, C, and the number of individuals, n, which minimize the variance of the treatment estimator, given the budget constraint are given as[76–78]

$$\mathrm{C} = \frac{T}{(\sigma_w/\sigma_b)\sqrt{c_1 c_2} \ + \ c_2} \ , \ n = \frac{\sigma_w}{\sigma_b}\sqrt{\frac{c_2}{c_1}} \tag{27}$$

A similar approach can be used with the inclusion of covariates.[76,79,80] Alternatively, power-based calculations are provided by Moerbeek, assuming a mixed model.[81] The total number of clusters is calculated as

$$N \geq 4\frac{\sigma_w^2(1 - \frac{n}{n-1}\rho_W^2) + n\sigma_b^2 \quad (1 - \rho_B^2 + \frac{1}{n-1}\rho_W^2)}{n} \\ \times (\frac{z_{1-\alpha/2} + z_{1-\beta}}{\Delta})^2 \tag{28}$$

where ρ_W^2 and ρ_B^2 are the within-cluster and between-cluster residual correlations between the outcome and the covariate. $\rho_W = 0$ for a cluster level covariate.

The additional cost to measure a covariate at the individual level is c_1^* and the additional cost of measuring a covariate at the cluster level is c_2^*. Therefore the total cost function for individual level covariates becomes

$$T = n\mathrm{C}(c_1 + c_1^*) + \mathrm{C}c_2$$

and for cluster level covariates

$$T = n\mathrm{C}c_1 + \mathrm{C}(c_2 \ + c_2^*)$$

The costs associated with and without the covariate can be estimated and compared. The inclusion of covariates is more cost effective when the cost of measurement is small and the correlation between covariates and outcome is large. The formula presented by Moerbeek assumes the covariates are uncorrelated with the treatment condition. When the number of clusters is small, this can be achieved via matching on this covariate, particularly recommended for covariates that vary at the cluster level.[79]

Inclusion of repeated measurements

Multiple time points introduce additional components of correlation, as the observations for each cluster will be correlated over time. In a longitudinal cluster randomized trial we have a three-level structure with outcomes measured at specific time points within subjects, within clusters. A three-level mixed effects regression model therefore contains additional fixed effect terms for time and the treatment by time interaction. The sample size methods for these designs are more complex than others and the required estimates may be difficult to find. The hypothesis of interest in these trials is the effect of the intervention over time. Assuming a mixed model, the calculation by Koepsell *et al.*[82] is based on the non-central-t distribution, with the treatment effect adjusted by a design constant allowing for different hypothesized paths of the intervention effect over time. A formula based upon the Wald test of the interaction term for the number of clusters per arm has been proposed[83]

$$n_3 = \frac{2\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2 (1-\rho_1)}{n_2 n_1 \Delta^2 \sum_{k=1}^{n_1} (T_k - \overline{T})^2 / n_1} \qquad (29)$$

where n_x is the number of units at level x $(x = 1, 2, or\ 3)$, T represents the equally spaced time variable and $\rho_1$is the correlation among level-one units (see later section on three-level trials for definition).

Roy's iterative method similarly proposes a test of the treatment by time interaction from a mixed effect model but additionally allows incorporation for a differential drop-out across treatment groups and over time.[66] Murray proposed that a mixed model with random coefficients is a more appropriate analysis for explicitly modelling more than two time points in the analysis.[84] The additional random effects make this method more complex than others and, although the authors have provided parameter estimates to aid planning for some outcomes, investigators will likely need to spend time and money sourcing suitable estimates. Sample size formulae for assessing change over time assuming an analysis by GEE have been derived by Liu.[85] However, except under certain correlation structures, the calculations involved in this method are substantial.

If the effect of treatment is expected to diverge over time, sample size can be calculated for testing the treatment effect at the final time point with incorporation of information from the entire study period assuming a compound symmetry structure and mixed model. This produces smaller sample sizes than an assessment at the final time point only, but the assumptions underpinning this method may limit its widespread application.[86]

Alternative designs

The above methods are described for the parallel group trial and small variations to this standard design. We now consider methodology for alternative design choices. Table 3 summarizes the available sample size methodology for alternative designs.

Table 3. Sample size methodology for alternative designs

Trial design	Outcome measure	Analysis	Reference
Matched/stratified	Continuous	Cluster-level	27,32,109
		Mixed model	89
		Bayesian	92
	Binary	Cluster-level	27,32,41,87,109
		Mixed model	89
		Adjusted test	91
	Rate	Cluster-level	27,90
Cross-over	Continuous	Cluster-level	93,106,107
		Mixed model	94
	Binary	Cluster-level	106
	Count	Cluster-level	106
Stepped-wedge	Continuous	Mixed model	95,96
Three-level	Continuous	Mixed model	77,98,100,101
		GEE	99
	Binary	GEE	99

Stratification and matching

Cluster randomized trials in general recruit a smaller number of units than an individually randomized trial. This can potentially lead to baseline imbalances in cluster characteristics across treatment groups. Matching or stratification can be used to improve similarity in clusters across treatment groups. In a matched-pair design, similar clusters are paired, or matched. One cluster from the pair is allocated to the intervention and the other to the control and a cluster-level analysis conducted. Similarity may be defined on cluster-level characteristics that are thought to affect the outcome, such as size or geographical location. Matching reduces the variance between clusters (within strata or within matched pair) and hence can provide efficiency in sample size. The efficiency gains depend upon the effectiveness of the matching. The sample size for an unmatched cluster randomized trial must be inflated by the following DE in order to have the same precision as the matched study[87]

$$\mathrm{DE} = 1/(1 - \rho_x) \qquad (30)$$

Its calculation requires knowledge of the correlation in the outcome between matched pairs, ρ_x. This correlation can be estimated from previous studies or from the corresponding correlation for a surrogate variable observed prior to randomization, if any exist, otherwise a range of plausible values can be considered.

In planning a matched trial, it is worth noting that any potential gain in efficiency can be lost if clusters drop out of the study, rendering the matched pair unuseable in the analysis. However, ignoring matching and including all clusters in an unmatched analysis of a matched design has been shown to be valid and efficient in trials that recruit a small number of relatively large clusters.[88]

The required number of cluster pairs, m', is calculated using the following formula assuming analysis at the cluster level

$$m' = \frac{\sigma^2 (t_{\alpha/2;m'-1} + t_{\beta;m'-1})^2}{d^2} \qquad (31)$$

This is the familiar formula for the paired t-test, where d is the expected difference within pairs, σ^2 the variance of this difference and $t_{x;m'-1}$ percentage points of the t distribution with $m'-1$ degrees of freedom.

For continuous outcomes the variance is calculated as

$$2\left(\sigma_b^2 + \frac{\sigma_w^2}{n}\right) \tag{32}$$

where σ_b^2 is the between-cluster variance within a matched pair and σ_w^2 the within-cluster component of variability.[32,89]

For binary outcomes the variance is calculated as

$$\frac{P_1(1-P_1)+P_2(1-P_2)}{n} + 2\sigma_b^2 \tag{33}$$

where P_1 the expected proportion in the control arm and P_2 the expected proportion in the intervention arm.[41]

The methods by Hayes which use the coefficient of variation in outcome for unmatched trials (equations 7, 8 and 14) can be used for matched trials with two modifications.[27] Two, rather than one, cluster should be added to account for the use of the normal approximation and k should be replaced with k_m, the coefficient of variation between clusters within the matched pair. The Hayes method for rates can be shown to be equivalent to an earlier approach by Shipley.[90]

Stratification is similar to matching, in that we potentially now have several clusters within each stratum, rather than two as we have in a pair-matched study. This has been addressed for binary outcomes with a straightforward calculation.[91] For continuous outcomes, Kikuchi and Gittins[92] follow the less common Bayesian approach to design and analysis. However, as the impact of stratification is difficult to ascertain in advance, recommendations are to ignore it in the sample size calculation, for a more conservative estimate.[1]

Cross-over designs

Cross-over designs require a smaller number of clusters than a parallel-group trial and are therefore useful when the availability of clusters is limited. A simple design effect for cluster-level analysis has been presented for the cross-over design in which entire clusters switch treatments during the course of the trial[93]

$$DE = \left(1 + \left(\frac{1}{2}n_1 - 1\right)\rho_2\right) - \frac{1}{2}n_1\eta \tag{34}$$

where n_1 is the number of participants recruited within each cluster across both time periods, ρ_2 is the correlation between subjects in the same cluster at the same time point and η is the inter-period correlation. In this design, different subjects from each cluster are included in separate periods of the trial (a cross-sectional sample). The treatment effect is calculated within clusters and therefore between-cluster variance is removed and the design is more efficient than the parallel-group.

Alternatively, each subject could be included in both periods within a cluster (a cohort sample). Here a mixed model is assumed. The treatment effect is calculated within subjects, within clusters, so both between-cluster and between-subject variations are eliminated, making this the most efficient cross-over design with cluster level randomization. The relative efficiency (RE) of the cross-over design with cross-sectional sample over the parallel-group cluster randomized design has been quantified by Rietbergen[94]

$$RE = \frac{\left(1 + \left(\frac{1}{2}n_1 - 1\right)\rho_2\right) - \frac{1}{2}n_1\eta}{1 + (n_1 - 1)\rho_2} \tag{35}$$

and similarly for the cohort sample

$$RE = \frac{1}{2}\frac{1 - \rho_1 - \rho_2}{1 + (n_1 - 1)\rho_2} \tag{36}$$

where ρ_1 is the intrasubject correlation.

Although cross-over designs can improve efficiency, the nature of the intervention or condition under study may make them inappropriate, as occurs in individually randomized trials.

Stepped-wedge design

The stepped-wedge design is similar to the cross-over design, except that the cross-over of treatments is all in one direction and staggered over time. All clusters receive the control intervention at baseline. At various points during the trial (referred to as steps), one or more clusters will cross over to receive the treatment intervention, with all clusters receiving treatment by the end of the trial. The point at which a cluster, or group of clusters, will cross over is randomly determined at the beginning of the trial.

The main criteria for use of a stepped-wedge design is when the implementation of the intervention can only be performed sequentially across clusters, perhaps due to resource constraints, and when the intervention is believed to do more good than harm and so it would be considered unethical for some clusters to not receive the intervention at some point during the trial. Although these designs are increasing in popularity, there is little published research describing best practice in their design and analysis. Hussey in 2005[95] provides the first guidance on sample size, which has been further developed by Woertman and assumes analysis by mixed model.[96]

This recently developed sample-size approach for the stepped-wedge design with continuous outcomes supposes that, between each step, one or more cross-sectional sampling waves of the clusters occur and outcome measurements are taken. The total number of individuals required under individual randomization is multiplied by a DE to give the number of individuals to be sampled across all clusters at each sampling wave

$$N_{sw} = \frac{(\mathrm{Z}_{1-\alpha/2} + \mathrm{Z}_{1-\beta})^2 \quad 4\sigma^2}{\Delta^2} \times \left[\frac{1+\rho(ktn+bn-1)}{1+\rho\left(\frac{1}{2}ktn+bn-1\right)} \times \frac{3(1-\rho)}{2t\left(k-\frac{1}{k}\right)} \right] \tag{37}$$

where k is the number of steps, b the number of pre-randomization sampling waves, t the number of sampling waves between each step, n the number sampled from each cluster at each sampling wave and ρ is the ICC. N_{sw} is the total number of individuals required at each time point, the required number of clusters is calculated as N_{sw}/n, the number of clusters switching treatment at each step is calculated by dividing the number of clusters by k and the total number of individuals required across the entire trial is N_{sw} multiplied by (b + kt).

Three-level cluster randomized trials

Additional levels of clustering may occur due to the choice of cluster. For example, three-level cluster randomized trials are fairly common in educational research where pupils (level 1 units) are sampled within classrooms (level 2 units) and randomization takes place at the level of the school (level 3 units). The total variance is now made up of the variance between schools, σ_3^2, the variance between classrooms within schools, σ_2^2, and the variance associated with students within classrooms and schools, σ_1^2 . We can define two ICCs,[97] for students within schools

$$\rho_2 = \sigma_3^2/(\sigma_3^2 + \sigma_2^2 + \sigma_1^2) \tag{38}$$

and for students within classrooms

$$\rho_1 = \sigma_3^2 + \sigma_2^2/(\sigma_3^2 + \sigma_2^2 + \sigma_1^2) \tag{39}$$

In a three-level trial, the required sample size is calculated as

$$n_3 n_2 n_1 = DE \times m \tag{40}$$

where m is the number of individuals required in each group in an individual randomized controlled trial (RCT) and n_x is the number of units at level x $(x = 1, 2, or\ 3)$.

The Design effect for three levels of clustering is

$$DE = 1 + n_1(n_2 - 1)\rho_2 + (n_1 - 1)\rho_1 \tag{41}$$

This DE can be used for continuous outcomes with equal cluster size analysed with either a mixed effects model or GEE assuming exchangeable correlation, as these methods are equivalent under equal cluster size.[98–100] The design effect in the original paper by Teerenstra[100] has been re-expressed for the purpose of this paper to use the Pearson correlations (38 and 39), as these are more familiar quantities and published estimates are more likely than the variance components described in the original paper.

Following Raudenbush,[76] optimization of the sample sizes at each level can be performed based upon cost constraints.[101,102]

Discussion

Sample size calculations for individually randomized trials must be inflated in order to be used for cluster randomized trials, to account for the inefficiency introduced by the correlation of outcomes between members of a cluster. A simple design effect described by Donner, Birkett and Buck[12] can be used for parallel-group trials when the cluster size is assumed constant and the outcome is continuous, binary, count or time-to-event.

Design effects have been derived for more complex designs including: variable cluster sizes; individual level attrition; cross-over trials; stepped-wedge designs; inclusion of baseline measurements; analysis by GEE; and three levels of clustering. These design effects are relatively straight forward to calculate. However, the opportunity to use them may depend upon the availability and quality of estimates of the parameters required for the calculation. When incorporating variable cluster size, the choice of methods depends upon whether every cluster size is known in advance, or just information on cluster size distribution. In the case of incorporating stratification, the only method available requires knowledge about the proportion of individuals in the stratum as well as the success probabilities in each, information which is unlikely to be available at the beginning of the trial. These other parameters, required to assist others planning future trials, are not currently reported as part of a trial's findings, but we hope will become routinely published in time.

The intracluster correlation coefficient featured more frequently as a measure of within-cluster correlation than the coefficient of variation, in our assessment of the sample size literature. This may be due to the wide availability of published reviews of ICC estimates[5,23–26] and patterns in ICCs.[18–22]

The majority of papers specify binary or continuous outcomes; few deal with other types of outcome. Simple approaches for alternative outcomes data potentially warrant future development.

Sample size by simulation is an alternative to using an analytical formula. Although the procedure may be computationally intensive, in some cases it may be preferable to complex numerical procedures and was used in four papers identified in the literature.[103–106] Many of the methods proposed recommend validation of the sample size calculated with a formula through simulation, particularly for time-to-event outcomes or where the number of clusters is small. However, the type I error is often inflated when the number of clusters is small, the cluster size is variable and for particular analyses such as the frailty model, and this should be taken into consideration during the planning and interpretation of simulations.

We have provided a comprehensive description of sample size methodology for cluster randomized trials, presented in a simple way to aid researchers designing future studies.

With the increasing availability of more advanced methods to incorporate the full complexity that can arise in the design of a cluster randomized trial, the researcher may feel overwhelmed by the volume of methods presented. However it should be noted that in some situations a simple formula may perform reasonably well in comparison with a more complex methodology. For example, when the coefficient of variation in cluster size is less than 0.23, it is not deemed necessary to adjust the sample size and the standard design effect obtained assuming fixed cluster sizes would suffice.[50]

For continuous outcomes with equal cluster sizes, the cluster-level and individual-level analyses are equivalent. Therefore a sample size calculation assuming either of these with the same measure of correlation should produce equivalent results. When cluster size is variable, an individual-level analysis is more efficient than a cluster-level analysis weighted by cluster size; therefore a sample size calculation based upon cluster-level analyses will be somewhat conservative if an individual analysis is then conducted.

For binary outcomes, if the intervention is designed to reduce the outcome proportion use of the coefficient of variation[27] will produce marginally smaller sample sizes than using the ICC.[12] When the intervention aims to increase the outcome proportion, the sample sizes using the coefficient of variation will be larger. When several methods may be used, the choice between them is also a question of practicality. The distribution of the outcome and whether required estimates are available should be considered. Further work is required to formally compare the resulting sample sizes calculated under competing methods, when alternative analyses are conducted, and to evaluate the situations in which the simple methods can provide reasonable results over the more complex. This was beyond the scope of this paper.

A limitation of this paper is that a full critique and comparison of the sample size methods were difficult due to the lack of consistency in reporting across the papers. No guidelines exist at present to judge the quality of methodological papers and guide authors in clear and transparent reporting. We hypothesize that the way in which these methods are reported can also be a barrier to their uptake. We hope that their presentation in this article will improve uptake and research in the performance of these methods. We are planning further work looking at developing guidelines for the reporting of methodology papers.

There is often a large amount of uncertainty associated with the estimate of the ICC, and the appropriateness of any of the methods described here will depend upon the level of uncertainty. In the case of a large amount of uncertainty, we recommend that at a minimum the sample size sensitivity to a range of ICC values be explored. We recommend that, at the design stage, an appropriate simple formula be used in the first instance to provide the researcher with a benchmark figure upon which the impact of incorporating further complexities can be assessed.

Funding

This work was supported by the Medical Research Council.

Acknowledgments

We would like to thank Allan Donner, Obi Ukoumunne, Mike Campbell, Steven Tereenstra and Gerard Van Breukelen for their feedback regarding the coverage of the review. We would also like to thank Sally Kerry and two anonymous reviewers for their comments on this paper, which significantly improved its development.

Conflict of interest: None declared.

References

1. Eldridge S, Kerry S. *A Practical Guide to Cluster Randomised Trials in Health Services Research*. Chichester, UK: Wiley, 2012.
2. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Chichester, UK: Wiley, 2000.
3. Murray D. *Design and Analysis of Group-Randomized Trials*. Oxford, UK: Oxford University Press, 1998.
4. Hayes R, Moulton L. *Cluster Randomised Trials*. Boca Raton, FL: Chapman & Hall, 2009.

5. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess* 1999;**3**:iii–92.
6. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomization-based inference for community intervention trials. *Stat Med* 1996;**15**:1069–92.
7. Kirkwood B, Cousens S, Victora C, deZoysa I. Issues in the design and interpretation of studies to evaluate the impact of community-based interventions. *Trop Med Int Health* 1997;**2**: 1022–29.
8. Campbell MJ. Cluster randomized trials in general (family) practice research. *Stat Methods Med Res* 2000;**9**:81–94.
9. Koepsell TD, Wagner EH, Cheadle AC *et al*. Selected methodological issues in evaluating community-based health promotion and disease prevention programs. *Annu Rev Public Health* 1992;**13**:31–57.
10. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health* 2004;**94**:423–32.
11. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 1978;**108**:100–02.
12. Donner A, Birkett N, Buck C. Randomization by cluster- sample size requirements and analysis. *Am J Epidemiol* 1981;**114**: 906–14.
13. Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. *J Clin Epidemiol* 1996;**49**:435–39.
14. Rosner B, Glynn RJ, Lee ML. Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach. *Biometrics* 2003;**59**:1089–98.
15. Austin PC. A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Stat Med* 2007;**26**:3550–65.
16. Donner A. A review of inference procedures for the intraclass correlation-coefficient in the one-way random effects model. *Int Stat Rev* 1986;**54**:67–82.
17. Ridout MS, Demetrio CG, Firth D. Estimating intraclass correlation for binary data. *Biometrics* 1999;**55**:137–48.
18. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol* 2004;**57**:785–94.
19. Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clin Trials* 2005;**2**: 99–107.
20. Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the Health Survey for England 1994. *Am J Epidemiol* 1999;**149**:876–83.
21. Pagel C, Prost A, Lewycka S *et al*. Intracluster correlation coefficients and coefficients of variation for perinatal outcomes from five cluster-randomised controlled trials in low and middle-income countries: results and methodological implications. *Trials* 2011;**12**:151.
22. Taljaard M, Donner A, Villar J *et al*. Intracluster correlation coefficients from the 2005 WHO Global Survey on Maternal and Perinatal Health: implications for implementation research. *Paediatr Perinat Epidemiol* 2008;**22**:117–25.
23. Hannan PJ, Murray DM, Jacobs DR Jr, McGovern PG. Parameters to aid in the design and analysis of community trials: intraclass correlations from the Minnesota Heart Health Program. *Epidemiology* 1994;**5**:88–95.
24. Kelder SH, Jacobs DR Jr, Jeffery RW, McGovern PG, Forster JL. The worksite component of variance: design effects and the Healthy Worker Project. *Health Educ Res* 1993;**8**:555–66.
25. Murray DM, Catellier DJ, Hannan PJ *et al*. School-level intraclass correlation for physical activity in adolescent girls. *Med Sci Sports Exerc* 2004;**36**:876–82.
26. Murray DM, Short B. Intraclass correlation among measures related to alcohol use by young adults: estimates, correlates and applications in intervention studies. *J Stud Alcohol* 1995;**56**: 681–94.
27. Hayes R, Bennett S. Simple sample size calculation for cluster-randomized trials. *Int J Epidemiol* 1999;**28**:319–26.
28. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and statistics in medicine. *Stat Med* 2007;**26**: 2–19.
29. Shih W. Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimating equations. *Biometr J* 1997; **39**:899–908.
30. Kerry S, Bland J. Trials which randomize practices II: sample size. *Fam Pract* 1998;**15**:84–87.
31. Connelly LB. Balancing the number and size of sites: an economic approach to the optimal design of cluster samples. *Control Clin Trials* 2003;**24**:544–59.
32. Hsieh F. Sample-size formulas for intervention studies with the cluster as unit of randomisation. *Stat Med* 1988;**7**: 1195–201.
33. Rosner B, Glynn R. Power and Sample size estimation for the clustered Wilcoxon test. *Biometrics* 2011;**67**:646–53.
34. Amatya A, Bhaumik D, Gibbons RD. Sample size determination for clustered count data. *Stat Med* 2013;**32**:4162–79.
35. Kim HY, Williamson JM, Lyles CM. Sample-size calculations for studies with correlated ordinal outcomes. *Stat Med* 2005; **24**:2977–87.
36. Campbell M, Walters S. *How to design, analyse and report cluster randomised trials in medicine and health related research*. Wiley, Chichester, 2014.
37. Whitehead J. Sample size calculations for ordered categorical data. *Stat Med* 1993;**12**:2257–71.
38. Schoenfeld D. Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983;**39**:499–503.
39. Gangnon R, Kosorok M. Sample-size formula for clustered survival data using weighted log-rank statistics. *Biometrika* 2004; **91**:263–75.
40. Jahn-Eimermacher A, Ingel K, Schneider A. Sample size in cluster-randomized trials with time to event as the primary endpoint. *Stat Med* 2013;**32**:739–51.
41. Byar DP. The design of cancer prevention trials. *Recent Results Cancer Res* 1988;**111**:34–48.
42. Xie T, Waksman J. Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. *Stat Med* 2003;**22**:2835–46.

43. Manatunga A, Chen S. Sample size estimation for survival outcomes in cluster-randomized studies with small cluster sizes. *Biometrics* 2000;**56**:616–21.
44. Spiegelhalter D. Bayesian methods for cluster randomized trials with continuous responses. *Stat Med* 2001;**20**:435–52.
45. Turner R, Thompson S, Spiegelhalter D. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clin Trials* 2005;**2**:108–18.
46. Turner R, Prevost A, Thompson S. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med* 2004;**23**:1195–214.
47. Feng Z, Grizzle JE. Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculation. *Stat Med* 1992;**11**:1607–14.
48. Byng R, Jones R, Leese M, Hamilton B, McCrone P, Craig T. Exploratory cluster randomised controlled trial of shared care development for long-term mental illness. *Br J Gen Pract* 2004; **54**:259–66.
49. Mukhopadhyay S, Looney S. Quantile dispersion graphs to compare the efficiencies of cluster randomized designs. *J Appl Stat* 2009;**36**:1293–305.
50. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol* 2006;**35**:1292–300.
51. Kerry S, Bland J. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Stat Med* 2001;**20**:377–90.
52. Pan W. Sample size and power calculations with correlated binary data. *Control Clin Trials* 2001;**22**:211–27.
53. Liu G, Liang K. Sample size calculations for studies with correlated observations. *Biometrics* 1997;**53**:937–47.
54. Kang S, Ahn C, Jung S. Sample size calculation for dichotomous outcomes in cluster randomization trials with varying cluster size. *Drug Inform J* 2003;**37**:109–14.
55. Manatunga A, Hudgens M, Chen S. Sample size estimation in cluster randomized studies with varying cluster size. *Biometr J* 2001;**43**:75–86.
56. van Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Stat Med* 2007;**26**:2589–603.
57. Candel MJ, Van Breukelen GJ. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Stat Med* 2010;**29**:1488–501.
58. Lake S, Kammann E, Klar N, Betensky R. Sample size re-estimation in cluster randomization trials. *Stat Med* 2002;**21**:1337–50.
59. Yin G, Shen Y. Adaptive design and estimation in randomized clinical trials with correlated observations. *Biometrics* 2005;**61**:362–69.
60. Liu X. Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit of randomization. *J Educ Behav Stat* 2003;**28**:231–48.
61. Hoover D. Power for t-test comparisons of unbalanced cluster exposure studies. *J Urban Health* 2002;**79**:278–94.
62. Snedecor GW, Cochran WG. *Statistical Methods*. Ames, IA: Iowa State University Press, 1980.
63. Donner A. Some aspects of the design and analysis of cluster randomization trials. *J R Stat Soc C Appl Stat* 1998;**47**:95–113.
64. Lui K, Chang K. Test non-inferiority and sample size determination based on the odds ratio under a cluster randomized trial with noncompliance. *J Biopharm Stat* 2011;**21**:94–110.
65. Taljaard M, Donner A, Klar N. Accounting for expected attrition in the planning of community intervention trials. *Stat Med* 2007;**26**:2615–28.
66. Roy A, Bhaumik D, Aryal S, Gibbons R. Sample size determination for hierarchical longitudinal designs with differential attrition rates. *Biometrics* 2007;**63**:699–707.
67. Lui K, Chang K. Sample size determination for testing equality in a cluster randomized trial with noncompliance. *J Biopharm Stat* 2011;**21**:1–17.
68. Thorpe KE, Zwarenstein M, Oxman AD *et al*. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol* 2009; **62**:464–75.
69. Neuhaus JM, Segal MR. Design effects for binary regression models fitted to dependent data. *Stat Med* 1993;**12:** 1259–68.
70. Teerenstra S, Eldridge S, Graff M, de Hoop E, Borm GF. A simple sample size formula for analysis of covariance in cluster randomized trials. *Stat Med* 2012;**31**:2169–78.
71. Murray DM, Hannan PJ. Planning for the appropriate analysis in school-based drug-use prevention studies. *J Consult Clin Psychol* 1990;**58**:458–68.
72. Preisser JS, Reboussin BA, Song EY, Wolfson M. The importance and role of intracluster correlations in planning cluster trials. *Epidemiology* 2007;**18**:552–60.
73. Preisser JS, Young ML, Zaccaro DJ, Wolfson M. An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat Med* 2003;**22:** 1235–54.
74. Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. *Stat Med* 1994;**13**:61–78.
75. McKinlay S. Cost-efficient designs of cluster unit trials. *Prev Med* 1994;**23**:606–11.
76. Raudenbush S. Statistical analysis and optimal design for cluster randomized trials. *Psychol Methods* 1997;**2**:173–85.
77. Moerbeek G, Van Breukelen G, Berger M. Design issues for experiments in multilevel populations. *J Educ Behav Stat* 2000; **25**:271–84.
78. Moerbeek M, Van Breukelen G, Berger M. Optimal experimental designs for multilevel logistic models. *J R Stat Soc D Stat* 2001;**50**:17–30.
79. Moerbeek M, Van Breukelen G, Berger M. Optimal experimental designs for multilevel models with covariates. *Commun Stat Theor Stat* 2001;**30**:2683–97.
80. Moerbeek M, Maas C. Optimal experimental designs for multilevel logistic models with two binary predictors. *Commun Stat Theor Stat* 2005;**34**:1151–67.
81. Moerbeek M. Power and money in cluster randomized trials: When is it worth measuring a covariate? *Stat Med* 2006;**25:** 2607–17.
82. Koepsell T, Martin D, Diehr P *et al*. Data-analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs – a mixed-model analysis of variance approach. *J Clin Epidemiol* 1991;**44**:701–13.

83. Heo M, Leon A. Sample size requirements to detect an intervention by time interaction in longitudinal cluster randomized clinical trials. *Stat Med* 2009;**28:**1017–27.
84. Murray DM, Blitstein JL, Hannan PJ, Baker WL, Lytle LA. Sizing a trial to alter the trajectory of health behaviours: methods, parameter estimates, and their application. *Stat Med* 2007;**26:** 2297–316.
85. Liu A, Shih W, Gehan E. Sample size and power determination for clustered repeated measurements. *Sta Med* 2002;**21:** 1787–801.
86. Heo M, Kim Y, Xue X, Kim M. Sample size requirement to detect an intervention effect at the end of follow-up in a longitudinal cluster randomized trial. *Stat med* 2010;**29**:382–90.
87. Freedman LS, Green SB, Byar DP. Assessing the gain in efficiency due to matching in a community intervention study. *Stat Med* 1990;**9:**943–52.
88. Donner A, Taljaard M, Klar N. The merits of breaking the matches: a cautionary tale. *Stat Med* 2007;**26:**2036–51.
89. Thompson S, Pyke S, Hardy R. The design and analysis of paired cluster randomized trials: An application of meta-analysis techniques. *Stat Med* 1997;**16**:2063–79.
90. Shipley M, Smith P, Dramaix M. Calculation of power for matched pair studies when randomization is by group. *Int J Epidemiol* 1989;**18:**457–61.
91. Donner A. Sample size requirements for stratified cluster randomization designs. *Stat Med* 1992;**11:**743–50.
92. Kikuchi T, Gittins J. A behavioural Bayes approach for sample size determination in cluster randomized clinical trials. *J R Stat Soc C Appl Stat* 2010;**59:**875–88.
93. Giraudeau B, Ravaud P, Donner A. Sample size calculation for cluster randomized cross-over trials. *Stat Med* 2008;**27:** 5578–85.
94. Rietbergen C, Moerbeek M. The design of cluster randomized crossover trials. *J Educ Behav Stat* 2011;**36:**472–90.
95. Hussey M, Hughes J. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;**28:** 182–91.
96. Woertman W, de Hoop E, Moerbeek M, Zuidema SU, Gerritsen DL, Teerenstra S. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol* 2013;**66:**752–58.
97. Siddiqui O, Hedeker D, Flay B, Hu F. Intraclass correlation estimates in a school-based smoking prevention study – outcome and mediating variables, by sex and ethnicity. *Am J Epidemiol* 1996;**144:**425–33.
98. Heo M, Leon A. Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics* 2008;**64:**1256–62.
99. Teerenstra S, Lu B, Preisser J, van Achterberg T, Borm G. Sample size considerations for GEE analyses of three-level cluster randomized trials. *Biometrics* 2010;**66:**1230–37.
100. Teerenstra S, Moerbeek M, van Achterberg T, Pelzer B, Borm G. Sample size calculations for 3-level cluster randomized trials. *Clin Trials* 2008;**5:**486–95.
101. Konstantopoulos S. Incorporating cost in power analysis for three-level cluster-randomized designs. *Eval Rev* 2009;**33:** 335–57.
102. Moerbeek M, van Breukelen G, Berger M. Design issues for experiments in multilevel populations. *J Educ Behav Stat* 2000; **25:**271–84.
103. Jung S. Sample size calculation for weighted rank tests comparing survival distributions under cluster randomization: a simulation method. *J Biopharm Stat* 2007;**17:** 839–49.
104. Hendricks S, Wassell J, Collins J, Sedlak S. Power determination for geographically clustered data using generalized estimating equations. *Stat Med* 1996;**15:**1951–60.
105. Braun T. A mixed model formulation for designing cluster randomized trials with binary outcomes. *Stat Modelling* 2003; **3:**233–49.
106. Reich NG, Myers JA, Obeng D, Milstone AM, Perl TM. Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. *PLoS One* 2012;**7:**e35564.
107. Harrison DA, Brady AR. Sample size and power calculations using the noncentral t-distribution. *Stata J* 2004;**4:** 142–53.
108. Tu X, Kowalski J, Zhang J, Lynch K, Crits-Christoph P. Power analyses for longitudinal trials and other clustered designs. *Stat Med* 2004;**23**:2799–815.
109. Feng Z, Thompson B. Some design issues in a community intervention trial. *Control Clin Trials* 2002;**23:**431–49.