 doi:10.20944/preprints201808.0521.v1

Choosing the target difference ("effect size") for a randomised controlled trial - DELTA^2 guidance

Guidance for researchers and funder representatives

Version 5.0: 20-8-2018
[FINAL PREPRINT VERSION]

AUTHORS

Jonathan A Cook,[1] jonathan.cook@ndorms.ox.ac.uk
Associate Professor
Centre for Statistics in Medicine
Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences
University of Oxford
Botnar Research Centre
Nuffield Orthopaedic Centre
Windmill Rd
Oxford, OX3 7LD

Steven A Julious, s.a.julious@sheffield.ac.uk
Professor
Medical Statistics Group, ScHARR
The University of Sheffield

Regent Court, 30 Regent Street
SHEFFIELD S1 4DA

William Sones William.sones@ndorms.ox.ac.uk
Statistician
Centre for Statistics in Medicine
Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences
University of Oxford
Botnar Research Centre
Nuffield Orthopaedic Centre
Windmill Rd
Oxford, OX3 7LD

Lisa V Hampson, lisa.hampson@novartis.com
Associate Director
Statistical Methodology & Consulting, Novartis, Basel, Switzerland

&

Department of Mathematics and Statistics
Lancaster University
Lancaster
UK, LA1 4YF

Catherine Hewitt, catherine.hewitt@york.ac.uk
Professor
Department of Health Sciences, Seebohm Rowntree Building
University of York, Heslington, York, YO10 5DD, UK

Jesse A Berlin, jberlin@its.jnj.com
Vice President and Global Head of Epidemiology
Johnson & Johnson
1125 Trenton-Harbourton Road
Titusville, New Jersey 08933
United States

Deborah Ashby, deborah.ashby@imperial.ac.uk
Co-Director, Imperial Clinical Trials Unit
Deputy Head, School of Public Health
Imperial College London
Stadium House, 68 Wood Lane
London, W12 7RH

Richard Emsley, Richard.Emsley@kcl.ac.uk
Professor
Department of Biostatistics and Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King's College London
De Crespigny Park
Denmark Hill
London, SE5 8AF

Dean A Fergusson, dafergusson@ohri.ca
Senior Scientist & Director,
Clinical Epidemiology Program,
Ottawa Hospital Research Institute, Ottawa, ON, Canada

Stephen J Walters, s.j.walters@sheffield.ac.uk
Professor
Medical Statistics Group, ScHARR
The University of Sheffield
Regent Court, 30 Regent Street
Sheffield, S1 4DA

Edward CF Wilson, ed.wilson@medschl.cam.ac.uk
Senior Research Associate in Health Economics
Cambridge Centre for Health Services Research & Cambridge Clinical Trials Unit
University of Cambridge
Institute of Public Health
Forvie Site, Robinson Way
Cambridge, CB2 0SR

Graeme Maclennan, g.maclennan@abdn.ac.uk
Director & Professor
The Centre for Healthcare Randomised Trials (CHaRT)
Health Sciences Building
University of Aberdeen
Foresterhill, Aberdeen, AB25 2ZD

Nigel Stallard N.Stallard@warwick.ac.uk
Professor
Warwick Medical School - Statistics and Epidemiology
University of Warwick
Coventry, CV4 7AL

Joanne C Rothwell, j.c.rothwell@sheffield.ac.uk
PhD student
Medical Statistics Group, ScHARR
The University of Sheffield
Regent Court, 30 Regent Street, Sheffield, S1 4DA

Martin Bland, martin.bland@york.ac.uk
Professor
Department of Health Sciences, Seebohm Rowntree Building
University of York,
Heslington, York, YO10 5DD, UK

Louise Brown l.brown@ucl.ac.uk
Senior Statistician
MRC Clinical Trials Unit at UCL
Institute of Clinical Trials & Methodology
2nd Floor, 90 High Holborn, London, WC1V 6LJ

Craig R Ramsay, c.r.ramsay@abdn.ac.uk
Director & Professor
Health Services Research Unit
University of Aberdeen
Health Sciences Building
Foresterhill
Aberdeen, AB25 2ZD

Andrew Cook, andrewc@soton.ac.uk
Consultant in Public Health Medicine and Fellow in Health Technology Assessment
Wessex Institute, University of Southampton
Alpha House, Enterprise Road, Southampton, SO16 7NS

David Armstrong, david.armstrong@kcl.ac.uk Kings College London
Professor
School of Population Health & Environmental SciencesFaculty of Life Sciences and Medicine
Addison House, Guy's Campus, London, SE1 1UL

Doug Altman, doug.altman@csm.ox.ac.uk
Professor
Centre for Statistics in Medicine
Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences
University of Oxford
Botnar Research Centre
Nuffield Orthopaedic Centre
Windmill Rd, Oxford, OX3 7LD

Luke David Vale, luke.vale@newcastle.ac.uk
Professor
Health Economics Group
Institute of Health & Society
Newcastle University
Newcastle upon Tyne, UK, NE2 4AX

1 Corresponding author

Keywords

Target difference, clinically important difference, sample size, guidance, randomised trial, effect size, realistic difference

Funded by the Medical Research Council UK and National Institute for Health Research

Acknowledgments

This project was funded by the Medical Research Council-National Institute for Health Research Methodology Research Programme in the UK in response to a commissioned call to lead a workshop on this topic in order to produce guidance. The members of the original DELTA (Difference ELicitation in TriAls)[2] group were:

Associate Professor Jonathan Cook, Professor Doug Altman, Dr Jesse Berlin, Professor Martin Bland, Professor Richard Emsley, Dr Dean Fergusson, Dr Lisa Hampson, Professor Catherine Hewitt, Prof Craig Ramsay, Miss Joanne Rothwell, Dr Robert Smith, Dr William Sones, Professor Luke Vale, Professor Stephen Walters, and Professor Steve Julious.

As part of the process of developing the guidance, a two-day workshop was held in Oxford in September 2016. The workshop participants were:

Professor Doug Altman, Professor David Armstrong, Professor Deborah Ashby, Professor Martin Bland, Dr Andrew Cook, Professor Jonathan Cook, Dr David Crosby, Professor Richard Emsley, Dr Dean Fergusson, Professor Andrew Grieve, Dr Lisa Hampson, Professor Catherine Hewitt, Professor Steve Julious, Professor Graeme MacLennan, Professor Tim Maughan, Professor Jon Nicholl, Dr José Pinheiro, Professor Craig Ramsay, Robert Smith, Miss Joanne Rothwell, Dr William Sones, Professor Nigel Stallard, Professor Luke Vale, Professor Stephen Walters, and Dr Ed Wilson.

The authors would like to acknowledge and thank the participants in the Delphi exercise and the one-off engagement sessions with various groups, including the Society for Clinical Trials, PSI, and Joint Statistical Meeting conference session attendees, along with the other workshop participants who kindly provided helpful input and comments on the scope and content of this document. We would also like to thank in particular Dr Robert Smith in his role as a member of the public who provided a helpful public perspective during the workshop and in the development and revision of the guidance document. Finally, the authors would like to thank Stefano Vezzoli for in-depth comments that helped to refine this document, helpful feedback from the MRC Methodological Research Programme Advisory Group, and representatives of the Medicines and Healthcare Products Regulatory Agency (MHRA) and Health and Social Care, Northern Ireland (HSCNI).

Authors

The authors of this guidance document were:

Jonathan A Cook, Steven A Julious, Lisa Hampson, Catherine Hewitt, Jesse Berlin, Deborah Ashby, William Sones, Richard Emsley, Dean Fergusson, Stephen Walters, Ed Wilson, Graeme MacLennan, Nigel Stallard, Joanne Rothwell, Martin Bland, Louise Brown, Craig Ramsay, Andrew Cook, David Armstrong, Doug Altman, and Luke Vale.

 doi:10.20944/preprints201808.0521.v1

CONTENTS PAGE

 doi:10.20944/preprints201808.0521.v1

Preface

The aim of this document is to provide guidance for researchers and funder representatives about choosing the target difference ("effect size") for a randomised controlled trial (RCT). As an aid to the reader, the key information is provided in a short two-page executive summary immediately after this preface on the next two pages. Some readers will need to focus on a particular aspect to address a specific query or topic. For researchers new to choosing the target difference for a RCT sample size calculation, reading the whole of the main document is recommended. Individuals with a public and patient involvement role in research and research commissioning will find Sections 2 and 3.1 are most pertinent. For more experienced readers, who may be interested in the subject or to help to peer review an RCT grant application, Sections 3 and 4 might well be sufficient.

The structure of the main guidance is set out as follows. The background and aim are given in Section 1. Section 2 considers the specification of the target difference. Methods for informing the specification of the target difference are covered in Section 3. In Section 4, reporting of sample size calculations, including the specification of the target difference, in key trial documents is covered. Finally, in Section 5, a number of case studies of carrying out sample size calculations in practice are provided, which give a flavour of the decisions made and how the final sample size calculation (and in particular the choice of target difference) was decided. The appendices cover the more technical/statistical details related to sample size calculation, including providing some formulae, alternative approaches to the sample size calculation, and how the target difference relates to alternative trial designs. A list of abbreviations and glossary of key terms are also provided.

 doi:10.20944/preprints201808.0521.v1

Executive summary for researchers and funder representatives

Specifying the target difference for a randomised controlled trial

The randomised controlled trial (RCT) is widely considered the gold standard study for comparing the effectiveness of health interventions. Central to its design is a calculation of the number of participants needed (the sample size). This provides reassurance that the study will be able to achieve its primary aim. It is typically done by specifying the magnitude of the difference between the intervention effects in the key (primary) outcome for the population of interest that can reliably be detected for a given sample size. This difference is called the study's "target difference" and should be appropriate for the primary estimand of interest (i.e., the combination of population, outcome, and intervention effects), as determined by the primary aim of the study.

There are two main bases for specifying a target difference: a difference that is considered to be *important* to one or more stakeholder groups (e.g., patients) and/or one that is *realistic* (plausible), based on existing evidence and/or expert opinion. Seven broad types of methods can be used to justify the choice of a particular value as the target difference: anchor, distribution, health economic, opinion-seeking, pilot study, review of the evidence base, and standardised effect size (SES). Different statistical and health economic approaches can be taken to justify the sample size, but the general principles are mostly the same. An exception is the relatively new technique of value of information analysis, which seeks to explicitly incorporate the opportunity cost of conducting research. As such, the appropriate sample size is one that maximises the return on investment in the trial, dispensing with the need to define a target difference. The use of alternative approaches is currently limited, with the conventional (Neyman-Pearson) approach the most commonly used.

To aid those new to the topic and to encourage better practice regarding the specification of the target difference for an RCT, the following *recommendations* are made when the conventional approach to the sample size calculation is used:

1. Begin by searching for relevant literature to inform the specification of the target difference. Relevant literature can:
    a. relate to a candidate primary outcome and/or the comparison of interest, and;
    b. inform what is an important and/or realistic difference for that outcome, comparison, and population (estimand of interest).
2. Candidate primary outcomes should be considered in turn, and the corresponding sample size explored. Where multiple candidate outcomes are considered, the choice of primary outcome and target difference should be based on consideration of the views of relevant stakeholders groups (e.g., patients), as well as the practicality of undertaking such a study and the required sample size. The choice should not be based solely on which yields the minimum sample size. Ideally, the final sample size will be sufficient for all key outcomes, although this is not always practical.
3. The importance of observing a particular magnitude of a difference in an outcome, with the exception of mortality and other serious adverse events, cannot be presumed to be self-evident. Therefore, the target difference for all other outcomes requires additional justification to infer importance to a stakeholder group.
4. The target difference for a definitive (e.g., Phase III) trial should be one considered to be important to at least one key stakeholder group.
5. The target difference does not necessarily have to be the minimum value that would be considered important if a larger difference is considered a realistic possibility or would be necessary to alter practice.
6. Where additional research is needed to inform what would be an *important* difference, the anchor and opinion-seeking methods are to be favoured. The distribution method should not be used. Specifying the target difference based solely on an SES approach should be considered a last resort, although it may be helpful as a secondary approach.
7. Where additional research is needed to inform what would be a *realistic* difference, the opinion-seeking and review of the evidence base methods are recommended. Pilot trials are

typically too small to inform what would be a realistic difference and primarily address other aspects of trial design and conduct.

8. Use existing studies to inform the value of key "nuisance" parameters that are part of the sample size calculation. For example, a pilot trial can be used to inform the choice of standard deviation (SD) value for a continuous outcome or the control group proportion for a binary outcome, along with other relevant inputs such as the amount of missing outcome data.
9. Sensitivity analyses that consider the impact of uncertainty around key inputs (e.g., the target difference and the control group proportion for a binary outcome) used in the sample size calculation should be carried out.
10. Specification of the sample size calculation, including the target difference, should be reported according to the guidance for reporting items (see below) when preparing key trial documents (grant applications, protocols, and result manuscripts).

Recommended core reporting items

A set of core items should be reported in all key trial documents (protocols, grant applications, and main results papers) to ensure reproducibility of the sample size calculation. Recommended core reporting items when the conventional sample size approach has been used are as follows:

1. Primary outcome (and any other outcome on which the calculation is based)
    a. If a primary outcome is not used as the basis for the sample size calculation, state why.
2. Statistical significance level and power
3. Express the target difference according to outcome type:
    a. Binary – state the target difference as an absolute and/or relative effect, along with the intervention and control group proportions. If both an absolute and a relative difference are provided, clarify if either takes primacy in terms of the sample size calculation.
    b. Continuous – state the target mean difference on the natural scale, the common SD, and the SES (mean difference divided by the SD).
    c. Time-to-event – state the target difference as an absolute and/or relative difference; provide the control group event proportion; the planned length of follow-up; and the intervention and control group survival distributions and the accrual time (if assumptions regarding them are made)). If both an absolute and relative difference are provided for a particular time point, clarify if either takes primacy in terms of the sample size calculation.
4. Allocation ratio
    a. If an unequal ratio is used, the reason for this should be stated.
5. Sample size based on the assumptions as per above:
    a. Reference the formula/sample size calculation approach, if standard binary, continuous, or survival outcome formulae are not used. For a time-to-event outcome, the number of events required should be stated.
    b. If any adjustments (e.g., allowance for loss to follow-up, multiple testing, etc.) that alter the required sample size are incorporated, they should also be specified, referenced, and justified along with the final sample size.
    c. For alternative designs, any additional inputs should be stated and justified. For example, for a cluster RCT (or individually randomised trials with potential clustering), state the average cluster size and intra-cluster correlation coefficient(s). Justification for the values chosen should be given. Variability in cluster size should be considered and, if necessary, the coefficient of variation should be incorporated into the sample size calculation.
    d. Provide details of any assessment of the sensitivity of the sample size to the inputs used.

Trial results papers should always reference the trial protocol. Additional items to give more explanation of the rationale should be provided where space allows (e.g., grant applications and trial protocols). When the calculation deviates from the conventional approach, whether by research question or statistical framework, this should be clearly specified. The reporting items would correspondingly need appropriate modification.

 doi:10.20944/preprints201808.0521.v1

Summary for Patient and Public involvement contributors to research projects and funding panels

This DELTA^2 guidance aims to provide a brief overview of the role of the target difference in a randomised controlled trial, how to choose it, and how to document what was done. The number of people needed in a study is based on a calculation of the number of people needed for the analysis to be informative ("the sample size"). Choosing a "target difference" is part of the typical way to work out how many people need to be included in a study. The target difference is the amount of a difference in the participants' response to the treatments that we wish to detect. It is probably the most important piece of information used in the sample size calculation. Appropriate selection of the target difference is essential to give us confidence in the conclusions of the study. There are other ways to determine the sample size, but these are not currently commonly used.

To determine what target difference to use, we need to think about what we are going to measure in our study to help us decide whether one treatment works better than another. An "outcome" is how we measure the effect of treatment. For example, if we are evaluating a treatment for hypertension, the outcome could be blood pressure. If more than one outcome is available, then we would consider what would be an important difference for each of them before deciding the overall number of people needed for the trial. We can also think about what would be a realistic value based on similar studies and what we think is possible.

There are a number of different approaches to determine the target difference. The guidance document describes seven broad types. Two of the easiest to use to decide what would be an important difference to detect are the "anchor" and "opinion-seeking" approaches. The "anchor" approach simply uses someone's view (usually the patient's) to determine what would be a meaningful difference in an outcome. As is implied with the name, when using the "opinion-seeking" method a researcher asks experts in their field of work what they think the target difference should be. In addition to the "opinion-seeking" approach, a "review of the evidence base" is the easiest to use to understand what would be realistic. For the "review of the evidence base" approach, a researcher will look at other similar research studies and see what their results have been. There are strengths and weaknesses to each of these approaches.

The focus of this guidance document is to provide assistance for researchers on how to estimate the target difference when they are designing studies and applying for funding. As well as providing assistance for researchers, it is hoped that the document will be used by those who decide whether a proposed study should be funded, helping them to assess if the target difference the researcher says they will expect to see is reasonable. The guidance document highlights the value of properly documenting the rationale behind the sample size and choice of target difference. This is so other researchers and interested readers can understand what has been done, the thinking behind it, how to interpret the results, and ultimately can have confidence in the conclusions of the study. Clarifying what the study is aiming for and getting the right sample size is very important, as research can have a big impact on not only those directly involved as participants, but also future patients.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Abbreviations

A&F Audit and feedback
ACL Anterior Cruciate Ligament
ACL SNNAP (ACL Surgery Necessity in Non Acute Patients)
ART Arterial Revascularisation Trial
CACE Complier Average Causal Effect
CRS Chronic Rhinosinusitis
CHART Continuous Hyperfractionated Accelerated Radio Therapy
CONSORT Consolidated Standards of Reporting Trials
CV Coefficient of Variation
DELTA Difference Elicitation in TriAls
ENGS Expected Net Gain of Sampling
ESS Endoscopic Sinus Surgery
ETDRS Early Treatment Diabetic Retinopathy Study
EQ-5D-3/5L Euroqol 5 Dimensions-3/5 Level instrument
EVSI Expected Value of Sample Information
FILMS Full-thickness macular hole and Internal Limiting Membrane peeling Study
HR Hazard Ratio
ICC Intra-Cluster Correlation
INB Incremental Net (monetary) Benefit
ITT Intention To Treat
KOOS Knee Injury and Osteoarthritis Outcome Score
MACRO Management for Adults with Chronic RhinOsinusitis
MAMS Multi-Arm Multi-Stage
MAPS Men After Prostate Surgery
MCDC/D Minimal clinically detectable change/difference
MCID Minimum clinically important difference
Medical expulsive therapy (MET)
MSDD Minimal statistically detectable difference
MID Minimally important difference
MRC Medical Research Council
MYPAN MYcophenolate mofetil for childhood PAN
NHS National Health Service
NICE National Institute for Health and Care Excellence
NIHR National Institute for Health Research
OPTION-DM Optimal Pathway for TreatIng neurOpathic paiN in Diabetes Mellitus
OR Odds Ratio
PICO(T) Population Intervention Control Outcome (Timeframe)
PPI Patient and Public Involvement
RAPiD Reducing Antibiotic Prescribing in Dentistry
RCT Randomised Controlled Trial
RR Risk Ratio
SD Standard Deviation
SEm Standard Error of the measurement
SES Standardised Effect Size
SNOT-22 Sinonasal Outcome Test 22 itrmd
SPIRIT Standard Protocol Items: Recommendations for Interventional Trials
SUSPEND Spontaneous Urinary Stone Passage Enabled by Drugsd
TOST Two One-Sided Test
UK United Kingdom

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Glossary of key terminology

Estimand is the intended effect to be estimated to address a trial objective. It can be defined in terms of the population of interest, the outcome measure, how intercurrent events (those which preclude observation of the outcome or potentially affect its measurement, e.g., death or participant withdrawal from the study) are dealt with, and how the outcome is expressed (e.g., mean difference).

Important difference is a difference in an outcome that is considered to be important to one or more stakeholder groups (e.g., patients).

Minimum (clinically) important change/difference (MCIC/D) is the smallest value that is judged to be important. The adjective "clinically" is often added to refer to the context of medical care. In shortened form, the acronym MCID is probably most often used in the literature. Minor variants in the term, such as minimal instead of minimum, are commonplace. The use of the word "change" instead of "difference" implies it was premised on a within-person change (e.g., from before to after treatment).

Minimum clinically detectable change/difference (MCDC/D) is the smallest value that is judged to be detectable in the sense that is greater that measurement error for a measure. It is premised on the rationale that a difference smaller than this is not likely to be important. Most commonly, such an approach is used for quality of life measures where the construct of interest cannot be directly measured. As such, this approach only indirectly addresses the issue of importance of a particular difference. The adjective "clinically" is used here to differentiate it from a *minimum statistically detectable change/difference (MSDC/D)*. Accordingly, while the shortened acronym MDC/D is often used in the literature, here MCDC/D is used to differentiate it. There are minor variants in the terminology, such as using minimal instead of minimum and the exact definition.

Minimum statistically detectable change/difference (MSDC/D) is the smallest value that is expected to be statistically detectable at the pre-specified Type I error rate. If the required sample size is achieved, the target difference is one that can reasonably be expected to be statistically detected should it exist. It is not, however, the only value nor the smallest value that could lead to a statistically significant change or difference. The latter is the MSDC/D. The adjective "statistically" is used here to differentiate it from a *minimum clinically detectable change/difference*. While the shortened acronym MDC/D is often used in the literature, here MSDC/D is used accordingly.

Statistical power is the probability that, for the given assumptions, the statistical analysis would correctly detect a given difference and produce a statistically significance result. It is the complement of the Type II error (i.e., the probability of a Type II error not occurring). Achieving 80% or 90% power are commonly accepted levels that the sample size is chosen to meet, although they are arbitrary choices.

Target difference is the value that is used in the sample size calculation of a randomised trial that expresses the difference between the intervention groups that is sought to be detected. There are no theoretical constraints on its value beyond those imposed by the outcome and the planned analysis. For example the proportion of participants with an adverse event can range from 0 to 1.0. A target difference may or may not be one that could be considered important and/or realistic.

Type I error is the probability of falsely rejecting the null hypothesis (typically the null hypothesis is usually that there is no difference between the treatments) and concluding the alternative hypothesis (corresponding, typically that there is a difference between the treatments). The Type I error is typically set to the 0.05 level and applied to a statistical analysis to infer the occurrence or not of a statistically significance finding.

Type II error is the probability of failing to reject the null hypothesis (typically that there is no difference) when there is a real difference between interventions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

1 Introduction

1.1 Aim

The aim of this document is to provide practical guidance on the choice of target difference used in the sample size calculation of a randomised controlled trial (RCT). Guidance is provided with a definitive trial, one that seeks to provide a useful answer, in mind and not those of a more exploratory nature. The term "target difference" is taken throughout to refer to the difference that is used in the sample size calculation (the one that the study formally "targets"). Please see the glossary for definitions and clarification with regards other relevant concepts. In order to address the specification of the target difference, it is appropriate, and to some degree necessary, to touch on related statistical aspects of conducting a sample size calculation. Generally the discussion of other aspects and more technical details is kept to a minimum, with more technical aspects covered in the appendices and referencing of relevant sources provided for further reading.

The main body of this guidance assumes a standard RCT design is used; formally, this can be described as a two-arm parallel-group superiority trial. Most RCTs test for superiority of the interventions, that is, whether or not one of the interventions is superior to the other (See Box 1 for a formal definition of superiority, and of the two most common alternative approaches). Some common alternative trial designs are considered in Appendix 3. Additionally, it is assumed in the main body of the text that the conventional (Neyman-Pearson) approach to the sample size calculation of an RCT is being used. Other approaches (Bayesian, precision and value of information) are briefly considered in Appendix 2 with reference to the specification of the target difference.

1.2 Background

An RCT is widely considered to be the optimal study design to assess the comparative clinical efficacy and effectiveness along with the cost implications of health interventions.[1] RCTs have been widely used to evaluate a range of interventions and have been successfully used in a variety of healthcare settings. An *a priori* sample size calculation ensures that the study has a reasonable chance to achieve its pre-specified objectives.[2]

A number of statistical approaches exist for calculating the required sample size.[1, 3, 4] However, a recent review of 215 RCTs in leading medical journals identified only the conventional (Neyman-Pearson) approach in use.[5] This approach requires establishment of the statistical significance level (Type I error rate) and power (1 minus the Type II error rate), alongside the target difference ("effect size"). Setting the statistical significance level and power represents a compromise between the possibility of being misled by chance, when there is no true difference between the interventions, and the risk of not identifying a difference, when one of the interventions is truly superior, whilst the target difference is the magnitude of difference to be detected between sample sets. The required sample size is very sensitive to the target difference. Halving it quadruples the sample size for the standard RCT design.[1]

A comprehensive review conducted by the original DELTA group[6, 7] highlighted the available methods for specifying the target difference. Despite there being many different approaches available, few appear to be in regular use.[8] Much of the work on identifying important differences has been carried out on patient-reported outcomes, specifically those seeking to measure health-related quality of life.[9, 10] In practice, the target difference often appears not to be formally based on these concepts and in many cases appears, at least from trial reports, to be determined based on convenience or some other informal basis.[11] Recent surveys among researchers involved in clinical trials demonstrated that the practice is more sophisticated than trial reports suggest.[8] The original DELTA group developed initial guidance, but this was restricted to a standard superiority two-arm parallel-group trial design and limited consideration of related issues.[12] Accordingly, there is a gap in the literature to address this and thereby help improve current practice which this guidance seeks to address.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

2 General considerations for specifying the target difference

2.1 Introduction

RCT design begins with clarifying the research question and then developing the required design to address it. Commonly the PICO(T) framework has been used for this purpose.[13] All of the relevant aspects of trial design (population, intervention, control, outcome, and timeframe) should reflect the research questions of interest. Selection of the primary outcome is considered in Section 2.3, given its key role in trial design and its relationship with the target difference. More recently, the need for greater clarity in trial objectives has been noted, reflecting the existence of multiple intervention (or treatment) effects of potential interest even for the same outcome.[14] These can differ subtly in the population of interest, the role for additional treatment or "rescue" medication, and how the effect is expressed. The concept of estimands has been proposed as a way to bring such distinctions to the fore. An estimand is a more specific formulation of the comparison of interest being addressed. This thinking is reflected in a recent addendum to international regulatory guidelines for clinical trials of pharmaceuticals. Five main strategies are proposed.[15] Of particular note is the treatment policy strategy, which is consistent with what has often been described as an intention-to-treat (ITT)-based analysis.[14, 16, 17] That is, the ITT analysis addresses the difference between a *policy* of offering treatment with a given therapy compared to the *policy* of offering treatment with a different therapy, regardless of which treatments are received. Different stakeholders can have somewhat differing perspectives on the comparison of interest and therefore the estimand of primary interest.[14] Corresponding methods of analyses to address estimands that deviate from traditional conventional analyses are an active area of interest[18] (See also, for example, Section A1.7 for brief consideration of causal inference methods for dealing with non-compliance).

The target difference used in the sample size calculation should be one that at least addresses the trial's primary objective and therefore the intended estimand of primary interest (with the corresponding implications for the handling of the receipt of treatment and population of interest). In some cases, ensuring the sample size is sufficient for more than one estimand may be appropriate, which might imply multiple target differences to address all key objectives. Different estimands may focus on different populations or subpopulations. Estimands will differ in their implications for the magnitude of missing data anticipated (see Section A1.5 for how missing data can be taken into account in the sample size calculation in simple scenarios). Whatever the estimand of interest, the target difference is a key input into the sample size calculation.

2.2 Perspectives

2.2.1 Governmental/charity funder

Funders vary in the degree to which they will specify the research question. The primary concern is that the study provides value for money by addressing a key research question in a robust manner and at reasonable cost to the funder's stakeholders. This is typically an implicit consideration when the sample size and the target difference are determined. However, a very different approach, value of information (See Appendix 2), allows such wider considerations to be formally incorporated. The sample size calculation and the target difference, if well specified, provide reassurance that the trial will provide an answer to the primary research question, at least in terms of comparing the primary outcome between interventions. The specific criteria that proposals are invited to address, and are assessed against, vary among funders and individual schemes within a funder, as does the degree to which the research question may be *a priori* specified by the funder.

One particular aspect that varies substantially among funding schemes and funders is the extent to which they take into account the cost and cost-effectiveness of the interventions under consideration. Some funding schemes require the consideration of costs to come from a particular perspective; this might be the society as a whole, or the health system alone. Whilst other schemes focus solely on clinical and patient perspectives, to greater or lesser extents.

All funders expect an RCT to have a sample size justification.[19] Typically, although not necessarily, this would be via a sample size calculation, most commonly based on the specification of a target difference. The specified target difference would be expected to be one that is of interest to their stakeholders; this is typically patients and health professionals, and sometimes the likely funder of the healthcare, e.g., the National Health Service (NHS) in the UK. For industry-funded trials, the considerations are different, and these are outlined in the next section.

The practical implications of an overly large trial are perhaps mostly financial (the funder has paid more than necessary to get an answer to the research question and thus there is less available for other trials). However, it is also ethically important to avoid more patients than necessary receiving the possibility of a suboptimal treatment, or simply to avoid unnecessary burden on further individuals and to avoid losing the opportunity to devote scarce resource funds to other desirable research. What is and is not sufficient in statistical and more general terms is often very difficult to differentiate except in extreme scenarios. A trial that is too small is at risk of missing an effect. The funder could also later use the target difference in the context of evaluating (formally or informally) whether to close a study due to the probability (or lack thereof) of providing a useful answer in the face of substantially slower progression partway through a trial's recruitment period.

2.2.2 Industry, payers and regulator

Industry-funded trials are typically (but not always) conducted as part of a regulatory submission for a new drug or medical device, or to widen the indications of an existing drug or device. Generally, an active intervention is compared to a placebo control, as this addresses the regulatory question of whether the intervention "works". The main exception would be situations in which a new drug is intended to replace an established effective drug, in which case the established drug would be the control. An example is the evaluation of the newer oral anticoagulants, which have been compared to active comparators such as warfarin or low molecular weight heparin in the submissions for approval.

From an industry perspective, the target difference is often one chosen so that it is important to regulators and healthcare commissioners. The key aspects of interest tend to be safety, including tolerability of treatment and consideration of side effects, whether the treatment is stopped due to a lack of effect, and the effect within those who complete treatment. This has corresponding implications for the estimand(s) of interest.[14, 15] Increasingly, payers (health insurance companies and governmental reimbursement agencies) are interested in comparisons with other active therapies, reflecting the need to inform treatment choices in actual clinical practice and considerations of affordability and cost-effectiveness. A new product will be more likely to be reimbursed if there are clinical advantages over existing therapies, either in terms of efficacy or adverse effect profiles, that are provided at an "acceptable" cost. When an intervention is compared to an active control, the treatment effect between them will almost certainly be smaller and the sample size larger than for a placebo-controlled trial, all other things being equal. One common distinguishing feature between a definitive trial (e.g., Phase III) conducted in an industry setting, versus an academic one, is that all of the evidence pertinent to planning such a trial of a new drug agent will often be readily available within the same company. It is also likely that at least some of the individuals involved will have been involved in a related earlier phase trial of the same drug.

2.2.3 Patient, service users, carers, and the public

From the perspective of patients, service users, carers, and the public[20], where a formal sample size calculation is performed, the target difference should be one that would be viewed as important by a key stakeholder group (such as health professionals, regulators, healthcare funders, and preferably patients). A specific point of interest for those who serve as public and patient involvement (PPI) contributors on research boards that make funding recommendations and/or assess trial proposals is likely to be ensuring the study has considered the most patient-relevant outcome (e.g., a patient-reported outcome), even if it is not the primary outcome. In some

situations, the most appropriate primary outcome may be a patient-reported outcome (e.g., comparing treatments for osteoarthritis where pain and function are the key measures of treatment benefit). It is highly desirable that a patient, service user, and carer perspective feeds into the process for choosing the primary outcome in some way and, where possible, the chosen target difference reflects one that would have a meaningful impact on patient health, according to the research question. Some funders now require at least some PPI in the development of trial proposals, and this perspective forms part of the assessment process.[20] It is also increasingly part of the assessment process for assessing existing evidence.[21]

2.2.4 Research ethics

Fundamental to the standard ethical justification of the conduct of an RCT, which is a scientific experiment on humans, is that it will a) contribute to scientific understanding, and also b) that the participant is aware of what the study entails and whenever possible provides consent to participate.[22, 23] Commonly a third condition, that the participant has the potential to benefit, is also appropriate; this is particularly the case where there may be some risk to the participant. Whatever the specifics of the trial in terms of population, setting, interventions, and assessments, it is important that the sample size for a study is appropriate to achieve its aim. There is a need for justification of some form for the number of participants required. As noted earlier, no more participants than "necessary" should be recruited to avoid unnecessary exposure to a suboptimal treatment and/or the practical burden of participation in a research study. Such a sample size justification may take the form of informal heuristics or, more commonly, a formal sample size calculation.

Clarifying what the study is aiming to achieve and determining an appropriate target difference and sample size is very important as the research can have a big impact, not only on those directly involved as participants but also on future patients. As far as possible, it is also relevant to consider key patient subgroups or subpopulations of individuals in terms of relevance of findings to them. This could be taken into account when undertaking the sample size calculation (See Appendix 1).

2.3 The primary outcome

2.3.1 The role of the primary outcome

The standard approach to an RCT is for one outcome to be assigned as the primary outcome. This is done by considering the outcomes that should be measured in the study.[24] The outcome is "primary" in the sense of it being more important than the others, at least in terms of the design of the trial, although preferably it is also the most important outcome to the stakeholders with respect to the research question being posed. The study sample size is then determined for the primary outcome. As noted earlier, it is important to consider how the primary outcome relates to the population of interest and intervention effects to be estimated (the estimand of interest). Choosing a primary outcome (and giving it prominence in the statistical analysis of the estimands of interest) performs a number of functions in terms of trial design, but it is clearly a pragmatic simplification to aid the interpretation and use of RCT findings. It provides clarification of what the study primarily aims to use to identify the intervention effects. The statistical precision with which this can be achieved is then calculated according to the analysis of interest. Additionally, it clarifies the initial basis on which to judge the study findings. Specification of the primary outcome in the study protocol (and similarly reporting it on a trial registry) helps reduce over-interpretation of findings. This arises from testing multiple outcomes and selectively reporting those that are statistically significant (irrespective of their clinically relevance). This multiple testing, or multiplicity[25, 26], is particularly important given the high likelihood of chance leading to spurious statistically significant findings when a large number of outcomes are analysed. Pre-specification of a primary outcome, along with the use of a statistical analysis plan and transparent reporting (e.g., making the trial protocol available), limits the scope for manipulating (intentionally or not) the findings of the study. This prevents *post hoc* shifting of the focus (e.g., in study reports) to maximise statistical significance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Box 1 Superiority, equivalence, and non-inferiority trials

Superiority trial

In a superiority trial with a continuous primary outcome, the objective is to determine whether there is evidence of a difference in the desired outcome between intervention A and intervention B with mean response μ_A and μ_B, respectively.[1] The null (H_o) and alternative (H_1) hypotheses typically under consideration are:

H_o: The means of the two intervention groups are not different, i.e., $\mu_A = \mu_B$,
H_1: The means of the two intervention groups are different, i.e., $\mu_A \neq \mu_B$.

For a superiority trial, the null hypothesis can be rejected if $\mu_A>\mu_B$ or if $\mu_A<\mu_B$ based on a statistically significant test result.[1, 27] This leads to the possibility of making a Type I error when the null hypothesis is true (i.e., there is no difference between the interventions). The statistical test is referred to as a two-tailed test, with each tail allocated an equal amount of the Type I error ($\alpha/2$, typically set at 2.5%). The null hypothesis can be rejected if the test of $\mu_A< \mu_B$ is statistically significant at the 2.5% level or the test of $\mu_A> \mu_B$ is statistically significant at the 2.5% level. The sample size is calculated on the basis of applying such a statistical test given the magnitude of a difference that is desired to be detected (the *target difference*) and the desired Type I error rate and statistical power. Consideration of a difference in only one direction (one-sided test) is also possible.

Equivalence trial

The objective of an equivalence trial is not to demonstrate superiority of one treatment over another, but to show that two interventions have no clinically meaningful difference, i.e., that they are clinically equivalent (or not different).[28] The corresponding hypotheses for an equivalence trial (continuous primary outcome) take the form:

H_0: There is a difference between the means of the two groups (i.e., they are not "equivalent")

$$\mu_A - \mu_B \leq -d_E \text{ or } \mu_A - \mu_B \geq d_E,$$

H_1: There is a no difference between the means of the two groups (i.e., they are "equivalent")

$$-d_E < \mu_A - \mu_B < d_E,$$

where d_E equates to the largest difference that would be acceptable while still being able to conclude that there is no difference between interventions. It is often called the equivalence margin. μ_A and μ_B are defined as before.

To conclude equivalence, both components of the null hypothesis need to be rejected. One approach to performing an equivalence trial is to test both component, which is called the two one-sided test (TOST) procedure.[1, 28] This can be operationally the same as constructing a (1-α)100% confidence interval (CI) and concluding equivalence if the CI falls completely within the interval $(-d_E, d_E)$. For example, d_E could be set to 10 (on the scale of interest). After conducting the trial, a 95% CI for the difference between interventions could be (-3, 7). As the CI is wholly contained within (-10, 10), the two interventions can be considered to be equivalent.

Non-inferiority trial

A non-inferiority trial can be considered a special case of an equivalence trial. The objective is to demonstrate that a new treatment is not clinically inferior to an established one. This can be formally stated under null (H_0) and alternative (H_1) hypotheses for a non-inferiority trial (continuous primary outcome) that take the form:

H_o: Treatment A is inferior to B in terms of the mean response $\mu_B - \mu_A > d_{NI}$,
H_1: Treatment A is non-inferior to B in terms of the mean response $\mu_B - \mu_A \leq d_{NI}$,

where d_{NI} is defined as the difference that is clinically acceptable for us to conclude there is no difference between interventions, and a higher score on the outcome is a better outcome. Non-inferiority trials reduce to a simple one-sided hypothesis and test, and correspondingly are usually operationalised by constructing a one-sided (1-$\alpha/2$)100% CI. Non-inferiority can be concluded if the lower end of this CI is greater than d_{NI}. No restriction is made regarding whether the new intervention is the same as or better than the other intervention. A mean difference far from d_{NI}, in the positive direction, is not a negative finding, whereas for an equivalence trial it could rule out equivalence.

Equivalence and non-inferiority margins

The setting of an equivalence (and non-inferiority) margin, or limit, is a controversial topic. There are regulatory guidelines on the topic though practice has varied.[29, 30] It has been defined more tightly, and arguably appropriately, as the "largest difference that is clinically acceptable, so that a difference bigger than this would matter in practice".[31] A natural approach would be for d_E to be same as the target difference for a superiority trial comparing the same estimand (see Section 3.2). In the context of replacement pharmaceuticals, this target difference has been suggested to "[be no] greater than the smallest effect size that the active (control) drug would be reliably expected to have when compared with placebo in the setting of the planned trial".[32] An acceptable margin can therefore be chosen via a retrospective comparison to placebo that shows the new treatment is non-inferior to the standard treatment, and thereby indirectly shows the new treatment is superior to placebo.[33] It may also be desirable to demonstrate no substantive non-inferiority, leading to a narrower margin, similar to the approach above.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

2.3.2 Choosing the primary outcome

A variety of factors need to be considered when choosing a primary outcome. First, in principle, the primary outcome should, as noted above, be a "key" outcome, such that knowledge of its result would help answer the research question. For example, in an RCT comparing treatment with eye drops to lower ocular pressure with a placebo for patients with high eye pressure (the key treatable risk factor for glaucoma, a progressive eye disease that can lead to blindness), loss of vision is a natural choice for the primary outcome.[34] However, it would clearly be important to consider other outcomes (e.g., side effects of the eye drop drug). Nevertheless, knowing that the eye drops reduced the loss of vision due to glaucoma would be a key piece of knowledge. In some circumstances, the preferable outcome will not be used because of other considerations. In the above glaucoma example, a surrogate might be used (intraocular pressure, i.e., pressure in the eye) because of the time it takes to measure any change in vision noticeable to a patient, and also because this may enable prevention or at least a reduction in the degree of vision loss. Indeed, intraocular pressure is sometimes the primary outcome of RCTs in this area instead of vision or the visual quality of life.

Consideration is also needed of the ability to measure the chosen primary outcome reliably and routinely within the context of the study. Missing data are a threat to the usefulness of an analysis of any study, and RCTs are no different. The optimal mode of measurement may be impractical or even unethical. The most reliable way to measure intraocular pressure is through manometry;[35] however, this requires invasive eye surgery. Subjecting participants to clinically unnecessary surgery for the purpose of an RCT is only ethical with very strong mitigating circumstances, particularly as an alternative, even if less accurate, way of measuring intraocular pressure exists. Furthermore, invasive measurements may dissuade participants from consenting to take part in the RCT.

Calculating the sample size varies depending on the outcome and the intended analysis. In some situations, ensuring the sample size is sufficient for multiple outcome is appropriate[36]. The three most common outcomes are binary, continuous, and survival (time-to-event) outcomes; they are briefly considered below and in greater depth in Appendix 1. Other outcome types are not considered here, although it should be noted that ordinal, categorical, and count outcomes can be used, though a more complex analysis and corresponding sample size calculation approach is likely to be needed. Continuous outcomes (or a transformed version of them) are typically assumed to be normally distributed, or at least "approximately" so, for ease and interpretability of analysis and for the sample size calculation. This assumption may be inappropriate for some outcomes, like operation time, hospital stay, and costs, which often have very skewed distributions. From a purely statistical perspective, a continuous outcome should not be converted to a binary outcome (e.g., converting a quality-of-life score to high/low quality of life). Such a dichotomisation would result in less statistical precision and lead to a larger sample size being required.[37] If it is viewed as necessary to aid interpretability, the target difference (and corresponding analysis) used in the continuous measure can also be represented as a dichotomy in addition to being expressed on its continuous scale. Some authors, although acknowledging that this should not be routine, would make an exception in some circumstances when a dichotomy is seen as providing a substantive gain in interpretability, even if it is at a loss of statistical precision.[38] For example, the severity of depression may be measured and analysed on a latent scale but the proportion of individuals meeting a pre-specified threshold for depression or improvement might also reported and potentially analysed.[39]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Box 2 Outcome types

The three most common outcome types (binary, continuous, and time-to-event) are briefly described below.

Binary

A binary outcome is one with only two possible values, e.g., cured or not, and dead or alive. In terms of trials, they are usually time-bound, i.e., whether a participant is alive or not at 6 months post-randomisation. Use of the date of the change in status (e.g., time of death) would lead to a survival or time-to-event outcome. Other common trial binary outcomes are the occurrence of an adverse event (e.g., surgical complication or a pharmacological event such as dryness of mouth).

Continuous

Continuous outcomes refer to those that have a numeric scale. True continuous measures (such as blood pressure measurements) have an infinite number of possible values. For example, a value of 125.2334456 mmHg for the systolic blood pressure is theoretically possible, even if is difficult to measure it with such precision. Ordinal outcomes (with a sufficient number of discrete values) are often analysed as if they were continuous due to the difficulties of both calculating the required sample size and also interpreting the result from a more formal, statistically appropriate analysis of an ordinal outcome. This is often done when analysing quality of life measures,[40] where a latent summary scale is produced by applying a scoring algorithm to responses to a set of items, even though there are a fixed number of discrete states (e.g., there are 243 for the EQ-5D-3L with values from -0.594 to 1.0, using the UK population weights). The difficulty of calculating the sample size for an ordinal variable increases quickly as the number of responses increases.[41]

Time-to-event

Time-to-event data are often called "survival" data; a common application is for recording the time to death. However, the same statistical methodology can be used to analyse the time to any event. Examples include disease progression, readmission to hospital, wound healing, and positive ones such as time to full recovery.

Time-to-event data present two special problems in their analysis and hence in sample size estimation:

1. Not all participants have an event; and
2. Participants are observed for varying amounts of time.

If all participants experience an event within the follow-up period, the data could be analysed as a continuous variable. In clinical studies, including RCTs, it is natural for participants to be observed for varying lengths of time. There are two reasons for this:

1. Some participants drop out before the end of follow-up; and
2. Participants are recruited at different times.

Some participants drop out before the end of follow-up, because they decline to take further part in the trial or because they experience some other event, which means that they can no longer be followed up. For example, in a trial where the event of interest is death from a cardiovascular cause, a participant who died in a road traffic accident would become unavailable for further follow-up and would be censored at the time of death.

If participants are followed up from recruitment to the final analysis, some will have been observed for a much longer time than others. In most clinical studies, this is the most frequent reason for varying durations of follow-up. The varying time of follow-up is the main reason why simply analysing the proportion of participants who experience an event, i.e., analyse it as if it were a binary outcome, is not appropriate.

3 Specifying the target difference

3.1 General considerations

3.1.1 Introduction

Despite its key role, the specification of the target difference for an RCT has received surprisingly little discussion in the literature and in existing guidelines for conducting clinical trials.[2, 6] As noted above, the target difference is the difference between the interventions in the primary outcome used in the sample size calculation that the study is designed to reliably detect. If correctly specified, it provides reassurance (should the other assumptions be reasonable and the sample size met) that the study will be able to address the RCT's main aim in terms of the primary outcome, the population of interest, and the intervention effects. It can also aid interpretation of the study's findings, particularly when justified in terms of what would be an important difference. The target difference

therefore should be one that is appropriate for the planned principal analysis (i.e., the estimand that is to be estimated and the analysis method to be used to achieve this).[15, 17, 42, 43] This is typically (for superiority trials) what is known as an ITT-based analysis, i.e., according to the randomised groups irrespective of subsequent compliance with the treatment allocation. Other analyses that address different estimands[14, 17, 43] of interest could also inform the sample size calculation (see also Appendix 1, Section A1.7 for a related topic). How the target difference can be expressed will depend also on the planned statistical analysis. A target difference for a continuous outcome could be expressed as a difference in means, medians or even as a difference in distribution. Binary outcomes could be expressed as an absolute difference in proportions or as a relative difference (e.g. odds or risk ratio). Irrespective of the outcome type, there are two main bases for specifying the target difference, one that is considered to be:

- *important* to one or more stakeholder groups (e.g., health professionals or patients); or
- *realistic* (plausible), based on either existing evidence (e.g., seeking the best available estimates in the literature), and/or expert opinion.

Recommendations on how to go about specifying the target difference are provided in Box 3. A summary of the seven methods that can be used for specifying the target difference is provided in Section 3.2 below.

A very large literature exists on defining a (clinically) important difference, particularly for quality-of-life outcomes.[44-46] Much of the focus has been on estimating the smallest value that would be considered clinically important by stakeholders (the "minimum clinically important difference – (MCID)")[44-47]. In a similar manner, discussion of the relevance of estimates from existing studies are also common occurrences. It should be noted that it has been argued that a target difference should always meet both of the above criteria.[48] This would seem particularly apt for a definitive Phase III RCT. There is some confusion in the reporting of sample size calculations for trials in the literature and what the use of a particular approach justifies. For example, using data from previous studies (see Sections 3.2.5 and 3.2.6) cannot by itself inform the importance, or lack thereof, of a particular difference.

The subsequent Sections (3.1.2 and 3.1.3) consider two special topics, individual- and population-level important difference, and reverse engineering of the sample size calculation, respectively.

3.1.2 Individual- versus population-level important differences

In an RCT sample size calculation, the target difference between the treatment groups strictly relates to the difference between the underlying populations. In a similar manner, the health economic consideration refers to how to manage a population of individuals in an efficient manner. However, the difference in an outcome that is important to an individual is not necessarily the same difference that might be viewed as important at the population level. Rose[49] grappled with the meaning and relationships between individual- and population-level differences, and their implications, in the context of disease prevention. He noted that, based on data from the Framingham study, an average 10 mmHg lowering of blood pressure could potentially result in a 30% reduction in attributable mortality. Whilst a 10 mmHg change in an individual might seem small, if a treatment could achieve that average difference, it would be very beneficial. 10 mmHg could therefore be justified as an appropriate and important target difference for a trial in a similar population. An individual may wish a greater impact, particularly if the intervention they are to receive is burdensome or carries some risk.

More recently, researchers in other clinical areas have also distinguished between what is "important" at an individual level and what is "important" at a group level for quality of life measures.[50-52] In an RCT sample size calculation, the parameters assumed for the outcome in the intervention groups in the sample size calculation, including the target difference, should reflect the

population level values, e.g., the mean difference in Oxford Knee Score (OKS), even though individual values can vary substantially.[53] When considering the importance of and/or how

Box 3 Recommendation for specifying the target difference in a randomised controlled trial sample size calculation

The following are recommendations for specifying the target difference in an RCT's sample size calculation when the conventional approach to the sample size calculation is used. Recommendations on the use (or not) of individual methods are made. More detailed advice on the application of the individual methods can be found elsewhere.[7]

Recommendations

1. Begin by searching for relevant literature to inform the specification of the target difference. Relevant literature can:
    a. relate to a candidate primary outcome and/or the comparison of interest, and;
    b. inform what is an important and/or realistic difference for that outcome, comparison, and population (estimand of interest).
2. Candidate primary outcomes should be considered in turn, and the corresponding sample size explored. Where multiple candidate outcomes are considered, the choice of primary outcome and target difference should be based on consideration of the views of relevant stakeholders groups (e.g., patients), as well as the practicality of undertaking such a study and the required sample size. The choice should not be based solely on which yields the minimum sample size. Ideally, the final sample size will be sufficient for all key outcomes, although this is not always practical.
3. The importance of observing a particular magnitude of a difference in an outcome, with the exception of mortality and other serious adverse events, cannot be presumed to be self-evident. Therefore, the target difference for all other outcomes requires additional justification to infer importance to a stakeholder group.
4. The target difference for a definitive (e.g., Phase III) trial should be one considered to be important to at least one key stakeholder group.
5. The target difference does not necessarily have to be the minimum value that would be considered important if a larger difference is considered a realistic possibility or would be necessary to alter practice.
6. Where additional research is needed to inform what would be an *important* difference, the anchor and opinion-seeking methods are to be favoured. The distribution should not be used. Specifying the target difference based solely on an SES approach should be considered a last resort, although it may be helpful as a secondary approach.
7. Where additional research is needed to inform what would be a *realistic* difference, the opinion-seeking and review of the evidence base methods are recommended. Pilot studies are typically too small to inform what would be a realistic difference and primarily address other aspects of trial design and conduct.
8. Use existing studies to inform the value of key "nuisance" parameters that are part of the sample size calculation. For example, a pilot trial can be used to inform the choice of standard deviation (SD) value for a continuous outcome or the control group proportion for a binary outcome, along with other relevant inputs such as the amount of missing outcome data.
9. Sensitivity analyses that consider the impact of uncertainty around key inputs (e.g., the target difference and the control group proportion for a binary outcome) used in the sample size calculation should be carried out.
10. Specification of the sample size calculation, including the target difference, should be reported according to the guidance for reporting items (see below) when preparing key trial documents (grant applications, protocols, and result manuscripts).

realistic a specific difference is, the intended trial population must be born in mind. The difference that would be considered important by patients may well vary between populations (e.g., according to the severity of osteoarthritis).[54] For example, the importance of a 5-point increase

(improvement) in the OKS for a relatively healthy population with a mean baseline level of 30 points (out of 48) could well differ from that for a population that has severe osteoarthritis with a mean baseline level of 10 points. Similarly, in terms of population risk, e.g. of a stroke, a small reduction at a population level might be considered very important, whereas for a group of high-risk patients, a more substantial reduction may be required.[49]

Work has shown that individuals differ in what magnitude of difference they consider important, at least in part due to their varying baseline levels.[10, 44] This general issue has implications when selecting a target difference, as it should be a difference that reflects the analysis at the group (and intended population) level and the comparison at hand. Care is therefore needed when using values from external studies to infer an important difference.

3.1.3 Reverse engineering

The difference that can be detected for a given sample size is often calculated. It can be apparent that this has been done, e.g., when one sees a spuriously precise target difference that leads to a round sample size without any other justification. For example, a target difference of 16.98 for a trial with pooled SD of 30, statistical power of 80% at two-sided 5% significance level, and two treatment groups of 100 participants has clearly been reverse-engineered.

It is important to distinguish calculating the target difference for a prospective trial from calculating the target difference on the basis of the recruited sample size once the trial has been completed (*post hoc* power calculation). The former has a useful role in the process of planning and deciding what is feasible; the latter is unhelpful and uninformative.[55]

Case study 6 describes a situation where a fixed (and complete) number of observations were expected without loss due to consent or attrition-driven subsampling, but the corresponding target difference was calculated and deemed to be an important and realistic difference to use.

3.2 Methods for specifying the target difference

The methods for specifying the target difference can be broadly grouped into seven types. These are briefly described below.

3.2.1 Anchor

The quantification of a target difference or effect size for a sample size calculation is not straightforward for an established endpoint or outcome measure.[56] For a new outcome, especially a patient-reported health-related quality-of-life measure, it is even more difficult, as clinical experience with using the new outcome may not been sufficiently long to evaluate what a clinically meaningful or important difference might be. Additionally, for a measure such as a quality-of-life outcome, the scale has no natural meaning and is completely a function of the scoring method (i.e., a 1 point difference does not have any naturally interpretable value).

The outcome of interest can, however, be “anchored” by using someone’s judgement, typically a patient or a health professional, to define what an important difference is.[45-47] This is typically achieved by comparing a patient’s health before and after a recognised treatment, and then linking the change to participants who showed improvement and/or deterioration according to the judgement of changes (e.g., on a 5-point Likert scale from “substantial deterioration” through to “substantial improvement”). Alternatively, a more familiar outcome (for which patients or health professionals more readily agree on what amount of change constitutes an important difference) can be used. In this way, one outcome is anchored to another outcome about which more is known. Contrasts between patients (such as individuals with varying severity of a disease) can also be used to determine a meaningful difference, e.g., via patient-to-patient assessments.[12, 57]

The FDA has described a variety of methods for determining the minimum important difference, including the anchor approach.[32] Changes in quality-of-life measures can be mapped to clinically

relevant and important changes in non-quality-of-life measures of treatment outcome in the condition of interest (although they may not correlate strongly). [58] There are a multitude of minor variations in the approach (e.g., the anchor question and responses, or how the responses are used), although the general principles are the same.[7, 45-47]

3.2.2 Distribution

Distribution approaches are not recommended for use to inform the choice of the target difference given their inherently arbitrary nature in this context. The rationale for this recommendation is set out below.

Two distinct distribution approaches can be grouped under this heading[7, 44]: measurement error and rule of thumb. The measurement error approach determines a value that is larger than the inherent imprecision in the measurement and that is therefore likely to be consistently noticed by patients. This is often based on the standard error of measurement (SEm). The SEm can be defined in various ways, with different multiplicative factors suggested as signifying a non-trivial (important) difference. The most commonly used alternative to the SEm method (although it can be thought of as an extension of this approach) is the reliable change index proposed by Jacobson and Truax[59], which incorporates confidence around the measurement error.

The rule-of-thumb approach defines an important difference based on the distribution of the outcome, such as using a substantial fraction of the possible range without further justification. An example would be viewing a 10 mm change on a 100-mm visual analogue scale measuring symptom severity as a substantial shift in outcome response.

Measurement error and rule-of-thumb approaches are widely used in the area of measurement properties of quality of life, but do not translate straightforwardly to an RCT target difference. For measurement error approaches, this is because the assessment is typically based on test-retest (within-person) data, whereas most trials are of parallel-group (between-person) design. Additionally, measurement error is not sufficient rationale as the sole basis for determining the importance of a particular target difference. More generally, the setting and timing of data collection may also be important to the calculation of measurement error (e.g., results may vary between pre- and post-treatment).[60] Rule-of-thumb approaches are dependent on the outcome having inherent value (e.g., the Glasgow coma scale), where a substantial fraction of a unit change (e.g., one-third or one-half) can be viewed as important. In this situation any reduction is arguably also important, and the issue is more one of research practicality (as per mortality outcome) than detecting a clinically important difference.

3.2.3 Health economic

Approaches to using economic evaluation methodology to inform the design of RCTs have been proposed since the early 1990s.[61, 62] These earlier approaches sought to identify threshold values for key determinants of cost-effectiveness, and are akin to determining an important difference in clinical outcomes, albeit on a cost-effectiveness scale. However, uptake has been very low. A recent review by Hollingworth and colleagues[63] identified only one study that considered cost-effectiveness in the sample size calculation. They also showed that trials powered on clinical end points were less likely to reach definitive conclusions of cost-effectiveness compared to clinical effectiveness.

Despite the lack of use, further development of methods has continued. A strand in the development of these methods has been to focus on a variation in the standard frequentist approach to sample size estimation. The most recent exposition of this was by Glick.[64, 65] Glick focused on a particular economic metric, the incremental net (monetary) benefit (INB) statistic. A key aspect of the INB is that it monetarises a unit of health effect by multiplying it by the decision-maker's willingness to pay for that unit of health effect. Power is taken to be the chance that the

lower limit of the CI calculated from the future trial exceeds 0. An important difference is then *any* difference in INB that is ≥0, and the size of the trial can be set so as to detect this. However, Glick notes that willingness to pay is not known for certain (e.g., in England and Wales, the National Institute for Health and Care Excellence (NICE)[66] currently specifies a range of between £20,000 and £30,000 per quality adjusted life year gained), and that other things being equal, increasing the decision-maker's willingness to pay for a unit of health effect reduces the sample size. An alternative economics-based approach, value of information, is summarised in Appendix 2.

3.2.4 Opinion-seeking

The opinion-seeking method determines a value, a range of plausible values, or a prior distribution for the target difference by asking one or more "experts" to state their opinion on what value(s) for a particular difference would be important and/or realistic.[67, 68] Eliciting opinions on the relative importance of the benefits and risks of a medicine may also be used to inform the choice of non-inferiority or equivalence margins for such trials.[69, 70]

The definition of an expert (e.g., clinician, patient, or trialist) must be tailored to the quantity on which an opinion is sought. Various approaches can be used to identify experts (e.g., key opinion leaders, literature search, mailing list, or conference attendance). Other variations include the approach used to elicit opinion (e.g., group and/or individual interviews, questionnaires, email surveys, or workshops)[71-73], the complexity of the data elicited (from a single value[74] to multiple assessments incorporating uncertainty[75] and/or sensitivity to key factors such as baseline level[76]), and the method used to consolidate the results into an overall value, range of values, or distribution.[67]

Many elicitation techniques have been developed in the context of Bayesian statistics to establish a prior distribution quantifying an expert's uncertainty about the true treatment difference.[67] The expert will be asked a series of questions to elicit a number of summaries of their prior distribution. The number and nature of these summaries will depend on the nature of the treatment difference (i.e., whether this is a difference in means, risk ratio, etc.) and what parametric distribution (if any) will be used to model the expert's prior. Typically, more summaries are elicited than are strictly necessary to enable model checking. Feedback of the fitted prior is an essential part of the elicitation process to ensure it adequately captures the expert's beliefs. Examples of prior elicitation include the CHART and MYPAN trials.[75, 77, 78] When the opinions of several experts are elicited, several priors may be used to capture a spectrum of beliefs (e.g., sceptical, neutral, or enthusiastic). Priors may be used to inform the design of a conventional trial, e.g., when setting the sample size or an early stopping rule[77, 79] to ensure the study would convince a prior sceptic. Alternatively, priors may be incorporated into the interpretation of a Bayesian trial to reduce uncertainty, which may be appropriate in cases such as rare diseases when a conventionally powered study is infeasible.[80] Bayesian approaches to sample size calculations are discussed in more detail in Appendix 2.

An advantage of the opinion-seeking method is the relative ease with which it can be carried out in its simpler forms.[76] However, the complexity increases substantially when undertaken as a formal elicitation.[75] Whatever the approach used, it should ideally match, as closely as possible, the intended trial research question.[71, 76, 81] Findings will vary according to the patient population and comparison of interest. Additionally, different perspectives (e.g., patient versus health professional) may lead to very different opinions on what is important and/or realistic.[81] The views of individuals who participate in the elicitation process may not represent those of the wider community. Furthermore, some methods for eliciting opinions have cost or feasibility constraints (e.g., those requiring face-to-face interaction). However, alternative approaches better able to capture the views of a larger number of experts require careful planning to ensure questions are clearly understood. Care is needed with these approaches, as they may be subject to low response rates[76] or may produce priors with limited face validity.

3.2.5 Pilot studies

Pilot studies come in various forms.[82] A useful distinction can be made between pilot studies, *per se*, and the subset of pilot trials that can be defined as an attempt to *pilot* the study methodology prior to conducting the main trial. As such, data from a pilot trial are likely to be directly relevant to the main trial. This section therefore focuses on pilot trials, although the considerations are relevant to other pilot studies that have not been designed with a particular trial design in mind. It should be noted that some Phase II trials can be viewed in a similar manner as preparing for a Phase III trial and therefore can inform sample size calculations.

Pilot trials are not well suited to quantifying a treatment effect as they usually have a small sample size, and are not typically large enough to quantify with much certainty what a realistic difference would be.[83] Accordingly, avoiding conducting formal statistical testing and focusing instead on descriptive findings and interval estimation is recommended.[82, 84] In terms of specifying the target difference for the main trial, pilot trials are most useful in providing estimates of the associated "nuisance" parameters (e.g., standard deviation (SD) and control group event proportion. See Section 5.3 for more details).[82, 85] Like any quantity, these parameters will, however, be estimated with uncertainty, which has implications for the sample size of both a pilot trial and a subsequent main trial.[86]

Another use of a pilot trial is to assess the plausibility (at a less exacting level of statistical certainty than would be typically required for a main trial) of a given difference considered to be important through the calculation of a CI.[85] Pilot trial-based CIs can be considered investigative and can be used to help with informing decision making. If an effect of this size is not ruled out by the CI of the estimated effect from the pilot trial, then results could be deemed sufficiently promising to progress to the main trial.[84, 87]

3.2.6 Review of the evidence base

An alternative to conducting a pilot trial is to review existing studies to assess a realistic effect and therefore inform the choice of target difference for the main trial.[7] This can be called a review of the evidence base. Pre-existing studies for a specific research question can be used (e.g., using the pooled estimate of a meta-analysis) to determine the realistic difference.[1] It has been argued strongly and persuasively that this should be routine prior to embarking on a new trial.[88] Extending this general approach, Sutton and colleagues[89, 90] derived a distribution for the effect of a treatment from a meta-analysis, from which they then simulated the effect of a "new" study; the result of this study was added to the existing meta-analysis data, which was then re-analysed. Implicitly this adopts a realistic difference as the basis for the target difference and therefore makes no judgement about the value of the effect should it truly exist. Using the same target difference as a previous trial, while heuristically convenient, does not provide any real justification as it may or may not have been appropriate when used in the last study.

It is likely that existing evidence is often informally used (indeed research funders typically require a summary of existing evidence prior to commissioning a new study), although little research has addressed how it should formally be done. Estimates identified from existing evidence may not necessarily be appropriate for the population and estimand under consideration for the trial, so the generalisability of the available studies and susceptibility to bias should be considered. Indeed, the planning of a new study implies some perceived limitation in the existing literature. Imprecision of the estimate is also an important consideration, and publication bias may also be an issue if reviews of the evidence base consider only published data. If a meta-analysis of previous studies is used to inform the sample size calculation for a new trial, additional evidence published after the search used in the meta-analysis was conducted may require the updating of the sample size calculation during trial conduct to maintain a realistic difference. The control group proportion or the SD (as well as other inputs that influence the overall sample size) can be estimated using existing evidence.

3.2.7 Standardised effect size

The magnitude of the target difference on a standardised scale (standardised effect size, SES) is commonly used to infer the value of detecting this difference when set in comparison with other possible standardised effects.[7, 83] Overwhelmingly, the practice for RCTs, and in other contexts in which the (clinical) importance of a difference is of interest, is to use the guidelines suggested by Cohen[91] for Cohen's d metric, i.e., 0.2, 0.5, and 0.8 for small, medium, and large effects, respectively, as de facto justification. These values were given in the context of a continuous outcome for a between-group comparison (akin to a parallel-group trial) with the caveat that they are specific to the context of social science experiments. Despite this, due in part to having some face validity and in part to the absence of a viable or ready alternative, justification of a target difference on this basis is widespread. Colloquially and rather imprecisely, Cohen's d value is often described as the trial "effect size".

Other SES metrics exist for continuous (e.g., Dunlap's d), binary (e.g., odds ratio), and survival (hazard ratio) outcomes, and a similar approach can be readily adapted for other types of outcomes.[91, 92] The Cohen guidelines for small, medium, and large effects can be converted into equivalent values for other binary metrics (e.g., 1.44, 2.48, and 4.27, respectively, for odds ratios).[93]. Guidelines for other effect sizes exist (including some suggested by Cohen).[91] Informally, a doubling or halving of a ratio is sometimes seen as a marker of a large relative effect. However, no equivalent guideline values are in widespread use for any of the other effect sizes. In case of relative effect metrics (such as the risk ratio), this probably reflects the difficulty in considering a relative effect apart from the control group response level.

The main benefit of using an SES method is that it can be readily calculated and compared across different outcomes, conditions, studies, settings, and people; all differences are translated into a common metric. It is also easy to calculate the SES from existing evidence if studies have reported sufficient information. Where calculated, the SD (or equivalent inputs) used should reflect the intended estimand (i.e., the population and outcome).

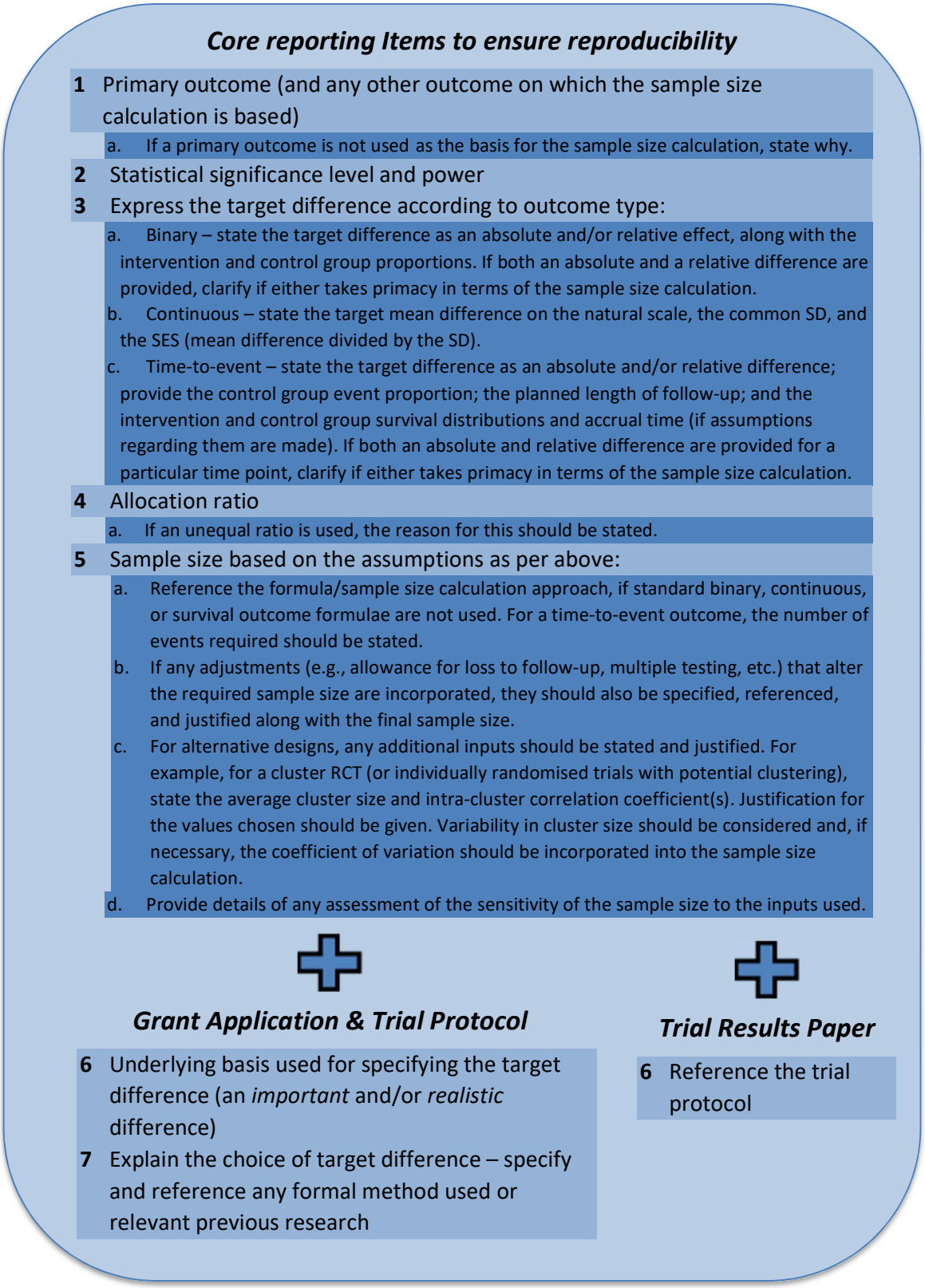
It is important to note that SES values are not uniquely defined, and different combinations of values on the original scale can produce the same SES value. For example, different combinations of mean and SD values produce the same Cohen's d statistic SES estimate. A mean (SD) of 5 (10) and 2 (4) both give a standardised effect of 0.5 SD. As a consequence, specifying the target difference as an SES alone, although sufficient in terms of sample size calculation, can be viewed as insufficient in that it does not actually define the target difference for the outcome measure of interest in the population of interest. A further limitation of the SES is the difficulty in determining why different effect sizes are seen in different studies: for example, whether these differences are due to differences in the outcome measure, intervention, settings, or participants in the studies, or study methodology. Using this approach should be viewed as, at best, a last resort. It is perhaps more useful (for a continuous outcome) to provide a benchmark to assess the value from another method. Preferably some idea of effect sizes for an accepted treatment in the specific clinical area of interest would be available.[94]

4 Reporting of the sample size calculation for a randomised controlled trial

The approach taken and the corresponding assumptions made in the sample size calculation should be clearly specified as well as all inputs and formula so that the basis upon which the sample size was determined is clear. This information is critical for reporting transparently, allows the sample size calculation to be replicated, and clarifies the primary (statistical) aim of the study. A recommended list of reporting items for recording in key trial documents (grant applications, protocols, and results paper) is provided in Figure 1 for when the conventional approach to sample size calculation has been used. Where another approach has been used, appropriate items should be reported sufficient to ensure transparency and allow replication.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Figure 1 Recommended sample size calculation reporting items for key trial documents



Core items sufficient to replicate the sample size calculation should be provided in all key documents. Under the conventional approach with standard (1:1 allocation two-arm parallel-group) trial design and unadjusted statistical analysis, the core items that should be stated are the primary

outcome, target difference appropriately specified according to the outcome type, associated nuisance parameters, and statistical significance and power. Specification of the target difference in the sample size calculation section varies according to the type of primary outcome.

When the calculation deviates from the conventional approach (see Appendix 1), whether by research question or statistical framework, this should be clearly specified. Formal adjustment of the significance level for either multiple outcomes, comparisons or interim analyses should be specified[25, 26, 36, 95]. Justification for all input values assumed should be provided.

BOX 4 Protocol sample size calculation section: binary primary outcome example - Men After Prostate Surgery (MAPS) trial[96]

The primary outcome is urinary continence. The sample size is based on a target difference of 15% absolute difference (85% vs. 70%) at 12 months post-randomisation. This magnitude of target difference was determined to be both a realistic and an important difference from discussion between clinicians and the project management group, and from inspection of the proportion of urinary continence in the trials included in a Cochrane systematic review.[97] The control group proportion (70%) is also based on the observed proportion in the RCTs in this review. Setting the statistical significance to the two-sided 5% level and seeking 90% power, 174 participants per group are required, giving a total of 348 participants. Allowing for just under 15% missing data leads to 200 per group (400 participants overall).

We recommend that trial protocols and grant applications report additional information explicitly clarifying the basis used for specifying the target difference and the methods/existing studies used to inform the specification of the target difference. Examples of a trial protocol sample size section under a conventional approach to the sample size calculation for a standard trial and unadjusted analysis are provided in Boxes 4-6 for binary, continuous, and time-to-event primary outcomes, respectively.

BOX 5 Protocol sample size calculation section: continuous primary outcome example – Full-thickness macular hole and Internal Limiting Membrane peeling Study (FILMS)[98]

The primary outcome is Early Treatment Diabetic Retinopathy Study (ETDRS) distance visual acuity.[99] A target difference of a mean difference of five letters with a common SD of 12 at 6 months post-surgery is assumed. Five letters is equivalent to one line on a visual acuity chart and is viewed as an important difference by patients and clinicians. The SD value is based on two previous studies – one observational comparative study[100] and one RCT.[101] This target difference is equivalent to an SES of 0.42. Setting the statistical significance to the two-sided 5% level and seeking 90% power, 123 participants per group are required, giving 246 participants (274, allowing for 10% missing data) overall.

Due to space restrictions, in many publications the main trial paper is likely to contain less detail than is desirable. Nevertheless, a minimum set of reporting items is recommended for the main trial results paper along with full specification in the trial protocol. The trial results paper should reference the trial protocol, which should be made publicly available. The recommended list of items given in Figure 1 for the trial paper (as well as for the protocol) is more extensive than that in the CONSORT (including the 2010 version)[102] and SPIRIT[103] statements.

BOX 6 Protocol sample size calculation section: survival primary outcome example - Arterial Revascularisation Trial (ART)[104]

The primary outcome is all-cause mortality. The sample size was based on a target difference of 5% in 10-year mortality with a control group mortality of 25%. Both the target difference and control group mortality proportions are realistic based on a systematic review of observational (cohort) studies.[105] Setting the statistical significance to the two-sided 5% level and seeking 90% power, 1464 participants per group are required, giving a total of 2928 participants (651 events).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

5 Case studies of sample size calculations

A variety of case studies are provided for different trial designs, including varying types of primary outcomes, availability of evidence to inform the target difference, and level of complexity. A short description is provided in Table 1 below.

Table 1 Case studies

No	Description	Trial
1	Standard (two-arm parallel-group) trial where the opinion-seeking and review-of-the-evidence-base methods were used to inform the target difference for a binary outcome	MAPS
2	Two-arm parallel-group trial where the anchor, distribution and SES methods were used to inform the target difference for a continuous quality-of-life outcome	ACL-SNNAP
3	Crossover trial where the opinion-seeking and review-of-the-evidence-base methods were used to inform the target difference for a binary patient-reported outcome	OPTION-DM
4	Three-arm parallel-group trial where the opinion-seeking method was used to inform the target difference for a binary clinical outcome	SUSPEND
5	Three-arm/two-stage parallel-group trial where the anchor and review-of-the-evidence-base methods were used to inform an important and realistic difference in a continuous quality-of-life outcome	MACRO
6	Two-arm cluster-cluster trial where the opinion-seeking and review-of-the-evidence-base methods were used to inform an important and realistic difference in a continuous cluster-level process measure outcome	RAPiD

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Case study 1 MAPS trial

Radical prostatectomy is carried out for men suffering from early prostate cancer. The operation is usually carried out through an open incision in the abdomen, which may damage the urinary bladder sphincter, its nerve supply, and other pelvic structures. Urinary incontinence occurs in around 90% of men initially, but the long-term prognosis varies from 2% to 60%, depending on how incontinence is measured and time after surgery. Successive Cochrane systematic reviews have shown that, although conservative treatment based on pelvic floor muscle training may be offered to men with urinary incontinence after prostate surgery, there is insufficient evidence to evaluate its effectiveness and cost-effectiveness. Men After Prostate Surgery (MAPS) was a multicentre RCT that aimed to assess the clinical effectiveness (primarily by looking at the presence of urinary incontinence post-treatment) and cost-effectiveness of active conservative treatment delivered by a specialist continence physiotherapist or a specialist continence nurse compared with standard management in men receiving a radical prostatectomy at 12 months after surgery.

The primary outcome was the presence of urinary continence. No other outcomes were considered. The sample size was based on a target difference of 15% absolute difference (85% specialist treatment versus 70% control). A Cochrane systematic review [84] suggested the current control group proportion was 70% (average across relevant control groups). This magnitude of target difference was determined to be both a realistic and an important difference from discussion between clinicians and the project management group, and from inspection of the proportion of patients with urinary continence in the trials included in the Cochrane systematic review.[84] Setting the statistical significance to the two-sided 5% level and seeking 90% power, 174 participants per group were required, giving a total of 348 participants prior to considering missing data. Allowing for just under 15% missing data increased the overall sample size to 400. The power (77%) should the control group response turn out to be 40% (i.e., using 55% for the treatment and 40% for the control) was calculated as a sensitivity analysis. As the power was still reasonably high and this was considered a less plausible scenario, the overall sample size was not changed.

BOX C1 Sample size calculation in published paper: MAPS [96]

The primary outcome is urinary continence. The sample size was based on a target difference of 15% absolute difference (85% versus 70%). This magnitude of target difference was determined to be both a realistic and an important difference from discussion between clinicians and the project management group, and from inspection of the proportion of urinary continence in the trials included in a Cochrane systematic review.[97] The control group proportion (70%) was also based on the observed proportion in the RCTs in this review. Setting the statistical significance to the two-sided 5% level and seeking 90% power, 174 participants per group are required, giving a total of 348 participants. Allowing for just under 15% missing data leads to 200 per group (400 participants overall).

Case study 2 ACL SNNAP trial

Anterior cruciate ligament (ACL) rupture is a common injury, mainly affecting young, active individuals. ACL injury can have a profound effect on knee kinematics (knee movement and forces), with recurrent knee instability (giving way) the main problem. In the UK, a surgical management strategy has become the preferred treatment for ACL-injured individuals. However, the preference for surgical management (reconstruction) of the ACL-deficient knee had been questioned by a Scandinavian trial, which suggested that rehabilitation can reduce the proportion of acute patients requiring surgery by up to 50%.[106]

A two-arm RCT, ACL SNNAP (ACL Surgery Necessity in Non Acute Patients) was planned to compare a strategy of non-surgical management with the option of surgery if required (the rehabilitation group) with a strategy of surgical management only (the reconstruction group) in the UK NHS setting of treating non-acute patients. The main outcome of interest was the Knee Injury and Osteoarthritis Outcome Score (KOOS)-4 score, which excludes the activities of daily living component of the KOOS. This decision reflected belief about the impact of ACL rupture and the aim of treatment).[106, 107] KOOS-4 seemed to be the most appropriate of the available condition-relevant quality-of-life measures.

Limited work had assessed what would be a minimally important difference in the overall KOOS score and the KOOS4 variant. The KOOS user guide recommended 8 to 10 points as the (current) best estimate of a minimally important change.[108] This was based on an anchor method approach using clinical judgement about the recovery time scale applied to a small cohort of ACL reconstruction patients.[109] Differences that occurred within the recovery period were 7 points or less, whereas those that occurred afterwards were 8 points or more for three of the four KOOS domain scores. Given the limited data on what would constitute an important difference, estimates from a distribution-based approach (minimal clinically detectable difference (MCDC)) were also considered. The MCDC was around 6-12 for individual domains.[110] A value of 8 points was taken as a reasonable value for the minimal clinically important difference in the KOOS-4 *overall* score.

A standard sample size calculation for comparing two means using an SD of 19 gave a required sample size of 120 in *each* group for 90% power at a two-sided 5% significance level. This is how many patients would be required for an individually randomised trial in the absence of any clustering of outcome. The impact of clustering of outcome by the main intervention deliverer (surgeon and/or physiotherapist) was also considered possible. Given the time of outcome (quality of life at 6 month), previous evidence suggested any clustering effect to be low: circa 0 to 0.06 for intra-class correlation (ICC) effect estimates from a database of previous surgical trials.[111] Clustering was assumed to occur to the same degree in both arms. Two surgeons from at least 13 sites were anticipated, whereas *a priori* more physiotherapists (at least 50% more, i.e., around 40) were anticipated to be involved in the study. Credible SDs for the cluster sizes were informally assessed using mock scenarios. Equal allocation was planned.

The sample size was estimated to be 130 patients per group to achieve just over 80% power, based on assuming an ICC of 0.06. With 26 surgeons, the number of patients per surgeon in the surgery management arm was expected to be 5 on average. With 40 physiotherapists, the number of patients per physiotherapist was expected to be 3 on average. Some allowance for variance in the number per health professional was also made. Given the anticipated challenges in recruiting to the study, keeping the sample size as small as possible was considered critical. As clustering was not certain, the sample size was increased to ensure at least 80% power if clustering occurred. In the absence of clustering, the power would be over 90%.

To allow for missing data, the sample size was set at 320 (allowing approximately 15% loss to follow-up). The total required sample size was therefore 320. As the funding agency requested an interim

check on the degree of clustering, a single planned interim check was set once data for 100 patients had been collected. This planned interim assessment would only assess the ICC magnitude and other sample size assumptions such as cluster size. A formal interim analysis comparing treatments was not planned. See Box C2 for the corresponding sample size explanation in the trial protocol.

BOX C2 Protocol sample size calculation example: ACL SNNAP

320 participants will be recruited to the study. The minimal clinically important change (MIC) for the KOOS score is 8-10 points.[108] Estimates of the minimal clinically detectable change (MCDC) for the two KOOS subscales most relevant for ACLD vary between 5 and 12 points (Symptoms 5-9, and Sport/Rec6-12).[108] Conservatively, a target difference of 8 points and SD of 19 (the highest value observed in a trial of acute patients at baseline amongst the KOOS subscales) is assumed. Given these assumptions, 120 participants per group are required (240 in total) to achieve 90% power at two-sided 5% significance level in the absence of any clustering of outcome.

To ensure sufficient power, clustering (clsampsi Stata command [112]) has been allowed for by conservatively assuming an intra-cluster correlation (ICC) of 0.06 [111] and cluster size n, mean (SD) of 26, 5 (12) and 43, 3 (5) for the ACL reconstruction and rehabilitation groups, respectively. Therefore 130 participants are required per group (260 participants overall) to ensure just over 80% power. Given the conservative nature of the assumed values and the anticipated gain in precision from adjusting for the baseline scores and other randomisation factors, actual power is likely to be higher even in the presence of clustering.

To allow for just over 15% missing data (response in a similar trial [107]), 320 participants will be needed. An interim analysis will be carried out to estimate the magnitude of clustering for the 6 months KOOS4 outcome once data is available for 100 participants. A decision as to whether the sample size should be increased to allow for a greater level of clustering than anticipated will be made based on the interim analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Case study 3 OPTION-DM trial

A common comorbidity for patients with diabetes is neuropathic pain. Although there are some pharmacologic treatments for this pain, it is unclear which is best. As the first-line treatment often does not work, patients may get second-line treatments as part of a care pathway. In the OPTION-DM (Optimal Pathway for TreatIng neurOpathic paiN in Diabetes Mellitus) trial, three care pathways were to be compared in a three-period crossover study. All patients would receive all three patient pathways. Each care pathway reflected a form of clinical practice that a patient might receive for their neuropathic pain. The main candidate primary outcome was the 7-day average 24 h pain after 16 weeks of treatment, measured on a numeric rating scale.

There was some experience of using such a pain score within the study team and in the published literature

- A recent placebo-controlled crossover trial observed a 0.5-point average difference between the active comparator and placebo;[113]
- Patients in this population on the active treatment were expected to improve from baseline by on average 2 points; and
- A 1-point improvement within an individual patient was viewed as a clinically important difference, based on an existing study that used an opinion-seeking approach.[114]

These criteria were used to inform the choice of a clinically important difference. The wish was to increase the proportion of patients improving by 1 point or more. The proportion of individuals improving can be calculated given the assumed reduction and difference between the groups. We expected a mean improvement of 2 points from baseline. Assuming the change from baseline followed a normal distribution, 66% of patients were anticipated to improve by 1 point (see Table C3).

Table C3 Cumulative proportion of individual patient differences from baseline, estimated from a normal distribution assuming mean differences of 2 and 2.5

Cumulative proportion of individuals improving	Anticipated improvements from baseline	
	2 points reduction	*2.5 points reduction*
0.50	-2.00	-2.50
0.52	-1.88	-2.38
0.54	-1.77	-2.27
0.56	-1.65	-2.15
0.58	-1.53	-2.03
0.60	-1.41	-1.91
0.62	-1.29	-1.79
0.64	-1.16	-1.66
0.66	**-1.04**	**-1.54**
0.68	-0.91	-1.41
0.70	-0.78	-1.28
0.72	-0.64	-1.14
0.74	**-0.50**	**-1.00**

If, for example, a clinically important mean difference of 0.5 between treatments was the target, this would equate to a mean change from baseline of 2.5 and 74% of patients showing a clinical improvement of 1 or more in the active group. These calculations suggested a clinically important mean difference of 0.5, which we can equate to the proportional of individual patients showing individual clinical improvements of 1 point.

The calculation was then adjusted for multiplicity. Each care pathway was planned to be compared to each of the other pathways at the end of the trial. As three formal comparisons were planned, the Bonferroni adjustment was used to adjust the significance level to maintain an overall two-sided 5% significance level. The sample size was calculated for 90% statistical power. See Box C3 for the corresponding sample size explanation presented in the protocol.

BOX C3 Protocol sample size calculation example: OPTION-DM

An individual showing a 1-point change in the numeric rating scale is considered a minimum clinically important difference.[114] Hence, the proportion of people improving by at least 1 point would seem a suitable outcome. However, we have based the sample size calculation on a continuous outcome, the mean change between groups, to maintain power. We have chosen a mean change between groups of 0.5 points based on the mean difference previously reported for a comparison of two active interventions for neuropathic pain in a crossover study.[113] We estimate this would equate to an 8% difference between groups in the proportion of people improving by at least 1 point.[56] Using a within-patient SD of 1.65,[113] an alpha of 0.0167 (0.05/3) to allow for three comparisons, and 90% power, we require 294 evaluable patients.[115]

536 patients in total will be screened for participation in the study. Assuming a 25% drop-out rate, 392 patients will be randomised to ensure 294 patients are expected to complete the study.

If the proportion of patients with an improvement of 1 point or more had been used as the primary outcome (i.e., dichotomising the pain score) and analysed accordingly as a binary outcome, for an effect of 8%, the corresponding sample size would have required a much larger sample size of 884 analysable patients (1179, allowing for drop out).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Case study 4 SUSPEND trial

Ureteric colic describes episodic severe abdominal pain from sustained contraction of ureteric smooth muscle as a kidney stone passes down the ureter into the bladder. It is a common reason for people to seek emergency healthcare. Treatments that increase the likelihood of stone passage would benefit patients with ureteric colic, as they will reduce the need for an interventional procedure.

At the time of planning, two smooth muscle relaxant drugs, tamsulosin (an alpha-adrenoceptor antagonist, or alpha blocker) and nifedipine (a calcium channel blocker), known as medical expulsive therapy (MET), were considered potentially beneficial treatments. The Spontaneous Urinary Stone Passage Enabled by Drugs (SUSPEND) trial was designed to inform the treatment choice. A three-arm RCT was planned to compare tamsulosin and nifedipine to a placebo control to facilitate spontaneous stone passage.

A head-to-head comparison of the two MET agents, nifedipine and tamsulosin, was considered vital. A comparison of the two active arms (combined) to the placebo arm (MET versus placebo) was also planned due to uncertainty about the strength of the existing evidence of clinical efficacy. The key outcome of interest was the presence or absence of a stone at 28 days. It was defined as the lack of any further intervention (or planned intervention) to resolve the index ureteric stone.

A review-of-the-evidence-base approach was used. Data were available from two systematic reviews[116, 117] that included RCTs comparing alpha blockers, calcium channel blockers, and a variety of controls (placebo, treatment as usual, or prescribed painkillers). Only three RCTs compared tamsulosin and nifedipine directly, although there were a number of other trials that compared them to another treatment or a placebo. RCT data from both reviews were combined in a network meta-analysis to maximise the available data to inform the sample size calculation.

The estimated risk ratio (RR) effects are shown in Table C4. For simplicity, the uncertainty around the estimates is not shown. The RRs of being stone-free, comparing nifedipine and tamsulosin to the mixed control group, were estimated to be 1.50 and 1.70, respectively. Of particular note, the RR of being stone free-for tamsulosin compared to nifedipine was estimated to be 1.15.

Table C4 Risk ratios (RRs) of comparison from network meta-analysis of RCTs assessing the use of tamsulosin, nifedipine, and various other treatments

RR of A vs B		**Treatment B**		
		Tamsulosin	Nifedipine	Other
Treatment A	Tamsulosin	1.00	1.15	1.70
	Nifedipine	0.87	1.00	1.50
	Other	0.59	.67	1.00

An estimate of the anticipated control (placebo) group event rate for being stone-free was needed before the sample size could be calculated. This was estimated to be 50% using a random effects estimate of the pooled proportion of the control arms of the RCTs from the two systematic reviews. This was then used as the placebo control group response in the sample size calculation, in lieu of better evidence that might be more relevant to the anticipated population. Using this and applying the corresponding RRs from the network meta-analysis, the stone-free level was anticipated to be 75% and 85% in the nifedipine and tamsulosin groups, respectively.

The study sample size was based on the comparison of the nidedipine and tamsulosin treatments. A standard sample size (for a two-sided 5% significance level and 90% power with a continuity

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

correction) for comparing two proportions gave a required n of 354 in *each* group. This sample size was inflated to 400 per group to account for an approximate 10% loss to follow-up. The total required sample size was 1200 (applying this size to the placebo group as well).

The placebo control group size was kept at 400 for the planned comparison to any MET (nifedipine and tamsulosin combined), which provided greater than 90% power. The size of the placebo group could have been reduced using an uneven allocation ratio, but was instead kept an equivalent size to the two active treatment arms. The funding agency strongly supported the inclusion of a placebo arm, given concerns about the potential risk of bias and the relatively small size of the existing placebo-controlled trials. No adjustment was made to the alpha level for multiple treatment comparisons (and therefore no inflation to the standard sample size), as the different comparisons were considered independent research questions: 1. MET versus placebo control and 2. nifedipine versus tamsulosin. See Box C4 for the corresponding sample size explanation from the protocol.[118]

BOX C4 Protocol sample size calculation example: SUSPEND

Combining the data from the two recent meta-analyses[116, 117] suggests a risk ratio of approximately 1.50 when comparing MET (either an alpha or calcium channel blocker) against "standard care" on the primary outcome. These reviews indicate a spontaneous stone passage proportion of approximately 50% in control groups of included RCTs. Only three of the included RCTs directly compared a calcium channel blocker with an alpha blocker. They suggested that alpha blockers are likely to be superior to calcium channel blockers. Combining information from Singh 2007[117] and Hollingsworth 2006[116] gives anticipated stone passage of approximately 85% in the alpha blocker group and approximately 75% in the calcium channel blocker group.

The most conservative sample size is required to detect superiority between the two active treatments and to this end will power the trial. To detect an increase of 10% in the primary outcome (spontaneous stone passage) from 75% in the nifedipine group to 85% in the tamsulosin group, with a two-sided Type I error rate of 5% and 90% power requires 354 per group. Adjusting for 10% loss to follow-up inflates this sample size to 400 per group. No adjustment for multiplicity has been made.

Recruiting 1200 participants (randomising 400 to each of the three treatment groups: tamsulosin, nifedipine, or placebo) will provide sufficient power (>90%) for the MET versus placebo comparison.

Case study 5 MACRO trial

Chronic rhinosinusitis (CRS) is a common condition affecting around 10% of the population that can lead to chronic respiratory disease or impaired quality of life. Initial management of CRS in the UK is in the family doctor setting, followed by referral to a hospital setting for medical treatment. Initial management fails to deliver sufficient relief for around one in three patients who attends hospital ear, nose, and throat clinics.[119, 120] The role of antibiotics for CRS is unclear, although they are commonly used in clinical practice. Endoscopic sinus surgery (ESS) is a commonly conducted operation. Its use varies from centre to centre due to an insufficient evidence base. Two Cochrane systematic reviews of treatment of CRS with medical and surgical treatments highlighted the need for new randomised trials.[121, 122] Two main research questions related to treatment for patients with CRS were apparent to the investigators:

1) The relative benefits of surgical versus medical treatment; and
2) The role of antibiotics.

Given the lack of clarity about current practice in the UK, two possible trial designs were considered potentially appropriate:

a) Two-stage trial incorporating two linked randomised comparisons:
   Stage 1: Antibiotic versus placebo for 3 months.
   Stage 2: Proceeding to receive ESS or continued medical therapy for those without significant benefit.
b) Three-arm randomised trial comparing antibiotic, placebo, and ESS.

The relative merits of the study designs are not considered here. Instead, the focus is on specifying the target difference. The Management for Adults with Chronic RhinOsinusitis (MACRO) trial was designed to have a sample size sufficient for whichever of the two designs was ultimately chosen.

The primary outcome was the Sinonasal Outcome Test (SNOT-22), a validated disease-specific quality-of-life instrument.[123] An anchor approach was used to estimate the minimally important difference (MID). A “medium” standardised effect size (according to Cohen) was also calculated and used as there is evidence to suggest that 0.5 SD would be a reasonable estimate of the MID for this type of outcome.[124, 125]

Data from an existing study were used to infer what might be realistic to observe.[123] Limited work had assessed what would be a minimally important difference in the SNOT-22 score. Based on a large existing study of around 2000 patients receiving surgery for CRS with/without nasal polyps that used the SNOT-22, an SD of 20 seemed plausible (group change score SDs were in the range of 19 to 20). An analysis adjusting for baseline was planned. A 10-point difference in the SNOT-22 (0.5 SD with SD of 20.0[123]) could be considered an important difference to detect.

The anchor method study suggested that the MID could be slightly smaller at 8.9 points. This estimate was derived by calculating the average difference between those who stated post-treatment that they were “a little better” (9.5 point reduction) and those who stated they were “about the same” (0.6 reduction): 9.5-0.6 = 8.9 points mean difference. In the aforementioned study of surgical patients, the overall mean change score was 16.2 substantially larger the aforementioned MID estimates. It is not realistic to expect all of this to be observed in a comparison of surgery versus another treatment. If 25% (arbitrary, but based on the judgement of the team) of the response were attributed to regression to the mean or the process of receiving treatment of some kind, this would suggest a difference of 13.8 might be plausible for surgery versus an essentially non-effective treatment. A similar value for 15% of the effect was also considered.

The four mean values (8.9, 10.0, 12.2, and 13.8 points) reflected what might be clinically important differences (8.9 to 10) and realistic target differences (12.2) to use in the sample size calculations to look at various potential sample sizes under the two designs. A range of standard sample size calculations for a two-arm trial was produced, looking for 80 or 90% statistical power at the two-sided 5% level, using a pooled SD of 20.0, and assuming around 10% missing data (which was plausible based on two previous studies of patients in this area).[126, 127]

Three-arm design

For 90% power and 8.9 target difference, 107 per group would be required. Applied to the three-arm design and allowing for 10% missing data, this would lead to 120 per group and 360 overall. In the presence of clustering by surgeon in the surgery arm only, this would still be sufficient to achieve just under 80% power (additionally assuming an ICC of 0.05, 10 clusters of cluster size 12, and similar levels of missing data) for the relevant comparisons. No allowance for uneven cluster sizes was made. However, the actual number of clusters that such a trial would use was thought likely to be somewhat higher, offsetting any potential loss due to uneven cluster sizes.

Two-stage design

What could be achieved with this sample size (360) was then considered for the two-stage design. The full sample would be available for stage 1, barring any missing data. In the absence of clustering, this would be more than sufficient to detect the 8.9 point difference (power of 98%). However, the stage 2 comparison drives the overall sample size calculation in a two-stage design, as the sample size must be inflated to deal with a loss of randomised participants after stage 1. This loss was assumed to be 50%, based on limited prior evidence and erring towards more conservative estimates. Assuming too low a loss between stages would have a huge impact on the precision of the stage 2 comparison.

BOX C5 Grant application sample size calculation example: MACRO

The trial will recruit 360 patients from 10 UK centres. Sample size justification is based on achieving at least 80% statistical power at the two-sided 5% significance level. No adjustment for multiple comparisons for a two-stage or three-arm design were made as each of the treatment comparisons are distinct.[25]

Two-stage design

The minimum clinically important difference (MCID) in the SNOT-22 based on an anchor study has been estimated to be 8.9 (SD of 20.0)[123]. However, previous evidence suggests an effect size as large as 13.8 for surgery against alternative treatment is plausible.[123] Assuming a more conservative (smaller) difference of 12.2 as both an important and realistic target difference, and allowing for an ICC of 0.05[111] (10 surgeons) leads to a sample size of 80 per group (90 allowing for just over 10% missing data) for stage 2 (surgery versus on-going medical treatment) for 90% power and thus 180 participants in total. Assuming a 50% non-response rate after the first line of treatment (stage 1) requires doubling the stage 2 number to ensure sufficient participants progress to stage 2, leading to a size of 360 overall. This size will readily allow (>95% power) a difference of 8.9 to be detected at stage 1. The assumed non-response rate was based on our recent feasibility study, where symptomatic improvement in the SNOT-22 scores, greater to or equal to the MCID, was seen in 50% of patients at 3 months[127]. If the two-stage design is used, the non-response rate and corresponding numbers progressing through to stage 2 will be assessed in the pilot phase. If necessary, the stage 1 recruitment target will be adjusted.

Three-arm design

To detect the estimated MCID difference (8.9), 107 participants per group are required for 90% power. Allowing for 10% missing data leads to 120 per group (360 overall). Even in the presence of clustering (ICC) of 0.05 with 10 clusters in the surgical arm, the power for this comparison would be around 80%.

The 10% missing data level assumed above for both designs is consistent with the two previous trials of macrolide antibiotics in CRS and our feasibility study[126, 127].

An overall sample size of 360 would lead to 180 available at stage 2 (90 per group). Using a target difference of 10.0, 360 would achieve 90% power. In the presence of clustering, a large target

 doi:10.20944/preprints201808.0521.v1

difference is needed to have 80 or 90% power. The sample size would be sufficient if a target difference of 12.2 was used, after allowing for clustering in the same manner as in the three-arm design. See Box C5 for the corresponding sample size explanation presented in the grant application.

Note. The final sample size was inflated at the request of the funder to allow the subgroup with and without nasal polpys to be analysed. For simplicity, this is not considered in the text above.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Case study 6 RAPiD trial

Increased use of antibiotics is a major contributor to the spread of antimicrobial resistance. Dentists are responsible for approximately 10% of all antibiotics dispensed in UK community pharmacies. Despite clear clinical guidance, evidence demonstrates that dentists often prescribe antibiotics inappropriately in the absence of clinical need. The effectiveness of strategies to change the behaviour of health professionals is variable, but audit and feedback (A&F) has been shown to lead to small but important improvements in behaviour across a range of contexts and settings. The RAPiD (Reducing Antibiotic Prescribing in Dentistry) trial[128] randomised all dental practices with responsibility for prescribing in Scotland (n=795) using routinely collected Scottish NHS dental prescribing and treatment claim data (available through the PRISMS prescribing information system) to compare the effectiveness of different individualised (to dentist with practices) A&F interventions for the translation into practice of national guidance recommendations on antibiotic prescribing.

795 practices were randomly allocated to an intervention or the control (no A&F). 632 intervention group dental practices were subsequently evenly allocated to one of eight A&F groups in a 2x2x2 factorial design. The three factors were (i) feedback with or without a written behaviour change message; (ii) providing the graph of monthly practice prescribing levels with or without health board prescribing levels in the graph; and (iii) receiving feedback reports twice (0 and 6 months) or three times (0, 6, and 9 months). This lead to a total of eight equal-sized intervention groups of 79 practices. The remaining 163 practices in Scotland formed the no intervention control group. The addition of this independent no intervention control group lead to a "partially" factorial design rather than fully factorial.

The RAPiD trial sample size calculation was unusual as the population of sample units was fixed by the size of the country, i.e., every dental practice with responsibility for prescribing (n=795) in Scotland was expected to take part as part of a national policy to participate in dental service delivery research. The sample size calculation was therefore based on identifying whether adequate statistical power could be achieved for the primary comparisons for target differences that were considered theoretically plausible (realistic) for a fixed size. The cost implications of a larger sample size were nominal and therefore the full population was always going to be used. The analysis was intended to be at the dentist level, adjusted for dental practice. However, the sample size calculation was carried out at the practice-aggregated level and was therefore conservative.

Box C6 Sample size calculation in published trial results paper: RAPiD Trial[116]

The required sample size to achieve 80% power (with two-sided alpha of 2.5% allowing for the multiple comparisons to allow for the two main research questions) to detect a 10% mean difference in overall antibiotic prescribing between intervention groups was 316 per group. This applied to the comparison between A&F only and A&F with an additional written behaviour change message, the comparison between those with and without a health board comparator, and the comparison between A&F at 0, 6, and 9 months versus 0 and 6 months only. Therefore, 632 practices were required to receive an A&F intervention with 79 practices in each of the eight sub-level experimental units. There were 795 practices eligible to be included in the trial, which left 163 practices in the control arm. The comparison between the control group (n = 163) and the intervention group (n = 632) had 80% power to detect a 12% mean decrease in overall antibiotic prescribing. The study was not powered to detect realistic two-way interaction effects between behavioural components.

A systematic review[130] demonstrated that the interquartile range of effects of A&F across different settings was 0.5% to 16%. The study team therefore determined that a 10% reduction (or less) would be both plausible and important. The routine prescribing data indicated that the mean number of antibiotic items prescribed per list was 141.1, with a SD of 140.9. Given that past prescribing behaviour is highly predictive of future prescribing data (correlated) both theoretically and empirically ($\rho = 0.91$, observed for the two most recent pre-intervention years), correction for

 doi:10.20944/preprints201808.0521.v1

the anticipated baseline correlation was used to reduce the precision.[129] A baseline prescribing data adjusted analysis was correspondingly planned.

With the sample size calculation for the A&F comparisons estimated, the study sample size for the comparison of A&F versus no A&F was fixed by the number of dental practices left in Scotland that could be randomised to no A&F intervention. The detectable difference was 12%, which was still considered both plausible (realistic) and important should it be observed. Given that the intervention group was being modelled twice within the two main hypotheses (intervention versus no A&F; and intervention factors versus no intervention factors), Bonferroni's adjustment was used to adjust the significance level to 2.5% to maintain an overall two-sided 5% significance level. See Box C6 for the sample size explanation presented in the trial results paper.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Appendix 1 Conventional approach to a randomised controlled trial sample size calculation

A1.1 Sample size calculations for a randomised controlled trial

Statistical sample size calculation is not an exact, or pure, science.[24, 131] First, investigators typically make assumptions that are a simplification of the anticipated analysis. For example, the impact of controlling for prognostic factors is very difficult to quantify and even though the analysis is intended to be adjusted (e.g., when randomisation has been stratified or minimised),[132] the sample size calculation is often based on an unadjusted analysis. Second, the calculated sample size can be very sensitive to the values of the inputs. In some circumstances a relatively small change in the value of one of the inputs (e.g., the control group event proportion for a binary outcome) can lead to a substantial change in the calculated sample size. However, the value used for one of the inputs (e.g., control group event proportion) may not accurately reflect the actual value that will be observed in the study. It is prudent to undertake sensitivity calculations to assess the potential impact of misspecification of key inputs (e.g., SD for a continuous outcome, level of missing data, etc.). This would also help inform decision making about the continuation of a trial where accumulating data suggests the parameter will be substantially different from the one assumed in the main sample size calculation.

The role of the sample size calculation is to determine how many observations are required in order that the planned main analysis of the primary outcome, that is the one chosen to address the primary estimand of interest, is likely to provide a useful result. The sample size may also be chosen with reference to further key analyses (e.g., those focussing on other outcomes and subpopulations that address alternative estimands of interest). Most simply, this can be done by choosing the RCT's sample size so as to maximise the number of participants required across the various analyses under consideration.

A variety of statistical approaches are available, although overwhelmingly, current practice is to use the conventional Neyman-Pearson approach. This is so much the case that the specification of "effect size", "significance level", and "power" are common parlance. The Neyman-Pearson approach is explain in Section A2, and the rest of this appendix assumes this approach is being used. Alternative approaches to the sample size calculation are briefly considered in Appendix 2 (see Sections A2.2-A2.4 for precision, Bayesian, and value of information approaches, respectively).

Often a simple formula can be used to calculate the required sample size[133]. The formula varies according to the type of outcome and, somewhat implicitly, the design of the trial and the planned analysis. Some of the simpler formulae are given in Sections A1.3-A1.6 for the standard RCT design, i.e., a two-arm parallel-group RCT, and for the most common outcome types (binary, continuous, and time-to-event).

A1.2 Neyman-Pearson approach

The most common approach to the sample size calculation for an RCT is based on what can be described as the Neyman-Pearson, or conventional, approach. In essence, this approach involves adopting a statistical hypothesis testing framework and calculating the sample size required given the specification of two statistical parameters (the power and significance level – see below for definitions). This approach is sometimes referred to as carrying out a "power calculation". This is a frequentist (as opposed to Bayesian) approach to answering the research question (see Appendix 2).

Although it is often not explicitly stated, this approach involves assuming a null hypothesis for which evidence to reject in favour of an alternative hypothesis is assessed. For a superiority trial with a standard design, the null hypothesis is that there is no difference between the interventions and the alternative hypothesis is that there is a difference between them (i.e., one is superior to the other with respect to the outcome of interest). This leads to four possible scenarios once the trial is conducted and the data have been collected and analysed (Table A1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Table A1 Possible scenarios following the statistical analysis of a superiority trial

		Statistical analysis result	
		Statistically significant	*Not-statistical significant*
Truth	*There is a genuine difference between the interventions*	Correctly concluding there is a difference (true positive)†	Wrongly concluding there is a not a difference when there is; Type II error (false negative)ϕ
	There is not a genuine difference between the interventions	Wrongly concluding there is a difference when there is not; Type I error (false positive)‡	Correctly concluding there is no difference (true negative)

† The probability of this occurring (assuming a difference of a particular magnitude exists) is the statistical power.

‡ The probability of this occurring (assuming a difference of a particular magnitude exists) is set by the significance level.

ϕ Often the most that can be concluded from a non-statistically significant result is that there is no statistical evidence of a difference, i.e., a difference cannot be conclusively ruled out.

There are two scenarios where a correct conclusion is made and two where an incorrect conclusion is made. The chance of these two errors is controlled by the statistical parameters, the significance level and the statistical power. Typically the probability of the Type I error (α) is controlled to be 0.05 (or 5%), which is achieved by using this level as the one with which it is concluded that the result is statistically significant (i.e., a probability of ≤ 0.05 is "statistically significant" and > 0.05 is not). Additionally, this is usually a two-sided significance level, in that it is not prescribed a-priori which direction a difference might be found. In a similar manner, we can also control the Type II error rate (β) by ensuring the statistical power (which is simply 1 minus the Type II error rate, i.e., 1- β) is sufficiently large. Typical values are 0.8 or 0.9 (i.e., 80% or 90% statistical power).

It is worth noting that the presence or absence of a statistically significant result cannot be used to decide whether there is an important difference or not. Often the most that can be concluded from a non-statistically significant result is that there is no statistical evidence of a difference, i.e., a difference cannot be conclusively ruled out. Additionally, it is possible to have a statistically significance result even when the observed difference is smaller than the target difference assumed in the sample size calculation.[134, 135] This value can be readily calculated for a continuous outcome. Here this is described as the minimal statistically detectable difference (MSDD). It should not be confused with the minimal clinically detectable change/difference (MCDC/MCDD), which is an entirely different concept (see the Glossary for brief descriptions). Some recommend calculating and reporting the MSDD as well as the target difference and the required sample size.[135]

Both the use of the 5% significance level and 80% or 90% power are arbitrary and have no theoretical justification, but are widely used. However, as excluding the possibility of either error is impossible, and the required sample size increases at a greater rate the closer either error rate is set to zero, these values have become the *de facto* standards. If well chosen, the target difference is a valuable aid to the interpretation of the analysis result, irrespective of whether it is statistically significant or not.

Given the assumed research hypothesis, the design, the statistical parameters, and the target difference, the sample size can be calculated. Formulae vary according to the type of outcome (see Appendix 1, Sections A1.3, A1.4, and A1.6), study design (see Appendix 3 for some common alternative designs), and the planned statistical analysis (see Appendix 1, Section A1.7). The general approach is similar across study designs. In more complex situations, the frequentist properties (e.g., the Type I and Type II error rates) can be estimated using simulations of data and consequential analysis of simulated results for scenarios where there is and is not a genuine difference between interventions.[136]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

This appendix presumes the conventional approach is to be used for the sample size calculation for a 2-arm trial with 1:1 allocation. Immediately below, simple formulae for the most common outcome types are provided. For completeness, Appendix 2 briefly summarises alternative approaches to calculating the sample size for an RCT. Statistical issues related to conducing a re-assessment of the sample size under a conventional and a Bayesian approach are considered elsewhere.[1, 137-139] Adaptive trial design (see Appendix 3) seek to formally incorporate potential changes to the design due to interim data into the trial design.

A1.3 Binary outcome sample size calculation for a superiority trial

There are a number of commonly used formulae for calculating the sample size for a binary outcome for a superiority trial, i.e., for a study in which two proportions are to be compared.[115] One formula for the required number of participants per arm, n, for a standard trial (assumed equal allocation and therefore group sizes) is presented below and is relatively straightforward to calculate:

$$n = \frac{\left(Z_{1-\beta}+Z_{1-\frac{\alpha}{2}}\right)^2 \left(\pi_A(1-\pi_A)+\pi_B(1-\pi_B)\right)}{(\pi_B-\pi_A)^2},$$

where n is the required number of observations in each of the two randomised groups. Z_{1-x} is the value from the standardised normal distribution for which the probability of exceeding it is x. π_A and π_B are the anticipated probability of an event in groups A and B. α is the statistical significance level (i.e., the Type 1 error rate), and β is the Type 2 error rate and is chosen so that $1-\beta$ is equal to the desired statistical power. The formula assumes even allocation between the treatment arms and a two-sided comparison.

The target difference can be expressed in multiple ways. It can be expressed as the absolute risk difference $(\pi_B - \pi_A)$ or as a ratio, typically the risk ratio $\left({}^{\pi_B}/_{\pi_A}\right)$ or odds ratio $\left(\frac{{}^{\pi_B}/_{(1-\pi_B)}}{{}^{\pi_A}/_{(1-\pi_A)}}\right)$. Different combinations of π_A and π_B can lead to the same odds or risk ratio, although they may produce very different absolute risk differences. For example, a proportion of 0.4 compared to one of 0.2 represents a risk ratio of 2 and a risk difference of 0.2. Proportions of 0.1 and 0.05 also represent a risk ratio of 2, but the risk difference of 0.05 is far smaller and will require a far larger sample size. Whenever the target difference is expressed as a ratio, the anticipated control (reference) group risk, π_A, should also be provided.

The value assumed for π_A greatly influences the sample size.[1] In this context the control group proportion can be considered as a nuisance parameter with the target difference, δ, fixed regardless of what the control group proportion is. Estimates of this parameter may come from a pilot trial or existing literature (see Sections 3.2.6 and 3.2.7). There needs to be an evaluation of the observed response dependent on the study design, population, and analysis in the study from which it is being estimated from. The planned analysis, particularly the summary measure used, is important for the calculation as adjusted and unadjusted analyses can be estimating different estimands.[140]

A1.4 Continuous outcome sample size calculation for a superiority trial

For ease of presentation, a slightly simplified formula[115] to estimate the sample size per arm for a superiority trial with a continuous outcome is:

$$n = \frac{2(Z_{1-\beta}+Z_{1-\alpha/2})^2\sigma^2}{\delta^2} + \frac{Z_{1-\alpha/2}^2}{4},$$

where $Z_{1-\beta}$ and $Z_{1-\alpha/2}$ are defined as before, σ is the population SD, and δ is the target mean difference. As before, the formula presented here assumes even allocation between the treatment arms and a two-sided test comparison.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

In practice, σ is typically assumed to be known, with an estimate from an existing study, S, used as if it were the population value. The formula can be further simplified by replacing δ by δ/σ, the Cohen's d standardised effect (d_{SES}):

$$n = \frac{2(Z_{1-\beta} + Z_{1-\alpha/2})^2}{d_{SES}^2} + \frac{Z_{1-\alpha/2}^2}{4}.$$

Specifying the effect on the standardised scale, d_{SES}, is therefore sufficient to calculate the required n for a given significance level and power. However, it should be noted that different combinations of mean and SD values produce the same standardised effect size (Cohen's d). See Section 3.2.7 for further discussion. While sufficient for the sample size calculation, specifying the target difference as a standardised effect alone can be viewed as an insufficient specification as it does not define the target difference in the original scale.

A key component in the sample size calculation of a continuous measure is the assumed magnitude of variance. An estimate of this parameter (usually expressed as a SD) may come from a pilot trial or existing literature (see Sections 3.2.5 and 3.2.6). It is possible to get into a "Gordian knot" when looking for an estimate of the variance. Ideally, an estimate of the variance taken from a large clinical study in the intended trial population with the same interventions would be available. However, if such a study was available, a new trial would likely not be necessary. If a new trial is truly needed, that need implies some limitations in the existing evidence. To decide on the relative utility of the variance estimates, various aspects of the study need to be considered (e.g., study design, population, outcome, analysis conducted, etc.) in a similar manner to the control group proportion and any estimate of a realistic target difference (see Sections 3.1, 3.2.5, and 3.2.6).[1, 115] The accuracy of the variance estimate will obviously influence the sensitivity of the trial to the assumptions made about the variance and will also influence the strategy of an individual clinical trial.

A more accurate, although computationally more demanding calculation if performed by hand, will give a slightly different result from the formula above and is used in various sample size software.[141] The difference between the simple and more complicated formulae is that the simple calculation assumes the population variance, σ^2, is known for the design and analysis of the trial. The more complicated calculation recognises that, in practice, the sample variance estimate, s^2, will be used when analysing the trial. The more accurate formula can be found elsewhere.[1]

A1.5 Dealing with missing data for binary and continuous outcomes

In most studies involving humans, it is likely that withdrawals, losses to follow-up, and missing data will occur during the trial.[142] Individuals in a trial could decide they no longer want to take part and completely withdraw from the trial, they could move during the study and not update the study team, and/or they could decide they do not want to answer a particular question on a questionnaire. Even in the most well designed and executed trial, some losses to follow-up are inevitable. Additionally, there are intercurrent events (e.g., death or change in treatment) that may preclude the possibility of an outcome under the conditions implied by the trial's aim and corresponding estimand of interest.

Irrespective of the reasons for missing data, sample sizes are frequently inflated to account for a degree of missing data during the study. The estimate of the extent of missing data is often gathered from a pilot trial, previous studies of the intervention, or trials in a similar population. In the presence of missing data, the power of a trial to detect the same target difference is reduced, hence the need for inflation of the sample size. To inflate the sample size to account for missing data, the overall sample size required, $2n$, is divided by the proportion of data anticipated to be available for analysis (p_{ob}):

$$2n/p_{ob}$$

For example, if 20% attrition is anticipated, then the target sample size is divided by 0.8. A more complex and accurate approach can be used to deal with loss-to-follow-up over time, which is particularly pertinent for time-to-event outcomes.

A1.6 Time-to-event sample size calculation for a superiority trial

Due to varying time of follow-up across study participants, it is not appropriate to analyse the proportion of participants who experience an event using logistic regression or a similar method. The analysis, and therefore the calculation of the sample size, for time-to-event data is also complicated by the fact that not all individuals will have the event of interest. As a consequence, it is not appropriate to simply compare mean observation times directly between groups. There are three main approaches to the sample size calculation for this type of outcome:

1. Compare Kaplan Meier survival curves, using the logrank test or one of several other similar methods;
2. Assume a particular model form without specifying the survival distribution (e.g., the Cox (proportion hazards) regression approach); or
3. Use a mathematical model for the survival times and hence for the survival curve, such as the exponential or the Weibull distributions.

For ease of discussion, the term "survival" is used to refer to the non-occurrence of the event by a specific time point and does not imply restriction of the methods to looking at mortality. The first two sample size methods are much more common than the third. For either a logrank or Cox regression based analysis, the analysis does not imply a specific distribution for the survival curve. The proportion surviving at any time-point during the follow-up can be estimated to avoid having to assume one for the purpose of the sample size calculation. A target difference is inferred, explicitly or implicitly, for all of the methods. It is commonly expressed as a hazard ratio.[143] Similarly to a binary outcome, adjusted and unadjusted analyses can estimate different estimands.[144]

The difference between the two groups can be expressed as a difference between the survival probabilities at a specified time-point. The data can be analysed accordingly, using the Greenwood standard errors to compare survival proportions.[145] However, this is statistically not a good way to compare groups, as it depends on the chosen time-point and does not use the data on survival beyond that point. A method that takes all of the observed survival times into account, such as the logrank test, is more convenient and statistically efficient. This is a test of significance that has no explicit associated estimate of the treatment effect. Despite this, a power calculation can be performed by characterising the two survival curves by their median survival time, the time when half of the population is estimated to have experienced an event in the respective group.

To infer information about the survival curve from the median survival time, it must be assumed that the survival curve follows a known mathematical pattern, even though this assumption may not be used in the analysis. For example, the survival curve can be (and commonly is) assumed to be an exponential decay curve. The survival proportion (π_A) for treatment A at some time t can then be used to estimate the median survival time m as follows:

$$m = t\left(\frac{\log_e(1/2)}{\log_e(\pi_A)}\right).$$

Instead of a difference between mean times and the SD of times that would have occurred if we were comparing the average survival time where all participants had reached the event, there are two median survival times or, equivalently, the median survival time in one group and the difference between medians, which can be considered the target difference. This is an implicit treatment effect size, although no such estimate is produced by the logrank test.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Alternatively, an assumption about the difference between the survival curves, the proportional hazards assumption, can be made. This is the assumption that the ratio of the risk of an event in one group over a given short time interval, to the risk of an event in the other group over the same time interval is a constant over the follow-up period. This ratio is the hazard ratio and is the parameter that we estimate in Cox proportional hazards regression. This hazard ratio can be considered to represent the target difference (albeit on a relative range). However, another parameter is still needed to characterise the survival curve, such as the median survival time in one group.

It is possible to characterise the target difference as either the difference between median survival times or the hazard ratio, or by comparing events as an absolute difference in the event rate at a specific time-point. Whichever approach is taken, the median survival in the control group or some similar parameter is needed to fully and uniquely specify the target difference. The statistical power of the comparison will depend on the total number of events rather than the total number of participants. A high number of events will imply high power. Participants who do not experience an event contribute little to the power. The median survival time and the planned follow-up time enable the number of events that will occur to be estimated.

Things become more complex if participants are recruited over a time period and then all followed up to the same calendar date. This results in widely varying follow-up times for censored cases. To allow for this, the recruitment period needs to be accounted for in the sample size calculation. If each participant will be followed for the same length of time, such as one year, the calculation is as if all were recruited simultaneously.

Methods for estimating the sample size usually rely on the number of events that need to be observed. The additional assumption of an exponential survival curve is typically made. Under these circumstances, the hazard, the instantaneous risk of an event, is a constant over time. The proportional hazards assumption is thus automatically satisfied. The hazard ratio (HR) can then calculated as:

$$\log_e(\pi_B) \Big/ \log_e(\pi_A) = \mathrm{HR} = m_A/m_B.$$

Again under the assumption of an exponential survival distribution for both interventions, we can estimate the required number of events, e_1, in one group by:

$$e_A = \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\log_e(\mathrm{HR})\right)^2}.$$

This can be doubled for the other group, or the HR can be used to calculate the number of events in the other group. Having calculated the number of events needed, the number of participants required to produce this number of events can be calculated. To do so requires making further assumptions regarding the survival distributions for each group, the length of follow-up, and any censoring of data. A varying length of follow-up according to the accrual pattern is also typically assumed to make maximum value of those recruited early in the recruitment period and avoid unnecessarily extending the follow-up of the final participants. This issue is beyond the scope of this guidance, but further discussion can be found here.[1] Sample size calculations which allow for non-proportional hazards are also possible.[146]

The target difference for this type of outcome can be variously expressed as a difference between the median survival times ($m_B - m_A$), the difference between the proportions surviving at a particular point in time ($\pi_B - \pi_A$), or the HR (which might vary over time). It is worth noting that, as for the other outcome types, how intercurrent events (e.g., change in treatment) are dealt with needs careful consideration, and similarly the reasons for censoring need assessing as not all may be viewed equally. For example, if death is not the event of interest, then the occurrence of a death leads to censoring of the outcome of interest. However, it may be viewed as indicative of the

likelihood of such an event occurring or as precluding the event from occurring with no impact on the likelihood (e.g., the death of someone who has had a knee joint replacement precludes the failure of the device due to wear and tear). The handling of such occurrences in the analysis and the corresponding impact on the sample size (and, in this context, the anticipated target difference and event rate) should be considered with reference to the estimand of interest.

A1.7 Other topics of interest

Adjusting the sample size calculation of a continuous outcome for baseline correction

For a continuous outcome measure, full specification of the target difference requires both the mean difference and the corresponding SD to be stated. If a baseline measure of the same continuous measure is also collected, then it is possible to adjust the comparison of means for the baseline value (at the individual participant level) and thereby to incorporate the correlation between the baseline and follow-up measure into the sample size calculation. A simple formula has been proposed to account for this correlation and is sometimes called a design effect or variance (deflation) factor.[129] The sample size, accounting for the correlation between the baseline and follow-up values, for a comparison of two means for a specified target difference, is:

$$(1 - q^2)n + 2,$$

where n is as before (the number needed per group from a calculation without adjustment for baseline), and q is the correlation between the baseline and subsequent outcome measure.

For example, consider a parallel two-arm superiority trial with a primary outcome measure of the SF-36 mental component score at 6 months, which is also taken at baseline. A target sample size of 266 subjects is calculated, assuming 90% power, 5% alpha, an 8-point difference, and an SD of 20. Previous studies have shown that the correlation between repeated measures of the SF-36 can be as high as 0.8 and as low as 0.6, which would lead to quite different numbers of required participants. If a more conservative choice is made, and a correlation of 0.6 is assumed, the required sample size is then:

$$\left(266 \times (1 - 0.66^2)\right) + 2 = 152.$$

If baseline/change score corrections are made, then it is vital to make a credible assumption regarding the correlation. Similar to specifying the target difference, the anticipated correlation value may be estimated from previous trials or observational data. The key advantage of incorporating the correlation between these two measures into the sample size calculation is that fewer subjects are required for the RCT. The key disadvantages are that another assumption is being made, and if the observed correlation is much lower than anticipated, then the trial will be underpowered to detect the target difference specified.

Compliance-adjusted sample size

An ITT-based analysis is the widely accepted default analysis for RCTs. It estimates the average treatment effect in the full randomised cohort irrespective of compliance with the treatment allocation.[16] Such a focus can alternatively be expressed as the desire to assess the “effectiveness” of the treatment as opposed to the “efficacy”.[14, 15] More specifically, it may be said to imply a treatment-policy-based estimand.

The impact of compliance on the anticipated average treatment effect can be taken into account by down-weighting the anticipated causal effect to allow for departures from randomised treatment (also referred to as non-compliance or non-adherence).[147, 148] A natural additional aspect of interest is the treatment effect among those who “comply” (receive the treatment as allocated); this is often described as the complier average causal effect (CACE). This can be viewed as leading to an “efficacy” focus analysis[147] and, more specifically, a principal stratum-based estimand.[15, 149]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

The most simplistic compliance scenario is a standard trial design with all-or-nothing compliance, i.e., each participant either does or does not comply, for which the impact of compliance can be relatively straightforwardly accounted for. For this setting, the relationship between the target difference for a full trial population (irrespective of compliance) and among compliers only can be readily expressed.

For a binary outcome, a corresponding approach for an RR is:

$$RR_{ITT} = \frac{RR_{CACE}(1 - p_{cA}) + p_{cA}}{(1 - p_{cB}) + RR_{CACE}p_{cB}},$$

where p_{cA} and p_{cB} are the proportion of non-compliance among the randomised intervention groups A and B, and RR_{CACE} and RR_{ITT} are the RR ratio among compliers and the ITT population, respectively.

For a continuous outcome, the impact of non-compliance in the intervention arm can be taken into account by multiplying the treatment effect between the groups for compliers (δ_{CACE}) by the level of non-compliance in the intervention (p_c), to get an overall (δ_{ITT}) treatment effect:

$$\delta_{CACE} = {}^{\delta_{ITT}}/_{p_c}.$$

An alternative formula is needed if non-compliance can occur in both arms.[148, 150] Trial sample size calculations are typically for an analysis that will estimate an ITT-based effect (or a treatment policy estimand). Compliance is not often explicitly considered in RCT sample size calculations. From one perspective, if the chosen target difference is one that is considered to be important to stakeholders for the population of interest, then compliance does not need to be part of the formal calculation. Instead, the presence of non-compliance is merely one of a number of reasons that may explain why this target difference might not be observed or that may lead to missing data. Alternatively, if the treatment effect in a compliant population can be specified (e.g., from a previous study), compliance could be taken into account as shown above to show that an effect that is realistic in an ITT population (or treatment policy estimand) is still detectable (and still of a magnitude that would be considered important).

It is worth noting that harms are typically analysed according to the treatment-received groups and therefore the above calculations are not appropriate. More complex analysis approaches for exploring compliance are possible. Compliance analyses often suffer from lack of precision (particularly CACE analyses) and this is an active area of research.[147, 151]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Appendix 2 Alternative approaches to the sample size calculation for a randomised controlled trial

A2.1 Introduction

Three main alternative (Precision, Bayesian and Value of information) approaches to sample size calculations are briefly considered in turn below. Other approaches exist though at present they are rarely used.[152, 153]

A2.2 Precision

The limitations of the conventional approach to the sample size calculation of an RCT are well known.[24, 154] One alternative is to base the sample size on the precision of the estimate of the interest, the treatment difference. This can be expressed through the CI, and the sample size can be chosen to achieve a CI of a particular interval width (i.e., difference between the upper and lower limits). The width of the 95% CI for a standard trial design will be:

i. for a binary outcome:

$$n = \frac{8 \times 1.96^2 \times p(1-p)}{w^2},$$

where w is the width of the CI, which could be chosen to exclude the magnitude of difference desired to be detected, and n is the number in each group. This formula makes use of the large sample binomial approximation. More complex calculations for the CI can be used instead, although with limited additional value for many situations.

ii. for a continuous measure:

$$n = \frac{8 \times 1.96^2 \times S^2}{w^2},$$

where w is the width of the CI for the mean difference, which could be chosen to exclude an important difference, S is the population SD (assumed to be known), and n is the number in each group.

For both of the above formulae, two-sided CIs with a different confidence level can be calculated by substituting the corresponding $Z_{1-\alpha/2}$ in place of 1.96. For example a 90% CI would use 1.645 instead of 1.96. These calculations implicitly do not take into account statistical power and will lead to a smaller sample size, given equivalent assumptions. This type of approach is increasingly used in the context of pilot trials (e.g., for ensuring the width of the CI for the group proportion or consent rate is sufficiently narrow).[155-158] However, use in the context of a definitive trial is limited to date.[5, 154, 159] The issue of the magnitude of a difference that is valuable to be observed is still present.

A2.2 Bayesian

The Bayesian concept of assurance,[4] also referred to as “average power”[160]or a “hybrid” Bayesian-frequentist method,[3, 161] can be used to inform the sample size calculation for a trial that is to be analysed within a conventional (Neyman-Pearson) framework. In this context, assurance is the unconditional probability that a trial will yield a statistically significant result, calculated by averaging the statistical power across a joint prior distribution for the treatment difference and any unknown relevant nuisance parameters (such as the response variance or control response rate for continuous or binary outcomes, respectively). High assurance implies the trial is adequately powered to detect a continuum of possible effects. This increases the robustness of the design but typically leads to larger than conventional sample sizes.[161]

When performing assurance calculations, a prior distribution for the treatment difference can be determined from expert opinion (see Section 3.2.4) or a synthesis of existing data (see Sections 3.2.5 and 3.2.6). Adjustment for between-trial heterogeneity[162, 163] and the bias inherent in existing

effect estimates can be made.[164] The latter arises because Phase II trials may be more at risk of internal biases and confirmatory trials are commissioned only after observing promising early phase results. A careful choice of prior (possibly truncated to support only alternative values of the treatment effect) is needed to ensure sample sizes are not unreasonably large and that assurance approaches complete certainty as the sample size becomes infinitely large.[161] A related approach avoiding this last subtlety is "conditional expected power", defined as the average power calculated, assuming an advantage for the novel intervention must exist.[165] In this setting, one can set the target difference (δ) to the value δ*, at which the trial, with sample size calculated to have high frequentist power to detect δ*, also has high conditional expected power.[166]

A wide variety of alternative Bayesian methods for sample size calculations also exist. The average power of trials with Bayesian final decision rules can be calculated.[167] Alternatively, the sample size of a Bayesian trial can be chosen to ensure there is a high prior predictive probability of the trial concluding with definitive levels of evidence either supporting adoption or abandonment of the novel intervention, thus reducing the region of indecision.[168] However, precision-based approaches calibrate a trial's sample size on the basis of the expected length of a 100(1 – α)% posterior credibility interval or the expected coverage of an interval of fixed width.[169] Judgements about what constitutes an acceptable length or coverage level will depend on how the trial results will feed into subsequent decision-making.

A2.3 Value of information approach

Bayesian decision-theoretic designs exist that choose the sample size to maximise the expected utility of the trial.[170, 171] This is implemented in health technology assessments as a value of information analysis. An efficient sample size is determined by comparing the (expected) cost of conducting a study of sample size n with the expected value of the information that the study will yield.[172, 173] As such, it offers a radically different approach to determining the sample size for an RCT from the conventional (Neyman-Pearson) power calculation approach. A key element of the decision-theoretic approach is the focus on expected values rather than hypothesis testing for making decisions.[173]

The cost of collecting information is simply the budget for a proposed clinical trial of sample size n. The information the trial yields is valued in terms of its ability to reduce uncertainty; all else being equal, larger trials will yield more information than smaller ones. The value of the information is the "expected reduction in the expected loss" from that study.

The logic is as follows: A decision must be made whether to adopt or reject a new treatment. As the decision is made under conditions of uncertainty, the "wrong" decision could be made. The expected loss associated with the decision is the probability of making the wrong decision multiplied by the loss (foregone health gain) if the wrong decision is made. More research (i.e., information in the form of a clinical trial or other data-gathering exercise) reduces the probability of error and hence reduces the expected loss. This expected reduction in expected loss is the expected value of sample information (EVSI). EVSI can be measured in terms of health gain (e.g., life years or quality adjusted life years) foregone, or it can be expressed in monetary terms. For example, NICE in England and Wales values a quality adjusted life year at between £20,000 and £30,000.[66] The expected net gain of sampling (ENGS) is the difference between the EVSI and the anticipated cost of the study. The most efficient sample size for the study is that which maximises the ENGS. Use of a value of information approach is an active area of research and various modifications to the basic approach have been proposed.[172, 174, 175]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Appendix 3 Specifying the target difference for alternative trial designs

A3.1 Introduction

Five types of alternative trial (multi-arm, cluster, cross-over, biomarker and adaptive) designs are considered in turn below in terms of their implications for specifying the target difference. A huge number of variations in trial designs (e.g. split-plot[176] and stepped wedge designs[177]) exist though in terms of the implications for specifying the target difference, the relevant issues are typically similar to those addressed below.

A3.2 Multi-arm

There are many different designs and aims of multi-arm trials, but the one thing they have in common is that they all include more than two trial arms. For example, this could involve comparing multiple treatments against a common active control or comparing two or more treatments against a placebo. Specifying the target difference in multi-arm trials is more complicated than a parallel two-arm trial as multi-arm trials aim to answer multiple research questions. The sample size has to be sufficient to address each research question and therefore multiple target differences could be appropriate.

The more trial arms there are, the more complicated the process becomes. In a trial with three intervention arms comparing treatments A, B, and C, there are seven theoretically possible comparisons that could be made: A versus B; B versus C; A versus C; AB versus C; AC versus B; BC versus A; and A versus B versus C (using a global test). A key aspect of the design of multi-arm trials is specifying what the estimand of interest is, which comparisons are of most interest and which hypotheses will be tested. The selection of comparisons may become simpler if one of the arms (say C) is a control arm, such as usual care or placebo. In this instance, what would be of primary interest would be treatment A compared to the control C, and treatment B compared to C. The simplistic approach to sample size calculation would be to considers each of these pairwise comparisons as if they were separate trials. The target difference for each would require specification and justification in the same manner as a standard trial, even though these might well be the same for both. It might also be of interest to compare A to B directly, although specifying that difference might depend on whether treatments A and B are different types of interventions, minor variations of the same intervention (e.g., doses of the same drug), or an experimental treatment and an active comparator, etc.

BOX 7 Example of one important key hypothesis - Cervical collar or physiotherapy versus wait and see policy for recent onset cervical radiculopathy trial [178]

We calculated the sample size for this three-arm trial on the basis of the comparison treatment (cervical collar or physiotherapy) versus a wait and see policy, with equal allocation in the treatment arms and three repeated measurements (at entry and at 3 and 6 weeks' follow-up), with an estimated correlation coefficient of the measurements of ρ=0.7 and a difference in the mean value of the visual analogue scale for arm pain of 10 mm, as a clinical relevant difference with an estimated SD in each treatment group of 30 mm. As arm pain is the main complaint in cervical radiculopathy, we chose this outcome for calculating the sample size. The total sample size needed to detect this difference at a 5% level of significance with a power of 90% was 240 (80 per group) patients.

Multiple arms allow more than one hypothesis to be explored; however, if it is appropriate to specify the most important hypothesis under study and use this to drive the sample size. In the example described above, it could be that treatment arm A compared to the placebo is of most importance and the comparisons of the treatment arms can thus be done in a hierarchy.[179] This would have consequent effects on specifying the target difference, as it would be the same as a parallel two-arm trial as previously outlined, with the comparison of A to the placebo primarily determining the sample size. If the aim is to look at the specified comparisons simultaneously, then there will be multiple target differences that could be used, depending on the comparison being made. In the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

example above, it would mean both treatment arm A and treatment arm B must be statistically different from the placebo for the study to be declared a success. Such studies are termed multiple must-win trials.[180]

When there are multiple target differences, the smallest target difference will be the one that has the biggest influence on the sample size calculation.[180] Where appropriate, the use of global comparison tests, pairwise comparisons, and/or statistical adjustments can be used to account for multiple comparisons.[181]

A3.3 Cluster randomised

Cluster RCTs involve randomising groups or clusters of individuals to trial arms rather than the individuals themselves. If cluster randomisation is used, then this needs to be accounted for within the design and analysis of the study, including the sample size calculation.[182] Individuals within each cluster will be more alike than they are like individuals within other clusters and cannot be considered independent of each other. The ICC is a measure of this similarity, and for a particular outcome represents the amount of variance that can be explained by the variation between clusters. Sample size calculations for cluster trials have been developed and involve inflating the sample size for an equivalent individually randomised trial by a design effect (also called a variance inflation factor)[183]:

$$1 + (m - 1)\rho \,,$$

where m is the average cluster size and ρ is the ICC. This formula can be used for binary and continuous outcomes.

For example, in an individually randomised trial of an exercise intervention for low back pain, the target difference was 1.57 points with a SD of 4. Assuming 90% statistical power and 5% two-sided significance level, a target sample size would be 274. If this trial were undertaken as a cluster RCT, then the target sample size, assuming an average cluster size of 20 and ICC=0.03, would be:

$$274 \times (1 + [(20 - 1) \times 0.03]) = 274 \times 1.57 = 432.$$

With a sample size of 432 and 20 individuals per cluster, this would require 22 clusters to be randomised (i.e., 440 individuals in total). This increases the sample size substantially compared with individual randomisation, for which a target sample size of 274 would be required.

The ICC, as a ratio, is a difficult quantity to estimate precisely.[184] Pilot trials and most clinical studies are too small to achieve this. Instead, databases of estimates from other data sources, which include similar RCTs, exist, so that the same or at least a similar outcome can be used to provide or inform the choice of a more reliable value. Existing databases of ICC values cover implementation science, organisational interventions, and surgical interventions.[111, 184, 185]

The calculations above for the design effect do not take into account variation in cluster sizes and assume that the same, or approximately the same, number of individuals per cluster are recruited.[186] If there is variation in cluster sizes, then the formula above will underestimate the adjustment required. An additional factor that needs to be considered in trials that anticipate variation in cluster sizes is the coefficient of variation (CV). The CV is the ratio of the SD of cluster sizes to the mean cluster size. The design effect is expressed as:

$$1 + \{(CV^2 + 1)m - 1\}\rho \,,$$

where m is the average cluster size and ρ is the ICC.

The maximum increase in sample size when accounting for variation in cluster size is $CV^2 + 1$. The choice of CV is important to ensure the appropriate inflation factor is estimated. A number of different methods for different scenarios have been proposed to estimate this coefficient[187]: knowledge of CVs observed in previous studies; investigating and modelling sources of cluster size variation; estimating likely minimum and maximum cluster sizes; when all individuals in each

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

recruited cluster participate in a trial; when cluster sizes are identical; and when cluster size follows a roughly normal distribution. It has been recommended that the potential impact of variation in cluster sizes should be explored when planning sample size calculations for cluster trials. In particular, it becomes important to assess this when large variations in cluster size, large ICC values, or large mean cluster sizes are anticipated.

It is not always necessary to incorporate the CV into the cluster trial sample size. If the CV is estimated to be <0.23, then the sample size does not need to be adjusted for the variation in cluster size as the impact on the sample size is negligible.[186]

Clustering can also potentially arise in individually randomised trials, when interventions are delivered within a group setting or when individual therapists deliver the intervention on an individual basis to a group of individuals.[188] The same methods outlined above apply in these situations but, depending on the type of interventions and nature of the clustering, this may only be required in one of the trial arms.

A3.4 Crossover trial

Crossover (randomised) trials involve randomising individuals to a *sequence* of interventions rather than a *single* intervention.[189] In the simplest form, they involve two treatments and two periods (a 2x2 crossover trial, also known as an AB/BA trial). More complex designs are possible with three or more interventions and/or periods.[1, 189] Here consideration is restricted to the 2x2 crossover design, and specifically the most common implementation where AB and BA sequences are used equally. For a binary outcome, the sample size can be calculated in terms of the conditional odds ratio (anticipating an analysis using McNemar's test) and approximated as:

$$OR_c \approx \frac{\pi_B(1-\pi_A)}{\pi_A(1-\pi_B)},$$

where p_A and p_B are the probability of an event under treatment A and B, respectively. The number of participants (with outcomes for both treatments) is:

$$n = \frac{(Z_{1-\alpha/2}(OR_c+1) + 2Z_{1-\beta}\sqrt{OR_c})^2}{(OR_c-1)^2},$$

with $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ defined as before.

It should be noted that the standardised effects (here expressed as odds ratios) for a parallel-group and crossover trial are not equivalent. In terms of expressing the target difference, the absolute difference along with the control group proportion is the most transparent, i.e., $(\pi_B - \pi_A)$ and π_A.

For a continuous outcome, a similar formula to the parallel-group superiority trial can be used:

$$n = \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2\sigma^2}{\delta^2} + \frac{Z_{1-\alpha/2}^2}{2},$$

with δ, $Z_{1-\alpha/2,}$ and $Z_{1-\beta}$ defined as before. However, unlike before, here σ refers to the anticipated within-person variance, and not the pooled variance of two independent treatment groups. Here, n refers to the overall sample size, as well as the number receiving each treatment. The crossover design enables the individual variance to be stripped out, leading to a more precise estimate and smaller sample size. In terms of target difference on the absolute level, it remains as before. As for the binary outcome situation above, it is noteworthy that expressing the target difference as a standardised effect size, when calculated simply by using the inputs to the sample size calculation for a crossover, is not directly comparable to a parallel-group trial even though the absolute target difference is the same. They will only be equivalent in the improbable event that the within-person correlation is zero. For example, given a target absolute difference of 10 points, anticipated intervention group SD of 30, and within-subject variance of 19, we would obtain markedly different

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

standardised effect sizes of 0.33 and 0.63, respectively. The pooled individual group SD is preferred when expressing an effect size as a standardised mean difference, even where a crossover trial is planned.

It should be noted that the impact of missing data on precision is more marked for a crossover trial than for a similarly sized parallel-group trial.

A3.5 Biomarkers

The sample size considerations required for biomarker-stratified trials do not differ substantially from those required for non-stratified trials, but there are some additional considerations that are important. The common components of a sample size calculation still apply, e.g., for a binary outcome, the significance level, statistical power, event proportion in the control arm, and the target difference being sought (δ). However, these components tend to be considered separately within each of the proposed biomarker stratum. Other considerations are briefly covered below.

Type and prevalence of the biomarker

There are various types of biomarkers but in the main, they can be classified into two predominant types: prognostic or treatment-selection biomarkers. Prognostic biomarkers stratify patients on the basis of the prognosis of the disease in the absence of treatment, and thus, they relate to the natural history of the disease. Treatment-selection biomarkers (also known as predictive biomarkers) stratify patients on the basis of their expected response (or not) to a particular treatment. Some biomarkers demonstrate both prognostic and predictive qualities. When designing a biomarker-stratified trial and performing sample size calculations, it is important to be aware of any existing data that describe the discriminatory performance of the biomarker in question, whether it be prognostic, predictive, or both. In particular, if a biomarker is prognostic, then the event proportion in the control arm will differ between strata, which may influence the sample size needed for each group.

The prevalence of the biomarker in question will affect the availability of patients for a particular biomarker stratum and, if rare, this could limit not only the recruitment rate but also the power with which an intervention can be tested in that group. Trials that use an enrichment strategy (where the new intervention is only tested in the biomarker-positive group first) can be an efficient way of testing for benefit. This approach can be used in trials testing targeted drug therapies that have been designed to act on a specific molecular pathway, such that if insufficient benefit is seen in the biomarker group with that molecular aberration, there is very little likelihood of a benefit being seen in the group with a normal molecular pathway. However, the trial can be expanded to test biomarker specificity by including the biomarker-negative group later if adequate activity is seen in the biomarker-positive group.

Testing for interaction

In some cases, it may be advisable to power the trial on the basis of detecting a statistically significant interaction between biomarker strata. In this case, the target difference is the difference in treatment effects rather than the overall treatment difference in outcome between randomised groups. However, attaining adequate statistical power for a test of interaction can lead to a potentially unfeasible sample size.

Although the presence of a statistically significant interaction may be compelling, it does not follow that a stratified medicine approach will be the recommendation from a stratified medicine trial. For example, it is possible that both biomarker stratified groups could derive benefit from the new intervention, and even though one group could derive statistically more benefit than the other, the intervention would still be recommended for all patients rather than taking a stratified approach (see scenario A in Table A3.1). However, under scenarios B and C in Table A3.1, it would be advisable to assess the extent of power available to test for an interaction between the biomarker and intervention.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 August 2018 doi:10.20944/preprints201808.0521.v1

Deciding whether to power for a significant test of interaction is dependent on how strongly the investigators feel that the biomarker groups can be treated as separate populations or whether they are inherently one population. Some suggested methods for determining sample size for interaction are described in the literature.[190-193]

Parameters to consider

Table A3.1 presents the parameters that are required (in addition to specification of significance level and statistical power) when determining sample sizes for a stratified medicine study. For simplicity, we consider a binary biomarker. For sample size calculations, investigators need to agree on reasonable values for the biomarker prevalence (X), the event proportion in the control arms of each biomarker group (E1 and E2), and the target difference required for each biomarker group (δ_1 and δ_2). The scenarios in Table A3.1 provide a guide on how the trial may be interpreted depending on the treatment effects observed in each biomarker group. Ideally, the evidence on which these conclusions are drawn would be based on adequately powered tests of interaction. The selection of δ_1 and δ_2 are challenging, as is the case with specifying any target difference. The choice is often complicated by other considerations based on secondary outcomes such as side effects or high costs associated with the new treatment. This is particularly true in oncology, where these designs are most commonly used. δ_2 may be selected on the basis of a difference below which the treatment would not be recommended, which could be close to the value under the null hypothesis.

Table A3.1 – Parameters required for sample size determination for a biomarker-stratified trial for a single two-level biomarker - possible scenarios

	Biomarker-positive group *Prevalence = X% Control group event proportion = E1*	**Biomarker-negative group** *Prevalence = 100-X% Control group event proportion = E2*	**Potential conclusions from the trial** *Is a stratified medicine approach recommended?*	*Is the new treatment recommended?*
Scenario A	$\geq \delta_1$ observed	$\geq \delta_2$ observed	No	Yes to all
Scenario B	$\geq \delta_1$ observed	$< \delta_2$ observed	Yes	Only to the biomarker-positive group
Scenario C	$< \delta_1$ observed	$\geq \delta_2$ observed	Yes	Only to the biomarker-negative group
Scenario D	$< \delta_1$ observed	$< \delta_2$ observed	No	No to all

A3.6 Adaptive designs

Adaptive designs for clinical trials[194] enable the analysis of data as they accumulate during a trial at one or more interim analyses, with the results of these analyses used to modify the trial design in some way. A wide range of design adaptations have been suggested, but perhaps the most common involve either early stopping if the intervention under investigation appears particularly promising or particularly unpromising, or a change in the planned sample size based on early estimates of nuisance parameters or treatment effects. Such designs are considered appealing because of the opportunities they give for increasing flexibility and efficiency. A recent and extensive summary of methodology in the area is given by Wassmer and Brannath.[195]

Like almost all confirmatory RCTs, trials with an adaptive design are usually designed to have a fixed Type I error rate and power for some specified target difference. The increased flexibility afforded by an adaptive design can, however, have implications for the choice of the target difference used

 doi:10.20944/preprints201808.0521.v1

in the construction of the design. In a conventional trial, there is often a compromise between ensuring the trial has sufficient power to detect a small clinically meaningful difference and minimising the sample size if a larger treatment difference is anticipated or hoped for. With an adaptive design, the final sample size can depend on the observed interim data. It can thus be possible to design the study to ensure statistical power is maintained to detect a small difference but to also allow the trial to stop with a smaller sample size if a larger treatment difference is observed. In a similar way, if there is uncertainty regarding the SD of the primary outcome at the planning stage, the final sample size can be adjusted depending on interim data to maintain power for a target difference specified on an absolute scale when this is considered desirable.

Much recent interest in adaptive designs for clinical trials has focussed on multi-arm multi-stage [196, 197] trial designs, in which more than one experimental treatment is initially compared with a control arm. Less effective treatments are then dropped as the trial progresses. In this case, issues relevant to the specification of target differences in multi-arm studies, as described in Section A3.2, should also be considered.

References

1. Julious, S., *Sample sizes for clinical trials*. 2010: Chapman and Hall/CRC Press, Boca Raton, FL.
2. ICH, *Harmonised tripartite guideline ICH. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group.* Stat Med, 1999. **18**(15): p. 1905 - 1942.
3. Spiegelhalter, D.J., K.R. Abrams, and J.P. Myles, *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. 1st ed. Statistics in Practice. 2004, Chicester: John Wiley & Sons.
4. Chuang-Stein, C., *Sample size and the probability of a successful trial.* Pharm Stat, 2006. **5**(4): p. 305-9.
5. Charles, P., et al., *Reporting of sample size calculation in randomised controlled trials: review.* BMJ, 2009. **338**(b1732).
6. Cook, J., et al., *Assessing methods to specify the targeted difference for a randomised controlled trial - DELTA (Difference ELicitation in TriAls) review.* Health Technol Assess, 2014. **18**: p. 28.
7. Hislop, J., et al., *Methods for specifying the target difference in a randomised controlled trial: the Difference ELicitation in TriAls (DELTA) systematic review.* PLoS Med, 2014. **11(5)**(e1001645).
8. Cook, J.A., et al., *Use of methods for specifying the target difference in randomised controlled trial sample size calculations: Two surveys of trialists' practice.* Clin Trials, 2014. **11**(3): p. 300-308.
9. Jaeschke, R., J. Singer, and G.H. Guyatt, *Measurement of health status. Ascertaining the minimal clinically important difference.* Control Clin Trials, 1989. **10**(4): p. 407-15.
10. Hays, R. and J. Woolley, *The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it?* Pharmacoeconomics, 2000. **18**: p. 419 - 423.
11. Chan, K.B., et al., *How well is the clinical importance of study results reported? An assessment of randomized controlled trials.* CMAJ, 2001. **165**(9): p. 1197-202.
12. Cook, J., et al., *Specifying the target difference in the primary outcome for a randomised controlled trial: guidance for researchers.* Trials, 2015. **16**(12).
13. Rios, L.P., C. Ye, and L. Thabane, *Association between framing of the research question using the PICOT format and reporting quality of randomized controlled trials.* BMC Med Res Methodol, 2010. **10**: p. 11.
14. Akacha, M., F. Bretz, and S. Ruberg, *Estimands in clinical trials - broadening the perspective.* Stat Med, 2017. **36**(1): p. 5-19.
15. Committee for Human Medicinal Products, *ICH E9 (R1) addendum on estimands and Sensitivity Analysis in Clinical Trials to the guideline on statistical principles for clinical trials EMA/CHMP/ICH/436221/2017*. 2017. p. 1-23.
16. Leuchs, A.K., et al., *Disentangling estimands and the intention-to-treat principle.* Pharm Stat, 2017. **16**(1): p. 12-19.
17. Phillips, A., et al., *Estimands: discussion points from the PSI estimands and sensitivity expert group.* Pharm Stat, 2017. **16**(1): p. 6-11.
18. Mallinckrodt, C.H., et al., *A structured approach to choosing estimands and estimators in longitudinal clinical trials.* Pharm Stat, 2012. **11**(6): p. 456-61.
19. Billingham, S.A., A.L. Whitehead, and S.A. Julious, *An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database.* BMC Med Res Methodol, 2013. **13**: p. 104.
20. National Institute for Health Research. *Involve*. 2017 4/5/2017]; Available from: http://www.invo.org.uk/.
21. National Institute for Health and Care Excellence. *Public Involvement*. 2017 [cited 2017 4/5/2017]; Available from: https://www.nice.org.uk/about/nice-communities/public-involvement.
22. World Medical Association. *WMA Declaration of Helsinki - ethical principles for medical research involving human subjects*. 2013 [cited 2017 4/5/2017]; Available from:

https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/.

23. Edwards, S.J., et al., *Why "underpowered" trials are not necessarily unethical.* Lancet, 1997. **350**(9080): p. 804-7.
24. Schulz, K.F. and D.A. Grimes, *Sample size calculations in randomised trials: mandatory and mystical.* Lancet, 2005. **365**(9467): p. 1348-53.
25. Cook, R. and V. Farewell, *Multiplicity considerations in the design and analysis of clinical trials.* Journal of the Royal Statistical Society, 1996. **159**(1): p. 93-110.
26. Schulz, K.F. and D.A. Grimes, *Multiplicity in randomised trials I: endpoints and treatments.* Lancet, 2005. **365**(9470): p. 1591-5.
27. Flight, L. and S.A. Julious, *Practical guide to sample size calculations: superiority trials.* Pharm Stat, 2016. **15**(1): p. 75-9.
28. Flight, L. and S.A. Julious, *Practical guide to sample size calculations: non-inferiority and equivalence trials.* Pharm Stat, 2016. **15**(1): p. 80-9.
29. Julious, S.A., *The ABC of non-inferiority margin setting from indirect comparisons.* Pharm Stat, 2011. **10**(5): p. 448-53.
30. Lange, S. and G. Freitag, *Choice of delta: requirements and reality--results of a systematic review.* Biom J, 2005. **47**(1): p. 12-27; discussion 99-107.
31. Committee for medicinal products for human use (CHMP), *Guideline on the choice of non-inferiority margin*. 2006.
32. *Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance.* Health Qual Life Outcomes, 2006. **4**: p. 79.
33. Jones, B., et al., *Trials to assess equivalence: the importance of rigorous methods.* BMJ, 1996. **313**(7048): p. 36-9.
34. Kass, M.A., et al., *The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma.* Arch Ophthalmol, 2002. **120**(6): p. 701-13; discussion 829-30.
35. Burr, J.M., et al., *Surveillance for ocular hypertension: an evidence synthesis and economic evaluation.* Health Technol Assess, 2012. **16**(29): p. 1-271, iii-iv.
36. Sugimoto, T., T. Sozu, and T. Hamasaki, *A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints.* Pharm Stat, 2012. **11**(2): p. 118-28.
37. Senn, S. and S. Julious, *Measurement in clinical trials: a neglected issue for statisticians?* Stat Med, 2009. **28**(26): p. 3189-209.
38. Wittes, J., *Commentary on 'Measurement in clinical trials: a neglected issue for statisticians?'.* Stat Med, 2009. **28**(26): p. 3220-2; discussion 3223-5.
39. Sharpe, M., et al., *Integrated collaborative care for comorbid major depression in patients with cancer (SMaRT Oncology-2): a multicentre randomised controlled effectiveness trial.* Lancet, 2014. **384**(9948): p. 1099-108.
40. Walters, S.J., *Quality of life outcomes in clinical trials and health-care evaluation : a practical guide to analysis and interpretation*. Statistics in practice. 2009, Chichester: Wiley.
41. Walters, S.J., *Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36.* Health Qual Life Outcomes, 2004. **2**: p. 26.
42. Hollis, S. and F. Campbell, *What is meant by intention to treat analysis? Survey of published randomised controlled trials.* BMJ, 1999. **319**(7211): p. 670-4.
43. Rosenkranz, G., *Estimands-new statistical principle or the emperor's new clothes?* Pharm Stat, 2017. **16**(1): p. 4-5.
44. Copay, A., et al., *Understanding the minimum clinically important difference: a review of concepts and methods.* Spine J, 2007. **7**: p. 541 - 546.

45. Wells, G., et al., *Minimal clinically important differences: Review of methods.* J Rheumatol, 2001. **28**: p. 406 - 412.
46. Beaton, D., M. Boers, and G. Wells, *Many faces of the minimal clinically important difference (MICD): A literature review and directions for future research.* Curr Opin Rheumatol, 2002. **14**: p. 109 - 114.
47. Engel, L., D.E. Beaton, and Z. Touma, *Minimal Clinically Important Difference: A Review of Outcome Measure Score Interpretation.* Rheum Dis Clin North Am, 2018. **44**(2): p. 177-188.
48. Fayers, P., et al., *Sample size calculation for clinical trials: the impact of clinician beliefs.* Br J Cancer, 2000. **82**: p. 213 - 219.
49. Rose, G., *Sick individuals and sick populations.* Int J Epidemiol, 2001. **30**(3): p. 427-32; discussion 433-4.
50. Guyatt, G.H., et al., *Methods to explain the clinical significance of health status measures.* Mayo Clin Proc, 2002. **77**(4): p. 371-83.
51. de Vet, H.C., et al., *Three ways to quantify uncertainty in individually applied "minimally important change" values.* J Clin Epidemiol, 2010. **63**(1): p. 37-45.
52. Cella, D., et al., *Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life.* Mayo Clin Proc, 2002. **77**(4): p. 384-92.
53. Murray, D.W., et al., *A randomised controlled trial of the clinical effectiveness and cost-effectiveness of different knee prostheses: the Knee Arthroplasty Trial (KAT).* Health Technol Assess, 2014. **18**(19): p. 1-235, vii-viii.
54. Beard, D.J., et al., *Meaningful changes for the Oxford hip and knee scores after joint replacement surgery.* J Clin Epidemiol, 2015. **68**(1): p. 73-9.
55. Walters, S.J., *Consultants' forum: should post hoc sample size calculations be done?* Pharm Stat, 2009. **8**(2): p. 163-9.
56. Julious, S.A. and S.J. Walters, *Estimating effect sizes for health-related quality of life outcomes.* Stat Methods Med Res, 2014. **23**(5): p. 430-9.
57. Brant, R., L. Sutherland, and R. Hilsden, *Examining the minimum important difference.* Stat Med, 1999. **18**(19): p. 2593-603.
58. Whitehead, J., et al., *Using historical lesion volume data in the design of a new phase II clinical trial in acute stroke.* Stroke, 2009. **40**(4): p. 1347-52.
59. Jacobson, N.S. and P. Truax, *Clinical significance: a statistical approach to defining meaningful change in psychotherapy research.* J Consult Clin Psychol, 1991. **59**(1): p. 12-9.
60. Newnham, E.A., K.E. Harwood, and A.C. Page, *Evaluating the clinical significance of responses by psychiatric inpatients to the mental health subscales of the SF-36.* J Affect Disord, 2007. **98**(1-2): p. 91-7.
61. Detsky, A.S., *Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials.* Stat Med, 1990. **9**(1-2): p. 173-84.
62. Torgerson, D.J., M. Ryan, and J. Ratcliffe, *Economics in sample size determination for clinical trials.* QJM, 1995. **88**(7): p. 517-21.
63. Hollingworth, W., et al., *Cost-utility analysis conducted alongside randomized controlled trials: are economic end points considered in sample size calculations and does it matter?* Clin Trials, 2013. **10**(1): p. 43-53.
64. Glick, H.A., *Sample size and power for cost-effectiveness analysis (part 1).* Pharmacoeconomics, 2011. **29**(3): p. 189-98.
65. Glick, H.A., *Sample size and power for cost-effectiveness analysis (Part 2): the effect of maximum willingness to pay.* Pharmacoeconomics, 2011. **29**(4): p. 287-96.
66. National Institute for Health and Care Excellence, *NICE Process and Methods Guides*, in *Guide to the Methods of Technology Appraisal*. 2013, National Institute for Health and Care Excellence (NICE): London.
67. O'Hagan, A., *Uncertain judgements : eliciting experts' probabilities*. Statistics in practice. 2006, Hoboken, NJ: John Wiley & Sons.
68. Gosling, J.P., *Methods for eliciting expert opinion to inform health technology assessment*. 2014.

69. Ryan, M., K. Gerard, and M. Amaya-Amaya, *Using Discrete Choice Experiments to Value Health and Health Care*. The Economics of Non-Market Goods and Resources. 2008: Springer Netherlands.
70. Mt-Isa, S., et al., *Balancing benefit and risk of medicines: a systematic review and classification of available methodologies.* Pharmacoepidemiol Drug Saf, 2014. **23**(7): p. 667-78.
71. Barrett, B., et al., *Sufficiently important difference: expanding the framework of clinical significance.* Med Decis Making, 2005. **25**: p. 250 - 261.
72. Bellamy, N., et al., *Rheumatoid arthritis antirheumatic drug trials. III. Setting the delta for clinical trials of antirheumatic drugs--results of a consensus development (Delphi) exercise.* J Rheumatol, 1991. **18**(12): p. 1908-15.
73. Devilee, J. and A. Knol, *Software to support expert elicitation. An exploratory study of existing software packages. RIVM Letter Report 630003001/2011.* 2011.
74. Howard, R., et al., *Determining the minimum clinically important differences for outcomes in the DOMINO trial.* Int J Geriatr Psychiatry, 2011. **26**(8): p. 812-7.
75. Hampson, L.V., et al., *Elicitation of expert prior opinion: application to the MYPAN trial in childhood polyarteritis nodosa.* PLoS One, 2015. **10**(3:e0120981).
76. Kirkby, H.M., et al., *Using e-mail recruitment and an online questionnaire to establish effect size: A worked example.* BMC Med Res Methodol, 2011. **11**: p. 89.
77. Parmar, M.K., et al., *Monitoring of large randomised clinical trials: a new approach with Bayesian methods.* Lancet, 2001. **358**(9279): p. 375-81.
78. Parmar, M.K., D.J. Spiegelhalter, and L.S. Freedman, *The CHART trials: Bayesian design and monitoring in practice. CHART Steering Committee.* Stat Med, 1994. **13**(13-14): p. 1297-312.
79. Chaloner, K. and F.S. Rhame, *Quantifying and documenting prior beliefs in clinical trials.* Stat Med, 2001. **20**(4): p. 581-600.
80. Hampson, L.V., et al., *Bayesian methods for the design and interpretation of clinical trials in very rare diseases.* Stat Med, 2014. **33**(24): p. 4186-201.
81. Allison, D.B., et al., *Sample size in obesity trials: patient perspective versus current practice.* Med Decis Making, 2010. **30**(1): p. 68-75.
82. Arain, M., et al., *What is a pilot or feasibility study? A review of current practice and editorial policy.* BMC Med Res Methodol, 2010. **10**(67).
83. Kraemer, H.C.M., J.;Noda, A., *Caution Regarding the Use of Pilot Studies to Guide Power Calculations for Study Proposals.* Arch Gen Psychiatry, 2006. **63**(5): p. 484-89.
84. Cocks, K. and D.J. Torgerson, *Sample size calculations for pilot randomized trials: a confidence interval approach.* J Clin Epidemiol, 2013. **66**(2): p. 197-201.
85. Teare, M.D., et al., *Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study.* Trials, 2014. **15**(264).
86. Whitehead, A.L., et al., *Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable.* Stat Methods Med Res, 2016. **25**(3): p. 1057-73.
87. Hippisley-Cox, J. and C. Coupland, *Predicting risk of upper gastrointestinal bleed and intracranial bleed with anticoagulants: cohort study to derive and validate the QBleed scores.* BMJ, 2014. **349**(g4606).
88. Clarke, M., S. Hopewell, and I. Chalmers, *Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting.* Lancet, 2010. **376**(9734): p. 20-1.
89. Sutton, A.J., N.J. Cooper, and D.R. Jones, *Evidence synthesis as the key to more coherent and efficient research.* BMC Med Res Methodol, 2009. **9**(29).
90. Sutton, A.J. and J.P. Higgins, *Recent developments in meta-analysis.* Stat Med, 2008. **27**(5): p. 625-50.
91. Cohen, J., *Statistical power analysis for the behavioral sciences*. Rev. ed. 1977, New York: Academic Press.

92. Higgins, J.P.T. and S. Green, *Cochrane handbook for systematic reviews of interventions version 5.1.0*. 2011, London: The Cochrane Collaboration.
93. Chinn, S., *A simple method for converting an odds ratio to effect size for use in meta-analysis.* Stat Med, 2000. **19**(22): p. 3127-31.
94. Vist, G.E., et al., *Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate.* Cochrane Database Syst Rev, 2008. **3**(Mr000009).
95. Schulz, K.F. and D.A. Grimes, *Multiplicity in randomised trials II: subgroup and interim analyses.* Lancet, 2005. **365**(9471): p. 1657-61.
96. Glazener, C., et al., *Urinary incontinence in men after formal one-to-one pelvic-floor muscle training following radical prostatectomy or transurethral resection of the prostate (MAPS): two parallel randomised controlled trials.* Lancet, 2011. **378**: p. 328 - 337.
97. Hunter, K., K. Moore, and C. Glazener, *Conservative management for postprostatectomy urinary incontinence.* Cochrane Database Syst Rev, 2007. **2**(CD001843).
98. Lois, N., et al., *Internal limiting membrane peeling versus no peeling for idiopathic full-thickness macular hole: a pragmatic randomized controlled trial.* Invest Ophthalmol Vis Sci, 2011. **52**(3): p. 1586-92.
99. *Early Treatment Diabetic Retinopathy Study design and baseline patient characteristics. ETDRS report number 7.* Ophthalmology, 1991. **98**(5 Suppl): p. 741-56.
100. Brooks, H., *Macular hole surgery with and without internal limiting membrane peeling.* Ophthalmology, 2000. **107**: p. 1939 - 1948.
101. Paques, M., et al., *Effect of autologous platelet concentrate in surgery for idiopathic macular hole: results of a multicenter, double-masked, randomized trial. Platelets in Macular Hole Surgery Group.* Ophthalmology, 1999. **106**: p. 932 - 938.
102. Schulz, K., D. Altman, and D. Moher, *CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials.* BMJ, 2010. **340**: p. c332.
103. Chan, A., et al., *SPIRIT 2013 statement: defining standard protocol items for clinical trials.* Ann Intern Med, 2013. **158**: p. 200 - 207.
104. Taggart, D., et al., *Protocol for the Arterial Revascularisation Trial (ART). A randomised trial to compare survival following bilateral versus single internal mammary grafting in coronary revascularisation.* Trials, 2006. **7**(7).
105. Taggart, D., R. D'Amico, and D. Altman, *Effect of arterial revascularisation on survival: a systematic review of studies comparing bilateral and single internal mammary arteries.* Lancet, 2001. **358**: p. 870 - 875.
106. Frobell, R.B., et al., *Treatment for acute anterior cruciate ligament tear: five year outcome of randomised trial.* BMJ, 2013. **346**(f232).
107. Frobell, R.B., et al., *A randomized trial of treatment for acute anterior cruciate ligament tears.* N Engl J Med, 2010. **363**(4): p. 331-42.
108. Roos, E.M., *KOOS User Guide*. 2012.
109. Roos, E.M. and L.S. Lohmander, *The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis.* Health Qual Life Outcomes, 2003. **1**: p. 64.
110. Collins, N.J., et al., *Measures of knee function: International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS).* Arthritis Care Res (Hoboken), 2011. **63 Suppl 11**: p. S208-28.
111. Cook, J.A., et al., *Clustering in surgical trials--database of intracluster correlations.* Trials, 2012. **13**(2).
112. E. Batistatou, C.R., and S. Roberts, *Sample size and power calculations for trials and quasi-experimental studies with clustering.* Stata Journal, 2014. **14**(1): p. 159-175.
113. Gilron, I., et al., *Morphine, gabapentin, or their combination for neuropathic pain.* N Engl J Med, 2005. **352**(13): p. 1324-34.

114. Dworkin, R.H., et al., *Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations.* J Pain, 2008. **9**(2): p. 105-21.
115. Julious, S.A., *Sample sizes for clinical trials with normal data.* Stat Med, 2004. **23**(12): p. 1921-86.
116. Hollingsworth, J.M., et al., *Medical therapy to facilitate urinary stone passage: a meta-analysis.* Lancet, 2006. **368**(9542): p. 1171-9.
117. Singh, A., H.J. Alter, and A. Littlepage, *A systematic review of medical therapy to facilitate passage of ureteral calculi.* Ann Emerg Med, 2007. **50**(5): p. 552-63.
118. McClinton, S., et al., *Use of drug therapy in the management of symptomatic ureteric stones in hospitalized adults (SUSPEND), a multicentre, placebo-controlled, randomized trial of a calcium-channel blocker (nifedipine) and an alpha-blocker (tamsulosin): study protocol for a randomized controlled trial.* Trials, 2014. **15**(238).
119. Baguley, C., et al., *The fate of chronic rhinosinusitis sufferers after maximal medical therapy.* Int Forum Allergy Rhinol, 2014. **4**(7): p. 525-32.
120. Young, L.C., et al., *Efficacy of medical therapy in treatment of chronic rhinosinusitis.* Allergy Rhinol (Providence), 2012. **3**(1): p. e8-e12.
121. Rimmer, J., et al., *Surgical versus medical interventions for chronic rhinosinusitis with nasal polyps.* Cochrane Database Syst Rev, 2014(Cd006991).
122. Sharma, R., et al., *Surgical interventions for chronic rhinosinusitis with nasal polyps.* Cochrane Database Syst Rev, 2014. **11**(Cd006990).
123. Hopkins, C., et al., *Psychometric validity of the 22-item Sinonasal Outcome Test.* Clin Otolaryngol, 2009. **34**(5): p. 447-54.
124. Norman, G.R., J.A. Sloan, and K.W. Wyrwich, *Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation.* Med Care, 2003. **41**(5): p. 582-92.
125. Norman, G.R., J.A. Sloan, and K.W. Wyrwich, *The truly remarkable universality of half a standard deviation: confirmation through another look.* Expert Rev Pharmacoecon Outcomes Res, 2004. **4**(5): p. 581-5.
126. Erskine, S.E., et al., *Managing chronic rhinosinusitis and respiratory disease: a qualitative study of triggers and interactions.* J Asthma, 2015. **52**(6): p. 600-5.
127. Bewick, J.C., et al., *Preliminary Findings: The Feasibility Study for a Randomized Controlled Trial of Clarithromycin in Chronic Rhinosinusitis.* Otolaryngol Head Neck Surg Endosc, 2014. **151**(S1): p. 125.
128. Elouafkaoui, P., et al., *An Audit and Feedback Intervention for Reducing Antibiotic Prescribing in General Dental Practice: The RAPiD Cluster Randomised Controlled Trial.* PLoS Med, 2016. **13**(8:e1002115).
129. Borm, G.F., J. Fransen, and W.A. Lemmens, *A simple sample size formula for analysis of covariance in randomized clinical trials.* J Clin Epidemiol, 2007. **60**(12): p. 1234-8.
130. Ivers, N., et al., *Audit and feedback: effects on professional practice and healthcare outcomes.* Cochrane Database Syst Rev, 2012. **6**(Cd000259).
131. Senn, S., *Statistical issues in drug development*. 2nd ed. Statistics in practice. 2007, Chichester, England: John Wiley & Sons.
132. Senn, S., *Controversies concerning randomization and additivity in clinical trials.* Stat Med, 2004. **23**(24): p. 3729-53.
133. Machin, D., *Sample size tables for clinical studies*. 3rd ed. Ebook central. 2009, Chichester, UK: Wiley-Blackwell.
134. Chuang-Stein, C., et al., *The role of the minimum clinically important difference and its impact on designing a trial.* Pharm Stat, 2011. **10**(3): p. 250-6.
135. Carroll, K.J., *Back to basics: explaining sample size in outcome trials, are statisticians doing a thorough job?* Pharm Stat, 2009. **8**(4): p. 333-45.
136. Royston, P., et al., *Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit.* Trials, 2011. **12**(81).
137. Proschan, M.A., *Sample size re-estimation in clinical trials.* Biom J, 2009. **51**(2): p. 348-57.

138. Bauer, P. and F. Koenig, *The reassessment of trial perspectives from interim data--a critical view.* Stat Med, 2006. **25**(1): p. 23-36.
139. Dallow, N. and P. Fina, *The perils with the misuse of predictive power.* Pharm Stat, 2011. **10**(4): p. 311-7.
140. Kent, D.M., T.A. Trikalinos, and M.D. Hill, *Are unadjusted analyses of clinical trials inappropriately biased toward the null?* Stroke, 2009. **40**(3): p. 672-3.
141. Flight, L. and S.A. Julious, *Practical guide to sample size calculations: an introduction.* Pharm Stat, 2016. **15**(1): p. 68-74.
142. Schulz, K.F. and D.A. Grimes, *Sample size slippages in randomised trials: exclusions and the lost and wayward.* Lancet, 2002. **359**(9308): p. 781-5.
143. Curran, D., R.J. Sylvester, and G. Hoctin Boes, *Sample size estimation in phase III cancer clinical trials.* Eur J Surg Oncol, 1999. **25**(3): p. 244-50.
144. Ford, I. and J. Norrie, *The role of covariates in estimating treatment effects and risk in long-term clinical trials.* Stat Med, 2002. **21**(19): p. 2899-908.
145. Pokhrel, A., T. Dyba, and T. Hakulinen, *A Greenwood formula for standard error of the age-standardised relative survival ratio.* Eur J Cancer, 2008. **44**(3): p. 441-7.
146. Royston, P. and M.K. Parmar, *An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect.* Trials, 2014. **15**(1): p. 314.
147. White, I.R., *Uses and limitations of randomization-based efficacy estimators.* Stat Methods Med Res, 2005. **14**(4): p. 327-47.
148. Wittes, J., *Sample size calculations for randomized controlled trials.* Epidemiol Rev, 2002. **24**(1): p. 39-53.
149. Emsley, R., G. Dunn, and I.R. White, *Mediation and moderation of treatment effects in randomised controlled trials of complex interventions.* Stat Methods Med Res, 2010. **19**(3): p. 237-70.
150. Dunn, G., M. Maracy, and B. Tomenson, *Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods.* Stat Methods Med Res, 2005. **14**(4): p. 369-95.
151. Dunn, G., et al., *Evaluation and validation of social and psychological markers in randomised trials of complex interventions in mental health: a methodological research programme.* Health Technol Assess, 2015. **19**(93): p. 1-115, v-vi.
152. Cesana, B.M. and P. Antonelli, *Sample size calculations in clinical research should also be based on ethical principles.* Trials, 2016. **17**(149).
153. Bacchetti, P., *Current sample size conventions: flaws, harms, and alternatives.* BMC Med, 2010. **8**(17).
154. Bland, J.M., *The tyranny of power: is there a better way to calculate sample size?* BMJ, 2009. **339**(b3985).
155. Jain, A., et al., *Nail bed INJury Assessment Pilot (NINJA-P) study: should the nail plate be replaced or discarded after nail bed repair in children? Study protocol for a pilot randomised controlled trial.* Pilot Feasibility Stud, 2015. **1**(29).
156. Thabane, L., et al., *A tutorial on pilot studies: the what, why and how.* BMC Med Res Methodol, 2010. **10**(1).
157. Browne, R.H., *On the use of a pilot sample for sample size determination.* Stat Med, 1995. **14**(17): p. 1933-40.
158. Sim, J. and M. Lewis, *The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency.* J Clin Epidemiol, 2012. **65**(3): p. 301-8.
159. Bonati, L.H., et al., *Long-term outcomes after stenting versus endarterectomy for treatment of symptomatic carotid stenosis: the International Carotid Stenting Study (ICSS) randomised trial.* Lancet, 2015. **385**(9967): p. 529-38.
160. Gillett, R., *An average power criterion for sample size estimation.* Journal of the Royal Statistical Society: Series D (The Statistician), 1994. **43**(23): p. 389-394.
161. Gordon Lan, K.K. and J.T. Wittes, *Some thoughts on sample size: a Bayesian-frequentist hybrid approach.* Clin Trials, 2012. **9**(5): p. 561-9.

162. Neuenschwander, B., et al., *Summarizing historical information on controls in clinical trials.* Clin Trials, 2010. **7**(1): p. 5-18.
163. Schmidli, H., et al., *Robust meta-analytic-predictive priors in clinical trials with historical control information.* Biometrics, 2014. **70**(4): p. 1023-32.
164. Burke, D.L., et al., *Meta-analysis of randomized phase II trials to inform subsequent phase III decisions.* Trials, 2014. **15**(346).
165. Brown, B.W., et al., *Projection from previous studies: a Bayesian and frequentist compromise.* Control Clin Trials, 1987. **8**(1): p. 29-44.
166. Ciarleglio, M.M., C.D. Arendt, and P.N. Peduzzi, *Selection of the effect size for sample size determination for a continuous response in a superiority clinical trial using a hybrid classical and Bayesian procedure.* Clin Trials, 2016. **13**(3): p. 275-85.
167. Eaton, M.L.M., R. L.; Soaita, A. I., *On the limiting behaviour of the "probability of claiming superiority" in a Bayesian context.* Bayesian Analysis 2013. **8**(1): p. 221-232.
168. Whitehead, J., et al., *Bayesian sample size for exploratory clinical trials incorporating historical data.* Stat Med, 2008. **27**(13): p. 2307-27.
169. Joseph, L. and P. Belisle, *Bayesian Sample Size Determination for Normal Means and Differences Between Normal Means.* Journal of the Royal Statistical Society. Series D (The Statistician), 1997. **46**(2): p. 209-226.
170. Pezeshk, H., et al., *The choice of sample size: a mixed Bayesian / frequentist approach.* Stat Methods Med Res, 2009. **18**(2): p. 183-94.
171. Stallard, N., et al., *Determination of the optimal sample size for a clinical trial accounting for the population size.* Biom J, 2016. **59**(4): p. 609-625.
172. Wilson, E.C., *A practical guide to value of information analysis.* Pharmacoeconomics, 2015. **33**(2): p. 105-21.
173. Claxton, K., *The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies.* J Health Econ, 1999. **18**(3): p. 341-64.
174. Steuten, L., et al., *A Systematic and Critical Review of the Evolving Methods and Applications of Value of Information in Academia and Practice.* PharmacoEconomics, 2013. **31**(1): p. 25-48.
175. Eckermann, S. and A.R. Willan, *Expected Value of Sample Information with Imperfect Implementation: Improving Practice and Reducing Uncertainty with Appropriate Counterfactual Consideration.* Med Decis Making, 2016. **36**(3): p. 282-3.
176. Zhu, H., S. Zhang, and C. Ahn, *Sample size considerations for split-mouth design.* Stat Methods Med Res, 2017. **26**(6): p. 2543-2551.
177. Barker, D., et al., *Stepped wedge cluster randomised trials: a review of the statistical methodology used and available.* BMC Med Res Methodol, 2016. **16**(69).
178. Kuijper, B., et al., *Cervical collar or physiotherapy versus wait and see policy for recent onset cervical radiculopathy: randomised trial.* BMJ, 2009. **339**(b3883).
179. Julious, S.A., D. Machin, and S.B. Tan, *An introduction to statistics in early phase trials*. 2010, Oxford: Wiley-Blackwell.
180. Julious, S.A. and N.E. McIntyre, *Sample sizes for trials involving multiple correlated must-win comparisons.* Pharm Stat, 2012. **11**(2): p. 177-85.
181. Fernandes, N. and A. Stone, *Multiplicity adjustments in trials with two correlated comparisons of interest.* Stat Methods Med Res, 2011. **20**(6): p. 579-94.
182. Kerry, S.M. and J.M. Bland, *Sample size in cluster randomisation.* BMJ, 1998. **316**(7130): p. 549.
183. Donner, A. and N. Klar, *Design and analysis of cluster randomization trials in health research*. 2000, London: Arnold.
184. Ukoumunne, O.C., et al., *Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review.* Health Technol Assess, 1999. **3**(5): p. iii-92.
185. Campbell, M.K., P.M. Fayers, and J.M. Grimshaw, *Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research.* Clin Trials, 2005. **2**(2): p. 99-107.

186. Eldridge, S.M., D. Ashby, and S. Kerry, *Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method.* Int J Epidemiol, 2006. **35**(5): p. 1292-300.
187. Rutterford, C., A. Copas, and S. Eldridge, *Methods for sample size determination in cluster randomized trials.* Int J Epidemiol, 2015. **44**(3): p. 1051-1067.
188. Roberts, C. and S.A. Roberts, *Design and analysis of clinical trials with clustering effects due to treatment.* Clin Trials, 2005. **2**(2): p. 152-62.
189. Senn, S., *Cross-over trials in clinical research*. 2nd ed. Crossover trials in clinical research. 2002, Chichester: Wiley.
190. Brookes, S.T., et al., *Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test.* J Clin Epidemiol, 2004. **57**(3): p. 229-36.
191. Chen, D.T., et al., *Strategies for power calculations in predictive biomarker studies in survival data.* Oncotarget, 2016. **7**(49): p. 80373-80381.
192. Gonen, M., *Planning for subgroup analysis: a case study of treatment-marker interaction in metastatic colorectal cancer.* Control Clin Trials, 2003. **24**(4): p. 355-63.
193. Mackey, H.M. and T. Bengtsson, *Sample size and threshold estimation for clinical trials with predictive biomarkers.* Contemp Clin Trials, 2013. **36**(2): p. 664-72.
194. Kairalla, J.A., et al., *Adaptive trial designs: a review of barriers and opportunities.* Trials, 2012. **13**(145).
195. Wassmer, G. and W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer Series in Pharmaceutical Statistics. 2016: Springer.
196. Sydes, M., et al., *Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial.* Trials, 2009. **10**(39).
197. Wason, J.M. and T. Jaki, *Optimal design of multi-arm multi-stage trials.* Stat Med, 2012. **31**(30): p. 4269-79.