Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review

Jonathan A Cook, Jennifer Hislop, Temitope E Adewuyi, Kirsten Harrild, Douglas G Altman, Craig R Ramsay, Cynthia Fraser, Brian Buckley, Peter Fayers, Ian Harvey, Andrew H Briggs, John D Norrie, Dean Fergusson, Ian Ford and Luke D Vale
Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAlS) review

Jonathan A Cook,¹* Jennifer Hislop,¹ Temitope E Adewuyi,¹ Kirsten Harrild,² Douglas G Altman,³ Craig R Ramsay,¹ Cynthia Fraser,¹ Brian Buckley,⁴ Peter Fayers,⁵ Ian Harvey,⁶ Andrew H Briggs,⁷ John D Norrie,¹ Dean Fergusson,⁸ Ian Ford⁹ and Luke D Vale¹⁰

¹Health Services Research Unit, University of Aberdeen, Aberdeen, UK
²Medical Statistics Team, University of Aberdeen, Aberdeen, UK
³Centre for Statistics in Medicine, University of Oxford, Oxford, UK
⁴Department of General Practice, National University of Ireland, Galway, Ireland
⁵Population Health, University of Aberdeen, Aberdeen, UK
⁶Faculty of Medicine and Health Sciences, University of East Anglia, Norwich, UK
⁷Health Economics and Health Technology Assessment, University of Glasgow, Glasgow, UK
⁸Ottawa Hospital Research Institute, Ontario, Canada
⁹Robertson Centre for Biostatistics, University of Glasgow, Glasgow, UK
¹⁰Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK

*Corresponding author

Declared competing interests of authors: none

Published May 2014
DOI: 10.3310/hta18280

This report should be referenced as follows:


Health Technology Assessment is indexed and abstracted in Index Medicus/MEDLINE, Excerpta Medica/EMBASE, Science Citation Index Expanded (SciSearch®) and Current Contents®/Clinical Medicine.
Criteria for inclusion in the Health Technology Assessment journal

Reports are published in Health Technology Assessment (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in Health Technology Assessment are termed ‘systematic’ when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

The HTA programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. ‘Health technologies’ are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

For more information about the HTA programme please visit the website: http://www.hta.ac.uk/

This report

The research reported in this issue of the journal was funded by the HTA programme as project number 06/98/01. The contractual start date was in October 2010. The draft report began editorial review in April 2012 and was accepted for publication in August 2012. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors’ report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health.

© Queen’s Printer and Controller of HMSO 2014. This work was produced by Cook et al. under the terms of a commissioning contract issued by the Secretary of State for Health. This issue may be freely reproduced for the purposes of private research and study and extracts (or indeed, the full report) may be included in professional journals provided that suitable acknowledgement is made and the reproduction is not associated with any form of advertising. Applications for commercial reproduction should be addressed to: NIHR Journals Library, National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre, Alpha House, University of Southampton Science Park, Southampton SO16 7NS, UK.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).
Editor-in-Chief of Health Technology Assessment and NIHR Journals Library

Professor Tom Walley  Director, NIHR Evaluation, Trials and Studies and Director of the HTA Programme, UK

NIHR Journals Library Editors

Professor Ken Stein  Chair of HTA Editorial Board and Professor of Public Health, University of Exeter Medical School, UK

Professor Andree Le May  Chair of NIHR Journals Library Editorial Group (EME, HS&DR, PGfAR, PHR journals)

Dr Martin Ashton-Key  Consultant in Public Health Medicine/Consultant Advisor, NETSCC, UK

Professor Matthias Beck  Chair in Public Sector Management and Subject Leader (Management Group), Queen’s University Management School, Queen’s University Belfast, UK

Professor Aileen Clarke  Professor of Health Sciences, Warwick Medical School, University of Warwick, UK

Dr Tessa Crilly  Director, Crystal Blue Consulting Ltd, UK

Dr Peter Davidson  Director of NETSCC, HTA, UK

Ms Tara Lamont  Scientific Advisor, NETSCC, UK

Dr Tom Marshall  Reader in Primary Care, School of Health and Population Sciences, University of Birmingham, UK

Professor William McGuire  Professor of Child Health, Hull York Medical School, University of York, UK

Professor Geoffrey Meads  Honorary Professor, Business School, Winchester University and Medical School, University of Warwick, UK

Professor Jane Norman  Professor of Maternal and Fetal Health, University of Edinburgh, UK

Professor John Powell  Senior Clinical Researcher, Department of Primary Care, University of Oxford, UK

Professor James Raftery  Professor of Health Technology Assessment, Wessex Institute, Faculty of Medicine, University of Southampton, UK

Dr Rob Riemsma  Reviews Manager, Kleijnen Systematic Reviews Ltd, UK

Professor Helen Roberts  Professorial Research Associate, University College London, UK

Professor Helen Snooks  Professor of Health Services Research, Institute of Life Science, College of Medicine, Swansea University, UK

Please visit the website for a list of members of the NIHR Journals Library Board:
http://www.journalslibrary.nihr.ac.uk/about/editors

Editorial contact: nihredit@southampton.ac.uk
Abstract

Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review

Jonathan A Cook,1* Jennifer Hislop,1 Temitope E Adewuyi,1 Kirsten Harrild,2 Douglas G Altman,3 Craig R Ramsay,1 Cynthia Fraser,1 Brian Buckley,4 Peter Fayers,5 Ian Harvey,6 Andrew H Briggs,7 John D Norrie,1 Dean Fergusson,8 Ian Ford9 and Luke D Vale10

1Health Services Research Unit, University of Aberdeen, Aberdeen, UK
2Medical Statistics Team, University of Aberdeen, Aberdeen, UK
3Centre for Statistics in Medicine, University of Oxford, Oxford, UK
4Department of General Practice, National University of Ireland, Galway, Ireland
5Population Health, University of Aberdeen, Aberdeen, UK
6Faculty of Medicine and Health Sciences, University of East Anglia, Norwich, UK
7Health Economics and Health Technology Assessment, University of Glasgow, Glasgow, UK
8Ottawa Hospital Research Institute, Ontario, Canada
9Robertson Centre for Biostatistics, University of Glasgow, Glasgow, UK
10Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK

*Corresponding author

Background: The randomised controlled trial (RCT) is widely considered to be the gold standard study for comparing the effectiveness of health interventions. Central to the design and validity of a RCT is a calculation of the number of participants needed (the sample size). The value used to determine the sample size can be considered the ‘target difference’. From both a scientific and an ethical standpoint, selecting an appropriate target difference is of crucial importance. Determination of the target difference, as opposed to statistical approaches to calculating the sample size, has been greatly neglected though a variety of approaches have been proposed the current state of the evidence is unclear.

Objectives: The aim was to provide an overview of the current evidence regarding specifying the target difference in a RCT sample size calculation. The specific objectives were to conduct a systematic review of methods for specifying a target difference; to evaluate current practice by surveying triallists; to develop guidance on specifying the target difference in a RCT; and to identify future research needs.

Design: The biomedical and social science databases searched were MEDLINE, MEDLINE In-Process & Other Non-Indexed Citations, EMBASE, Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Methodology Register, PsycINFO, Science Citation Index, EconLit, Education Resources Information Center (ERIC) and Scopus for in-press publications. All were searched from 1966 or the earliest date of the database coverage and searches were undertaken between November 2010 and January 2011. There were three interlinked components: (1) systematic review of methods for specifying a target difference for RCTs – a comprehensive search strategy involving an electronic literature search of biomedical and some non-biomedical databases and clinical trials textbooks was carried out; (2) identification of current trial practice using two surveys of triallists – members of the Society for Clinical Trials (SCT) were invited to complete an
online survey and respondents were asked about their awareness and use of, and willingness to recommend, methods; one individual per trialist group [UK Clinical Research Collaboration (UKCRC)-registered Clinical Trials Units (CTUs), Medical Research Council (MRC) UK Hubs for Trials Methodology Research and National Institute for Health Research (NIHR) UK Research Design Services (RDS)] was invited to complete a survey; (3) production of a structured guidance document to aid the design of future trials – the draft guidance was developed utilising the results of the systematic review and surveys by the project steering and advisory groups.

Setting: Methodological review incorporating electronic searches, review of books and guidelines, two surveys of experts (membership of an international society and UK- and Ireland-based triallists) and development of guidance.

Participants: The two surveys were sent out to membership of the SCT and UK- and Ireland-based trialists.

Interventions: The review focused on methods for specifying the target difference in a RCT. It was not restricted to any type of intervention or condition.

Main outcome measures: Methods for specifying the target difference for a RCT were considered.

Results: The search identified 11,485 potentially relevant studies. In total, 1434 were selected for full-text assessment and 777 were included in the review. Seven methods to specify the target difference for a RCT were identified – anchor, distribution, health economic, opinion-seeking, pilot study, review of evidence base (RoEB) and standardised effect size (SES) – each having important variations in implementation. A total of 216 of the included studies used more than one method. A total of 180 (15%) responses to the SCT survey were received, representing 13 countries. Awareness of methods ranged from 38% (n = 69) for the health economic method to 90% (n = 162) for the pilot study. Of the 61 surveys sent out to UK trialist groups, 34 (56%) responses were received. Awareness ranged from 97% (n = 33) for the RoEB and pilot study methods to only 41% (n = 14) for the distribution method. Based on the most recent trial, all bar three groups (91%, n = 30) used a formal method. Guidance was developed on the use of each method and the reporting of the sample size calculation in a trial protocol and results paper.

Conclusions: There is a clear need for greater use of formal methods to determine the target difference and better reporting of its specification. Raising the standard of RCT sample size calculations and the corresponding reporting of them would aid health professionals, patients, researchers and funders in judging the strength of the evidence and ensuring better use of scarce resources.

Funding: The Medical Research Council UK and the National Institute for Health Research Joint Methodology Research Research programme.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of tables</td>
<td>xi</td>
</tr>
<tr>
<td>List of figures</td>
<td>xiii</td>
</tr>
<tr>
<td>List of boxes</td>
<td>xv</td>
</tr>
<tr>
<td>List of abbreviations</td>
<td>xvii</td>
</tr>
<tr>
<td>Scientific summary</td>
<td>xix</td>
</tr>
<tr>
<td>Background</td>
<td>xix</td>
</tr>
<tr>
<td>Aim</td>
<td>xix</td>
</tr>
<tr>
<td>Objectives</td>
<td>xix</td>
</tr>
<tr>
<td>Methods</td>
<td>xix</td>
</tr>
<tr>
<td>- Systematic review of methods for specifying a target difference for a randomised controlled trial</td>
<td>xix</td>
</tr>
<tr>
<td>- Identification of triallists’ current practice</td>
<td>xx</td>
</tr>
<tr>
<td>- Production of guidance on specifying the target difference for a randomised controlled trial</td>
<td>xx</td>
</tr>
<tr>
<td>Results</td>
<td>xx</td>
</tr>
<tr>
<td>- Systematic review of methods for specifying a target difference for a randomised controlled trial</td>
<td>xx</td>
</tr>
<tr>
<td>- Identification of triallists’ current practice</td>
<td>xx</td>
</tr>
<tr>
<td>- Guidance on specifying the target difference in a randomised controlled trial</td>
<td>xxi</td>
</tr>
<tr>
<td>Conclusions</td>
<td>xxi</td>
</tr>
<tr>
<td>Further research priorities</td>
<td>xxi</td>
</tr>
<tr>
<td>Funding</td>
<td>xxii</td>
</tr>
<tr>
<td><strong>Chapter 1</strong> <strong>Introduction and background</strong></td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Project summary</td>
<td>3</td>
</tr>
<tr>
<td>Background</td>
<td>3</td>
</tr>
<tr>
<td>- Why seek an important difference?</td>
<td>3</td>
</tr>
<tr>
<td>- Inherent meaning versus constructed scales</td>
<td>4</td>
</tr>
<tr>
<td>- How can an important difference be determined?</td>
<td>5</td>
</tr>
<tr>
<td>- How important differences relate to the design of randomised controlled trials</td>
<td>6</td>
</tr>
<tr>
<td>Summary</td>
<td>9</td>
</tr>
<tr>
<td><strong>Chapter 2</strong> <strong>Systematic review of methods for specifying a target difference</strong></td>
<td>11</td>
</tr>
<tr>
<td>Introduction</td>
<td>11</td>
</tr>
<tr>
<td>Methodology of the review</td>
<td>11</td>
</tr>
<tr>
<td>- Search strategy</td>
<td>11</td>
</tr>
<tr>
<td>- Inclusion and exclusion criteria</td>
<td>11</td>
</tr>
<tr>
<td>- Data extraction</td>
<td>12</td>
</tr>
<tr>
<td>- Method of analysis</td>
<td>12</td>
</tr>
<tr>
<td>Results</td>
<td>13</td>
</tr>
<tr>
<td>- Search results</td>
<td>13</td>
</tr>
<tr>
<td>- Anchor method</td>
<td>14</td>
</tr>
<tr>
<td>- Distribution method</td>
<td>20</td>
</tr>
</tbody>
</table>
## CONTENTS

- Health economic method 24
- Opinion-seeking method 28
- Pilot study method 31
- Review of evidence base method 32
- Standardised effect size 33
- Combination of methods 38

**Discussion** 42
  - Key findings 42
  - Strengths and limitations 44

**Chapter 3** Surveys of triallists’ current practice 45

**Introduction** 45

**Methodology of the surveys** 45
  - Survey 1: Society of Clinical Trials membership 45
  - Survey 2: UK- and Ireland-based triallists 45
  - Ethical review 46
  - Data analysis 46

**Results** 46
  - Survey 1: Society of Clinical Trials membership 46
  - Survey 2: UK- and Ireland-based triallists 48

**Discussion** 52
  - Key findings 52
  - Strengths and limitations 54

**Chapter 4** Guidance on specifying the target difference in a randomised controlled trial sample size calculation 55

**Sample size calculations for randomised controlled trials** 55
  - Background 55
  - Research question 56
  - Sample size calculation 57
  - The role of the primary outcome 58

**Specifying the target difference** 59

**Description and guidance on the use of individual methods for specifying the target difference** 62
  - Anchor method 62
  - Distribution method 63
  - Health economic method 63
  - Opinion-seeking method 64
  - Pilot study method 64
  - Review of evidence base method 65
  - Standardised effect size 66

**Reporting of the sample size calculation** 67

**Summary** 69

**Further research priorities** 70

**Acknowledgements** 71

**References** 73

**Appendix 1** Protocol 105

**Appendix 2** Literature search strategies 115
Appendix 3  List of included studies  121
Appendix 4  Survey form sent to UK- and Ireland-based triallists  167
List of tables

TABLE 1 Motivating examples: the Norwegian Spine Study and the Full-thickness macular hole and Internal Limiting Membrane peeling study (FILMS) trial 1

TABLE 2 Number of search results obtained from each included database 13

TABLE 3 Two worked examples based on the Torgerson and colleagues’ approach 26

TABLE 4 Opinion-seeking method example: Delphi method approach for an antirheumatic drug trial 30

TABLE 5 Use of additional method(s) 39

TABLE 6 Studies utilising more than one method: method combinations 39

TABLE 7 Survey 1: respondent characteristics ($n = 180$) 47

TABLE 8 Survey 1: awareness, usage and willingness to recommend methods 48

TABLE 9 Survey 2: respondent characteristics ($n = 34$) 49

TABLE 10 Survey 2: awareness, use and willingness to recommend methods 51

TABLE 11 Survey 2: most recent trial ($n = 33$) 51
List of figures

FIGURE 1 Statistically and clinically important difference 5
FIGURE 2 The screening process 13
FIGURE 3 Anchor method illustrative example 16
List of boxes

BOX 1 Methods for specifying an important and/or realistic difference 7

BOX 2 Anchor method example: the Clinical Global Impression anchor with a seven-point ordinal scale 15

BOX 3 Summary of Detsky’s approach (health economic method) 25

BOX 4 Standardised effect size example: goal attainment scaling 34

BOX 5 Conventional approach to the sample size calculation for a two-group parallel RCT 56

BOX 6 A realistic and/or important difference as the basis for the target difference: example based on the Men After Prostate Surgery (MAPS) trial 60

BOX 7 Protocol sample size calculation example: binary primary outcome [Men After Prostate Surgery (MAPS) trial] 68

BOX 8 Protocol sample size calculation example: continuous primary outcome (FILMS) 69

BOX 9 Protocol sample size calculation example: survival primary outcome [Arterial Revascularisation Trial (ART)] 69
List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>analysis of variance</td>
</tr>
<tr>
<td>CART</td>
<td>classification and regression trees</td>
</tr>
<tr>
<td>CENTRAL</td>
<td>Cochrane Central Register of Controlled Trials</td>
</tr>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>CoV</td>
<td>coefficient of variation</td>
</tr>
<tr>
<td>CTU</td>
<td>Clinical Trials Unit</td>
</tr>
<tr>
<td>DELTA</td>
<td>Difference Elicitation in TriAls</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>European Quality of Life-5 Dimensions</td>
</tr>
<tr>
<td>ERIC</td>
<td>Education Resources Information Center</td>
</tr>
<tr>
<td>FILMS</td>
<td>Full-thickness macular hole and Internal Limiting Membrane peeling Study</td>
</tr>
<tr>
<td>HSRU</td>
<td>Health Services Research Unit</td>
</tr>
<tr>
<td>ICH</td>
<td>International Conference on Harmonisation</td>
</tr>
<tr>
<td>MCID</td>
<td>minimum/minimal(ly) clinically important difference</td>
</tr>
<tr>
<td>MDC</td>
<td>minimum/minimal(ly) detectable change</td>
</tr>
<tr>
<td>MID</td>
<td>minimum/minimal(ly) important difference</td>
</tr>
<tr>
<td>MRC</td>
<td>Medical Research Council</td>
</tr>
<tr>
<td>NICE</td>
<td>National Institute for Health and Care Excellence</td>
</tr>
<tr>
<td>NIHR</td>
<td>National Institute for Health Research</td>
</tr>
<tr>
<td>NMB</td>
<td>net monetary benefit</td>
</tr>
<tr>
<td>PICOT</td>
<td>population, intervention, control, outcome and time frame</td>
</tr>
<tr>
<td>QALY</td>
<td>quality-adjusted life-year</td>
</tr>
<tr>
<td>RCI</td>
<td>reliable change index</td>
</tr>
<tr>
<td>RCT</td>
<td>randomised controlled trial</td>
</tr>
<tr>
<td>RDS</td>
<td>Research Design Service</td>
</tr>
<tr>
<td>ROC</td>
<td>receiver operating characteristic</td>
</tr>
<tr>
<td>RoEB</td>
<td>review of evidence base</td>
</tr>
<tr>
<td>SCT</td>
<td>Society for Clinical Trials</td>
</tr>
<tr>
<td>SD</td>
<td>standard deviation</td>
</tr>
<tr>
<td>SDC</td>
<td>smallest detectable change</td>
</tr>
<tr>
<td>SE</td>
<td>standard error</td>
</tr>
<tr>
<td>SEM</td>
<td>standard error of the measurement</td>
</tr>
<tr>
<td>SES</td>
<td>standardised effect size</td>
</tr>
<tr>
<td>SF-36</td>
<td>Short Form questionnaire-36 items</td>
</tr>
<tr>
<td>SRM</td>
<td>standardised response mean</td>
</tr>
<tr>
<td>UKCRC</td>
<td>UK Clinical Research Collaboration</td>
</tr>
<tr>
<td>VAS</td>
<td>visual analogue scale</td>
</tr>
</tbody>
</table>
Scientific summary

Background

The randomised controlled trial (RCT) is widely considered to be the gold standard study for comparing the effectiveness of health interventions. Central to the design and validity of a RCT is a calculation of the number of participants needed (the sample size). This provides reassurance that the trial will identify a difference of a particular magnitude if such a difference exists. The value used to determine the sample size can be considered the ‘target difference’. From both a scientific and an ethical standpoint, selecting an appropriate target difference is of crucial importance. Specifying too small a target difference could be a wasteful (and unethical) use of data and resources. Conversely, too large a target difference could lead to an important difference being easily overlooked because the study is too small. Furthermore, an undersized study may not usefully contribute to the knowledge base and could potentially have a detrimental impact on decision-making.

Determination of the target difference, as opposed to statistical approaches to calculating the sample size, has been greatly neglected. A variety of approaches have been proposed for formally specifying what an important difference should be [such as the ‘minimal clinically important difference (MCID)’], although the current state of the evidence is unclear, particularly with regard to informing RCT design by specifying a target difference.

Aim

The aim was to provide an overview of the current evidence on methods for specifying the target difference in a RCT sample size calculation.

Objectives

- To conduct a systematic review of methods for specifying a target difference.
- To evaluate current practice by surveying triallists.
- To develop guidance on specifying the target difference for a RCT.
- To identify future research needs.

Methods

The study comprised three interlinked components.

Systematic review of methods for specifying a target difference for a randomised controlled trial

A comprehensive search of both biomedical and some non-biomedical databases was undertaken. Additionally, clinical trial textbooks and guidelines were reviewed. To be included, a study had to report a formal method that could potentially be used to specify a target difference. The biomedical and social science databases searched were MEDLINE, MEDLINE In-Process & Other Non-Indexed Citations, EMBASE, Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Methodology Register, PsycINFO, Science Citation Index, EconLit, Education Resources Information Center (ERIC) and Scopus for in-press publications. All were searched from 1966 or the earliest date of the database coverage and searches were undertaken between November 2010 and January 2011.
Identification of triallists’ current practice

This involved two surveys:

- Members of the Society for Clinical Trials (SCT) were sent an invitation (followed by a reminder) to complete an online survey through the society’s email distribution list. Respondents were asked about their awareness and use of, and willingness to recommend, methods for determining a target difference in a RCT.

- Survey of leading UK- and Ireland-based triallists. The survey was sent to UK Clinical Research Collaboration (UKCRC)-registered Clinical Trials Units (CTUs), Medical Research Council (MRC) UK Hubs for Trials Methodology Research and National Institute for Health Research (NIHR) Research Design Services (RDS). One response per triallist group was invited. In addition to the information collected in the SCT survey, this survey included questions on the approach used for the most recent trial developed. The initial request was personalised and sent by post, followed by two reminders.

Production of guidance on specifying the target difference for a randomised controlled trial

The draft guidance was developed by the project steering and advisory groups utilising the results of the systematic review and surveys. Findings were circulated and presented to members of the combined group at a face-to-face meeting along with a proposed outline of the guidance document structure and list of recommendations. Both the structure and main recommendations were agreed at this meeting. The guidance was subsequently drafted and circulated for further comment.

Results

Systematic review of methods for specifying a target difference for a randomised controlled trial

The search identified 11,485 potentially relevant studies, of which 1434 were selected for full-text assessment, with 777 included in the review. Fifteen clinical trial textbooks and the International Conference on Harmonisation (ICH) tripartite guidelines were also reviewed. Seven methods were identified – anchor, distribution, health economic, opinion-seeking, pilot study, review of evidence base (RoEB) and standardised effect size (SES) – each with important variations. The most frequently identified methods used to determine an important difference were the anchor, distribution and SES methods. No new methods were identified by this review beyond the seven pre-identified methods described earlier; however, substantial variations in the implementation of each method were detected. It is critical when specifying a target difference to decide whether the focus is to determine an important and/or a realistic difference as the appropriate methods vary accordingly. Some methods for determining an important difference within an observational study are not appropriate for specifying a target difference in a RCT (e.g. statistical hypothesis testing approach). Multiple methods for determining an important difference were used in some studies although the combinations varied, as did the extent to which results were triangulated.

Identification of triallists’ current practice

The two surveys regarding formal methods to determine the target difference in a RCT provided insight into current practice among clinical triallists.

Society for Clinical Trials survey

Of the 1182 members on the SCT membership email distribution list, 180 responses were received (15%). Awareness ranged from 69 (38%) for the health economic method to 162 (90%) for the pilot study method. Usage was lower than awareness and ranged from 16 (9%) for the health economic method to 133 (74%) for the pilot study method. The highest level of willingness to recommend was for the RoEB method (n = 132, 73%) and the lowest was for the health economic method (n = 28, 16%). Willingness to recommend among those who had used a particular method was substantially higher than across all...
respondents: the lowest level was for the opinion-seeking method \( (n = 40, 56\%) \) and the highest level was for the RoEB method \( (n = 118, 89\%) \).

UK- and Ireland-based triallist survey
Of the 61 surveys sent out, 34 \( (56\%) \) responses were received. Awareness of methods ranged from 97\% \( (n = 33) \) for the RoEB and pilot methods to only 41\% \( (n = 14) \) for the distribution method. All respondents were aware of at least one of the different formal methods for determining the target difference. Usage ranged from 24\% \( (n = 8) \) for both the distribution and health economic methods to 94\% \( (n = 32) \) for the RoEB method. Usage was substantially less than awareness for all methods except for the pilot study, RoEB and SES methods. The highest level of willingness to recommend was for the RoEB method \( (76\%, n = 26) \) followed by the SES method \( (65\%, n = 22) \), with the distribution method having the lowest level of willingness to recommend \( (26\%, n = 9) \). Based on the most recent trial \( (n = 33) \), all but three groups \( (91\%, n = 30) \) used a formal method. The vast majority \( (91\%, n = 30) \) stated that the target difference was one that was viewed as important by a stakeholder group. Just over half \( (61\%, n = 20) \) stated that the basis for determining the target difference was to achieve a realistic difference given the interventions under evaluation.

**Guidance on specifying the target difference in a randomised controlled trial**
Guidance was developed for specifying the target difference in a RCT. Additionally, guidance on reporting the sample size calculation was developed which includes a minimum set of items for reporting the specification of the target difference in the trial protocol and main results paper. A minimum set of items for reporting the specification of the target difference in the trial protocol and main results paper was developed.

**Conclusions**
The specification of the target difference is a key component of a RCT design. There is a clear need for greater use of formal methods to determine the target difference and for better reporting of its specification. Although no single method provides a perfect solution to a difficult question, methods are available to inform specification of the target difference and should be used whenever feasible. Raising the standard of RCT sample size calculations and the corresponding reporting of them would aid health professionals, patients, researchers and funders in judging the strength of the evidence and ensure better use of scarce resources.

**Further research priorities**
1. A comprehensive review of observed effects in different clinical areas, populations and outcomes is needed to assess the generalisability of the Cohen’s interpretation for continuous outcomes, and to provide guidance for binary and survival (time-to-event) measures. To achieve this, an accessible database of SESs should be set up and maintained. This would aid the prioritisation of research and help researchers, funders, patients and health-care professional assess the impact of interventions.
2. Prospective comparison of formal methods for specifying the target difference is needed in the design of RCTs to assess the relative impact of different methods.
3. Practical use of the health economic approach is needed; the possibility of developing a decision model structure that reflects the view of a particular funder (e.g. the Health Technology Assessment programme) and incorporates all relevant aspects, should be explored.
4. Further exploration of the implementation of the opinion-seeking approach in particular is needed. The reliability of a suggested target difference that would lead to a change in practice should be explored. Additionally, the impact of eliciting the opinion of different stakeholders should also be evaluated.
5. The value of the pilot study for estimating parameters (e.g. control group event proportion) for a definitive study should be further explored by comparing pilot study estimates with the resultant definitive trial results.

6. Qualitative research on the process of specifying a target difference in the context of developing a RCT should be carried out to explore the determining factors and interplay of influences.

**Funding**

The Medical Research Council UK and the National Institute for Health Research Joint Methodology Research Programme.
Chapter 1 Introduction and background

Introduction

The randomised controlled trial (RCT) is widely considered to be the best method for comparing the effectiveness of health interventions.\textsuperscript{1} But simply detecting any difference in the effectiveness of interventions may not be sufficient or useful: if the interventions differ to a degree or in a manner that is of little consequence in patient, clinical or economic (or other meaningful) terms, then the two interventions might be considered equal. If RCTs are to produce useful information that can help patients, clinicians and planners make decisions about health care, it is essential that they are designed to detect differences between the interventions that are meaningful.

Specifying the ‘target difference’, the difference that a trial sets out to detect, is also necessary for calculating the number of participants who need to be involved. It is an essential component of an a priori sample size calculation. Performed before the trial starts, this calculation determines the number of participants needed for the trial to reliably detect a difference of predetermined magnitude between interventions. Assuming that the trial manages to recruit the number of participants determined by the sample size calculation, the sample size calculation provides reassurance that the trial is likely to detect such a difference to a predefined level of statistical precision.

From both a scientific and ethical standpoint, selecting an appropriate target difference is of crucial importance. If a small target difference is determined as the appropriate measure of a meaningful difference between one intervention and another, the sample size calculation will usually indicate that a large number of participants are needed for a trial. Selecting a target difference that is too small to be meaningful may result in a large study identifying a difference between interventions that may have limited patient, clinical or economic significance. Conversely, when a larger difference is targeted, fewer participants will be required, but if the target difference is too large then this may lead to a small study incapable of confirming a smaller important difference. Either would be a wasteful (and unethical) use of data and resources. Furthermore, an undersized study may not usefully contribute to the knowledge base and could detrimentally impact on decision-making.\textsuperscript{2} The importance and impact of the target difference chosen can be demonstrated using two motivating examples (Table 1).

<table>
<thead>
<tr>
<th>Trial</th>
<th>Target difference\textsuperscript{a}</th>
<th>Result</th>
<th>Triallists’ interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norwegian Spine Study\textsuperscript{d}</td>
<td>Mean difference of 10 points in the Oswestry Disability Index (SD 18 points)</td>
<td>\textit{–8.4 points, 95% CI –13.2 to –6.6 points}</td>
<td>‘As there is no consensus based agreement of how large a difference between groups must be to be of clinical importance it is impossible to conclude whether the effect found in our study is of clinical importance’</td>
</tr>
<tr>
<td>Full-thickness macular hole and Internal Limiting Membrane peeling study (FILMS)\textsuperscript{d}</td>
<td>Mean difference of six letters in distance visual acuity (SD 12 letters)</td>
<td>\textit{4.8 letters, 95% CI –0.3 to 9.8 letters}</td>
<td>‘There was no evidence of a difference in distance visual acuity after the internal limiting membrane peeling and no-internal limiting membrane peeling techniques. An important benefit in favor of no-internal limiting membrane peeling was ruled out’</td>
</tr>
</tbody>
</table>

\textsuperscript{a} For both trials the statistical power and two-sided significance level were 80% and 5% respectively.
The Norwegian Spine Study was a RCT comparison between surgery with disc prosthesis and conservative non-surgical treatment (multidisciplinary rehabilitation) for patients with chronic lower back pain. Sample size calculations indicated that 180 participants were needed for the study to reliably detect a difference at 2 years of 10 points in the primary outcome measure, the Oswestry Disability Index. Ultimately, 179 participants were recruited and the analysis demonstrated a statistically significant difference of –8.4 points between the treatments in favour of surgery, less than the prespecified value although such a magnitude was comfortably within a plausible range of uncertainty [95% confidence interval (CI) –13.2 to –6.6 points]. Should clinical practice now change given the study’s finding? Was the observed difference of a sufficient magnitude to warrant the risk that surgery entails over the conservative treatment? The study investigators concluded that ‘As there is no consensus on how large a difference between groups must be to be of clinical importance it is impossible to conclude whether the effect found in our study is of clinical importance … our study underlines the need for such a consensus agreement.’ Thus, the study demonstrated a statistically significant difference between the interventions, but not one that met the authors’ own predefined target difference. Because the investigators had opted for a target difference that was difficult to justify, an otherwise well-conducted study did not produce a clear answer to the clinical question. Had a clear and defensible decision been reached before the start of the trial on what difference in Oswestry Disability Index can be considered clinically important, the research could have contributed more effectively to treatment decision-making.

The Full-thickness macular hole and Internal Limiting Membrane peeling Study (FILMS) compared peeling, or not, of the internal limiting membrane as part of surgery for a macular hole. The primary outcome was visual acuity and the study was designed to detect a difference of six letters on an eye chart, although no clear justification for this choice was reported. Once completed and analysed, there was a mean difference of five letters (95% CI –0.3 to 9.8) between groups in favour of peeling although this was (just) not statistically significant at the 5% level. Five letters actually corresponds to one additional line on the eye chart and carries intuitive appeal and arguably should have been the target difference. Interpreting their findings in light of their selected target difference the authors concluded that ‘There was no [statistical] evidence of a difference in distance visual acuity after the internal limiting membrane peeling and no-internal limiting membrane peeling techniques. An important benefit in favor of no internal limiting membrane peeling was ruled out.’ Although the study did partially answer the research question, the findings were not as definitive as they could, and perhaps should, have been if a smaller and more widely accepted target difference had been used.

This study also included an economic evaluation, the primary outcome of which was the incremental cost per quality-adjusted life-year (QALY) saved. The evaluation found that, on average, internal limiting membrane peeling was less costly and more effective, although neither difference was statistically significant. Moreover, there was a 90% probability that internal limiting membrane peeling would be considered cost-effective compared with no internal limiting membrane peeling at a threshold that the UK NHS would generally consider acceptable. Although such findings suggest that the balance of evidence favours internal limiting membrane peeling, interpretation of the results based on the chosen target difference suggests an inconclusive result.

Both of the examples above illustrate that the statistical evidence alone is an insufficient basis on which to interpret the findings of a study, which is ultimately dependent on how the study was designed and the wider context. The target difference is the difference that the study is designed to reliably detect. For example, a trial of nutritional supplementation was designed to detect a reduction from 50% to 25% in reported infections for critically ill patients receiving parenteral nutrition. The trial seeks to inform a decision (do we adopt the new treatment or stay with the existing treatment?) and this involves not only identifying the differences for single measures but also weighing up the benefits, harms and (resource) costs of an alternative course of action.
Surprisingly, given its critical impact, the determination of the target difference, as opposed to statistical approaches to calculating the sample size, has been greatly neglected.\(^7\)\(^8\) A variety of approaches have been proposed for formally determining what an important difference should be [such as the ‘minimum clinically important difference (MCID)’]; however, the relative merits of the available options for informing RCT design are uncertain.

**Project summary**

The Difference ELicitation in TriAls (DELTA) project described in this monograph aimed to address this gap in the evidence base. The study was prompted by the observation that a number of methods have been proposed that could potentially be used. Beyond the medical area, it is possible that there are further methods that might also be adopted. Reviews of subsets of methods have been conducted,\(^2\)\(^9\) but there is a need for a comprehensive review of the methods available that considers their use explicitly for the design of RCTs. There is also uncertainty about current practice among the clinical trials community and about the usage of methods in practice. The aim of this research project was to provide an overview of the current evidence through three main components:

1. A systematic review of potential methods for identifying a target difference developed within and outside the health field to assess their usefulness.
2. Identification of current ‘best’ trial practice using two surveys of trialists [members of the international Society for Clinical Trials (SCT) and leading UK clinical trial researchers from UK Clinical Research Collaboration (UKCRC) Clinical Trials Units (CTUs)].
3. Production of a structured guidance document to aid the design of future trials.

The research was commissioned and funded by the UK Medical Research Council (MRC) Methodology Research Programme (MRC005885 and G0902147 respectively) and was undertaken by a collaborative group in which the majority of members have extensive experience of the design and conduct of RCTs (both as investigators and as independent committee members) and have conducted methodological research related to RCTs (e.g. quality-of-life measurement, statistical methodology, reporting, surgical trials and economic evaluation).

The project website can be found at [www.abdn.ac.uk/hsru/research/assessment/methodological/delta/](http://www.abdn.ac.uk/hsru/research/assessment/methodological/delta/).

**Background**

**Why seek an important difference?**

The RCT examples in the previous section illustrate the difference between statistical evidence of a difference and what might be described as an important difference. It is worth considering why there should be a distinction between the two. Conventional RCT sample size calculations provide reassurance that RCTs will be able to detect a difference of a particular magnitude while also controlling for the risk of falsely concluding a difference when there is none. In statistical hypothesis testing terminology, this is controlling the possibility of falsely accepting that there is a difference when there is not (a type I error) and failing to reject the hypothesis of no difference when there is a genuine difference (a type II error). The risk of such errors exists because of the variability in outcome and the need for sufficient data to achieve statistical precision. However, if these errors are controlled, why would we not act on the basis of an observed statistical difference? The reason lies in the consequences of actions. Changing from one treatment to another can result in benefits or harms, or in costs, to the patient or to the care provider. Returning to the Norwegian Spine Study, all major surgical procedures incur a very small risk of serious morbidity or mortality related to anaesthesia. From the perspective of patients and clinicians,
is this risk warranted for the anticipated benefit gained? Health-care providers wish to know that the not insubstantial costs (resource costs as much as financial costs) related to surgery are worthwhile. This underlying issue has a long history in the statistical literature. Gosset, the statistician of ‘t-test’ fame, wrote as early as 1939:

> When I first reported on the subject, I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment [emphasis added].

The ‘advantage’ and associated costs drive our desire for a particular degree of certainty in decision-making; in Gosset’s case, working for Guinness, it was with regards to the brewing industry. The context should drive how certain we need to be that we have come to the correct conclusion and therefore there is no universal value that can be applied in all scenarios. If ‘a lot’ is at stake then we wish to be more certain than when the consequences of an incorrect decision are minor. The parallel with regards to health care is clear, although in this setting ‘advantage’ and ‘costs’ can be broadened to include not just economic definitions (based on the use of resources and profit or loss) but also benefits and harms for patients; this broadening of scope is needed to allow an informed decision about treatment to be made. It can be argued that a full answer to the question will therefore require all relevant benefits, harms and costs to be considered simultaneously. This naturally leads to decision modelling approaches (see How can an important difference be determined? and Chapter 2 for further details) that seek to link both the current evidence and the decision (including potential consequences for costs and health of making the wrong choice). The desire for an (clinically) ‘important’ difference can be viewed as a middle ground seeking to ensure that any harms and costs are incurred for a good reason: the patient receives spinal surgery over alternatives because we can realistically expect benefits to justify the associated risks and costs. Focusing on a benefit (or harm) as the most important outcome is a natural and intuitive, if imperfect, way to guide our decision.

The implication of imposing an interpretation of clinical importance onto the statistical result (observed difference) can be seen in Figure 1. Eight possible (although not exhaustive) scenarios, with regards to interpretation when comparing two interventions, are shown; they reflect possible combinations of statistical and clinical importance. Drawing an interval beyond which clinical importance is determined is obviously a simplification: in reality, a gradation in merit is more likely. Clearly, this approach does not provide a full answer as it is focused on a single outcome (ignoring impact on other outcomes) and is a simplification even for that outcome; however, it is still a useful and intuitive way to interpret the result. Determining what constitutes an important difference, however, is not as straightforward as it seems. A number of factors come into play. These include the potential for improvement (beyond the control intervention in a RCT setting) in the anticipated population as well as the type of outcome and whether the focus is restricted to this outcome alone or whether the wider impacts (e.g. benefits, harms and costs) of the interventions are to be considered. Expectations about what can be viewed as a realistic difference may also be brought into consideration [e.g. the performance of a similar intervention against the same (control) intervention]. Specification of an important difference is a particular challenge for certain types of outcomes.

**Inherent meaning versus constructed scales**

The difficulty of interpreting outcomes varies. Some outcomes have an inherent and tangible meaning, whereas others do not and are much more difficult to interpret. If spine surgery were to lead to an additional 10 in every 100 people being able to walk as opposed to being bedridden without surgery, we would be comfortable in noting its benefit (at least in terms of mobility). We would most likely view any genuine difference in mortality, however small, as being important. However, it would be valuable to know the impact of the benefits or harms associated with spine surgery on the health of the living
by considering an array of less dramatic outcomes. Many medical conditions are chronic and not life-threatening, although they clearly impact on the health and well-being of an individual. The impact of treatment is not likely to be ‘miraculous’ and gains are typically small or moderate and may not be readily detected or summarised by function or symptom measurement. How can we evaluate this impact on health? Health-related quality-of-life instruments address this need by generating latent scores that represent well-being or quality of life. Typically, multiple responses by a patient to a series of carefully designed and tested questions are used to generate a score for a particular dimension or construct of health, such as social well-being or ability to perform everyday tasks. However, the range of possible values for such dimensions is dependent on the framing of the individual scores and on the calculation of the weights used to generate the overall score. As a result, the impact in real terms of a difference in 1 point on the scale is unclear, as is the interpretation of any of the possible scores. Although undoubtedly useful for measurement and comparison of otherwise unmeasurable health-related outcomes, determining the difference in these scores that represents meaningful benefit or harm in real terms is often uncertain; this difficulty of interpreting quality-of-life measures is well known.

Along with seeking to validate and assess the inherent reliability of such a measure, there is a need for some assessment of its ability to detect a meaningful or important difference. Evaluating this (often described as measuring responsiveness) is now a recognised part of the process of validating a quality-of-life instrument. Interpreting scales that purport to measure latent constructs which cannot be directly measured, such as generic and condition-specific health, is particularly challenging. Determining an important difference in such a score for a sample size calculation presents a particular challenge.

**How can an important difference be determined?**

An influential development has been the concept of the MCID as a rationale to define an important difference. Originally, this was defined as the ‘the smallest difference … which patients perceive as

---

**Figure 1** Statistically and clinically important difference.

1. Statistically and clinically important difference in favour of A
2. Statistically and possibly clinically important difference in favour of A
3. Statistically but not clinically important difference in favour of A
4. Neither statistically nor clinically important difference
5. Non-statistical but clinically important difference possible
6. Statistically but not clinically important difference in favour of B
7. Statistically and possibly clinically important difference in favour of B
8. Statistically and clinically important difference in favour of B

Favours A | No difference | Favours B
---|---|---
Observed difference

<table>
<thead>
<tr>
<th>Statistically and clinically important difference in favour of A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistically and possibly clinically important difference in favour of A</td>
</tr>
<tr>
<td>Statistically but not clinically important difference in favour of A</td>
</tr>
<tr>
<td>Neither statistically nor clinically important difference</td>
</tr>
<tr>
<td>Non-statistical but clinically important difference possible</td>
</tr>
<tr>
<td>Statistically but not clinically important difference in favour of B</td>
</tr>
<tr>
<td>Statistically and possibly clinically important difference in favour of B</td>
</tr>
<tr>
<td>Statistically and clinically important difference in favour of B</td>
</tr>
</tbody>
</table>
beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient’s management’, but has also been referred to as the ‘minimum difference that is important to a patient’. Variations have been suggested such as the ‘minimally clinically important improvement’, the ‘patient acceptable symptom state’ and the ‘sufficiently important difference’, which seeks to adopt a wider perspective by taking into account cost, risk and harms. A variety of economic approaches have also been suggested from both a conventional and a Bayesian perspective (see Bayesian approaches to the design and analysis of randomised controlled trials) that explicitly explore the trades-offs between benefits, harms and cost. With respect to defining an important difference most work has been carried out on patient-reported outcomes, reflecting the belief that patients find it more difficult than clinicians to specify an important difference and also the challenge of interpreting constructed quality-of-life measures. All seek to ascertain a cut-point for a scale (whether directly measurable or latent) on which an ‘important’ difference or change can be separated from an ‘unimportant’ one.

How important differences relate to the design of randomised controlled trials

Conventional (Neyman–Pearson) approach
Randomised controlled trials allow the direct comparison of alternatives. They are overwhelmingly what can be described as Phase III (to use the pharmacological terminology) or confirmatory trials, which seek to answer the research question based on assessing ‘real’ outcomes in a substantial sample of people. Under the most common design, patients are randomly allocated to one of two treatments and statistical evidence of a difference between groups is determined (a superiority trial). Alternative designs in which the aim of the study is to demonstrate that a new treatment is equivalent or not inferior to another treatment (described in the literature as equivalence or non-inferiority trials respectively), are also possible. Irrespective of the design question, an a priori sample size calculation is required to provide reassurance that the study will provide a meaningful answer to the research question. The vast majority of trials adopt the same basic (sometimes called the Neyman–Pearson or statistical hypothesis testing) approach to calculating the sample size. This general approach for RCTs is outlined below.

To calculate the sample size for a trial, the researcher must strike a balance between the possibility of being misled by chance and the risk of not identifying a genuine difference. Standard practice is to pick one (often the most important outcome for a key stakeholder) or a small number of primary outcomes for which the sample size is conducted. A null hypothesis that the two treatments have equal effects is defined for a superiority trial. It is this null hypothesis that the trial is attempting to assess evidence against. Rejecting the null hypothesis when it is true (type I error) would lead to a concluding that one treatment is superior to another when in reality there is no real difference in treatment effect. The significance level of the test (α) is the probability of the occurrence of a type I error, that is, falsely concluding that there is a difference when there is not. Failing to reject the null hypothesis when it is false (type II error) would lead to a trial concluding that one treatment is not superior to another when in reality one treatment is superior. The probability of the occurrence of a such an error is β, which is equal to 1 minus the (statistical) power of the test. The smaller the specified difference to be detected the lower the power for a given sample size and significance level.

Once these two criteria are set, and the statistical tests to be conducted during the analysis stage are chosen, the sample size is determined depending on the magnitude of difference to be detected. This ‘target’ difference is the magnitude of difference that the RCT is designed to reliably investigate. It is worth noting that in practice the value used as a target difference might be one that is considered important or arrived at from another basis, such as what would be a realistic difference as observed in previous studies or a difference that leads to an achievable sample size. Different methods can be used to determine this target difference (Box 1; see also Chapter 2 for details). This issue is considered in more depth in Chapter 4 (see Specifying the target difference).

For an equivalence (or non-inferiority) trial, as opposed to a superiority trial, a range of values around zero will be required within which the interventions are deemed to be effectively equivalent (or not inferior) in
order to establish the magnitude of difference that the RCT is designed to investigate. The limits of this range are points at which the differences between treatments are believed to become important and one of the treatments is considered superior; the difference between one of these points and zero can be viewed as defining a minimum difference between treatments that would be important and which the study is designed to reliably investigate. The sample size calculation can also incorporate the possibility of an expected ‘real’, although non-important, difference into the calculation, reflecting that two treatments are very unlikely to have a (statistically) identical effect. A hybrid approach exists between the superiority and the equivalence/non-inferiority designs, sometimes called ‘as good as or better’. Under this approach a ‘closed’ testing (i.e. ordered) procedure is used that allows both an inferiority and a superiority question to

Methods for specifying an important difference

- **Anchor**: Under such an approach, the outcome of interest is ‘anchored’ by using either a patient’s or a health professional’s judgement to define an important difference. This may be achieved by comparing before and after-treatment and then linking this change to participants who had an improvement/deterioration. Alternatively, a contrast between patients can be made to determine a meaningful difference.
- **Distribution**: This covers approaches that determine a value based on the inherent imprecision in the measurement and therefore likely to represent a minimal level for a meaningful difference.
- **Health economic**: This covers approaches that make use of the principles of economic evaluation and typically involves defining a threshold value for the cost of a unit of health effect that a decision-maker is willing to pay and using data on the differences in costs, effects and harms to make an estimate of relative efficiency. This can be based on a net benefit or value of information approach that seeks to take into account all relevant aspects of the decision and can be viewed as implicitly specifying a target difference.
- **Standardised effect size**: Under such an approach, the magnitude of the effect on a standardised scale is used to define the value of the difference. For a continuous outcome, the standardised difference (most commonly expressed as Cohen’s d ‘effect size’) can be used. Cohen’s cut-offs of 0.2, 0.5 and 0.8 for small, medium and large effects, respectively, are often used. Binary or survival (time-to-event) outcome metrics (e.g. an odds, risk or hazard ratio) can be utilised in a similar manner although no widely recognised cut-offs exist. Cohen’s cut-offs approximate to odds ratios of 1.44, 2.48 and 4.27 respectively. Corresponding risk ratio values vary according to the control group event proportion.

Methods for specifying a realistic difference

- **Pilot study**: A pilot (or preliminary) study may be carried out when there is little evidence, or even experience, to guide expectations and determine an appropriate target difference for the trial. The planned definitive study can be carried out in miniature to inform the design of the future study. In a similar manner, a Phase II trial could be used to inform a Phase III trial.

Methods for specifying an important and/or realistic difference

- **Opinion-seeking**: This includes formal approaches for specifying the target difference on the basis of eliciting opinion (often a health professional’s although it can be a patient’s or other’s). Possible approaches include forming a panel of experts, surveying the membership of a professional or patient body or interviewing individuals. This elicitation process can be explicitly framed within a trial context.
- **Review of evidence base**: The target difference can be derived using current evidence on the research question. Ideally, this would be based on a systematic review of RCTs, and possibly meta-analysis, of the outcome of interest, which directly addresses the research question at hand. In the absence of randomised evidence, evidence from observational studies could be used in a similar manner. An alternative approach is to undertake a review of studies in which an important difference was determined.

**BOX 1 Methods for specifying an important and/or realistic difference**

- Anchor: Under such an approach, the outcome of interest is ‘anchored’ by using either a patient’s or a health professional’s judgement to define an important difference. This may be achieved by comparing before and after-treatment and then linking this change to participants who had an improvement/deterioration. Alternatively, a contrast between patients can be made to determine a meaningful difference.
- Distribution: This covers approaches that determine a value based on distributional variation. A common approach is to use a value that is larger than the inherent imprecision in the measurement and therefore likely to represent a minimal level for a meaningful difference.
- Health economic: This covers approaches that make use of the principles of economic evaluation and typically involves defining a threshold value for the cost of a unit of health effect that a decision-maker is willing to pay and using data on the differences in costs, effects and harms to make an estimate of relative efficiency. This can be based on a net benefit or value of information approach that seeks to take into account all relevant aspects of the decision and can be viewed as implicitly specifying a target difference.
- Standardised effect size: Under such an approach, the magnitude of the effect on a standardised scale is used to define the value of the difference. For a continuous outcome, the standardised difference (most commonly expressed as Cohen’s d ‘effect size’) can be used. Cohen’s cut-offs of 0.2, 0.5 and 0.8 for small, medium and large effects, respectively, are often used. Binary or survival (time-to-event) outcome metrics (e.g. an odds, risk or hazard ratio) can be utilised in a similar manner although no widely recognised cut-offs exist. Cohen’s cut-offs approximate to odds ratios of 1.44, 2.48 and 4.27 respectively. Corresponding risk ratio values vary according to the control group event proportion.

Methods for specifying a realistic difference

- Pilot study: A pilot (or preliminary) study may be carried out when there is little evidence, or even experience, to guide expectations and determine an appropriate target difference for the trial. The planned definitive study can be carried out in miniature to inform the design of the future study. In a similar manner, a Phase II trial could be used to inform a Phase III trial.

Methods for specifying an important and/or a realistic difference

- Opinion-seeking: This includes formal approaches for specifying the target difference on the basis of eliciting opinion (often a health professional’s although it can be a patient’s or other’s). Possible approaches include forming a panel of experts, surveying the membership of a professional or patient body or interviewing individuals. This elicitation process can be explicitly framed within a trial context.
- Review of evidence base: The target difference can be derived using current evidence on the research question. Ideally, this would be based on a systematic review of RCTs, and possibly meta-analysis, of the outcome of interest, which directly addresses the research question at hand. In the absence of randomised evidence, evidence from observational studies could be used in a similar manner. An alternative approach is to undertake a review of studies in which an important difference was determined.
be potentially answered in a single study. As with the standard superiority and equivalence/non-inferiority designs, a definition of the difference to be detected is still needed.

Once the target difference, or for an equivalence/non-inferiority trial the limit(s) of equivalence, is determined then the method of estimating the sample size will depend on the proposed statistical analysis, trial design (e.g. cluster or individually randomised trial) and statistical properties specified (e.g. agreement for paired data). The general approach is similar across studies under the Neyman–Pearson approach. Implicitly, the study is designed as a ‘stand-alone’ evaluation (without the resort to any external data) to answer the research question.

Other statistical approaches to defining the required sample size are Fisherian, Bayesian and decision-theoretic Bayesian approaches, along with a hybrid of both the Bayesian and the Neyman–Pearson approaches. Economic-based methods tend to follow a decision-theoretic approach, with the types of benefits, harms and costs included reflecting the perspective of the research funder and the values used and the way that they are combined (typically a net benefit approach) reflecting the context of the study. Despite the existence of these different methods a recent review of RCT sample size calculations identified only the Neyman–Pearson approach in widespread usage. Regardless of the statistical method used the key issue is what magnitude of a difference is of practical interest and one that the study should be designed to detect. As will be described in The relevance of decision theory-based models, the intent of the economic-based approaches can be different as they tend to focus on maximising some measure of efficiency. Prior evidence may be informally guide the process though explicit incorporation of prior evidence in the sample size calculation for a RCT is also possible.

Bayesian approaches to the design and analysis of randomised controlled trials

Much has been written about the benefits of Bayesian approaches to the design and analysis of RCTs. They allow incorporation of current evidence (represented as a prior distribution) with new evidence to produce a summary of the overall evidence (posterior distribution). With substantive computing power now readily available, their use has grown because of the flexibility of the approach and intuitive appeal. One area in which the value of using Bayesian methods has been highlighted is for sample size determination of RCTs. Standard sample size calculations assume that the imputed values are ‘known’ (i.e. without any uncertainty); in reality, inputs such as event proportions or variances are estimates of what is anticipated to happen in the trial. A Bayesian framework provides a natural framework in which uncertainty about the inputs can be incorporated (as a prior distribution) and the impact of uncertainty on the estimates can be included to evaluate a predictive distribution of power. Following such an approach still requires specification of the difference to be detected. The use of Bayesian methods for design and analyses does not avoid the need for interpretation of the importance of the results. Although they allow a more comprehensive, although also complex, representation of evidence through posterior distributions, consideration of the relevance of the findings is still necessary. To state that the outcome (posterior) distribution differs between treatments, or that the distribution of the mean difference excludes zero, does not clarify whether we should act on the basis of the result. Correspondingly, ‘Bayesian p-values’ (e.g. the posterior predictive probability that the outcome is better for treatment A than for treatment B) are similar to frequentist p-values in that for the same reasons (as noted above) they are an insufficient basis on which to make a decision. One approach to this issue is to define an ‘indifference zone’ in the outcome distribution in a similar manner to that adopted in equivalence trials. The CHART trial provides an example of using clinicians’ opinion (both a range of equivalence and expectations regarding a realistic difference) regarding 2-year survival for trial design and data monitoring; this approach could be readily used a priori to determine the corresponding sample size within a Bayesian framework. A Bayesian approach can naturally be extended into a formal decision model, which enables consequences of the decision to be formally incorporated into the analysis.

Adaptive designs

Adaptive designs are studies that are set up to formally modify the trial design, according to accumulated data. They are used most commonly within the pharmaceutical setting for Phase II studies in which the
primary outcome can be assessed in the short term after randomisation. A decision about adapting the trial design can therefore be made while the study is still recruiting. Such adaptive designs can be viewed as a form of decision model. Typically, they need specification of what is viewed as important (e.g. outcome values that would lead to the intervention not being worthwhile to pursue) at the outset to control the process of adaptation. Common types of adaptation are discarding a treatment arm as it is unlikely to achieve the desired or optimal effect (e.g. Phase II dose-finding study) or stopping a trial if benefit/futility has been shown (sometimes called group sequential trial). Although the latter may have a decision rule that involves only a statistical rule of precision (e.g. p-value or Bayesian posterior probability), decision-making of Data Monitoring Committees and study investigators will involve more than the pure statistical results of accumulated data, with consideration of the consequences of making a decision such as stopping early. Some Phase II designs adapt based on a maximum allowable sample size and/or decision rules based on Bayesian posterior probability. Trend analysis models have been proposed that adjust the sample size according to the trend in the current data. They are not commonly used because of the risk of bias (similar to play-the-winner designs) with regards to inflated type I and type II errors and because knowledge of the adaptation reveals the direction, and possibly magnitude, of the observed difference at that time point. In contrast, Phase III trials will typically be designed to detect or exclude a particular (target) difference.

The relevance of decision theory-based models
As highlighted earlier in Why seek an important difference?, the decision that the study is designed to inform, such as treating patients with lower back pain with surgery or conservative treatment, should guide the design of the study. Decision theory-based models allow the decision and corresponding losses (consequences and costs) to be quantified and incorporated in an extended model that incorporates the clinical evidence and any other relevant data. The underlying rationale is that any decision should be made to maximise utility (value) given current evidence. There has been much discussion about whether RCT results should be analysed on their own or whether a decision modelling approach should be adopted that incorporates the decision (and corresponding consequences) which the study is seeking to inform. For example, should the spinal surgical treatment (from the Norwegian Spine Study) be analysed on its own or should prior evidence be incorporated into the model, all with estimation of the impact of changing the treatment. It has been noted that investigators who design and carry out a study are generally not the same as those who make decisions. As such, there is a practical challenge in implementing such an approach. However, models have been proposed that specifically address the design of RCTs: they seek to determine the optimal sample size given current evidence and the estimated consequences of the actions. Although clearly of a different nature to other approaches, such as the MICID and variants, these models may be viewed as informing the same decision (but taking into account other factors); as such, they are also considered in this monograph.

Summary
In practice, specification of the target difference is often not based on any formal concepts and in many cases (at least from trial reports) appears to be determined according to convenience or some other informal basis. A variety of methods have been proposed to formally determine a target difference (including those for the MCID and its variants). These methods take different approaches to determining the target difference: some seek to find the most realistic estimate of the effect based on current evidence, some focus on an important difference, whereas others seek to incorporate the cost and consequences into a single analysis.
Chapter 2 Systematic review of methods for specifying a target difference

Introduction

The aim of the systematic review was to identify methods for specifying a target difference for a RCT. In this chapter, the methods and results of the systematic review are described. For each method, important variations in approach, a summary of the included studies and practical considerations with respect to its use, are described. The findings across all methods are summarised in the discussion section of this chapter.

Methodology of the review

Search strategy

Both medical and non-medical literature were searched given that the underlying issue is relevant to both areas and useful methods could be reported only in the non-medical literature. Search strategies and databases searched were informed by preliminary scoping work. Full details of the databases searched and search strategies used can be found in Appendix 2. The biomedical and social science databases searched were MEDLINE, MEDLINE In-Process & Other Non-Indexed Citations, EMBASE, Cochrane Central Register of Controlled Trials (CENTRAL), Cochrane Methodology Register, PsycINFO, Science Citation Index, American Economic Association’s electronic bibliography (EconLit), Education Resources Information Center (ERIC) and Scopus for in-press publications. All were searched from 1966 or the earliest date of the database coverage and searches were undertaken between November 2010 and January 2011. There was no language restriction.

The search strategies aimed to be sensitive but, because of the paucity of relevant subject indexing terms available in all databases, relied mostly on text word and phrase searching using appropriate synonyms. It was anticipated that reporting of methods in the titles and abstracts would be of variable quality and that, therefore, a reliance on text word searching would be inadvisable. Consequently, several other methods were used to complement the electronic searching including checking of reference lists, citation searching for key papers using Scopus and Web of Science and contacting experts in the field.

In addition, textbooks covering methodological aspects of clinical trials were consulted. These were identified by searching the integrated catalogue of the British Library as well as the catalogues of publishers of statistical books, including Wiley, Open University Press and Chapman & Hall, for relevant books published in the last 5 years (2006 onwards). Additionally, key clinical trial textbooks as determined by the steering group were reviewed along with the International Conference on Harmonisation (ICH) tripartite clinical trials guidelines. The corresponding references are given in Appendix 2.

Inclusion and exclusion criteria

Studies reporting a method that could potentially be used to specify the target difference were included in this review. Methods may seek to determine an important and/or a realistic difference. All study designs were eligible for inclusion as it was considered unnecessary and potentially restrictive to limit them. In addition, studies using methods in hypothetical scenarios were eligible for inclusion, as were studies implicitly specifying a target difference by determining an optimal study sample size. All included studies were required to base their assessment on at least one outcome of relevance to a clinical trial or an outcome that could be used for this purpose.
Exclusion criteria were:

- studies failing to report a method for specifying a target difference (or equivalent)
- systematic reviews of methods for specifying the target difference
- studies reporting only on sample size statistical considerations (e.g. a new formula for sample size calculation)
- studies considering a metric (e.g. risk ratio or number needed to treat) without reference to how a difference could be determined.

Full-text papers were obtained for all titles and abstracts identified by the search strategy that were determined to be potentially relevant to the review. One of four reviewers (JH, TG, KH and TA) screened abstracts and data extracted information from the full-text papers of relevant studies. The reviewers undertook a practice sample of abstracts to ensure consistency in the screening process. When there was uncertainty regarding whether or not an article should be requested for full-text assessment or included in the review, the opinion of another member of the team (JC) was sought and the article was discussed until consensus on inclusion or exclusion was reached.

The titles and abstracts of all studies requested for further full-text assessment were provisionally categorised according to the known methods for eliciting a target difference. This categorisation process used the information provided in the abstracts of each study that was requested for further full-text assessment, to provisionally determine which known method (or methods) it might use. Based on the full-text assessment, final classification of the articles according to the methods used was carried out. A register of studies meeting the inclusion criteria was organised using Reference Manager bibliographic software version 12 (Thomson ResearchSoft, San Francisco, CA, USA), using the keyword facility to classify articles by type, reference source and methodology.

**Data extraction**

Data were extracted from all included studies to help summarise the variation and range of applicability of each method. Data extracted included (when reported):

1. what is being measured (e.g. a single measure of clinical effectiveness or safety, a composite measure of clinical effectiveness and/or safety, a measure of overall or disease-specific health or a cost-effectiveness measure)
2. the type of outcome measure (e.g. binary, ordinal, continuous or survival outcome)
3. the relevant summary measure reported (e.g. mean difference, risk ratio or absolute risk difference)
4. the size of the sample used to elicit a value for the difference
5. the perspective used to define the difference (e.g. patients’ or clinicians’).

Data were also extracted on the following details (when reported):

- the context in which the difference was elicited (e.g. real or hypothetical RCT)
- methodological details and noteworthy features (e.g. unique variations).

Some data items were relevant only to certain methods. Additional data were extracted on specific information relevant to each particular method. No generic data extraction form was used across all methods.

**Method of analysis**

A narrative summary description of each method found was produced by reporting extracted details from the data, categorising the key characteristics of each method and assessing the strengths and weaknesses of each method to aid the development of guidance.
Results

Search results
The search of databases identified 11,485 potentially relevant studies after deduplication. The number of studies obtained from each database is provided in Table 2. A depiction of the screening process itself and the number of studies pertaining to each method at key stages of the process is provided in Figure 2.

### Table 2

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of titles and abstracts identified by the search strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDLINE/MEDLINE In-Process &amp; Other Non-Indexed Citations/EMBASE</td>
<td>7189</td>
</tr>
<tr>
<td>CENTRAL</td>
<td>487</td>
</tr>
<tr>
<td>Science Citation Index</td>
<td>1898</td>
</tr>
<tr>
<td>EconLit</td>
<td>255</td>
</tr>
<tr>
<td>PsycINFO</td>
<td>1367</td>
</tr>
<tr>
<td>Cochrane Methodology Register</td>
<td>10</td>
</tr>
<tr>
<td>ERIC</td>
<td>158</td>
</tr>
<tr>
<td>Scopus</td>
<td>121</td>
</tr>
<tr>
<td>Total</td>
<td>11,485</td>
</tr>
</tbody>
</table>

*a Number of titles and abstracts after deduplication.

11,485 titles and abstracts identified from primary search

10,051 excluded at abstract screening stage

1434 selected for full-text assessment

667 articles excluded

10 articles identified from other sources

777 studies included
- 253 anchor method only
- 171 distribution method only
- 13 health economic method only
- 60 opinion-seeking method only
- 5 pilot study method only
- 22 reviews of the evidence base method only
- 37 standardised effect size method only
- 216 more than one method used*

Figure 2 The screening process. a. For a breakdown of studies in which more than one method was used in combination, please see Combination of methods.
Following the screening of titles and abstracts, a total of 1434 articles were retained for further full-text assessment and were provisionally categorised based on the different methods used, using the information available in the title and/or abstract. After full-text assessment and the identification of additional studies from other sources, including citations and consultation with experts, a total of 777 studies were included in the review. The full list of included studies is available in Appendix 3.

Hand-searching of specific journals was not undertaken. Fifteen clinical trials textbooks along with the ICH clinical trials guidelines were hand-searched (see Appendix 2). The full-text papers of five identified systematic reviews of methods for determining an important difference were retained to provide additional key references for the review. The output of one author related to a particular approach (decision-analytic economic evaluation framework) (Andy Willan, SickKids Institute, Toronto, ON, Canada) was searched after the main search strategy identified several articles on this method. Another researcher (Robert Gatchel, University of Texas, Arlington, TX, USA) contacted the principal investigators of this project following a letter published by the principal investigators on this subject. The titles and abstracts of the relevant articles by this expert had already been identified by the search strategy and requested by the reviewers for full-text assessment. One of these articles was subsequently included in the review. Eight further studies were identified from citations of full-text studies screened or were sent for assessment by members of the steering group. The included studies were categorised as using one (or more) of seven methods: anchor, distribution, health economic, opinion-seeking, pilot study, review of evidence base (RoEB) or standardised effect size (SES).

There was substantial variation in terminology across and within methods even when the same concept was being addressed. In particular, various terms were used to describe the concept of the M(C)ID. The terminology varied among (and sometimes within) papers, particularly for older papers. Common variations in terminology include studies identifying a ‘difference’ or a ‘change’, and whether this was ‘meaningful’, ‘significant’, ‘perceivable’ or ‘important’. Some studies specifically referred to ‘clinical’ importance or significance, whereas others did not. In one instance, justification was given for not including a reference to ‘clinically’ important difference [e.g. minimal important difference (MID)], arguing that the word ‘clinically’ should no longer be part of the term as it puts the focus on more objective ‘clinical’ measures of change rather than measures that are important to patients. The words ‘minimum’, ‘minimal’ and ‘minimally’ were also added in many cases to the terms used, with minimal being the most common. The definition of an ‘important’ difference also varied. Piva and colleagues defined the minimum detectable change (MDC) as ‘the amount of change needed to be certain, within a defined level of statistical confidence, that the change that occurred was beyond that which would be the result of measurement error’. Cousens and colleagues defined the ‘clinically important change’ as a change ‘that, by consensus, is deemed sufficiently large to impact on the patient’s clinical status’.

**Anchor method**

**Brief description of the anchor method**

In its most basic form, the anchor method evaluates the minimum (clinically) important change in score for a particular instrument. This is established by calculating the mean change score (post minus pre) for that instrument among a group of patients for whom it is indicated (using an additional measurement – the ‘anchor’) that a minimum clinically important change (or difference) has occurred. The word ‘minimum’ reflects the separation of patients according to the amount of change, with the smallest amount of change viewed as ‘important’ calculated. In evaluating a patient’s progress (e.g. before and after a given treatment), a clinician could use this value as an indication of the expected change in score of an instrument among patients who have indicated on the anchor measurement that they consider themselves (or are considered by someone who completes the anchor measurement on their behalf) to have undergone an improvement within a particular time frame.
Variations on the anchor method

An anchor method example is given in Box 2. However, various implementation approaches of the anchor method have been used (i.e. varying the number of points and labels attached) along with variations in how an anchor is used to determine an important difference [typically the M(C)ID]. One of the most common variations is not only to consider the mean change score for improved patients but also to take account of the relative change compared with another subset of those undergoing assessment (who have been tested using both the same instrument and the anchor) who reported no change over time.54–58 The results for both groups can be compared, potentially allowing the calibration of a resulting MCID value by adjusting for the mean value found for the ‘no change’ group (i.e. variation in score that is unimportant).

Yalcin and colleagues59 refer to the original method as ‘within-patient’ change and this variation as ‘between-patient’ change. However, these terms are not solely used for these variations and can therefore be confusing. The variations are depicted in Figure 3. Another very common variation is to consider the percentage change score in the instrument under consideration,60 rather than simply using the absolute change score. Including the percentage change score as well as the absolute score has been justified as possibly reducing variation, thereby ensuring that the MCID values obtained are more robust. Other variations reported used the median change rather than the mean change since baseline.61–63

Diagnostic accuracy methodology has also been used to define the MCID with an anchor method. The anchor response is used as the reference standard, and the sensitivity and specificity of a cut-point difference in the instrument under consideration is assessed against the anchor definition. Use of the cut-off that optimised together sensitivity (proportion of those who had experienced an important difference correctly identified) and specificity (proportion of those who did not have an important difference correctly identified) was almost universal. In a few cases this was not carried out, for example using a cut-off to maximise sensitivity over specificity, unless the point nearest the left-hand corner of the sensitivity versus (1-specificity) plot gave a particularly low value for sensitivity,64 or using an 80% specificity rule.65,66 In other studies, a receiver operating characteristic (ROC) curve analysis was undertaken but was not used to calculate the MCID.67,68 In other instances, diagnostic accuracy data were reported but a ROC curve approach was not used to determine the estimate.69 When multiple MCID estimates were generated (e.g. from different anchor definitions or by using different anchors), an average or triangulation of the results was sometimes used.70–72 Other less commonly used methods to derive a MCID value from an anchor-based method included various types of regression models [e.g. analysis of variance (ANOVA), logistic regression, Rasch analysis, classification and regression trees (CART) analysis, mixed-effects models

BOX 2 Anchor method example: the Clinical Global Impression anchor with a seven-point ordinal scale.94
### Study design

The study sample (30 patients) complete the instrument of interest (O) both before and after the treatment is received. The change score is calculated for each patient:

\[ O_{ch} \]

The anchor (A) is a five-point ordinal scale:

1 2 3 4 5

It is only completed after treatment for each patient:

\[ A \]

<table>
<thead>
<tr>
<th>Study design</th>
<th>Estimating the MCID – select minor variations (i–iv) in approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. Point 4 on the anchor (A) is chosen as the cut-off for denoting an important change (improvement only). The mean change score (( O_{ch} )) on the instrument of interest (O) for all those who selected point four on A, is the MCID for A.</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Anchor method illustrative example" /></td>
<td></td>
</tr>
<tr>
<td>ii. Anything above point 3 (i.e. points 4 and 5) on the anchor (A) is chosen as denoting an important change (improvement only). The mean change score on the instrument of interest (( O_{ch} )) for all those who selected points four or five on A, is the MCID for O.</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Anchor method illustrative example" /></td>
<td></td>
</tr>
<tr>
<td>iii. Point 4 on the anchor (A) is chosen as the cut-off for denoting an important change (improvement only) though the estimate is calibrated by adjusting for change in scores among the ‘no change’ patients (point 3). The mean change score on the instrument of interest (( O_{ch} )) for those who selected point 4 on A, minus the mean change score for those who selected point 3 on A, is the MCID for O.</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Anchor method illustrative example" /></td>
<td></td>
</tr>
<tr>
<td>iv. On the anchor (A), the minimum change in either direction (i.e. points 2 and 4) is chosen as denoting change (deterioration or improvement). This ‘change’ is calibrated by accounting for change in scores among the ‘no change’ patients (point 3). The mean score on the instrument of interest (( O_{ch} )) for all those who selected points 2 or 4 on A, minus the mean score of O for all those who selected point 3 on A, is the MCID for O.</td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Anchor method illustrative example" /></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3** Anchor method illustrative example.
and linear discriminant analysis. Harman and colleagues developed a logistic regression model for the probability of experiencing a negative life event based on prior mental health levels. A substantially different approach was proposed by Redelmeier and colleagues in which other patients were used as anchors on which a patient could rate his or her own health (or health improvement) in comparison to others.

Aside from variation in the method used to elicit a MCID value, there was variation in the anchor itself and in the ‘transition question’ used to determine whether patients had improved or not. A 15-point Likert scale was often used as the anchor, based on the method established by Jaeschke and colleagues, although five-point and seven-point scales were also widely used. A visual analogue scale (VAS) can be used (e.g. with <10 mm considered ‘unchanged’ and >30 mm considered more extreme than minimal change). Existing instrument questions were also adapted, including the Short Form questionnaire-36 items (SF-36) global change score, Disabilities of the Arm, Shoulder and Hand (DASH) and the (modified) Rankin score. For example, Khanna and colleagues used multiple anchors, including the SF-36 global change score and four additional items from the SF-36 relating to walking and climbing stairs. Several studies utilised an item from the instrument that was being assessed for MCID values as the anchor question for other items within the same instrument. For example, Colwell and colleagues used the pain frequency and pain severity items within the Gout Assessment Questionnaire (GAQ) as anchors by which to gauge the MCID values for other items within this outcome.

In many cases, the original anchor categories were later merged because of sample constraints (e.g. poor literacy among respondents or an insufficient number of patients in a particular category of improvement to enable analysis to be undertaken separately for that category). In some cases, if the anchor considered deteriorating states as well as improvement, the sign on the deterioration point of the Likert scale was changed and its participants were merged with the improving patients to give a larger number of patients in the anchor groups (although this makes the strong assumption that MCID values will be identical for improvement and deterioration). For ROC curve analysis to be used within the anchor method, the anchor question must be dichotomous (those experiencing a MCID vs. those who did not experience a MCID). There is therefore a need to split the anchor in some way. This often requires an arbitrary decision about what points on the Likert scale constitute important change in comparison with no real change. Consensus between two colleagues was used in one instance to derive a definition for important change, along with the Rasch method to find ‘natural’ cut-points for the data.

The anchor question was most often posed to patients alone although in some cases the clinicians’ views alone were used, even in instances when the instrument under consideration was a patient-reported outcome. The opinion of a clinician was also used instead of a patient-reported outcome, or in conjunction with a patient’s score, either to determine the extent of agreement between patient and clinician scores or to average the score (e.g. to reduce potential patient biases) by incorporating both patient and clinician views. Parents’ views were also occasionally used, especially for studies of paediatric outcomes. Anchors do not necessarily have to be based on individuals’ opinions; ‘objective’ measures of improvement can be used (e.g. ≥5 mm healthy toenail growth). Other studies used retrospective analysis, for example the mean change score among those who were later hospitalised compared with those who were not later hospitalised. Readmission and/or death was also considered in this way. If a treatment of known efficacy was the intervention that patients received, the mean change score was used as patients were expected to show clinically important improvements over time. It is worth noting that the anchor method was not always successful in deriving values for interpreting an important difference.

Summary of anchor studies
A total of 447 studies were found that used the anchor method; 194 of these also used another method (see Combination of methods for details). Several studies had split their sample into separate cohorts (e.g. if the data came from two separate clinical trials). In most instances, a rationale for calculating the MCID was not explicitly given. As many studies looked at instrument development, calculation of

© Queen’s Printer and Controller of HMSO 2014. This work was produced by Cook et al. under the terms of a commissioning contract issued by the Secretary of State for Health. This issue may be freely reproduced for the purposes of private research and study and extracts (or indeed, the full report) may be included in professional journals provided that suitable acknowledgement is made and the reproduction is not associated with any form of advertising. Applications for commercial reproduction should be addressed to: NIHR Journals Library, National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre, Alpha House, University of Southampton Science Park, Southampton SO16 7NS, UK.
the MCID often formed part of a wider array of reliability and validity testing of new instruments; other studies merely noted that for a particular instrument (or use of an established instrument for a particular disease/condition) the MCID had yet to be established. Future sample size calculation was very rarely cited in the rationale for calculating the MCID.116

Across all its different possible forms, the anchor method was used in a wide range of specialties including, but not limited to, surgery,117 orthopaedics,70 paediatrics,109 rheumatology,65 stroke rehabilitation,118 cancer,119 urology,120 respiratory medicine,75 mental health67 and emergency medicine.121 Almost all cases looked at quality-of-life measures, particularly with regard to pain,85 function and mobility.122,123 The method is not used exclusively for establishing the MCID for new or existing quality-of-life tools; non-scale-based outcomes were also considered. These usually related to patient physical functioning (in many cases in conjunction with a wide range of instruments for a particular condition being assessed at the same time), for example a 5-minute walk test, 1-minute stair climbing,94 comfortable gait speed77 and (more unusually) the number of palms of your hand it would take to cover up all of the psoriasis on your body.61 The method can be used for acute124 or chronic125 conditions.

Practical considerations for use of the anchor method
A number of common issues with regard to applying the anchor method were highlighted. In particular, the generalisability of a determined MCID value was often queried. In some instances this was simply in relation to the size of the sample,75,94,105,118,121,122 which in itself is not an problem exclusive to studies determining important difference values. However, in other instances the range of applicability of the resulting MCID value may be compromised.49,94,118,121,126 For example, a MCID value for a chronic progressive illness may not be applicable to a newly diagnosed population if the duration of illness among most participants surveyed to determine the MCID value is far longer.116,127–130 Although it is, in principle, possible to account for this in any future anchor method studies (e.g. by providing a breakdown of MCID values for subgroups based on duration of illness), in practice there may be more complex confounding issues. Long-term sufferers of a chronic condition may have different expectations of change than a newly diagnosed population. In addition, severity of illness may not necessarily be directly correlated with duration, and this may be a more important aspect to consider. In fact, the effect of severity of illness on identified MCID values was often examined by study authors; however, not all studies found a statistically significant effect.63,71,105,127,131

Several studies that had considered separately MCID estimates for improvement and deterioration noted differences in the values derived for these groups.75,88,95,121,126 Although this, as with other generalisability concerns, may be related to participants’ expectations or regression to the mean, it is important to consider whether an assumption of linearity, with no change centred on zero difference, is appropriate. The decision to merge both improvement and deterioration groups (e.g. because of a small sample size) to derive one singular ‘change’ value instead may be inappropriate if the size of an important change is different for improving participants and deteriorating participants.49,54,95,122

The validity of the anchor used was also cited as being potentially problematic from a methodological point of view. Likert scales were most commonly used as anchors to derive the MCID value, but there was considerable variation in the number of point options provided within the Likert scale. Differing sizes of anchor scales may be appropriate as the change experienced by participants, and their ability to differentiate, will vary between contexts and study populations. In addition, there may be other factors that influence the size of the Likert scale used. For example, one study noted that, if the population of interest has poor literacy, a Likert scale with a large number of options may be less suitable than one with fewer points for respondents to consider.132 Several studies using Likert scales with a large number of options later had to merge multiple options together because so few participants had selected each option, thereby retrospectively creating a smaller Likert scale that might have been more suitable in the first place.
A more fundamental consideration when using the anchor method is how to decide on an appropriate cut-off point to apply to the anchor responses. Although any formal approach for determining this cut-off was seldom mentioned as part of the study methods, several study authors pointed out the difficulty of choosing a particular value. For anchor studies using a ROC curve approach, choosing a cut-off point on the Likert scale to dichotomise the sample into those who did and those who did not experience change is essential, but the choice of cut-off is arbitrary and may compromise the extent to which an anchor method detects or fails to detect important change. In some instances, multiple cut-offs are analysed, or mean change values are reported for every different point on the Likert scale. This is useful as it allows those wishing to utilise MCID values to use the value most suitable for their own needs. Conversely, it could be argued that, by not choosing an appropriate cut-off and instead providing every possible value of clinical importance from the anchor responses, studies fail to provide guidance for which MCID values are more likely to be realistic overall. Some papers focused on only one choice of definition; this was particularly the case for papers that had used a ROC approach. Reporting of anchor methodology was sometimes insufficient to allow reanalysis of the data if an alternative anchor definition of important change is desired.

More generally, the validity of both the outcome under consideration and the anchor tool being used to define its MCID were noted as important. Validity issues apply more widely to the development of any new outcome for health measurement. For example, given an ambulatory population, the outcomes used must be able to detect small changes and be free from ceiling effects, whereas for a population with severe disease, the possibility of floor effects could be more relevant. Information on an outcome’s ability to detect change in the first place complements MCID values and helps determine whether or not a point estimate of change is suitable. The inability to achieve a high level of discrimination (e.g. good diagnostic performance if a ROC curve approach is used) would suggest that a MCID for a particular outcome cannot be reliably determined using this anchor. Possible reasons for poor performance include a disconnection between the anchor and the outcome (e.g. difference in perspective) as well as the properties of the outcome itself (e.g. low reliability).

With regard to the population of interest, the generalisability of the study population is not the only aspect of the study conduct that requires methodological consideration. One commonly cited difficulty in using the anchor method relates to the potential for participants to exhibit recall bias over the course of the study period. Although this concern is not exclusive to the calculation of a MCID, it is relevant to consider the possible impact on the study estimate, particularly in instances in which there may be additional concerns about study methodology (e.g. variation in length of participant follow-up). There are examples of methods used to counteract this bias. For example, Wyrwich and colleagues, at each data collection session, asked participants what they were doing that day. At the following session, to aid their memory, their answer from the previous session was relayed to them as a reference point to help them recall how they had felt at the previous point in time. Response shift may also be problematic, as participants’ perceptions of what amount of improvement they would be satisfied with may change during the course of their treatment, and their expectations may not be consistent over time.

The method proposed by Redelmeier and colleagues, in which other participants act as the reference point avoids recall bias because all data can be collected at the same time. However, the data analysis is more complex and this approach is not universally appropriate for calculating the MCID in all contexts. For example, participants might find it difficult to discuss particularly sensitive or private health issues with others. A more general consideration with regard to patient involvement is the value of their MCID scores in comparison to (most commonly) clinicians’ views of the extent to which important change had occurred. The development of methods to elicit an important difference has occurred alongside increasing value being placed on the experience of patients and their increasing autonomy as experts of their own illness. However, in some circumstances (e.g. paediatrics) it may be appropriate to consider the views of others in addition to (or even superseding) the views of patients. It has also been noted that clinicians’ views of whether or not important change has occurred may not be independent of patients’ own views on this. Patients’ own views on their progress may be shaped by the views of their clinician. It
is therefore important to note the methods used in data collection with regard to whether or not clinicians and/or patients were blinded to the views of others when responding to the question of whether or not the patient had experienced important change. This would establish whether or not the values elicited are potentially biased. Consideration of data collection methods is also important with regard to minimising non-completion. Using interviews rather than survey questionnaires for this purpose may, in addition to improving completeness of data collection, also minimise the potential for misunderstanding the meaning of a question being used. However, this is a resource-intensive approach leading to a small number of participants likely being involved.

Common psychological biases may be an issue (e.g. a ‘gratitude factor’ or halo bias) where anchor responses are more favourable than is realistic. Potential patient biases may also depend on the effect of a particular outcome on a patient’s own circumstances. For example, in a health-care system in which financial assistance requires proof of continuing ill health, a patient may be reluctant to express that an important improvement has occurred. It is also widely acknowledged that an overall average MCID estimate may not correctly quantify whether or not important change has taken place for an individual patient, as there will of course be variation between patients. Furthermore, the MCID estimate may vary between studies. The agreement or otherwise of the estimate with different studies should be explored.

Triallists wishing to use a MCID estimate derived using an anchor method should review the methodological quality of the anchor method study, specifically with regard to the above mentioned issues, to establish whether or not the MCID value is valid for the future trial population of interest. Researchers wishing to use an anchor method to establish the MCID of a particular outcome should consider these issues so that they can report variation and potential biases as explicitly as possible.

**Distribution method**

**Brief description of the distribution method**

The distribution method typically determines a value that is larger than the inherent imprecision in the measurement (e.g. MDC) and which is therefore likely to represent a meaningful difference or what can be statistically detected (e.g. minimal statistically detectable difference). Other approaches are based on the nature of the outcome (e.g. a fraction of the response range for a VAS).

**Variations of the distribution method**

**Measurement error-based approach**

The most common approach for defining an important difference was based on the standard error of the measurement (SEM) and specifically using the formula $1.96\sqrt{2}\text{SEM}$ for two measurements utilising a 95% confidence level [often called MDC or smallest detectable change (SDC)]. A significance level other than 95% can be readily used for calculating the MDC although this was not commonly done. The rationale for using $1.96\sqrt{2}\text{SEM}$ originates from the measurement error associated with the first and second (repeat) measurement from a test–retest scenario. The SEM is often defined as the standard deviation (SD) multiplied by $\sqrt{1 - r}$, where $r$ is a measure of reliability (in the simplest case Cronbach’s alpha can be used). Variants exist for how the formula is defined, such as replacing $r$ with a correlation coefficient (Pearson’s correlation coefficient or the intraclass correlation coefficient). The choice of SD can also vary (e.g. of both measurements, only the first measurement or the change score). Alternatively, the mean square error (Sw) can be calculated through an ANOVA model to estimate the within-person variance, the formula being $2.77\sqrt{Sw}$. A simplified formula, the SD of the differences, $SDw$, can be used in place of $\sqrt{Sw}$ if there is no real source of a difference between the first and the second measurements. For this to hold there would need to be no learning between measurements or any change in the conditions that would lead to a genuine mean difference. According to the study design, the SEM has been defined variously as the ‘standard error of the residuals between repeated trials’, the ‘intra-rater standard deviation’ or the square root of the ‘mean square error’. The SEM is not directly...
dependent on the sample size although the precision of the estimates used to calculate it are reliant on the sample size from which they were derived.\textsuperscript{151} The SEM for a particular measure may vary depending on the method used to estimate reliability and how outliers are dealt with. The value of the SEM can be reduced if there are systematic errors that can be minimised\textsuperscript{152} or if test–retest reliability can be improved by, for example, using a larger number of repetitions.\textsuperscript{153,154}

The difference (for two measurements) is typically taken to have clinical interpretability if it is > 2.77 SEM, that is, the difference is greater than that which would be expected by measurement error alone at the 95\% confidence level 1.96\textsuperscript{2}. Because the SEM is reported in the same units as the original instrument, the result can be expressed as a percentage of the total possible range of the instrument, to provide a value that aids clinical interpretation.\textsuperscript{155} Some have suggested that using a 95\% confidence level may be too conservative for clinical decision-making with regard to individual patients, and alternative levels have been suggested, including 1.0 SEM,\textsuperscript{156–160} 1.65 SEM,\textsuperscript{161} 2.0 SEM\textsuperscript{151,162,163} or even 2.58 SEM.\textsuperscript{164} Justification for using 1.0 SEM was given as the high level of association between 1.0 SEM and established clinically important differences for a disease-specific quality-of-life measure.\textsuperscript{156,157,159,165–167} Sometimes the value \sqrt{2} SEM is used (approximately a 68\% CI),\textsuperscript{168} or a 90\% CI is used in conjunction with it 1.65\sqrt{2} SEM.\textsuperscript{169–174} or the t-distribution value is used instead of the z-value.\textsuperscript{145,146} Another study calculated the standard error (SE) of the difference from the SEM using the formula:

\[ \sqrt{\text{SEM}_{\text{baseline}}^2 + \text{SEM}_{\text{follow-up}}^2} \]  \textsuperscript{175}

Other variations in implementing the general approach are possible.

A similar approach was proposed by Jacobson and colleagues.\textsuperscript{176–184} It was originally used to define the clinical significance based upon specifying ‘functional’ and ‘dysfunctional’ populations. The first approach involves the calculation of the reliable change index (RCI), which is the amount of change within an individual that is reliable, based on the difference in their pre- and post-test treatment scores divided by the SE of the difference between the two test scores. The RCI was defined as \sqrt{2(SE)\textsuperscript{2}} where SE = \sqrt{SD(1 – r)} and r is the reliability of the instrument used to measure change. A RCI value of > 1.96 indicates that change in the individual is important, akin to the measurement error approach outlined above. Variant in the RCI formula exist.\textsuperscript{2,176,183,184} A criticism of the above approach is that it does not account for the presence of regression to the mean; if individuals in the study sample are likely to be an extreme subset of patients prior to treatment they will be more susceptible to change over time that is not wholly attributable to the intervention under consideration. One solution is to seek to determine whether regression to the mean is present and to correct for its presence.\textsuperscript{181,183,184}

A second simpler approach, which could be used in conjunction with the RCI involves defining a range of agreement beyond which individuals can be concluded not to be in the ‘dysfunctional’ (or ‘patient’) population and likely to be ‘functional’ (‘non-patient’) score.\textsuperscript{177,178} This can be defined in different ways and depending on the data available. For example:

- Is the post-treatment score outwith 2 SDs of the mean score of the dysfunctional population?
- Is the post-treatment score within 2 SDs of the functional population?
- Is the post-treatment score closer to the mean of the functional population than the mean of the dysfunctional population?

Although 2 SDs from the population has been proposed as a cut-off, minor variations of 1.0 SD\textsuperscript{179} or 1.5 SDs\textsuperscript{180} have also been proposed, as has a weighted cut-point when data are not normally distributed.\textsuperscript{181,182} Of these, the first can be implemented without having to specify functional and dysfunctional populations in a pre- and post-treatment study by using the pre-treatment sample\textsuperscript{181} to calculate a plausible (95\% confidence) range of agreement beyond which a non-trivial or ‘important’ difference can be defined as having occurred.
It may be difficult to define what the normative or functional population is, or measure a non-patient population’s scores on a disease-specific instrument. The level of overlap could be high, making it difficult to find a criteria that differentiates functional from dysfunctional, or making it difficult for individuals undergoing treatment to show reliable change for a particular instrument. A sample could be divided into more than two categories (e.g. by severity) whereby there are multiple cut-offs indicating when an individual with a pre-test score in one distribution exhibits change that exceeds the cut-off for an adjacent category that is not necessarily the functional population. Nevertheless, with overlap, there may still be difficulties in identifying the separate distributions of each group. Minor variants include using Cohen’s d to calculate the percentage of non-overlap. Because this assessment of clinical change depends on a normative sample, rather than the magnitude of change in individuals, those with extremely poor scores at baseline who improve dramatically may not meet the criteria for clinically important change, depending on the cut-off used, whereas those experiencing comparably smaller changes in magnitude may cross the cut-off point. It is worth noting that these approaches has been used outside clinical settings, for example in marriage counselling, but it may be difficult to define functional and non-functional populations for many clinical and non-clinical groups. More generally, patient change scores may not reflect true change, as biases may be at play, and flooring and ceiling effects of the instrument used to measure change may also pose problems.

The RCI provides a basis for categorising an individual as having a meaningful or important change or not based on his or her scores. It should be noted that this method was proposed to specify a cut-off value between two populations and that it was developed to be applied at an individual level to determine the importance of an observed difference. In contrast, the target difference of a RCT is the magnitude of difference at the group (population) level that is desired to be detected (should it exist). A clinically important change at an individual level may not reach statistical significance at a group level. Implicitly the above measurement error approaches assume normally distributed data.

**Statistical test-based approach**

Under such an approach, a ‘minimal (statistically) detectable difference’ is calculated, the smallest difference that can be statistically detected within a particular study; it is used as a guide for interpreting an observed difference. Difference terms are sometimes used for the same basic approach. For example, van der Hoeven defined the ‘minimum significant difference’ as ‘for a given significance level \( \alpha \) is the smallest shift \( \delta \) such that \( P(\Delta = \delta) \leq \alpha \). The delta symbols used represent the sample (\( \delta \)) and population (\( \Delta \)) differences. Many variations exist on the general approach depending on the data collected and the planned statistical analysis. The ‘smallest detectable difference’ for a single group with two measurements was defined by Valk and colleagues and similarly by Pijls and colleagues as:

\[
\sqrt{(z_x + z_{\alpha})^2 + \frac{\alpha^2}{n}}
\]

where \( \alpha \) is the significance level, \( 1 - \beta \) is the statistical power, \( \sigma^2 \) is the variance of the within-person differences, \( z_x \) is the \( 100(1-\alpha) \) percentile of the standard normal distribution and \( n \) is the number of observations. The ‘minimum detectable difference’ was defined by Hanson and colleagues and Bridges and Farrar for two independent groups (equal group size and variance), and allowing for unknown variance, as:

\[
\sqrt{2(t_{\alpha}, v + t_{\beta}, v)} \times SD/\sqrt{n}
\]

where \( t_{\alpha} \) (and \( t_{\beta} \) correspondingly) is the \( 100(1-\alpha) \) percentile of the \( t \) distribution with \( v \) degrees of freedom, \( n \) is the number of observations per group and \( \alpha \) and \( \beta \) are as above. Anderson and colleagues used a similar formula to define the ‘minimum significant difference’, which allowed for unequal groups sizes although without regard to statistical power:

\[
t_{\alpha, v} \sqrt{\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}}}
\]
where \( n_1 \) and \( n_2 \) are the number of observations in the respective groups and \( t_{\alpha} \) and \( S_w \) are as before.

In the context of samples of soil, Hoss and colleagues\(^{201}\) defined the ‘minimal detectable difference’ for comparing two independent groups (unequal group size) in a similar manner although expressed as a coefficient of variation (CoV):

\[
100 \sqrt{\frac{(SD_c)^2}{n_c} + \frac{(SD_r)^2}{n_r}} \times \frac{t_{\alpha, v}}{x_r}
\]

(5)

where \( t \) is defined as before, \( SD_c \) and \( SD_r \) are the SDs in the control and reference groups, respectively, \( n_c \) and \( n_r \) are the corresponding number of observations per group and \( x_r \) is the mean in the reference group.

The mean square error estimate from an ANOVA model can be used when more than two groups or repeated measures within a group are used and other factors are accounted for within the analysis (see, for example, Fuchman and colleagues\(^{202}\), Kropmans and colleagues\(^{203}\) and Warren-Hicks and colleagues\(^{204}\)). Various minor modifications on the general approach exist. These include extending the approach to post hoc testing following ANOVA using Tukey’s honestly significant difference procedure\(^{205}\) and Dunnett’s multiple comparison test.\(^{206}\) In a similar manner, though without the formal statistical test framework, Ndlovu and colleagues\(^{207}\) calculated the minimum detectable difference based on precision (variance) and ‘intrinsic sensitivity’ in the context of bone mineral measurement. Other methods for defining an important difference using the statistical testing approach involved the CoV:

\[
2.8 \text{CoV},\quad 2 \sum \frac{\text{CoV}}{n_{\text{replications}}} \quad \text{and} \quad 2\sqrt{\text{precision error}}
\]

where precision error is a modified CoV for paired data.

**Rule of thumb-based approach**

A small number of studies stated that they defined an important difference based on the distribution of the outcome, such as using a substantial fraction of the possible range, for example using 10 mm on a 100-mm VAS measuring symptom severity.\(^{139}\) Similarly, Abrams and colleagues\(^{211}\) divided 100 by the number of response level changes that could possibly be achieved. For example, an instrument question with four possible responses available will have a maximum of three possible response level changes, and therefore the absolute minimum amount of change that could be achieved was 100/3 = 33.3.

**Summary of distribution studies**

A total of 324 studies were found that used the distribution method; 153 of these also used another method (see Combination of methods for details). The majority of the papers reported results from studies in a range of clinical areas, including back pain, diabetes, mobility, quality of life. Several papers reported on non-clinical areas, mainly toxicity testing. The outcome of interest varied and included biochemical markers, degree of flexion, depression, oxygen output, quality of life, time, radioactivity level and whole effluent toxicity.

**Practical considerations for use of the distribution method**

Of the three basic types of distribution method identified – the measurement error-based (including RCI approach), statistical testing-based and rule of thumb approaches – the first two are widely used in important difference assessments, but neither translates cleanly into the target difference context. Statistical testing-based approaches cannot be used for specifying the target difference in a sample.
size calculation; in an important difference setting the assessment is carried out after the data are collected and therefore the sample size is a known and fixed quantity, whereas in a RCT sample size calculation context the sample size is the object to be determined, which in turn depends on the target difference. The measurement error approach also does not translate straightforwardly; such an assessment is typically based on test–retest (within-person) data, whereas many trials are of a parallel group design (i.e. assessing a between-group difference). It may also be difficult to obtain consistent and reproducible measurements for some patient populations and/or conditions. The rule of thumb approach is dependent on the outcome having inherent value. It is akin to the approach often adopted for defining an equivalence/non-inferiority margin in which a substantial fraction (e.g. a third or a half) of the previously observed effect is used.

Health economic method

Brief description of the health economic method

The approaches included under this method make use of the principles of (health) economic evaluation and typically involve defining a threshold value for the cost of a unit of health effect that a decision-maker is willing to pay and using data on the differences in costs, effects and harms to make an estimate of relative efficiency. In some respects this might be thought of as an explicit analytical expansion on the sentiment expressed in the MCID.

Variations on the health economic method

The two earliest papers identified used simple manipulation of costs and effects data to identify the difference in effectiveness that would lead to the incremental cost per unit of health effect being no more than a given threshold that represents the decision-makers maximum willingness to pay for that health effect. As an example of the approach the steps involved in Detsky’s approach are outlined in Box 3. With Detsky’s approach there is an implicit assumption that the target differences compared in the cost-effectiveness analysis are meaningful. Furthermore, although the method attempts to weigh up the costs and benefits of conducting the research, the spectrum of costs considered is very narrow. For example, the consequences for costs following a change in treatment are not considered. Most of the other approaches can be viewed as an elaboration of this general approach.

An alternative approach is typified by Torgerson and colleagues. They likewise considered that the relevant issue to be the economic ‘significance’ (or importance) of a difference. Their definition of significance was based on the difference in effectiveness that would, for a given difference in cost, ensure that the incremental cost per unit of effectiveness was no greater than some prespecified threshold value or that the average cost per unit of effectiveness was equivalent between the two procedures. The sample size was determined using the conventional approach although the difference deemed important was based on the estimated cost differential and control group outcome. The key distinction between the approach of Detsky and that of Torgerson and colleagues is that, in the former, the difference in effectiveness between interventions is compared with the cost of the research whereas, in the latter, the difference in effectiveness between interventions is compared with the difference in costs of treatments between interventions. In both, only rudimentary consideration was given to what costs and effects should be, and neither considered the imprecision in cost estimates.

In Table 3 two simple scenarios are illustrated using data from the National Institute for Health and Care Excellence (NICE) appraisal comparing laparoscopic surgery with open surgery for inguinal hernia repair. In the first worked example, the costs of using both interventions have been estimated and these data are used to estimate the target difference in effectiveness (measured in terms of QALYs) if an incremental cost per QALY of £20,000 is used (a threshold value adopted by NICE in England). In the second example, which compares average (or, more correctly, marginal) costs per QALY, the target difference in effectiveness is estimated from information on the costs of the two intervention and information on the QALYs provided by open repair (the control intervention).
Following the same general principle as Torgerson and colleagues, Samsa and Matchar used an economic model to estimate the difference in effectiveness of a treatment for stroke that would be needed so that the costs of treatment would be offset by cost savings from disabilities avoided. The authors’ economic decision rule was that the introduction of the intervention should be at least cost neutral, but a cost-effectiveness threshold could have been readily used (and would have resulted in a smaller MCID being identified). Briggs and Gray took an approach that in its basic form was conceptually similar to that of Torgerson and colleagues. It assumes that a maximum value for incremental cost-effectiveness is defined. From this, a value for the difference in cost can be calculated such that the critical difference in cost ($\delta C$) is equal to the threshold value for the decision-maker’s willingness to pay for a unit of health effect ($R_c$) multiplied by a predefined difference in effectiveness ($\Delta \tilde{E}$). More formally this can be represented as:

$$\delta C = R_c \Delta \tilde{E}$$

(6)

It is unclear how $\Delta \tilde{E}$, the predefined difference in effectiveness, should be determined. $\Delta \tilde{E}$ might be defined as the MCID and one of the other methods (e.g. anchor) used to provide an estimate. When comparing two interventions the hypothesised difference in cost ($\Delta C$) should be less than the critical difference in cost ($\delta C$). If the hypothesised difference in cost is greater than the critical difference in costs then the trial should not be performed.

As a response to the approach outlined by Briggs and Gray, O’Hagan and Stevens proposed a Bayesian method of assessing sample size. Central to their approach and in common with the other approaches is the identification of a threshold value for the decision-maker’s willingness to pay for a unit of effectiveness. For any given threshold value the objective is to maximise the net monetary benefit (NMB), where the NMB is the product of the difference in effects between two interventions multiplied by the threshold value and minus costs. More formally it can be represented as:

$$\text{NMB} = R_c \Delta E - \Delta C$$

where $\Delta E$ and $\Delta C$ are the differences in effectiveness and costs, respectively, between interventions and $R_c$ is the decision-maker’s willingness to pay for a unit of effectiveness. The critical issue is that a decision in favour of one intervention would be taken if the NMB is >0. The Bayesian approach requires that prior

---

**BOX 3 Summary of Detsky’s approach (health economic method)**

1. Estimate the sample size for a range of ‘clinically important’ differences and also the expected (realistic) difference. These could be derived by different methods. In Detsky’s examples, the realistic difference is based on a meta-analysis.
2. Estimate the cost of the trial for each sample size estimated.
3. Estimate the benefits of the trial. This is based on the proportion of patients who might benefit from the intervention and who would receive it after the trial results are known, which is derived by:
   i. Estimating the effectiveness in the control group combined with information on the difference in effectiveness between the treatment and control groups.
   ii. Combining (i) with information on the total number of people who are likely to receive the new treatment, should it be shown to be worthwhile. The number of people likely to get the treatment is based on an estimate of the number of people who could benefit from the treatment adjusted by an implementation factor to reflect that not everyone will get the new treatment.
4. The costs and benefits of running a trial to detect each of the plausible target differences identified from (i) are calculated. For each identified target difference the associated costs and benefits are compared in an incremental cost-effectiveness analysis. The decision-maker then chooses the size of trial (and target difference) that has an incremental cost-effectiveness ratio no greater than their maximum willingness to pay for a unit of health effect.
distributions are specified for the set of unknown parameters used to derive the NMB. These parameters (in the simplest form) might be costs and effects for the interventions compared or they might be the parameters that are the determinants of costs and effects. The approach does not require that any comparison is made between the parameter estimates used to calculate the NMB for the interventions under investigation. However, this implies that any difference in values for a parameter between interventions is potentially important (as the decision rule is such that any value other than 0 for NMB informs the choice between interventions). It would be possible to include prior information about what is a meaningful difference between interventions for a given parameter. This would require the use of another (e.g., opinion-seeking) method.

Other analysts have considered the identification of optimal sample size from the perspective of a profit-maximising firm.\textsuperscript{218,219} These analysts define an expected net gain, that is, profit function, the objective of the approaches being to maximise the expected net gain from research. The implied minimum value of this function is 0. Both of the approaches outlined use a Bayesian decision-theoretic approach to sample size determination. How judgements are formed about differences in effectiveness, costs or cost-effectiveness (from the perspective of the payer of health care) is not specified. Gittins and Pezeshk\textsuperscript{218} suggest that the extent to which a therapy is adopted corresponded to a personal threshold for the difference between interventions. How this is arrived at is unclear.

Kikuchki and colleagues\textsuperscript{220} and Willan\textsuperscript{219} describe an extension of the approach outlined above by considering the optimal sample size from the perspective of some single payer system. In both approaches,
the objective is to maximise a net benefit function. The assumption with the net benefit approach is that any positive value is important. The approach incorporates the cost of research, the likelihood of adopting a more beneficial treatment and subsequent effects on costs and effects of changing management. Again, however, how judgements are made about whether differences in the input parameters are important is not outlined.

Willan and Eckermann\(^{58}\) provide an example of the use of modelling to determine the expected value of sampling information and the expected trial cost. The optimal position for the decision-maker is when the difference between the expected value of sampling information and the expected trial cost is maximised. The expected value of sampling information in simple terms describes the monetary value of removing uncertainty surrounding a decision to adopt an intervention. It is based on the net benefit statistic described above and as such requires the definition of a threshold value for a decision-maker’s willingness to pay for a unit of a health effect and an assumption that any positive value of net benefit is important. The results of the analysis are dependent on the starting values used in the model and the structure of the model (both of which make implicit or explicit assumptions) and important differences between interventions. For example, excluding an event from the model assumes that there is no meaningful difference between interventions. This assumption may be explicitly based on other information about what constitutes a MCID or may implicitly assume that there is no magnitude of a difference that is important. Furthermore, as with O’Hagan and Steven’s\(^{19}\) approach, any difference in values for a parameter between interventions is potentially important (unless some further decision rule on what constitutes a MCID is explicitly built into the analysis). In contrast to the above studies, Baccetti and colleagues\(^{221}\) argue that looking at the value of information is difficult and risky for investigators planning a trial because of the complexity of the method and the number of assumptions required make it easy for peer reviewers to raise concerns. They suggest that when deciding on the sample size of a trial the ‘cost-efficiency’ should be maximised, where cost-efficiency is defined as the ratio of the expected scientific/clinical/practical value (\(v\)) of the study for a given sample size (\(n\)) to the cost (\(c\)) of conducting the study for a given sample size (\(n\)), that is, cost-efficiency = \(v/n/c\). Given the difficulties in establishing value, which forms the basis of their criticism of the value of information approaches, they argue that it is possible to specify general properties of how sample size influences projected value and that this relationship, along with the cost of research, can be used to identify the sample size of the study. The decision rule is that cost-efficiency should be maximised and that any increase in cost-efficiency is potentially important. However, the definition of value used is uncertain.

**Summary of health economic studies**

Thirteen studies were identified that used a health economic method; none used another method as well. These studies are illustrative of a generally progressive development of methodology over time, with approaches becoming increasingly sophisticated. All have used some degree of modelling and all based conclusions on an economic decision rule; as such, the use of terminology such as ‘minimally important difference’ was infrequent and generally absent in the most recent studies, which is unsurprising given that later papers were based on explicit critiques of traditional methods for sample size determination, which require a target difference to be defined. All of the methods described made use of hypothetical examples and, based on discussions with experts in this area, the use of economic methods in practice has been very rare.

**Practical considerations for use of the health economic method**

The economic methods in this section describe increasingly complex approaches to weigh up the difference between interventions over multiple outcomes (different measures of effectiveness, harms, uptake, adherence, costs of interventions, costs of research, etc.). Which outcomes are included will be determined by the decision problem faced. The methods describe an increasingly comprehensive consideration of the outcomes and an increasingly sophisticated way of modelling their joint impact. Nonetheless, to do this they have to define a decision rule, for example, treatment \(x\) is better than treatment \(y\) when the NMB for \(x\) compared with \(y\) is >0. The perspective of the decision-maker is clearly critical in what should be considered. It is unlikely that one analysis would satisfy all relevant perspectives – those of the patient,
clinician, funder and health-care policy-maker. However, it may be easier to see such an approach being adopted by a research funder to provide reassurance that a particular study is worthy of the research cost incurred. The more sophisticated modelling approaches also are required to make implicit or explicit assumptions about the value of a target difference between treatments for the individual parameters that are used to estimate NMB. How this is carried out is not addressed by the health economic approaches. Such methods can be viewed as a comprehensive framework for decision-making that incorporates judgments about what would be clinically important and relevant empirical evidence.

A major issue with the use of such methods, as with any economic evaluation, concerns the data sources used to supply the required parameter inputs. This includes an understanding of the relationship between parameter input values, for which little robust data are often available. The expected value of sampling information approach outlined by Willan and Eckermann\textsuperscript{38} assumes that the current evidence is sufficiently summarised to be able to evaluate the incremental benefit of a future study in reducing uncertainty. This implies an up-to-date economic evaluation, which in turn implies the systematic assembly of current clinical and other evidence. Although the methods have intuitive appeal, the sophistication of the approach along with the underlying data demands perhaps explain their limited use to date. The more comprehensive the model structure the more challenging it is to populate the model. Assumptions for key aspects of the model, such as the time horizon, which defines the period over which the model will incorporate the costs and consequences of the intervention decision, may be difficult to predict and would therefore require sensitivity analyses to be conducted. Narrowing the scope with regard to which costs and benefits need to be considered, for example by adopting the perspective of a profit-maximising company, would simplify the approach.

Basing the target difference on a clinical outcome, as per the conventional approach, furthers the science of interventions by producing a result that directly assesses the intervention’s clinical outcome in isolation from other factors (e.g. costs). This also has the intuitive appeal of simplicity of interpretation for the patient, health-care professional and funder. Given this, it seems unlikely that the health economic method would currently be accepted as the sole basis for specifying the target difference. It should be noted that determining the study size (target difference) under the conventional approach alone, as is the typical current practice, may lead to a study that cannot reliably answer the corresponding health economic question.\textsuperscript{215}

**Opinion-seeking method**

**Brief description of the opinion-seeking method**

The opinion-seeking method determines a value (or plausible range of values) for the target difference by asking ‘expert(s)’ to give their view on what value or values would be reasonable. This requires participants to be presented with information (either real data or hypothetical scenarios) in order for them to provide their opinions. Methods for eliciting opinions and for establishing a consensus to consolidate these opinions of relevant groups of individuals vary. A target difference that is realistic, important or both could be sought.

**Variations on the opinion-seeking method**

There can be wide variation regarding the most relevant group of people to seek an opinion from (e.g. patients, clinicians, trialists); the method of selecting individual experts (e.g. literature search, mailing list, conference attendance); and the number of experts consulted. Other variations include the method used to elicit values (e.g. interview, survey), the complexity of the data elicited and the method used to consolidate results into an overall value or range of values for the difference.

Most studies solicited the opinion of clinicians (across specialties and/or settings and/or disciplines), although others asked the opinions of patients,\textsuperscript{124,222–225} the views of both clinicians and patients,\textsuperscript{135,226–229} or the views of multidisciplinary experts (e.g. including non-clinicians such as trialists).\textsuperscript{41,230–236} The methods used to recruit experts also varied. In most studies, opinions were sought from a sample of
study investigators,\textsuperscript{26,30} members of the general population or a membership list or mailing list or by using a search strategy to identify authors of published studies on a particular subject.\textsuperscript{134,136} Experts who were listed as members of relevant specialty organisations were recruited either through random selection\textsuperscript{222} or by contacting all members of the list (e.g. members of the Royal College of Psychiatrists).\textsuperscript{237} Convenience sampling (e.g. from patients attending an outpatient clinic\textsuperscript{227} or those who noticed an advertisement requesting participants)\textsuperscript{234,236} was also used. For face-to-face meetings, participants could be recruited from among conference delegates.\textsuperscript{238} Of the remaining studies, either experts were gathered from special interest groups (e.g. conference attendees or a relevant subcommittee) or the method of expert recruitment was not reported. The number of experts whose views were sought varied from 5\textsuperscript{239} to 1584,\textsuperscript{240} although these data were not always reported.\textsuperscript{9}

A survey was the most common mechanism by which an opinion was sought, although interviews or face-to-face meetings were also used occasionally. A combination of survey and face-to-face methods or survey and interview-based approaches was also used.\textsuperscript{26} In one case the method involved a face-to-face meeting and additional elicitation although the method was not reported.\textsuperscript{239} When studies involved face-to-face meetings, the method of deriving consensus for a resulting important difference value once opinions had been elicited was either not explicitly specified\textsuperscript{43} or was derived using established techniques for such processes (e.g. nominal group technique).\textsuperscript{238} Specified processes commonly involved some kind of quantitative mechanism for eliciting opinions to summarise agreement, for example through voting,\textsuperscript{230} or the compilation of individual attendees’ scoring results for a set of paper patients.\textsuperscript{241} Splitting the number of attendees into groups was used in some face-to-face meetings.\textsuperscript{229,238} If survey methods were used alongside face-to-face meetings, survey responses could be gathered before the meeting,\textsuperscript{238} during the meeting\textsuperscript{244} or after the meeting.\textsuperscript{241}

For group-based decision-making it was necessary to establish a threshold for acceptable consensus, and this varied from situation to situation. Delphi processes were often used in studies using an opinion-seeking approach. Most of the studies used survey and face-to-face methods for eliciting an important difference used a Delphi process (normally two ‘rounds’),\textsuperscript{134–136,241} although it should be noted that most of the studies doing so were written by the same group of authors. A Delphi process was also used by some of the studies that had used only a survey method for seeking opinions (although again many had been undertaken by the same authors).\textsuperscript{242–245} Several studies used an 80% consensus threshold,\textsuperscript{246–248} or a value very close to this level (‘all but two’ out of nine experts).\textsuperscript{136} An example opinion-seeking method study in which the Delphi approach was used is described in Table 4.

Other studies using a survey varied in terms of the complexity of the data collected. Some simply asked what value on a particular instrument respondents would consider to be clinically significant.\textsuperscript{232,237,249} Items or criteria can also be ranked in terms of their importance in helping participants form their judgement(s) regarding the extent to which a particular value represents an important difference.\textsuperscript{232,240} Opinion could also be sought by study investigators on participants’ preferences for particular hypothetical scenarios. For example, in the study by Allison and colleagues,\textsuperscript{223} respondents (obesity patients entering clinical trials) were asked to indicate on a VAS how much of an increased risk of a serious adverse event they would tolerate in order to lose a variety of expressed proportions of their current body weight (5%, 10%, 20%, 30%). Trade-off and standard gamble methods were also used in some of the studies using an opinion-seeking method that had interviewed participants (either face-to-face or by telephone).\textsuperscript{18,222,224,225,227,228,250} Another study using interview techniques\textsuperscript{226} asked participants (patients and clinicians) to rate the clinical importance (on a five-point Likert scale) of specific goals for the patient, rank the five most important from among these rated goals and indicate how much improvement would be required to consider that the patient had achieved a meaningful difference in quality of life or functional capacity. Man-Son-Hing and colleagues\textsuperscript{231} used a flip chart and booklet during interviews to graphically assist participants in conceptualising risk. In some instances, rather than being asked about their expectations of change over time with regard to a particular disease, treatment and outcome measure, respondents (clinicians and trialists) were asked about thresholds that would influence their practice.\textsuperscript{26,30,47} Fayers and colleagues\textsuperscript{26} asked surgeons about their expectations regarding a particular treatment. Similarly, the
CHART trials provide an example of eliciting opinion regarding equivalence and a realistic difference for a survival outcome to determine the target difference in a Bayesian framework (although the sample size was calculated using the conventional approach). Latthe and colleagues developed a distribution for doctors’ prior beliefs about the effectiveness of a particular treatment, using a survey method for seeking opinion. Oremus and colleagues additionally enquire in their study not only about thresholds for equivalence between treatments, but also about thresholds for the likelihood of a treatment to meet first-line therapy standards. The analysis of experts’ views also varied in complexity. For example, Rider and colleagues developed a large number of definitions of clinical improvement using CART analysis and logistic regression, to compare with values from an opinion-seeking method, identified (using diagnostic accuracy methodology) to determine a smaller number of definitions that had good (80–85%) sensitivity and specificity. These definitions were returned to the experts who were asked to rate their face validity. Similar approaches were used by Ruperto and colleagues and Giannini and colleagues.

Summary of opinion-seeking studies
A total of 81 studies were found that used the opinion-seeking method; 21 also used another method (see Combination of methods for details). Most commonly, opinion-seeking methods had been applied in the fields of rheumatology, respiratory-related outcomes or mental health, although there was a wide range of specialties noted and studies did not always relate to a specific clinical specialty (e.g. studies focusing on the opinions of triallists).

Practical considerations for use of the opinion-seeking method
One advantage of the opinion-seeking method is the ease with which it can be carried out (e.g. through a survey). Eliciting an opinion within a trial context has been carried out utilising a Bayesian framework, which can be more informative but also more difficult to implement. The estimate may not be representative of the desired population or a future trial and values may differ between experts and stakeholders (e.g. patients and clinicians). For non-survey approaches one major issue for implementation is the sample size. With face-to-face meetings there will be an upper limit on the number of people whose opinion can be sought if consensus on a value or a small range of values is the goal of the meeting. Results may not be generalisable to a larger population than the one envisaged and the sample may not be representative of the general population. Face-to-face meetings may be influenced by group processes. If this is felt to be a problem, a Delphi method could be implemented through a survey to compile the views of multiple participants anonymously. However, survey methods may be vulnerable to typical difficulties such as low response rate, missing data and technical problems if online and so careful planning is required. The Delphi method itself allows for variation in implementation of how consensus is reached.

The nature of the expertise itself may influence the results, as being involved in a trial could bias an expert’s judgement of the estimates they provide (e.g. being too enthusiastic about a treatment’s possible effects). Indeed, the specialty may influence values and the views of participants may change over time. In addition, a pre-existing definition of an important difference (e.g. anchor-based estimate) can be used to help participants formulate their own estimate. The wording of the question is important.

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using a mailed survey six experts were asked to recommend a MCID for the Doyle Index (a score of tenderness out of 144), to be used in a hypothetical trial of two drugs with stated inclusion/exclusion criteria. In this first ‘round’ responses were collated and anonymised. A second survey was sent to the experts, this time with the anonymised responses from the first round. Experts have a chance to consider the views of the other experts and are then offered the chance to modify their original responses. Responses from the ‘second’ round were again collated and anonymised. A third ‘round’ proceeded in the same way.</td>
<td></td>
</tr>
<tr>
<td>The median (range) estimates for the Doyle index MCID were 6.5 (28.5), 5.0 (7.5) and 5.5 (5.7) for rounds 1–3 respectively; 5.5 could therefore be used as the MCID and the target difference in a future trial</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 4** Opinion-seeking method example: Delphi method approach for an antirheumatic drug trial

| Methodology | Using a mailed survey six experts were asked to recommend a MCID for the Doyle Index (a score of tenderness out of 144), to be used in a hypothetical trial of two drugs with stated inclusion/exclusion criteria. In this first ‘round’ responses were collated and anonymised. A second survey was sent to the experts, this time with the anonymised responses from the first round. Experts have a chance to consider the views of the other experts and are then offered the chance to modify their original responses. Responses from the ‘second’ round were again collated and anonymised. A third ‘round’ proceeded in the same way. |
| Findings | The median (range) estimates for the Doyle index MCID were 6.5 (28.5), 5.0 (7.5) and 5.5 (5.7) for rounds 1–3 respectively; 5.5 could therefore be used as the MCID and the target difference in a future trial |
as it needs to be clear and specific enough to prevent different participants making varying assumptions about the situation they have been presented with. Interview-based methods (telephone or face-to-face) might affect estimates and questions may need to be rephrased, although these approaches allow for clarification, unlike survey formats. Trade-off approaches (commonly used for interview-based opinion-seeking approaches) may be problematic as hypothetical scenarios may not be adequate predictors of real behaviour, may be considered too artificial and may be difficult to comprehend, particularly if participants are not used to understanding risk concepts. Describing risk to participants in different ways can have an impact on the resultant estimate.

**Pilot study method**

**Brief description of the pilot study method**

A pilot study may be used as a method to determine a relevant value for the target difference in situations in which there is little evidence, or even experience, to guide expectations on an appropriate value for the target difference. One definition of a pilot study is running the intended study in miniature before conducting the actual trial. Similarly, a Phase II study in the regulatory drug setting could be viewed as a pilot for the Phase III study. By carrying out a pilot study that recruits an identical, or similar, population to the one that is expected to be recruited to a future trial, the resulting data can be used by triallists to estimate parameters for a sample size calculation for the main trial. This can be the effect size or one of the parameters needed (e.g. SD for a continuous outcome or the event proportion for a binary outcome/survival outcome). Pilot studies are typically small in size.

**Variations on the pilot study method**

The simplest approach would be to use the observed effect in the pilot study as the target difference in a RCT. Such an approach has been criticised. One paper modified this approach by calculating the baseline SD of the Roland–Morris Questionnaire (RMQ) in the pilot study, which was used in the sample size calculation for a future trial for a given difference in the outcome.

The study by Salter and colleagues used a similar approach but adjusted the estimate for the SD for the uncertainty by using the upper limit of a 90% CI of the variance to allow for the imprecision. This study also inflated the calculated sample size to account for similar levels of attrition as seen in the pilot study. This need to adjust the pilot study SD estimate was shown in the study by Browne, which looked at the likelihood of actual power being greater than or equal to planned statistical power in a future trial. A simulation was carried out for combinations of various pilot sample sizes (5, 10, 30, 50 and 100), mean differences between trial arms (10, 30 and 75), which gave standardised effects of 0.1, 0.3 and 0.75, respectively, with a prespecified SD of 100, and planned statistical power of 80% and 90%. Using one of five upper confidence limits (50%, 60%, 70%, 80% or 90%) as an estimator of SD (as opposed to the SD point estimate), Browne showed that the probability of the actual power exceeding the planned power can be substantially increased by taking account of uncertainty around the pilot estimate. A hypothetical example involving treatment for nephropathy was provided to illustrate the findings. Wang and colleagues undertook a simulation study of the expected sample sizes for a Phase III trial based on the observed difference in effect size from a Phase II trial (which was considered a pilot for this Phase III study). By simulating values for the Phase II sample size (50, 100 and 200) with planned power of 80%, and setting α to 0.025, this study showed the effect of using the point estimate of the effect size from the Phase II trial compared with using the lower limit of both 1 and 2 SDs from the point estimate (the standardised effect for a Phase III trial was expected to be lower than that of a Phase II trial). This study also anticipated that there would be a threshold effect size for a Phase II study below which investigators would not proceed to a Phase III trial because it would require such a large sample size as to be unfeasible.

**Summary of studies using the pilot study method**

Six studies used a pilot study method. Five of these studies used this method alone and one used it in combination with another method (see Combination of methods for details). Two studies had conducted pilot interventions designed to reduce pain, whereas in the study by Browne, a hypothetical
example involving renal patients was provided to illustrate the method used. The sample sizes of the simulated pilot studies ranged from 5 to 200 and for the real pilot studies ranged from 12 to 160.

**Practical considerations for use of the pilot study method**

Pilot studies are not merely useful for sample size calculations but can also provide additional information on feasibility and highlight challenges to participant recruitment that may need to be taken into account in a sample size calculation. There are practical difficulties in undertaking a pilot study that may limit the relevance of the result (e.g. changes in the population, intervention or outcome definition and timing of measurement), however, the most problematic is the inherent uncertainty in the study results because of the small sample size. The estimate of the effect size will likely be very imprecise. Use of a pilot study to inform either the SD or the control group event proportion is more useful although the imprecision of these values should also be acknowledged. Strictly speaking, an internal pilot, that is, assessing data from the participants in the first stage of the study, cannot be used to specify the target difference, or a related sample size component (e.g. control group event proportion), although it could be used as part of an adaptive strategy to preserve the prespecified target difference. A pilot study is most useful when it can be readily and quickly (e.g. rapid recruitment and short outcome follow-up) conducted.

**Review of evidence base method**

**Brief description of the review of evidence base method**

The RoEB method summarises existing evidence in order to identify an important or a realistic difference for a specific treatment comparison. The optimal implementation would be to meta-analyse results for the outcome of interest based on the studies found using a defined search method. An alternative approach is to review existing evidence for a specific outcome of interest to determine a difference that can be viewed as important.

**Variations on the review of evidence base method**

The most common approach in included studies involved implementing a prespecified strategy for reviewing the evidence base for either a particular instrument or a variety of instruments for an important difference. These studies identified, from the results across all studies reporting the same outcome, a plausible value for an important difference for the instruments or outcomes under consideration. Cranney and colleagues reviewed existing studies reporting an important difference for particular outcomes in osteoporosis. In addition, they distinguished between ‘individual’ and ‘group’ important differences. For this reason they retained a selection of osteoporosis RCTs that had been identified from their review search strategy to determine the level of an important difference that had been used in the RCTs. Blumenauer and colleagues used the method of identifying studies reporting an important difference for a particular outcome measure, although did not provide a conclusive value for an important difference in summary. However, the values identified using the RoEB method for particular outcome measures were then compared with the results of trials that reported those outcomes.

It was also common to review the level of observed differences (similar to the ‘group’ important differences used by Cranney and colleagues). Of note, Revicki and colleagues considered differences for mortality. Several studies required included studies to have reported sufficient detail to enable effect size estimates to be calculated by the reviewers. In the studies by Johnston and colleagues, Thomas and colleagues, and Woods and colleagues, this was done to determine the sample size of a future study. The last study used Cohen’s h statistic for a binary outcome (see Combination of methods). The study by Thomas and colleagues reported meta-analysis of data from identified reviews within the evidence on the subject of interest (sports medicine outcomes). The effect size used in the power calculation was therefore a pooled effect size. Pooled standardised effects were also used by Woods and colleagues. The study by Johnston and colleagues used the effect sizes identified from reviewing the evidence base to derive a sample size calculation for a future RCT (i.e. a realistic target difference). Others had used a similar approach in reviewing the observed effect size, but did not explicitly use the findings in a sample size calculation. One study by Schwartz and colleagues found some evidence to dispute Norman
and colleagues’ concept of the universality of half a SD. The study by Nietzel and colleagues calculated an effect size with reference to a normative population. RCI calculations comparing treated groups with a normative population were also found.

Julious provided worked examples of using a pre-existing meta-analysis result to determine the target difference for a new study and also considered the associated uncertainty around the estimated pooled effect. An example of assessing uncertainty regarding the SD based on previous studies is also considered. Extending this general approach, Sutton and colleagues derived a distribution for the effect of treatment from the 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 were then reanalysed. This process was repeated a large number of times until a power calculation was made using the proportion of times out of the total number of simulations that the null hypothesis was rejected. Implicitly this adopts a realistic difference as the basis of the target difference. Additionally, the primary analysis of the trial should involve the other studies given that the study was justified in this way. Zanen and Lammers reviewed studies in order to derive a CoV statistic to use in the sample size calculations for an equivalence trial.

Summary of review of evidence base studies
A total of 28 studies were found that used the RoEB method; six of these also used another method (see Combination of methods for details). Several studies reported using a systematic approach to reviewing the evidence base.

Practical considerations for use of the review of evidence base method
Reviewing the existing evidence base is valuable as it provides a rationale for choosing an important or realistic value for the target difference in the sample size calculation of a future clinical trial. However, an estimate identified from the existing evidence may not necessarily be directly appropriate for the population under consideration in a future clinical trial, that is, the generalisability of the identified values may be questionable because of their methodological risk of bias, the intervention implementation or the population studied. The extent to which it is possible to account for these factors depends on the level of detail provided in the previous studies. In addition, publication bias is a relevant issue to consider, particularly for reviews of the evidence base that do not consider alternative sources of information besides searching databases of published evidence sources. The formal use of meta-analysis results to determine the sample size implies that a future study will also be analysed using the current evidence and, arguably, that if a further study is published during its conduct then the sample size should be updated. A review of studies that have determined an important difference is also possible although each individual study (use of anchor method) would have the practical issues mentioned earlier for that type of method. For both approaches, similar to the use of a pilot study, the imprecision in the estimate (whether the target difference or an associated parameter, for example the SD) needs to be considered. Although a meta-analysis is likely to have greater precision than a pilot study, imprecision is still an important consideration.

Standardised effect size

Brief description of the standardised effect size
The SES method calculates the effect size on a standardised scale. For a continuous measure this is usually simply the difference in means (either between two groups of people or between two time points in the same group of people) divided by an appropriate SD (Cohen’s d effect size). The SD is usually the pooled SD of the groups when a comparison between groups (e.g. treatment) is being made, whereas the SD of either the first time point or the change score is often used for within-group comparisons (e.g. before and after treatment). The magnitude of this standardised effect is then used to assess whether an important difference has occurred. Guidelines of 0.2 SDs, 0.5 SDs and 0.8 SDs are widely used for interpreting the magnitude of the effect, and therefore an observed effect can be interpreted according to these values. Alternatively, by specifying the SD, an effect on the original scale of a particular magnitude can be determined. For binary outcomes a variety of SES metrics (e.g. odds and risk ratios or an absolute difference) can be determined. For binary outcomes a variety of SES metrics (e.g. odds and risk ratios or an absolute
difference in event proportion) exist. A hazard ratio or absolute difference in event proportion can be used for a survival (time-to-event) outcome.

**Variations on the standardised effect size**

Three type of standardised effects have been defined for a continuous outcome: between groups, within group and compared with a reference population. The between-groups standardised effect is the situation that mirrors the specification of the target difference in a parallel groups trial, in which a difference between two groups is tested. In the simplest case (equal group sizes) it is calculated for an outcome, $x$, as:

$$\frac{x_{\text{group}1} - x_{\text{group}2}}{\sqrt{SD_{\text{group}1}^2 + SD_{\text{group}2}^2}}$$

(7)

The formula can be modified for uneven sample sizes by calculating the (weighted) pooled SD.

In most of the studies identified, the use of standardised effects related to calculation of a within-group effect. For the within-group standardised effect, most commonly in the literature on important differences, the standardised effect is often used to assess changes in health over time for a quality-of-life measure. The most common situation is to measure subjects’ health on a particular outcome measure at two time points: baseline (before treatment) and follow-up (after treatment). First, the group mean change in score from baseline to follow-up (or equivalently the difference in the mean scores at the two time points) for individuals in a sample is calculated. This mean change in score is divided by the SD of the baseline score:

$$\frac{x_{\text{follow-up}} - x_{\text{baseline}}}{SD_{\text{baseline}}}$$

(8)

The SD used varied. An example is given in Box 4 and shows the possible impact of using different SDs to calculate the SES. Other authors divided the mean change in score by the SD of the change in score:

$$\frac{x_{\text{follow-up}} - x_{\text{baseline}}}{SD_{\text{change}}}$$

(9)

A number of studies used both approaches. Howard and colleagues used a different formula for the SD of the change scores, which accounted for within-person correlation.

**BOX 4 Standardised effect size example: goal attainment scaling**

Fifty-three nursing home patients received a specialist geriatric medicine consultation. The goal attainment scale was measured both pre and post consultation. The mean (SD) scores for pre consultation and post consultation and the corresponding change score were 37.3 (3.5), 45.7 (6.9) and 8.4 (6.5) respectively. Three variations on the (Cohen’s $d$) SES are:

1. using the preconsultation score SD: $8.4/3.5 = 2.4$ SDs
2. using the postconsultation score SD: $8.4/6.9 = 1.2$ SDs
3. using the change score SD: $8.4/6.5 = 1.3$ SDs.

Note: based on Cohen’s criteria all three would be classed as a large SES.
An alternative is to use the pooled SD ($SD_{pooled}$) of the two time points. In the simplest case (i.e. equal group sizes) it is calculated as:

$$ \frac{x_{\text{follow-up}} - x_{\text{baseline}}}{\sqrt{SD^2_{\text{follow-up}} + SD^2_{\text{baseline}}}} $$

In some studies, the SES was of the mean change between groups rather than within one group over time. A further variation on the choice of SD was described by Horton, who compared the difference in means between two groups, divided by the largest SD value.

The study by Andrew and Rockwood calculated the SES as a ‘modified Cohen’s $d$’. A standard Cohen’s $d$ is produced using the pooled SD except that this SES is ‘corrected’ for the (Pearson’s) correlation between the baseline and follow-up scores:

$$ \frac{x_{\text{follow-up}} - x_{\text{baseline}}}{SD_{\text{pooled}}} \sqrt{1 - r} $$

Fredrickson and colleagues utilised a more complex within-group SES for a crossover design. In this study each subject received all four different treatments over four time periods, and subjects’ health was measured at the end of each time period. The authors used the following formula (Dunlap’s $d$) to calculate the SES between each active treatment group and the placebo group at each follow-up time point:

$$ d = 2t \left( \frac{1 - r}{n} \right)^{1/2} $$

where $t$ is the SE (from a paired $t$-test) of the mean difference in the change of health between baseline and follow-up between the treatment groups, $r$ is the correlation between the pairs of measures and $n$ is the number of observations. Two studies compared study results with data from a reference population who served as normative data. Harris and colleagues subtracted the mean score of a reference population from the baseline and follow-up mean scores of a sample and then divided by the SD of the reference population score. The SES is calculated as $x_1 - x_2$, where

$$ x_1 = \frac{\mu_{\text{pre}} - \mu_{\text{norm}}}{\sigma_{\text{norm}}} \quad \text{and} \quad x_2 = \frac{\mu_{\text{post}} - \mu_{\text{norm}}}{\sigma_{\text{norm}}} $$

A lower value for the score used in this study indicated a better outcome. A similar method was used by Rajagopalan and colleagues. They divided the study sample into those with and those without a condition of interest. They then divided the mean difference between these two groups by the 'population SD'. However, it is not clear whether they are referring to a reference population (and, if so, which one) or all of the subjects in their study. The similarity of such an approach to the RCI distribution method (described in Distribution method) should be noted.

One other study used a similar approach for a categorical outcome with repeated measures. The study authors used the difference in proportions between baseline and follow-up. SES metrics are commonly used for binary (e.g. odds ratio, risk ratio) and survival outcomes (e.g. hazard ratio) in medical research and a similar approach can be readily adopted although this review identified no studies that formally carried this out. A doubling or halving of a ratio is sometimes seen as a marker of a large relative effect.
Summary of standardised effect size studies

A total of 166 studies were found that used the SES; 129 also used another method (see Combination of methods for details). The papers reported the results from studies in a range of clinical areas, including, but not limited to, Alzheimer’s disease, back pain, cardiac surgery, insomnia, schizophrenia, stroke and osteoarthritis. The most common outcome of interest was quality of life. Other outcomes of interest included disability, functional ability, goal attainment, speech function and cognitive ability. In considering change, some studies often used a clinician to assess subjects’ health, whereas others used the subjects’ self-assessment or both (not necessarily for the same instrument). The view of the parents of children who were being treated was often used in studies in paediatric care along with clinical and/or child assessment. Similarly, Rockwood and colleagues used the combined judgement of the patient and the caregiver as well as the clinician’s assessment. Basoglu and colleagues used one instrument that was administered by a clinician, but it was unclear whether the patient or the clinician completed other instruments.

Many studies used the method in order to inform the magnitude of a difference in an outcome that can be viewed as important. A standardised effect was often used to compare the health of different groups of people, either people with a disease with population norms or people with a condition of interest with those without the condition. Some papers were interested in assessing the effect of a treatment on health or comparing the effects of different treatments. The study by Fredrickson and colleagues included repeated measures over time, as well as different doses of a medication, as the authors were also interested in using standardised effects to investigate when changes in health occurred. Four papers specifically calculated effect sizes to inform the design of a future study. The larger a standardised effect, the fewer the number of subjects required to detect a difference between groups or time points. Pyne and colleagues went a step further and calculated effect sizes for different subgroups of people to assess whether any of the instruments performed better in certain subsets of people. Fredrickson and colleagues also made specific reference to the use of standardised effects for calculating required sample sizes, but did not perform any such calculations. van der Putten and colleagues made reference to the link between how responsive an instrument was (as measured by a SES) and the sample size required for, and power of, a study to detect a statistically significant result. In addition, Rockwood and colleagues study used the SES from an early study to determine the sample size necessary to power their RCT appropriately.

Overwhelmingly, studies used the values suggested by Cohen, whether explicitly or implicitly, where 0.2, 0.5 and 0.8 represent a small, medium and large effect respectively. Some studies had not merely utilised a SES as a post hoc indication of the magnitude of change for responsiveness purposes, but instead had selected a value a priori for the effect size level that would denote clinically meaningful change, and then calculated the corresponding value for the outcome of interest that would correspond to this level of change. The a priori value for clinically meaningful change varied between studies from 0.2 SДs to 0.5 SДs and both 0.4 SДs and 0.8 SДs. Alternative values for small, medium and large SEs (0.1, 0.5 and 1.0) have been suggested although without any identifiable usage. However, several of the studies that had used multiple methods, of which one was a SES approach, had used alternative proportions of the SD for small (minimal) change, including a quarter of a SD (0.25 SДs), a third of a SD (−0.33 SДs) or even 0.3 SДs. Occasionally, studies considered multiple levels as being of interest for interpretation of an important difference, for example both 0.5 SДs and 0.2 SДs. Sometimes larger values were used, and two studies used 1 SД as an effect size cut-off, of which the study by Hurst and Bolton had also used 0.6 SДs. When studies did not cite Cohen specifically, many still interpreted their SES estimates according to his guidelines. Even when the values were not explicitly stated, most studies had based their interpretation of the effect on Cohen’s criteria, concluding that effect sizes of change were ‘large’, ‘moderate’ or ‘small to moderate’.

Practical considerations for use of the standardised effect size

The main benefits of using the SES method are that it can be readily calculated and it allows comparison across different outcomes, conditions, studies, settings and people. The latter is possible because all of the
differences are translated into a common unit. Such an approach is frequently used in meta-analyses to summarise findings across studies. In addition, it is relatively easy to calculate effect sizes for published studies, provided that they have reported sufficient information. However, different combinations of values can produce the same value on the standardised scale. For the standard Cohen’s d statistic, different combinations of mean and SD values produce the same SES estimate; for example, a mean and SD of 5 and 10, respectively, as well as values of 2 and 4, respectively, give a standardised effect of 0.5 SDs. A larger effect on the original scale can therefore have the same effect on the standardised scale when there is greater variance (e.g. when there is more variability in response because of a more heterogeneous population or more imprecise measurement). As a consequence, specifying the target difference as a SES alone, although sufficient in terms of sample size calculation, can be viewed as an insufficient specification in that it does not define the target difference in the original scale. Furthermore, it can be difficult, if not somewhat futile to try, to explain why different effect sizes are seen in different studies, for example whether these differences are due to differences in the outcome measures, interventions, settings or subjects in the studies. Therefore, as Haymes and colleagues recommend, a comparison of effect sizes should be interpreted cautiously. Nevertheless, given its ready application, the SES approach (with Cohen’s interpretation) is useful when no more pertinent evidence is available; anecdotally, ‘large’ effects are not commonly observed in many settings.

Different SES metrics have been used for continuous (e.g. Cohen’s d and Dunlap’s d), binary (e.g. odds ratio) and survival (hazard ratio) outcomes, however, overall, Cohen’s d is by far the most commonly used metric. Although SESs are commonly used for binary outcomes (e.g. odds ratio, risk ratio) and survival outcomes (hazard ratio) we did not identify any studies that proposed similar guidelines/approaches for their use in determining an important difference, although they could in principle also be used in a similar manner. Cohen’s h was the only binary SES metric that was identified for which such a formal assessment of standardised effect magnitude was made, although this will commonly be carried out informally (e.g. halving or doubling of risk/odds as being an important effect) when conducting sample size calculations or assessing the impact of relationships in epidemiology for metrics such as the odds, risk and hazard ratios. The vast majority of the literature relates to within-group SESs for a continuous outcome. The SD used varied between studies (i.e. between baseline, final and change score values); this undermines comparability as standardised effects based on within-person changes will tend to be larger. Although a variety of SES metrics exist for binary outcomes, the commonly used metrics (such as the odds ratio) do not uniquely identify the sample size and need to be considered in conjunction with the control group proportion. SESs have been used to compare the responsiveness, discriminatory ability and sample size requirement of instruments. They can also be used to compare treatment effects or assess changes in health over time.

Standardised effect sizes are relatively easy to calculate and are objective and there is a widely accepted guideline for the interpretation of the effect sizes for a continuous outcome. Effect sizes of <0.2 are considered ignorable, effect sizes of ≥0.2 to <0.5 are considered small, those ≥0.5 to <0.8 are considered moderate and those ≥0.8 are considered large, based on Cohen’s guidelines. Approximate equivalent values for the odds ratio can be calculated using Cohen’s cut-offs giving 1.44, 2.48 and 4.27 for a small, medium and large effect respectively. Corresponding risk ratio values vary according to the control group event proportion. However, these guidelines may not be appropriate in some situations and were not intended by Cohen to become default values for all purposes. Cohen suggested that, in settings outside of a laboratory or when investigating new research areas, only small effects are to be expected. Some attempts to assess the generalisability of the guidelines have been undertaken and provide some reassurance. Gordon and colleagues suggested that, without the use of an established instrument as a reference criterion (or anchor) for assessing changes, it is still a judgement call as to what constitutes a clinically important difference in a particular situation. Both Cheung and colleagues and Pyne and colleagues made use of an anchor to categorise the health or changes in health of their study subjects. Dumas and colleagues called for further research using an external criterion to investigate clinically meaningful changes. Matza and colleagues specifically state that, although an effect size can be used to assess how small a change in health an instrument can detect, another method (such as an anchor or
distribution method) is needed to determine what would be considered a MCID. However, this view is not universally accepted. For example, Konst and colleagues²⁰¹ specify that a large effect size indicates a clinically important change in health.

Several studies measured the health of the same people over time although had not adjusted for the correlation between measurements recorded at baseline and measurements recorded at follow-up when calculating effect sizes.²⁹⁹,³⁰⁴ Making this adjustment will lead to a different magnitude of standardised effect, with larger correlations leading to larger effects unless the change score SD is used.²⁹⁹,³⁰⁴ As noted above, sometimes the SD at baseline is used.²⁸³–²⁸⁸ Fredrickson and colleagues³⁰⁴ argue that this approach is ‘flawed as it assumes that differences in performance between subjects [on different treatments] at baseline will provide a reliable estimate of differences within subjects over time thereby potentially reducing the estimate of the effect size [of one treatment compared with another]’. With regards to target difference, the SES based on a change score is not directly comparable to that based on a two-group parallel study. Only one study was identified that used a SES approach for a categorical outcome although it was analysed as a continuous outcome.²⁸⁹ Occasionally, studies did not give any details of how they calculated the standardised effect, apart from sometimes referencing other papers.²⁸²,³⁰⁹,³¹⁷,³¹⁸,³⁴³

A clear disadvantage of using the SES method to specify a target difference is that the SES is dependent on the group means and SDs. Different populations will have different means and SDs²⁸⁵ and therefore it is reasonable to expect effect sizes, and therefore what might be called small, medium and large effects, to vary between comparisons according to interventions, outcomes and conditions. Pyne and colleagues³⁰² speculated that, because experimental studies tend to have stricter inclusion criteria and greater control over any interventions than observational studies, changes in health might be larger in experimental studies. Experimental studies may overestimate actual effect sizes or provide an underestimate.³⁰² When determining a target difference for a RCT based on the SES method, the SD used should reflect the anticipated RCT population and, as far as possible, the statistical analysis.

**Combination of methods**

**Summary of methods used within studies combining more than one method**

Of the included studies, 216 used more than one method. A summary of how often each method was used with one or more additional methods is given in Table 5. Details of the numbers of studies that used each of the different combination of methods are available in Table 6.

**Combining results derived from multiple methods**

In terms of interpretation, the method for triangulating the results of multiple methods was not always clear, in some cases triangulation not been carried out and in other cases the rationale for the choice of method was not provided.

For most studies combining anchor and SES method estimates, the anchor method was used to derive a clinically important difference. In some studies, anchor values had been derived previously but an additional method (e.g. a SRM) was used in the new study.³⁴⁴ Wyrwich and colleagues³⁴⁵ chose to use their anchor method estimate as the overall result because of the close convergence of the results from both methods. Other studies similarly used anchor methods with the effect size results used to confirm the size of the change as ‘small’, ‘medium’ or ‘large’ based on Cohen’s criteria.³³⁸,³⁴⁶,³⁴⁷ Drossman and colleagues,³⁴⁸ along with Fairchild and colleagues,³³⁴ based the estimate on the corresponding ROC analysis. Others chose an anchor value even though it differed from the SES value.³⁴⁹–³⁵¹ For example, Gold and colleagues³⁴⁹ used their anchor results because the estimates were lower. On the other hand, Lasch and colleagues³⁵² chose to use the SES values (which were lower than the anchor values) ‘until further evidence is established’, whereas Arbuckle and colleagues³⁵³ chose to use their effect size values because they were the more conservative estimates. In other cases, the value used varied from method to method. For example, Barnes and colleagues³⁵⁴ assessed multiple instruments and used the most conservative estimate of clinically important difference for each outcome. Miskala and colleagues³⁵⁵ used both methods...
for the subscales of their instrument of interest, but took the most conservative value from both methods to determine the overall scale important difference value. Middel and colleagues\textsuperscript{356} utilised a thorough method of triangulation in which discordant results (important change using one method but not the other) were analysed in more detail – a sensitivity analysis of whether discordance was trivial or non-trivial.

Some studies used both the anchor and the SES estimates in deriving a final value for a clinically important difference, although in some cases this may have been due to the serendipity of the results being close in value anyway, rather than an a priori attempt to use the data from both methods. One study simply stated that effect size estimates were ‘in agreement’ with the anchor results.\textsuperscript{346} Swigris and colleagues\textsuperscript{357} determined their final value from the mean of all MCID point estimates found, and de Morton and

---

**TABLE 5** Use of additional method(s)

<table>
<thead>
<tr>
<th>Method</th>
<th>No. (%) of studies using one or more additional methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor</td>
<td>194 (43)</td>
</tr>
<tr>
<td>Distribution</td>
<td>153 (47)</td>
</tr>
<tr>
<td>Health economic</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Opinion-seeking</td>
<td>21 (26)</td>
</tr>
<tr>
<td>Pilot study</td>
<td>1 (17)</td>
</tr>
<tr>
<td>RoEB</td>
<td>6 (21)</td>
</tr>
<tr>
<td>SES</td>
<td>129 (78)</td>
</tr>
</tbody>
</table>

**TABLE 6** Studies utilising more than one method: method combinations

<table>
<thead>
<tr>
<th>Methods</th>
<th>No. of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Anchor ✓ Distribution ✓ SES</td>
<td>70</td>
</tr>
<tr>
<td>✓ Anchor ✓ Distribution ✓</td>
<td>46</td>
</tr>
<tr>
<td>✓ Anchor ✓ Distribution ✓ Opinion-seeking ✓ SES</td>
<td>8</td>
</tr>
<tr>
<td>✓ Anchor ✓ Distribution ✓ Opinion-seeking ✓</td>
<td>1</td>
</tr>
<tr>
<td>✓ Anchor ✓ Distribution ✓ SES</td>
<td>63</td>
</tr>
<tr>
<td>✓ Anchor ✓ Distribution ✓ RoEB ✓ SES</td>
<td>2</td>
</tr>
<tr>
<td>✓ Anchor ✓ Distribution ✓ SES</td>
<td>2</td>
</tr>
<tr>
<td>✓ Anchor ✓ Opinion-seeking ✓ SES</td>
<td>1</td>
</tr>
<tr>
<td>✓ Anchor ✓ RoEB ✓ SES</td>
<td>1</td>
</tr>
<tr>
<td>✓ Anchor ✓ SES</td>
<td>1</td>
</tr>
<tr>
<td>✓ Anchor ✓</td>
<td>13</td>
</tr>
<tr>
<td>✓ Anchor ✓</td>
<td>3</td>
</tr>
<tr>
<td>✓ Anchor ✓</td>
<td>1</td>
</tr>
<tr>
<td>✓ Anchor ✓</td>
<td>1</td>
</tr>
<tr>
<td>✓ Anchor ✓</td>
<td>1</td>
</tr>
<tr>
<td>✓ Anchor ✓</td>
<td>2</td>
</tr>
<tr>
<td>✓ Anchor ✓</td>
<td>1</td>
</tr>
<tr>
<td>✓ Anchor ✓</td>
<td>1</td>
</tr>
</tbody>
</table>

---

DOI: 10.3310/hta18280

HEALTH TECHNOLOGY ASSESSMENT 2014 VOL. 18 NO. 28
colleagues found their results to be identical for each of the methods used. When there was greater variation between the results found by the different methods, ranges tended to be reported, with values for each method falling within the range. Walters and Brazier used the anchor and SRM for data from eight studies on the European Quality of Life-5 Dimensions (EQ-5D) and Short Form questionnaire-6 Dimensions (SF-6D) measures, producing a weighted mean and range for the M(C)ID based on the anchor results, which was verified by comparing the SRM results with Cohen’s criteria. The study by Shulman and colleagues combined both anchor and effect size methods by using effect size cut-offs for the transition question on the anchor instrument.

For studies combining anchor and distribution methods, one study utilised an instrument as the anchor for which the MCID had already been established. In another case the anchor method estimate was already known but the study verified that the value was in agreement with the distribution method. McLeod and colleagues estimated a distribution value based on the instrument, noting that for each item in the instrument, a change of 0.5 units had been proposed as a minimum level for an important difference in previous studies. Therefore, the number of questions in the instrument is multiplied by 0.5 to give the value for which a MCID is likely to equal or exceed. Studies by Wynich and colleagues had noted the close proximity of SEM results to MCID results found using other methods and this was cited as the rationale for using this parameter. Some studies that had used 2.77 SEM (the two-measurement case) suggested that it is only when the resulting estimate exceeded the SD of the measurement in those categorised as ‘stable’ – those indicating that they experienced ‘no change’ in an anchor response – that the result can be interpreted as being clinically important. Another study used a similar approach in which the SD of only the ‘unchanged’/‘stable’ groups of patients, as defined by an anchor method, was used to calculate the distribution method estimates for clinical change; this was defined as the mean change plus 1.65 SDs of the change scores for this ‘unchanged’ group. Quinn and Wells anchored a VAS scale to another wound score to calculate the MCID. The mean difference in the VAS scale between ‘optimal and ‘suboptimal’ scores in the anchor was used; optimal and suboptimal score on the anchor was defined using a thumb of thumb type approach. Some studies reported the estimate found using both methods but did not analyse the values any further. Studies used both methods to identify a point estimate of important change and in at least one instance this was because there was agreement in the values found using both methods. More often, rather than reporting point estimates, studies reported a range of values for important differences, based on all methods used. Other studies used anchor-based values rather than distribution-based values to infer the level of an important difference. One study used the distribution method estimate rather than the anchor method estimate because the distribution value was higher and therefore a more conservative estimate. Other studies also used the largest estimate, which in the case of Kocks and colleagues was an average of two anchor values (as this method yielded higher results than the SEM), whereas Bols and colleagues specified that the point estimate had to be larger than the smallest ‘real’ change value found using a distribution method. Hsieh and colleagues used the larger of two anchor estimates, which exceeded the measurement error approach estimate.

The study by Lauridsen and colleagues had asked participants at baseline to rate the amount of change after treatment that they would find to be acceptable, and noted that this exceeded the minimum important change that was actually found. In one study, it was not possible to derive an estimate from the anchor method and the distribution estimate alone was used. Two studies examined the agreement between methods. As mentioned above, Wynich and colleagues considered the similarity between SEM values and anchor values, whereas Rejas and colleagues used formal methods (Cohen’s kappa and tau-b) to assess agreement. Both studies identified the closest agreement as being found with the anchor and 1.0 SEM methods. Polson and colleagues also did not choose a point estimate or range because of the cut-off for the anchor method being ‘much improved’ rather than ‘minimally improved’. Two studies used a ratio of minimum important change (i.e. anchor result) to MDC (distribution result) and reported findings based on whether or not the ratio exceeded 1. Two studies that had used a ROC curve to determine the anchor method estimate discussed the considerations for when sensitivity or specificity or both should be prioritised. A study by Bagó and colleagues also considered the
purpose of the research, noting that, for patient-reported outcomes, the anchor method may be preferred as it incorporates patients’ views. Of studies that used a SES method alongside distribution methods to calculate a target difference, a measurement error distribution method was used along with an effect size of 0.2 SDs, 0.5 SDs and/or 1.0 SD and then the sample difference was used to convert to a change.

All of the studies combining these methods used at least one method for calculating the SEM and also calculated a value for half a SD (0.5 SD). This corresponds to the value proposed by Cohen for a ‘moderate’-sized standardised effect, and also to the findings of Norman and colleagues and the value that they considered frequently yielded a MCID estimate similar to that derived by other methods (e.g. anchor-based methods). The ‘small’ SES proposed by Cohen (0.2 SDs) was also used in two studies. The remaining study used a higher effect size estimate of 1.0 SD, justifying this on the basis that individual change was being assessed rather than group changes. Kemmler and colleagues argued that clinical relevance requires a value that exceeds both the distribution and SES method estimates. Of the remaining studies, a range for clinical importance was provided based on values from all (distribution and SES) estimates used in three studies. Lemieux and colleagues rationale for using a range rather than a point estimate in their data was because they felt that the ranges were too wide to use a point estimate.

Of the studies that combined anchor, distribution and SES method estimates, most reported a range of values based on all methods used. Occasionally, point estimates were reported from within a range based on all methods used. Of those choosing the results of a particular method over others, most chose the results from the anchor method as the values reporting clinically important change; one study used the effect size (0.5 SDs) estimate on the basis of it being the most responsive to change and one study used a distribution estimate (2.77 SEM) as it exceeded the level of measurement error found by using 1.0 SEM and was close to the anchor estimate.

There were some notable variations across studies in the summarisation of estimates. A study by Broom and colleagues excluded mean change values found using anchor methods that had effect sizes not in the 0.2–0.6 SD range, arguing that these changes would be too trivial (<0.2 SDs) or too large (>0.6 SDs) to represent minimal change (e.g. change would be considered substantial rather than minimal if >0.6 SDs). This was also the case in two studies by Yost and colleagues, one excluding anchor estimates for which the SES was >0.8 SDs and the other excluding anchor estimates for which the SES was not within the 0.2–0.5 SD range. Some studies, in which a combination of methods were used to evaluate important change, also calculated other metrics that could have been used as part of the process but were not (e.g. SES). Yount and colleagues reported an overall range, but had also used point estimates for the anchor results and a range for the other methods. Cole and colleagues argued that anchor values should be used in considering individual change, and for group-level change distribution (SEM) estimates were preferable to SES (0.5 SDs) estimates because the latter does not account for the internal consistency of the measurement being used. A study by Williams and colleagues presented a wide range of values but then considered clustering of values within this range and reported the range within which values tended to cluster. Barber and colleagues calculated 95% CIs for each of the methods used and, finding overlap, used an anchor argument that the distribution method estimate supported it. Dubois and colleagues found a variable range of estimates across methods and suggested the use of cumulative distribution curves rather than single values. Brouwer and colleagues provided a point estimate across multiple instruments that used a 0–100 scale.

No studies used a health economic method in conjunction with any of the other methods. However, there is a natural role for other methods for populating the key parameters in a health economic method’s underlying model. For example, the anchor method could be used to define a range of clinically important differences, which is the first step in Detsky’s approach. Similarly, for the more complex health economic approach, judgement (explicit and implicit) about importance is needed.

Of the studies using the opinion-seeking method in conjunction with anchor methods, three by Wyrwich and colleagues used Delphi techniques followed by typical anchor methods for determining an
important difference. The study by Binkley and colleagues surveyed clinicians on the amount of change they would consider clinically important (both with and without stated options available for them to choose a particular level), and then used an anchor with additional ROC curves, with the anchor being each patient’s doctor’s rating the prospective change following forthcoming surgery. A study by Wells and colleagues used an anchor method in which patients compared themselves against fellow patients. From this, experts reviewed the values found using Delphi, and also used a 0.25 SD level, to determine the most appropriate level of minimal difference for importance. Many studies had used paper patients, or hypothetical patient scenarios, with numerous examples for experts being created from data sets, and it was unusual to find a study like that of Raj and colleagues who asked four experts to rate expected change for just two patient scenarios. The methodological distinction between an anchor and an opinion-seeking method can be narrow as both involve judgement. For example, clinicians could be asked to judge whether an important change had taken place or not among individual (hypothetical) patients for whom they are provided with the corresponding outcome values. An estimate of the MCID could be calculated using an anchor-based approach or a direct estimate of the MCID could be requested as opposed to a position on an anchor. Liang conducted similar surveys with experts to determine the probability that a particular patient has improved. The amount of change that had occurred when there was 70% agreement that change had occurred was chosen as the cut-off value.

When the RoEB method was reported alongside the opinion-seeking method, distinct estimates were sometimes not easily identified. It was difficult to determine instances in which reviewing of the evidence base was a method in its own right as opposed to being supplemental evidence to guide decision-making using opinion-seeking methods. Some studies providing estimates from a review of the evidence base to experts to aid their decision-making often also incorporated distribution or effect size estimates to assist this process. For example, Ornetti and colleagues calculated the MDC based on a 95% CI for mean differences found in the literature.

Synthesis of findings of different methods was sometimes straightforward. The RoEB method can be naturally used with other methods (e.g. SES) although determination of a final estimate of an important (and/or a realistic) difference may still be challenging. Mills and colleagues found that their opinion-seeking results were almost identical to their SEM results so either method would be suitable for calculating the MCID for comfort in footwear. Other methods of incorporating results from multiple methods used a point estimate from within a range across all methods used. One study by Samsa and colleagues reported using a pilot study to validate an estimate of the MCID, which was based on a RoEB method (see Review of evidence base method). The MCID estimate was derived using an anchor method. For two studies that used four methods, the opinion-seeking method was used to confirm or synthesise the findings of the other methods.

Discussion

Key findings

The most frequently identified methods used to determine an important difference were the anchor, distribution and SES methods. There are various reasons for the popularity of these methods compared with other identified methods for establishing an important difference. Such methods are commonly used in instrument validation studies, reflecting the need to assess the importance of different magnitudes of change in quality-of-life instruments. For example, the ease with which a distribution method in its simpler forms can be used is attractive, and when the anchor method is carried out, the other two can be readily carried out. No new methods were identified by this review beyond the seven pre-identified methods. However, multiple studies of each method were detected and substantial variations in implementation, even for relatively simple approaches such as the anchor method, were noted. Studies overwhelming focused on a continuous outcome as opposed to other types of outcomes, although other outcome types were evaluated in opinion-seeking and RoEB method studies.
A number of key issues were common across the methods. It is critical to decide whether the focus is to determine an important and/or a realistic difference. Some methods can be used for both (e.g. opinion-seeking methods) and some for only one or the other (e.g. anchor or pilot study methods). Comparison of the way that the difference was determined and the context of the target difference are important. Some approaches commonly used for determining an important difference either cannot be used for specifying a target difference (statistical test-based approach), or do not straightforwardly translate into the typical RCT context (measurement error approach). For methods that involve judgement (anchor, opinion-seeking and health economic methods), the perspective adopted (i.e. whose values and outlook) is a key consideration.

Some methodological issues are specific to particular methods. For example, the necessity of choosing a cut-off point to define an ‘important’ difference/change is specific to the anchor method. This approach is a widely recognised part of the validation process for new quality-of-life instruments for which the scale has no inherent meaning without reference to an outside marker (i.e. anchor). Distribution methods fell into three main categories: measurement error, statistical test and rule of thumb based. All have clear limitations, with the first not matching the setting of a standard RCT design (two parallel groups) as it is typically based on within-person measurement error. The second cannot be used to specify a target difference given that it is, in essence, a rearranged sample size formula. The last is dependent on the interpretability of the individual scale. The SES method was used in a substantial number of studies for a continuous outcome although very few formal usages for non-continuous outcomes were noted, even though informal use of such an approach is likely to be widespread. Cohen’s interpretation of effect size was typically relied on with a within-person effect size calculated as opposed to the standard trial context of between groups. No parallel for a binary outcome exists although equivalent approximate odds ratio values to Cohen’s $d$ values can be used. The validity of Cohen’s values is uncertain despite wide usage and some proposed modifications.\(^{23,341}\)

The opinion-seeking method was often used with multiple strategies involved in the process (e.g. questionnaires being sent to experts using particular sampling methods with an additional conference being organised to discuss findings in more detail). Delphi nominal group techniques for face-to-face meetings are increasingly used in research (e.g. instrument development through a Delphi process) and are potentially useful. In terms of planning a trial, the opinion-seeking method can be relatively easy to conduct but resulting estimates of a target difference may be of low value depending on the robustness of the approach used to elicit opinions.

The health economic or pilot study method was infrequently reported as a specific method although the latter has been commonly used, albeit not without problems. For the health economic method this is likely due to the complexity of the method and/or the resource-intensive procedures that are required, with some implementations being recent methodological developments. The use of pilot studies to determine the target difference is problematic and probably only useful for the control group event proportion or SD for a binary or continuous outcome respectively. Internal pilot studies may be incorporated into the start of larger clinical trials but are not useful for specifying the target difference, although they could be used to revise the sample size calculation. Pilot studies and the RoEB method may have been conducted informally by triallists (for the purpose of trial design), but have not ultimately been published. The RoEB method can be applied to identify both an important and a realistic difference, whereas a pilot study addresses only a realistic difference. In using both methods consideration is needed of the applicability to the anticipated study and the impact of statistical uncertainty on estimates.

A RoEB approach for a particular outcome measurement or study population may be incorporated into any of the other methods identified for establishing an important difference. However, the number of studies reporting a formal method for identifying an important difference using the existing evidence was surprisingly small. The wide variation in the extent to which reviews of the existing evidence base have been undertaken prospectively with a specific and formal strategy may be due to the lack of availability of multiple studies until recently.
Many of the identified difficulties in interpreting the resulting estimate for an important or a realistic difference (e.g. the effect of baseline severity on estimates, reliability of measurements) are not solely applicable to such studies. Nevertheless, they are important issues for triallists to consider when using such studies to determine a target difference. For example, an important difference estimate for improvement is arguably not identical to the threshold for deterioration. Recall bias was one of the most commonly cited limitations in studies which used an anchor approach and where patients rated their change over time, and the vulnerability of measuring perceived change or satisfaction over time to other potential biases (e.g. response shift) was noted.

Some methods can be readily used in conjunction with others, which could increase the robustness of their findings. The anchor and distribution methods were often used together within the same study and in a substantial number of studies were also used with the SES approach. Multiple methods for determining an important difference were used in some studies, although the combinations varied, as did the extent to which results were triangulated (if any form of triangulating results was used). The result found using one method may validate the result found using another method, but on the other hand the use of multiple methods seems to have created increasing uncertainty over the estimate of important difference in some studies.

**Strengths and limitations**
The strengths of this review are that it is a comprehensive systematic review of methods that could be used to specify the target difference in a sample size calculation of a future RCT. To our knowledge this is the first such review. A large number of databases were searched using detailed search methods to ensure that all relevant studies reporting this issue could be incorporated into the review. From the search results, it is clear that this review covers a large body of evidence on establishing an important and/or realistic difference that can be utilised in RCT design for specifying the target difference. As is natural when addressing such a research question in a single systematic review, there were several practical difficulties.

The absence of standardised terminology for the concepts of an important, realistic target difference necessitated a broad search strategy that identified a large number of titles and abstracts for screening. Some of the methods are also very different in their conception (e.g. health economic vs. anchor) and this prevented the development of a more specific and possibly sensitive search strategy; in addition, even within method types terminology can be variable. This required several compromises to be made with regard to the screening process. For example, titles and abstracts were screened by a single reviewer because of the large number of studies, except when they were classified as uncertain when a second reviewer also assessed them. However, there was regular discussion between reviewers about individual studies under consideration and use of a third arbitrator when necessary. Although it is clear that standardisation of terminology would be very useful, it may be difficult to achieve in practice given that terminology usage can reflect disciplinary background and slightly different purposes. For example, it is useful to note that patient-reported outcome advocates favour not including the words ‘clinical’ or ‘clinically’ in their terms as it has been argued that this increases the focus on the values of the clinician rather than the patient being the key participant in their own care.

The quality of data collected and reporting in included studies were variable and this caused difficulties. Quality and reporting of data may depend on a study’s purpose. In most included studies, sample size calculations were not the main reason for conducting the research to identify an important difference, whereas they often were for studies that addressed a realistic difference. The common usage of two methodologic approaches (pilot study and RoEB), although often for other purposes, was not reflected in many included studies. The full-text reports of such studies were not included unless some reference had been made to planning a future RCT or determining a realistic and/or important difference.
Chapter 3 Surveys of triallists’ current practice

Introduction

Although methods for specifying the target difference used in RCT sample size calculations exist, as identified in the systematic review of the literature reported in the previous chapter, it is unclear to what extent triallists are aware of and use these methods when designing clinical trials. Trial reports and protocols will typically report the sample size calculation and the values assumed therein.22 The process of determining inputs into the sample size calculation (including the target difference) typically lacks detail, particularly in reports of trials in which there are space restrictions. Arguably it is those with practical experience of designing RCTs who are the best placed to provide advice about the use of such methods. Furthermore, there may be other methods, or existing methods that are implemented in a way that has not been captured in the systematic review of the literature reported in Chapter 2. To address this the usage of methods among leading clinical triallists was assessed.

This was achieved using two related, although distinct, surveys, one of the membership of the international SCT419 and one of UK- and Ireland-based triallists.420–422 The aim of the surveys was to evaluate current practice among clinical triallists – specifically, which methods respondents were aware of, used and would be willing to recommend. These data were used to establish a baseline from which to consider the requirements for guidance and also to provide some information about perceived strengths and limitations of the different approaches for eliciting a target difference. Although the two surveys were essentially the same, that given to UK- and Ireland-based triallists was slightly more extensive (see later for details). The methodology and findings of the two surveys are reported in this chapter.

Methodology of the surveys

Survey 1: Society of Clinical Trials membership

Members of the SCT were surveyed through the email distribution list for this organisation.419 This is the only international society specifically supporting the conduct of RCTs that is not restricted to a specific clinical area. The survey asked generic questions about the respondent (position, affiliation, location and involvement in the design of RCTs), and his or her group’s awareness and usage of methods for determining the target difference, in addition to providing an opportunity to suggest an additional method. A brief summary of each of the seven previously identified methods was provided on the online form. Additionally, the respondent was asked whether or not they would be willing to recommend the use of any of the methods. Finally, an opportunity to comment on the topic was provided. Members received an email via the society’s email distribution list inviting them to complete the online survey. The invitation included a brief introduction to this issue and the aim of the survey. The online survey was implemented bespoke for this purpose by the Health Services Research Unit (HSRU) programming team, University of Aberdeen. Once potential participants received the email, they were able to access the survey by clicking on, or typing in, the website link provided and to complete and submit their responses. An email reminder was sent out 1 week after the initial email invitation. As it was not possible to tailor reminders to individuals who had not completed the survey, a general reminder was sent to the entire study sample.

Survey 2: UK- and Ireland-based triallists

The sample for the survey of UK- and Ireland-based triallists included three groupings who contribute to trial design: UKCRC-registered CTUs,421 the regional National Institute for Health Research (NIHR) Research Design Service (RDS) offices in England422 and the MRC UK Hubs for Trials Methodology Research420 (as of 24 August 2011). One individual (typically the director) from the CTUs, MRC Hubs and RDS offices was invited to complete the survey. When the same individual held a position at more than one entity, only one
survey was sent and a response on behalf of the relevant groups requested. Individuals were requested to forward the survey on to the appropriate member of their group if they were not personally able to complete it.

In addition to the information collected in the SCT survey, this survey requested information about the most recent trial developed by the group (see Appendix 4). These details included the underlying basis adopted for the target difference (e.g. realistic difference or important difference) and any methods used for determining the target difference. Additionally, respondents were asked if there is anything that would aid them in the design of RCTs and if they would be happy to be contacted for further details.

The initial request was personalised and sent by post. It included an invitation letter, paper version of the survey and description of the methods available for determining the difference. A paper reminder was sent 2 weeks after the initial notification of the survey. Following this, an additional (final) email reminder was sent a week later including an electronic invitation, version of the survey and description of the methods.

**Ethical review**
The surveys were approved by the University of Aberdeen’s College of Life Sciences and Medicine Ethics Review Board (CERB/2011/6/657). This project abided by the MRC’s guidance on good research practice and conformed to the University of Aberdeen’s research governance guidelines. We piloted the survey invitations and formats with members of the project team and local researchers. The responses to the online survey and submitted survey data are stored within a secure database on a secure server within the HSRU.

**Data analysis**
The surveys were analysed separately. The response rate was defined as the respective number of responding participants divided by the number of potential participants in the survey population. Data were summarised quantitatively or narratively as appropriate. No statistical analysis was carried out. Survey results were discussed with the steering and advisory groups and used to develop the guidance on methods for eliciting the target difference (see Chapter 4).

**Results**

**Survey 1: Society of Clinical Trials membership**
Of the 1182 members on the SCT membership email distribution list, 180 responses were received (15%). However, only 519 of the society’s members described themselves as statisticians, epidemiologists or health professionals, who might be viewed as most suited to completing the survey. Respondent characteristics are given in Table 7. Thirteen countries were represented although more than 75% of respondents were from North America (127 and 15 from the USA and Canada respectively). Eighteen respondents were based in the UK. The vast majority of respondents were statisticians/epidemiologists with 13 being health professionals. The majority were affiliated to an academic institution, with similar numbers from a contract research organisation, private industry and a regulatory authority. Of the 180 respondents, 162 (90%) stated that they were presently involved in trial design.

The responses regarding awareness, usage and willingness to recommend methods are given in Table 8. Awareness of methods ranged from 38% (n = 69) for health economic methods to 90% (n = 162) for pilot study methods. No additional method was reported. The use of adaptive designs such as the continual reassessment method for finding the optimal dose and treatment selection models was highlighted by one respondent although these typically have an arbitrary, although prespecified, sample size. The use of the ‘CI approach’ was highlighted by one respondent for non-inferiority studies in which the lower bound of the CI for the treatment difference of the accepted treatment compared with placebo is used to define the margin of non-inferiority; the aim of the non-inferiority study is to rule out a difference of a fraction (say 50%) of the previous effect.
As expected, usage was lower than awareness and ranged from 9% (n = 16) for the health economic method to 74% (n = 133) for the pilot study method. Awareness and usage of ‘reverse engineering’ was highlighted by one respondent in which a sample size is chosen (e.g. the largest thought feasible within recruitment and/or financial constraints) and the corresponding minimum difference calculated using a rearrangement of the appropriate sample size formula.

The highest level of willingness to recommend among respondents was for the RoEB method (73%), with the lowest being for the health economic method (16%). Willingness to recommend was lower than or equal to awareness and usage for all methods except for health economic method, although the recommendation for this method was still substantially lower than awareness (16% vs. 38% respectively). Willingness to recommend among those who had used a particular method is also shown in Table 8; levels of recommendation were substantially higher than across all respondents, ranging from 56% for the opinion-seeking method to 89% for the RoEB method.

**Table 7** Survey 1: respondent characteristics (n = 180)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n (% of respondents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location*</td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>127 (71)</td>
</tr>
<tr>
<td>UK</td>
<td>18 (10)</td>
</tr>
<tr>
<td>Other European country</td>
<td>10 (6)</td>
</tr>
<tr>
<td>Canada</td>
<td>15 (8)</td>
</tr>
<tr>
<td>China</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Japan</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Australia</td>
<td>3 (2)</td>
</tr>
<tr>
<td>African country</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Profession</td>
<td></td>
</tr>
<tr>
<td>Health professional</td>
<td>13 (7)</td>
</tr>
<tr>
<td>Statistician/epidemiologist</td>
<td>153 (85)</td>
</tr>
<tr>
<td>Other scientist (e.g. ethicist or behavioural scientist)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Trial staff</td>
<td>8 (4)</td>
</tr>
<tr>
<td>Other</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Institution*</td>
<td></td>
</tr>
<tr>
<td>Academic institution</td>
<td>103 (58)</td>
</tr>
<tr>
<td>Contract research organisation</td>
<td>23 (13)</td>
</tr>
<tr>
<td>Governmental agency</td>
<td>17 (9)</td>
</tr>
<tr>
<td>Health-care provider</td>
<td>6 (3)</td>
</tr>
<tr>
<td>Private industry</td>
<td>24 (13)</td>
</tr>
<tr>
<td>Other</td>
<td>6 (3)</td>
</tr>
<tr>
<td>Currently involved in trial design</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>162 (90)</td>
</tr>
</tbody>
</table>

a Based on n = 179 as one respondent did not complete.
A number of comments were made by respondents regarding the use of particular methods. One respondent recommended using a combination of anchor, distribution, RoEB and opinion-seeking methods. Another stated that they would recommend using a pilot study only for estimating variance in a sample size calculation and would use the opinion-seeking method only for effect size. One noted that a pilot study should not be used for estimating the effect size and another stated the need to consider variability in pilot estimates. One respondent stated that they would be least likely to recommend the SES approach. The limitations of all of the methods they had used were noted by one respondent, who desired to know more about the other approaches. Another stated that they did not favour any method over the others. One stated that they viewed the value of information approach (a health economic method) as the only one that gave the optimal sample size and one noted hesitancy about the anchor method based only on patient or clinician judgement. Two other respondents noted that the approach to be recommended was trial specific and dependent on the outcome, resources available to implement methods and current knowledge. Another stated wariness about relying on any one method/source of data. Finally, the need to undertake validation of clinician estimates of effect size that would change their practice is needed, that is, how often practice is changed by a smaller effect or how often an effect of the proposed size is not sufficient to alter practice.

Beyond the use of a particular method the following points were made in relation to the process of specifying a target difference and sample size. Use varies by the stage of research and the expectation for Phase III (confirmatory) studies was greater than that for Phase II studies, which may have little to base the target difference on. One respondent suggested that the specification of the difference should be consistent with the proposed analytical approach. Another noted that they considered a lack of data more of a concern than the methods used to determine the difference. One respondent queried how often the effect size is honestly chosen as opposed to being picked to make the sample size ‘affordable’. Another stated that a regulatory body had on one occasion ‘told’ the investigators what size of study to use. The lack of awareness of the issue among clinicians was highlighted. Finally, the possibility that two similar trials could differ with respect to the target difference because of differences between the clinical investigators’ aims and objectives was suggested.

Survey 2: UK- and Ireland-based triallists
Information on the groups represented is given in Table 9. Of the 61 surveys sent out, 34 (56%) responses were received, representing 25 (52%) CTUs, five (63%) Hubs and eight (80%) RDS offices (some respondents having more than one affiliation). Respondents were predominately directors of one of these bodies (76%, n = 26), with the remainder being statisticians (9%, n = 3) or other (15%, n = 5). The vast
### TABLE 9  Survey 2: respondent characteristics (n = 34)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n (% of respondents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representing†</td>
<td></td>
</tr>
<tr>
<td>CTU</td>
<td>25 (52% of CTUs)</td>
</tr>
<tr>
<td>Hub</td>
<td>5 (63% of Hubs)</td>
</tr>
<tr>
<td>RDS</td>
<td>8 (80% of RDS offices)</td>
</tr>
<tr>
<td>Position</td>
<td></td>
</tr>
<tr>
<td>Director</td>
<td>26 (76)</td>
</tr>
<tr>
<td>Statistician</td>
<td>3 (9)</td>
</tr>
<tr>
<td>Other</td>
<td>5 (15)</td>
</tr>
<tr>
<td>Intervention types in trial portfolio†</td>
<td></td>
</tr>
<tr>
<td>Pharmacological</td>
<td>29 (88)</td>
</tr>
<tr>
<td>Non-pharmacological</td>
<td>32 (97)</td>
</tr>
<tr>
<td>Phase of trial in portfolio‡</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>1 (3)</td>
</tr>
<tr>
<td>II</td>
<td>24 (73)</td>
</tr>
<tr>
<td>III</td>
<td>31 (94)</td>
</tr>
<tr>
<td>IV</td>
<td>19 (58)</td>
</tr>
<tr>
<td>Clinical area in trial portfolio‡</td>
<td></td>
</tr>
<tr>
<td>Blood</td>
<td>5 (15)</td>
</tr>
<tr>
<td>Cancer</td>
<td>22 (67)</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>18 (55)</td>
</tr>
<tr>
<td>Dementias and neurodegenerative diseases</td>
<td>15 (45)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>16 (48)</td>
</tr>
<tr>
<td>Ear</td>
<td>4 (12)</td>
</tr>
<tr>
<td>Eye</td>
<td>6 (18)</td>
</tr>
<tr>
<td>Genetics and congenital disease</td>
<td>6 (18)</td>
</tr>
<tr>
<td>Infection</td>
<td>12 (36)</td>
</tr>
<tr>
<td>Inflammatory and immune</td>
<td>7 (21)</td>
</tr>
<tr>
<td>Injuries and emergencies</td>
<td>10 (30)</td>
</tr>
<tr>
<td>Medication for children</td>
<td>14 (42)</td>
</tr>
<tr>
<td>Mental health</td>
<td>19 (58)</td>
</tr>
<tr>
<td>Metabolic and endocrine</td>
<td>8 (24)</td>
</tr>
<tr>
<td>Musculoskeletal</td>
<td>17 (52)</td>
</tr>
<tr>
<td>Neurological</td>
<td>11 (33)</td>
</tr>
<tr>
<td>Oral and gastrointestinal</td>
<td>14 (42)</td>
</tr>
<tr>
<td>Primary care</td>
<td>24 (73)</td>
</tr>
<tr>
<td>Renal</td>
<td>13 (39)</td>
</tr>
</tbody>
</table>

*continued*
majority stated that their group dealt with both pharmacological (88%, \( n = 29 \)) and non-pharmacological (97%, \( n = 32 \)) trials. With regard to RCT phases, the groups’ trial portfolio contained 24 (73%), 31 (94%) and 19 (58%) Phase II–IV trials respectively. One group (3%) reported also undertaking a Phase I study. All clinical areas under the NIHR UK portfolio categorisation were represented, with frequencies ranging from 12% (\( n = 4 \)) for ear-related research to 73% (\( n = 24 \)) for primary care-related research.

The responses regarding awareness, usage and willingness to recommend methods are given in Table 10. Awareness of methods ranged from 97% (\( n = 33 \)) for the RoEB and pilot study methods to 41% (\( n = 14 \)) for the distribution method. No other methods were suggested and all stated that they had used at least one of the methods. Use of each method was substantially less than awareness of the respective method except for the pilot study, RoEB and SES methods. The largest drop-off between awareness and use was for the opinion-seeking and health economic methods. Awareness of the opinion-seeking method was 88% (\( n = 30 \)) and use was 53% (\( n = 18 \)); for the health economic method the corresponding figures were 62% (\( n = 21 \)) and 24% (\( n = 8 \)). All respondents were aware of at least one of the different formal methods for determining the target difference. Almost all had used at least two methods (94%). One respondent stated that their group had not used any of the methods as they had only recently formed. The difficulty in differentiating between some of the methods without full definitions was noted by one respondent.

The highest level of willingness to recommend was for the RoEB method (76%) followed by the SES method (65%). The latter two had willingness to recommend levels that were slightly higher than levels for use. Willingness to recommend was around 50% for the other methods. The vast majority (88%) recommended more than one method. One specifically suggested using the anchor and RoEB methods in combination. Two (6%) stated that they would not recommend any method.

Data on the most recent trial that each group had been involved with are given in Tables 10 and 11. Based on the most recent trial, all bar three (91%) groups used a formal method and all the prespecified methods were in use. Two respondents reported that their group had used alternative informal methods: reverse engineering the study sample size to ensure that the research cost fell ‘within [the] funding range’, and basing it on the lead clinical applicant’s opinion, although this was not formally elicited. There was one further case which was a recently formed group that had not started designing RCTs. A total of 21 (64%) groups stated that they used more than one formal method for the most recent trial. The most common type of primary outcome was a clinical function measure (33%) followed by a mortality outcome (27%). Disease-specific (21%) and generic (12%) quality-of-life measures were also represented in multiple studies. A non-quality-of-life patient-reported outcome and a health economic measure were each reported twice as being used as the primary outcome. Other outcome types were non-mortality time-to-event (6%), cardiovascular events (6%), weight-related outcomes (6%), length of stay and violent events. In one case (3%) there was no primary outcome, and 11 (33%) had more than one primary outcome.
The vast majority stated that the chosen target difference was one that was viewed as important by a stakeholder group (91%). Just over half (61%) stated that the basis for determining the target difference was to achieve a realistic difference given the interventions under evaluation. Eleven (33%) stated that it was to determine a difference that gave an achievable sample size. Two other approaches for eliciting the target difference were reported. One considered what difference would be worthwhile detecting given

### TABLE 10 Survey 2: awareness, use and willingness to recommend methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Aware of, n (%)</th>
<th>Used, n (%)</th>
<th>Recommend, n (%)</th>
<th>Recommend if used, n (%)</th>
<th>Most recent trial, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor</td>
<td>22 (65)</td>
<td>15 (44)</td>
<td>16 (47)</td>
<td>13 (87)</td>
<td>6 (18)</td>
</tr>
<tr>
<td>Distribution</td>
<td>14 (41)</td>
<td>8 (24)</td>
<td>9 (26)</td>
<td>3 (38)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Health economic</td>
<td>21 (62)</td>
<td>8 (24)</td>
<td>11 (32)</td>
<td>5 (63)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Opinion-seeking</td>
<td>30 (88)</td>
<td>18 (53)</td>
<td>18 (53)</td>
<td>14 (78)</td>
<td>9 (27)</td>
</tr>
<tr>
<td>Pilot study</td>
<td>33 (97)</td>
<td>30 (88)</td>
<td>20 (59)</td>
<td>20 (67)</td>
<td>8 (24)</td>
</tr>
<tr>
<td>RoEB</td>
<td>33 (97)</td>
<td>32 (94)</td>
<td>26 (76)</td>
<td>25 (78)</td>
<td>17 (52)</td>
</tr>
<tr>
<td>SES</td>
<td>30 (88)</td>
<td>28 (82)</td>
<td>22 (65)</td>
<td>22 (79)</td>
<td>14 (42)</td>
</tr>
<tr>
<td>Other</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>NA</td>
<td>0 (0)</td>
</tr>
<tr>
<td>None</td>
<td>0 (0)</td>
<td>1 (3)</td>
<td>2 (6)</td>
<td>NA</td>
<td>3 (9)</td>
</tr>
</tbody>
</table>

NA, non-applicable.

a Based on n = 33 as one CTU had only recently started.

### TABLE 11 Survey 2: most recent trial (n = 33)*

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n (% of respondents)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome</td>
<td></td>
</tr>
<tr>
<td>Generic quality of life (e.g. EQ-5D)</td>
<td>4 (12)</td>
</tr>
<tr>
<td>Disease-specific quality of life (e.g. Oxford Knee Score)</td>
<td>7 (21)</td>
</tr>
<tr>
<td>Other patient-reported outcome (non-quality-of-life measure)</td>
<td>2 (6)</td>
</tr>
<tr>
<td>Mortality</td>
<td>9 (27)</td>
</tr>
<tr>
<td>Clinical functional measure (e.g. forced expiratory volume)</td>
<td>11 (33)</td>
</tr>
<tr>
<td>Economic outcome (e.g. incremental cost per QALY)</td>
<td>2 (6)</td>
</tr>
<tr>
<td>Other</td>
<td>8 (24)</td>
</tr>
<tr>
<td>There was no primary outcome</td>
<td>1 (3)</td>
</tr>
<tr>
<td>What was the underlying principle(s) adopted in determining the difference?</td>
<td></td>
</tr>
<tr>
<td>A realistic difference given the interventions under evaluation</td>
<td>20 (61)</td>
</tr>
<tr>
<td>A difference that would led to an achievable sample size</td>
<td>11 (33)</td>
</tr>
<tr>
<td>A difference that would be viewed as important by a relevant stakeholder group (e.g. clinicians)</td>
<td>30 (91)</td>
</tr>
<tr>
<td>Other</td>
<td>2 (6)</td>
</tr>
</tbody>
</table>

a Based on n = 33 as one CTU had only recently started.
the cost of the intervention, and the other considered what magnitude of a target difference (and hence size and cost of project) would likely be funded. All used at least one basis for determining the difference. Ten (30%), seven (21%) and two (6%) reported using two, three and four bases of consideration when determining the target difference respectively. In total, 16 of the 19 that used two or more bases stated that they sought both a realistic and an important difference. Two respondents stated that they used an additional basis in combination with all three prespecified approaches (a realistic, an important and an achievable difference); the additional bases were cost of the intervention and the ‘likelihood of securing funding’.

A number of general points about specifying the target difference were made by respondents. One respondent noted that they used a modification of the SES method in which a SES range of 0.4–0.5 SDs for a continuous outcome was used if supported by the evidence base. For a binary outcome, a 50% relative reduction in risk was used. However, if the intervention was low cost then a smaller SES (e.g. 0.25 SDs for a continuous outcome) was acceptable. The difficulty in defining ‘clinically important’ was noted and it was suggested that defining it at a population level was more useful than defining it at an individual level. The potential danger of basing the sample size on the current evidence if the available studies were small and/or of poor quality was noted. Additionally, publication bias may lead to an estimate based on available studies being an inflated estimate of the true effect. One respondent noted the value of up-front health economic modelling. The need to ‘factor in’ a realistic recruitment rate was also raised. Different studies might lead to different methods for specifying the target difference being preferred.

A number of factors were reported that would make it easier to determine the target difference for RCTs. Five (15%) highlighted the need for a central resource to clarify what a clinically important difference for common outcomes (one specifically stated from a patient perspective) is, and also how to utilise existing evidence and health economic approaches. Particular need for guidance for non-standard trials (multiarm non-inferiority, equivalence trials) was raised by two respondents. Education of clinical collaborators was highlighted by two respondents, one stating that a principal investigator should be expected to have carried out a systematic review before designing a new study and another noting that general education on the issues involved in determining the target difference and sample size was needed. The role of funders was raised by three (9%) respondents: the need for guidelines from funders, the need for educating funders that large and expensive trials may be necessary and the need for funding of pilot studies. Better reporting of outcomes and clinically relevant subgroups to enable summary statistics to be extracted was highlighted, as was the particular need for patient-reported outcome data.

**Discussion**

**Key findings**

The two surveys provide insight into current practice among clinical trialists regarding specification of the target difference. To our knowledge this has not been investigated before. Responses suggest that use of formal methods is greater than would appear from trial reports or, at least, the level of use is higher for the type of RCTs that the UK and Ireland trialist survey represents (trials conducted by leading experts). Variations in awareness, use and willingness to recommend between methods were substantial. The two surveys represent different groups: an international society of people involved in clinical trials and leading UK- and Ireland-based trialists. There were some differences in the absolute levels between the two groups, which might be expected given the more heterogeneous composition of the SCT sample. Nevertheless, the findings support the view that sample size calculation is a more complex process than would appear to be the case from trial reports and protocols.

Reported awareness of formal methods was high for most methods although it was substantially lower for the anchor, distribution and health economic methods; this was a common finding across both surveys. Awareness of the opinion-seeking method was lower in the SCT sample than among the UK and Ireland
trialists. This may reflect a greater focus among the SCT membership on pharmacological trials conducted for regulatory approval, with the Phase II trial typically informing the Phase III trial and the Phase II sample size influenced by convention/regulatory body expectations. The pilot study and RoEB methods were the most commonly used methods followed by the SES method. The health economic, anchor and distribution methods were the least commonly known and used methods. With regard to recommendations, the RoEB method consistently had the highest level of recommendation across the two surveys. In both surveys the use of an informal approach such as ‘reverse engineering’ to suit expected recruitment and associated research costs was mentioned. Slightly more respondents were willing to recommend the health economic method than had actually used it, perhaps reflecting both the intuitive appeal of this approach and the fact that cost considerations influence decisions even when not explicitly stated. Recommendation among users of the health economic method was substantially lower, although this was based on small numbers.

Multiple methods for determining the target difference were recommended by a substantial number of respondents and the need to use more than one method was highlighted by a number of respondents. Some caveats and recommendations were noted, such as using pilot data only to inform variability estimates, for example, the SD for a continuous outcome. Another respondent suggested that the SES is acceptable but only in the absence of better data. Provision of such information along with key issues to consider when conducting such a calculation would seem a useful addition to the current literature.

The basis for calculating the target difference was further explored in the survey of UK- and Ireland-based trialists by referring to the last trial that the group had been involved in. The details provided by respondents about the most recent trial conducted reflected a wide variety of outcome types, for which all of the methods were used to varying degrees. Although the vast majority stated that the basis for the target difference was a difference viewed as important by a stakeholder group, around half also used another basis and two used four separate principles for determining the target difference. Furthermore, the majority of respondents stated that they had used more than one method in determining the sample size calculation. Such complexity of considerations, in our experience, is rarely reported in the sample size calculation section either in the trial report or in the protocol and was not reflected in the literature considered for inclusion in the systematic review reported in Chapter 2. This view is supported by a review of RCT sample size calculations. Clearer and more explicit reporting of the basis for determining the target difference, including any formal methods used, is needed.

The importance of the views of funders and regulatory bodies was highlighted, as was how they shape the determination of sample size. Funders naturally consider the associated research cost of the study and the feasibility of the proposal. The desirable precision may be prohibitively high or impractical to achieve. It was noted that it can be difficult to secure funding for large and expensive trials, although a large (and by extension an expensive) trial may be needed to answer the research question. The expectation of regulatory bodies was also raised. How often the target difference is honestly chosen as opposed to being picked to make the sample size ‘affordable’ was queried. The need to ‘factor in’ a realistic recruitment rate was raised. For obvious reasons related to the perceived quality of research, such a consideration is not, typically, reported transparently. How cost and feasibility should be taken into account may be unclear to applicants and may lead to a reluctance to be explicit about the practical considerations. The use of some health economic models (specifically those based on the principles of economic evaluation) in which the research costs can be incorporated along with current evidence to determine ‘optimal’ study size could potentially provide a more transparent approach. Furthermore, such an approach could be tailored to a funder’s perspective and considerations.

The need for guidance through a central resource to clarify what is clinically important for common outcomes and guidance on using methods was highlighted. However, the approach used will likely be trial specific (e.g. it may relate to the phase of the trial) and dependent on the type of outcome, the resources (time, expertise, etc.) available to implement methods, current knowledge of the clinical area and also the proposed trial analysis (e.g. Bayesian or update of a meta-analysis).
Strengths and limitations
The response rates achieved were relatively low despite the surveys being short and well presented and despite utilising reminders. However, it seems unlikely that non-response has led to unrepresentative findings, and the low response rates probably reflect the difficulty of achieving a high level of response in certain population groups and when email distribution lists are used, as was the case for the SCT survey. A further factor may be the nature of the survey: it focused on RCT methodology, which made determining who should be invited to respond, and perhaps completion of the survey, less than straightforward. One respondent noted the difficulty of differentiating between some of the methods, despite a brief description of each method being provided; furthermore, reported awareness may be slightly higher than the true value as it was necessary to present the pre-identified methods with descriptions to achieve informed responses. There may also be a very small amount of overlap in respondents between the two surveys. Nevertheless, the surveys provided insightful information about the practice and views of trialists regarding determination of the target difference.

The survey of UK- and Ireland-based trialists included MRC Hubs for Trial Methodology and the RDS in addition to CTUs. This reflected our experience of the varied way in which the design of RCTs is dealt with in different parts of the UK and Ireland. For example, in Scotland, CTUs typically design trials ‘in house’ or have a very close affinity with a research group within their host institution. However, in England the NIHR RDS may take on this role or at least elements of trial design. Furthermore, across the UK and Ireland, the establishment of the MRC Hubs has altered the clinical trial landscape. They provide, at least in the primary location(s) of the hub, an additional grouping of trial expertise. Given this, all three groupings (CTUs, Hubs and RDS) were included in the sample although individuals were allowed to respond on behalf of more than one entity.

Overall, the surveys provide valuable information on the awareness and use of methods by trialists and their views. Variations in practice exist and a key requirement highlighted was the need for guidance documentation to inform the process of target difference determination. The need for more transparent reporting of the issues considered when specifying the target difference is also needed. Greater clarity and possibly formalisation of the judgement process of funders is needed.
Chapter 4  Guidance on specifying the target difference in a randomised controlled trial sample size calculation

Sample size calculations for randomised controlled trials

Background
A challenge for the prospective trialist is that from an ethical viewpoint no more participants should be recruited than are necessary to answer the research question. Furthermore, recruitment to RCTs can be extremely time-consuming and resource intensive and many trials fail to meet their recruitment target or have to extend beyond the original recruitment period. Given these considerations, adopting an appropriate sample size is of critical importance. Furthermore, the finite financial resources available should be used efficiently. The target difference can be viewed as the key input into a RCT sample size calculation as it quantifies what is sought. The aim of this chapter is to provide guidance (primarily for researchers) about how to specify this component of the sample size for a definitive RCT.

The calculation of the sample size for a RCT, or at least the reporting of it, is arguably as much an art form as a scientific endeavour. Nevertheless, it is important to try to estimate and report the sample size in as robust and transparent a fashion as possible. Surprisingly, little information is available on the process of determining the sample size and specifying the target difference in particular. This is true even within RCT textbooks and peer-reviewed articles, which overwhelmingly focus on the statistical aspect of the sample size calculation (see Chapter 2). This perhaps reflects the difficulty of providing a satisfactory answer to a difficult question, one that inherently requires judgement to address. Nevertheless, given that it is central to the primary utility of a RCT, there is merit in more explicit consideration and transparent reporting of how the value adopted for the target difference was derived. Uncertainty about what is being sought will be reflected in uncertainty in the interpretation of RCT results.

In this chapter, the scope is restricted to what might be termed the conventional, or standard, approach to sample size calculation (Box 5; see Boxes 6–9 for examples): a stand-alone trial utilising the conventional statistical framework for sample size calculation – namely a Neyman–Pearson framework. Additionally, the focus will be on the two-group parallel design, which is the design adopted for the majority of RCTs. Other designs, such as crossover and cluster trials, follow the same general process only requiring a different formulation for the sample size calculation and specification of different parameters (e.g. intracluster correlation coefficient). Alternative statistical approaches have been proposed and are briefly discussed below. Furthermore, the main focus will be on an assessment of superiority, a study that seeks to find whether there is a difference between two interventions in favour of either. Other research questions will be briefly considered. Finally, we focus primarily on what might be termed a definitive or Phase III/IV randomised trial that seeks to provide a clear answer to the research question. Many of the issues we discuss are, however, relevant to other types of RCTs (e.g. Phase I/II drug trials or pilot RCTs) although typically the justification of the target difference is less developed in those settings given that less is likely to be known about what difference would be appropriate. Given that a subsequent definitive (Phase III) study would follow successful early-phase trials and pilot studies, the necessity to determine accurately the target difference is markedly reduced. Nevertheless, some of the methods outlined below (e.g. opinion-seeking) can be and have been used for such studies, although in practice ‘rules of thumb’ or conventions (e.g. regulatory authority guidance) are often the key determinant.

Adaptive designs are not explicitly considered though it should be noted that they typically require specification of some type of target difference both in terms of the initial design but also to inform a formal adaptation process (e.g. dropping an intervention arm).
Specification of the research question will be considered in this section along with the corresponding statistical framework. The following sections will cover specification of the target difference and guidance on the use of the available methods. This will be followed by guidance on reporting the sample size calculation and a brief summary. The final section will propose areas for further research.

**Research question**

**Superiority, equivalence and non-inferiority**

Most studies are based on testing for superiority; they are designed to address whether an outcome differs (in either direction) between two interventions. Although this chapter primarily focuses on superiority studies, we note that the specification of a target difference also occurs for equivalence and non-inferiority studies, that is, an equivalence margin needs to be determined within which the interventions may be viewed as having an equivalent (or non-inferior) outcome.

**Stand-alone study or evaluated in conjunction with other evidence**

Although existing evidence is routinely available and often informally used, a RCT is conventionally designed as a stand-alone experiment. A decision needs to be made whether existing data should be formally incorporated into the study sample size calculation. It has been argued persuasively that a new trial should be designed on the basis of available evidence and once the trial is completed the data it provides should be incorporated into the evidence base. There are, however, good reasons for desiring a stand-alone methodologically robust study that is large enough to provide a definitive answer. It is reasonable to have concerns about reproducibility, applicability and consistency in methodology, which cannot always be fully addressed within a meta-analysis framework. Very large studies designed as a single ‘final’ study are sometimes referred to as ‘mega’ or ‘large simple’ trials, for example the CRASH trials, which investigated the impact of treating trauma patients with corticosteroids and tranexamic acid. However, there is a role for conducting a study that, when combined with current evidence, would have sufficient statistical precision to answer the research question. The desired level of precision may not be feasible within a single study; additionally, substantial evidence may already exist. It should be noted that following this approach to the calculation of the sample size carries the obligation that the trial is analysed on the same basis, that is, the main analysis should include all of the available evidence (existing evidence and the new trial). Another possible scenario is when multiple trials could be potentially conducted simultaneously with a view to formally combining once all data are available (see Review of evidence base method for further discussion). Nevertheless, if existing data are used in the sample size calculation, the basis of the target difference will need to consider the importance of any difference deemed realistic as opposed to solely achieving a statistical difference (of any magnitude).
Sample size calculation

General considerations
Statistical sample size calculation is not an exact science. 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 predict and even though the analysis is intended to be adjusted (e.g. when randomisation has been stratified or minimised), 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 substantial change in the calculated sample size. For example, a 10% difference between groups requires 313 per group for 80% power at the two-sided 5% significance level, if the assumed control group level is 80%; however, if the control group value used is 60% then 408 per group are needed. A further consideration is that sample size calculations are conducted a priori to provide reassurance that study results will be informative. 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. Although reassessing the assumed control group proportion or SD is not unusual during the conduct of a large study (whether utilising a formal statistical approach or a more simple check based on interim data), it is clearly not possible to know the final value until completion. Failure to fully understand this concept is reflected in the incorrect use of ‘post hoc’ sample size calculations. Performing a sample size calculation to determine what would be the likely result (i.e. statistical power at a particular significance level given the other inputs) when the actual analysis result is available is not sensible.

Statistical approaches
Different statistical approaches can be adopted to calculate the sample size. The vast majority of trials adopt the same basic methodology based on the Neyman–Pearson approach. In essence, this approach involves adopting a statistical testing framework and calculating the sample size required given the specification of two statistical parameters (the power and significance level – see below for definitions). Although different types of outcomes, study designs (e.g. cluster randomised trials) and analyses require modification of the formula, the general approach is similar across study designs. Other statistical approaches for defining the required sample size are Fisherian, Bayesian and decision-theoretic Bayesian approaches, along with a hybrid of both the Bayesian and Neyman–Pearson approaches. Although the sample size calculation process is different under a Bayesian statistical framework with regards to specification of the statistical aim, there is some similarity. A range in the posterior distribution of an outcome in a Bayesian statistical framework is often used in a similar manner to the equivalence limit in an equivalence trial under the Neyman–Pearson method. Health economic-based methods developed since the late 1990s tend to follow a Bayesian approach and sometimes also utilise a decision-theoretic framework (see Chapter 2 and Description and guidance on the use of individual methods for specifying the target difference). Bayesian methods readily allow incorporation of prior beliefs (which can be based on empirical evidence and/or opinion). However, methods other than Neyman-Pearson are rarely used, as confirmed by a review of RCT sample size calculations, which identified only the Neyman–Pearson approach in use in a sample of over 200 studies. As noted above, this document focuses on the Neyman–Pearson approach except when the method for specifying the target difference necessitates an alternative approach (see, for example, Health economic method).

To calculate the sample size for a superiority trial under the Neyman–Pearson approach, a compromise is required between the possibility of being misled by chance into concluding that there is a difference (in the intervention effect) when there is not and the risk of not identifying a genuine difference. Once all of the inputs have been specified, the required sample size can be determined. In a more complex situation, a simple formula may not be available and a simulation study of the hypothesised true intervention effect, under the assumptions and using the proposed statistical analysis, is needed to calculate the required sample size. The standard approach involves specifying a null hypothesis of no difference and seeking evidence to reject the null hypothesis in favour of the alternative hypothesis (there is a difference in
either direction). Two types of errors can be made (type I and type II) as noted in Chapter 1. Commonly used values are \( \alpha = 0.01 \) or 0.05 for a type I error (and commonly referred to as a 1% or 5% statistical significance level) and \( \beta = 0.1 \) or 0.2 for a type II error (90% or 80% power to detect a difference of the size specified) under the Neyman–Pearson statistical approach. Without much exception a two-sided significance level is adopted as it is rare that the possibility of a difference in both directions is not conceivable. Even when a one-sided test approach is adopted, such as would be the case for a non-inferiority study, it is conventional to adopt the same one-tailed significance as if it were a two-sided test (i.e. a one-sided 2.5% significance level for a non-inferiority study corresponds to a two-sided 5% test such as would be used for a superiority study). Once the two statistical criteria (the significance level and the power) are set and the statistical test to be conducted is chosen, the sample size is determined depending on the magnitude of difference to be detected (‘target’ difference) and associated component of the outcome (i.e. SD for a continuous outcome). The target difference is the magnitude of difference (intervention effect) that the RCT is designed to investigate in the primary outcome.

**The role of the primary outcome**

In the standard approach to a RCT, one outcome is chosen to be the primary outcome. This is done by consideration of the outcomes that should be measured in the study. 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 with respect to the research question being posed. The study sample size is then determined for the primary outcome. Choosing a primary outcome 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 aims to identify and the statistical precision with which it can be achieved. Additionally, it clarifies the initial basis on which to judge the study findings. Specification of the primary outcome in the study protocol helps prevent undue overinterpretation arising from testing multiple outcomes and reporting statistically significant (although often clinically irrelevant) outcomes. This multiple testing, or multiplicity, is particularly important given the likelihood of the play of chance leading to statistical significance when a large number of outcomes are under investigation. This, along with the use of a statistical analysis plan, limits the scope for manipulation of the definition of the primary outcome to maximise statistical significance (i.e. the lowest \( p \)-value). For example, a three-level ordinal outcome (e.g. low, medium and high) can be collapsed into a dichotomy (e.g. medium could be categorised as either a low or a high value) in two ways and the two approaches could give different statistical results.

Choosing a 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 a RCT comparing treatment with eye drops to lower ocular pressure with observation 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. However, it would clearly be important to consider other outcomes (e.g. side effects of the 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 because of the time it takes to measure any change in vision noticeable to a patient. The primary outcome should always be a key outcome for the comparison being undertaken. Second, consideration is needed about the ability to measure the chosen primary outcome reliably and routinely within the context of the study. Missing data are a threat to the 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 eye pressure (intraocular pressure) is through manometry; however, this requires invasive eye surgery. Subjecting participants to clinically unnecessary surgery for the purpose of a RCT is clearly not ethical without very strong mitigating circumstances, particularly as an alternative, even if less accurate, way of measuring intraocular pressure exists. Furthermore, in the context of manometry, an informed consent process would lead to a substantial number of people not consenting to the surgery required for the manometry, if not the study
overall. Third, the outcome needs to be one for which an appropriate difference can be detected with an achievable sample size. There are different bases for determining this difference, although whatever value is used, it will need to lead to an achievable sample size. Obtaining a sufficient sample size is in turn dependent on there being sufficient potential participants, a practical time frame for the outcome measurement and interventions that will provide an answer before any technology becomes obsolete, and which can be achieved with the financial and other resource constraints faced. When planning a study it is not unusual for a number of outcomes to be considered as potential primary outcomes, and a judgement has to be made as to which best meets these criteria. The development and use of core outcome sets across studies will help ensure that, even when a core outcome is not used in the sample size calculation, it is collected and reported.

In some circumstances more than one primary outcome may be used. For example, more than one outcome may be used to cover the different aspects of the purported difference. This is less common in a regulatory setting as formal statistical adjustment for the use of two (or more) outcomes may be required through adjustment of the significance level for multiple comparisons. Such adjustments are often overly conservative and can negatively impact on the ability to detect a difference for any of the chosen primary outcomes. Alternatively, the sample size calculation may also be conducted in a way that ensures sufficient precision for a secondary outcome. For clarity, the remainder of this chapter will be based on the premise of only one primary outcome, although we note that more than one might be appropriate in which case a judgement about the value of statistical correction for multiple outcomes will have to be made.

Types of outcome
The sample size formula varies depending on the outcome and the intended analysis. The most common outcomes are binary, continuous and survival (time-to-event) outcomes; continuous outcomes are typically assumed to be normally distributed, or at least ‘approximately’ so, for ease and interpretability of analysis and for the sample size calculation. All three types are considered below; we do not consider other types of outcome measure although we note that ordinal, categorical and rate outcomes can be used, for which a more complex analysis and corresponding sample size calculation approach may be needed. 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) and the sample size calculation based on the resultant binary outcome; such a dichotomisation would result in less statistical precision and lead to a larger sample size being required. If 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.

Specifying the target difference
The specification of the target difference has received surprisingly little discussion in the literature. As noted above, the target difference is the difference in the primary outcome value used in the sample size calculation that the study is designed to reliably detect. There are two main bases for specifying the target difference:

- a difference considered to be important (e.g. by a stakeholder group such as health professionals or patients)
- a realistic difference based on current evidence (e.g. seeking the best available estimates in the literature).

It should be noted that it has been argued that a target difference should always meet both of these bases. A large amount of literature exists on defining a (clinically) important difference.
common general approach is the MCID. This has been defined as ‘the smallest difference . . . which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient’s management’ or more simply as the ‘minimum difference that is important to a patient’. Variants exists on this basic approach. In the context of specifying a target difference for a typically two parallel group trial, it should be noted that the focus is on a difference at the group level and between two groups of different participants. This contrasts with the vast majority of the MCID (and variant) literature, which focuses overwhelmingly on within-patient change and whether an important difference can be said to have occurred. An alternative approach is to consider all relevant issues including the consequences of decision-making, whereby a difference of any magnitude can be viewed as important, and therefore a study’s size (and implicitly the target difference) is determined by reference to resource implications. This is considered in more depth in Description and guidance on the use of individual methods for specifying the target difference.

The other main basis for a target difference is to specify a realistic difference. For example, if a systematic review of RCTs on the research question is available, it can be used to specify what difference is supported by current evidence. In essence, it makes no claim regarding the clinical importance or otherwise of the difference. However, it is clear that some indication of the value or ‘importance’ of the study finding is needed to inform a decision. When a method that assesses a realistic difference is used, consideration of the importance of the difference is needed. For some outcomes the importance may be very clear (e.g. mortality) whereas for others (especially quality-of-life and surrogate outcomes) further justification is needed. Recruitment, study management and finance will naturally come into play when determining the sample size of a study. However, such considerations do not negate concerns about what is a realistic and/or important difference. Box 6 shows a simple example in which what is viewed as realistic and important could influence the target difference.

For a superiority trial it is generally accepted that the target difference should be a ‘clinically important’ difference or ‘at least as large as the MCID’. It should be noted that the target difference in a conventional sample size calculation is not the minimum difference that can be statistically detected. High statistical power ensures that a difference of a specified magnitude is likely to be detected. For a given sample size, the power increases as the magnitude of the target difference increases. Given this, it is possible that an observed difference slightly smaller than the target difference might lead to a statistically significant (even if not viewed as important) difference. For example, under a two-sided significance level of 5% and power of 90% for a continuous outcome, an observed difference of 0.65 of the target difference can be statistically detected. As a consequence, if the MCID were to be used as the target difference, the statistical analysis could achieve statistical significance for values smaller than the target difference (i.e. MCID).

BOX 6 A realistic and/or important difference as the basis for the target difference: example based on the Men After Prostate Surgery (MAPS) trial

Realistic difference: based on a systematic review of RCTs, a target difference of 20% (from 70% to 90%) is reasonable.

Important difference: based on clinical opinion, a target difference of 15% is chosen (from 70% to 85%).

Important and realistic difference: based on a systematic review of RCTs and clinical opinion regarding an important difference, a target difference of 15% is used.

Note: if the realistic difference is smaller than the important difference this implies that the trial should not be conducted as an important difference does not appear feasible.
For an equivalence (or non-inferiority) trial, as opposed to a superiority trial, a range of values around zero will be required within which the interventions are deemed to be effectively equivalent (or not inferior) in order to establish the magnitude of difference that the RCT is designed to investigate. The limits of this range, the equivalence margin (sometimes called zone of indifference), are points at which the differences between interventions are believed to become important and one of the interventions is considered not to be equivalent (or to be inferior). The difference between the end of the margin and zero could be defined as the MID between interventions. Alternatively, this range can be based on inherent variability (distribution method), preserving at least a fraction of the active interventions’s effect (‘CI approach’), or use of another important difference approach. Interestingly, some authors have been clearer in their definition that the equivalence margin in the sample size calculation of an equivalence trial has to reflect a MCID, whereas authors have in general been less specific regarding the basis of the mean difference for a superiority trial sample size calculation. This may reflect the context in which such designs are typically used within a regulatory framework, with the control being typically a placebo in a superiority trial, whereas for an equivalence study the control is an active treatment in clinical use; implicitly, more is seen to be at risk by changing the status quo in an equivalence study (the effect of the active control) than by seeking to identify whether a new treatment works in a superiority study (placebo-controlled trial). However, when such designs are used outwith this setting (e.g. HTA programme), this distinction cannot be justified and it is difficult to see why a lower requirement for a superiority trial than an equivalence or non-inferiority trial is desirable. It is noteworthy that the sample size calculation for an equivalence/non-inferiority study requires specification of both the difference desired to be detected (margin) and the difference viewed as realistic. This separates out the two bases for determining a target difference in a superiority study context.

The target difference is specified differently depending on the type of primary outcome. For a continuous outcome, the target difference on either the original or the standardised scale is often referred to as the ‘effect size’. Strictly speaking, this alone does not fully specify the target difference; the assumed variability of the outcome is also needed to convert the effect size between the original and the standardised scale. For a binary outcome the target difference will be conditional on the control group event proportion; to uniquely specify the sample size both the target difference and the control group event proportion are needed, which together imply a unique pair of absolute and relative target differences. Similarly, survival outcomes require the control group proportion or survival distribution and length of follow-up to be stated in addition to the target difference. This is necessary as the sample size required is sensitive to both the absolute and the relative difference. It is not uncommon for only one or the other to be specifically stated in trial reports. However, the corresponding control group proportion should also be stated to fully specify the target difference. For a continuous outcome, the target difference can be specified in two ways, as a mean difference or as an ‘effect size’ (SES), the latter conventionally being the mean difference divided by the (pooled) SD. Full specification of the target difference for a continuous measure requires both the difference and the corresponding SD to be stated (by specifying the mean difference and the SD or the mean difference and the SES).

A variety of formal methods are available to determine the target difference in the primary outcome (see Chapter 2):

- anchor
- distribution
- health economic
- opinion-seeking
- pilot study
- RoEB
- SES.

Each method is described briefly in the next section and some specific guidance on their use is provided. It should be noted that the health economic method, unlike the others, does not as such typically specify a
target difference for a clinical outcome; instead, it attempts to take all relevant considerations into account (including positive and negative consequences) to determine the optimal sample size. Nevertheless, an explicit definition of a threshold value for an economic measure may be required, for example a cost per QALY threshold.

**Description and guidance on the use of individual methods for specifying the target difference**

**Anchor method**

Under such an approach, the outcome of interest is ‘anchored’ by using a judgement (typically a patient’s or a health professional’s) to define what constitutes an important difference. Typically, this is achieved by measuring the outcome in a cohort of patients before and after receiving a treatment for the condition; the (within-person) change is then linked to the corresponding judgement to define which participants underwent an improvement (or deterioration) and which did not. The judgement may be made by the patient (self-assessment) or by another (e.g. health professional or parent). An event-based outcome (e.g. visit to a health-care professional) has been used as an alternative ‘objective’ type of anchor, or as a validation of a judgement-based anchor approach. Using between-patient contrasts (sometimes referred to as the social comparison method), in which patients whose disease is at varying stages are used and a judgement is made for each pair of patients, is also possible. From these paired judgments, the difference in outcome that is thought to be ‘important’ can be determined. A similar approach is to use an expert’s (e.g. health professional’s) judgement to choose between patients with responses for a quality-of-life outcome to determine the magnitude of difference that is viewed as important. From these judgments the value that is viewed as important can be determined.

Many variations exist on how the anchoring judgement is implemented (e.g. number of points on a Likert scale and variations in the point labels). Additionally, how the judgement is used to determine the MID varies. Positive (improvement) and negative (deterioration) change may be considered together or separately. Additionally, a ‘no change’ group can be used to ‘reset’ the difference to address regression to the mean; the differences between multiple levels of improvement/deterioration may be evaluated as opposed to assessing only no improvement compared with improvement; and the value used as the cut-off is typically the mean difference although the optimal cut-point may be determined using ROC curve methods. Finally, an outcome (e.g. quality-of-life measure) that has been ‘anchored’ can itself be used to anchor another outcome. This approach is similar in manner to ‘cross-walking’ (or ‘mapping’) in health economics in which the relationship of one outcome to another is estimated.

A distinction between what is ‘important’ at an individual level and what is ‘important’ at a group level has been made by some authors. However, with regards to specifying a target difference (a group-level difference), the use of an important difference based on individual-level judgments is legitimate as long as it was determined in a similar population to that in the anticipated RCT. Smaller (or perhaps larger) differences at a group, or population, level might be viewed as important if additional considerations are taken into account (e.g. intervention costs), although this is also true at an individual level.

**Key points for using the anchor method to specify the target difference**

- Suitable for continuous (or ordinal) outcome.
- Anchor implementation is critical; for example, the perspective and anchor adopted.
- Particularly suited to quality-of-life measures.
- The magnitude of the difference can be sensitive to the population group (e.g. ceiling/floor and disease severity effects may exist).
- Use of the most common anchor approach implies that a within-person (important) difference can be applied though a between-person approach is also possible.
**Distribution method**

This method is based on the statistical properties of an outcome; typically, it is based on determining a value that is larger than the inherent imprecision in the measurement and is therefore likely to represent a meaningful difference. A very common approach is to use test–retest data for an outcome to specify the magnitude of difference that could be due to random variation in the measurement. The limited usefulness for calculating the MCID, or an important difference, has been noted by a number of authors although this method has been used alongside other methods. An extension to the measurement error approach is the RCI, which uses the former approach but also involves reference to ‘normative’ and ‘abnormal’ populations and defining a cut-off between the two. Such an approach does not readily lend itself to calculating a target difference as specification of the ‘normative’ and ‘abnormal’ populations will be very difficult, particularly when two active groups are being compared. Additionally, some prior rule would be needed, even if the cut-point was defined, to determine the target difference.

Other distribution approaches exist. The use of the minimal (smallest) detectable difference approach is common for interpreting the clinical significance of results in individuals but cannot be used for specifying a target difference as it is based on what can be statistically determined in a particular sample and is therefore dependent on the sample size. Simple ‘range’ approaches, which might also be viewed as a distribution approach, have been used, for example moving at least one-tenth of the range on a pain VAS score or at least one level between points on an ordinal scale. Such approaches are based on the meaningfulness of the individual points on the scale and do not specify importance as such. They are reliant on the outcome having obvious meaning (e.g. Rankin score).

**Key points for using the distribution method to specify the target difference**

- Suitable for a continuous (or possibly ordinal) outcome.
- Use of the distribution method (i.e. measurement error approach) is of limited merit due to its weak justification of an ‘important’ difference.
- A simple range or levels approach should be a last resort if no more informative methods can be used and only when the outcome has clear meaning.

**Health economic method**

This method refers to those approaches to determining an important difference that are based on economic evaluation analysis. The use of the anchor method to determine an important difference for economic outcomes (e.g. QALYs) has been considered earlier (see Anchor method). It is worth noting that an economic evaluation is typically used to inform decisions at the group (population) level rather than to inform, for example, the treatment decision for an individual. Initial approaches used economic evaluation methods to determine threshold values for determinants of costs or cost-effectiveness. Such approaches required the specification of a decision rule such that treatment A would be preferred to treatment B if the cost of A was less than the cost of B, or treatment A would be preferred to treatment B if the incremental cost per QALY was below a given value. More recently, the health economic method has been based on a net benefit statistic because it simplifies the analysis compared with those required for the incremental cost per QALY (because this statistic is a ratio of two related measures). The net benefit statistic is used to determine (typically in a decision-analytic model) the value of future research and, by extension, through estimating the expected value of sampling information, the sample size. This approach requires the definition of a threshold value for the net benefit statistic and implicitly or explicitly determines target differences for those measures used as inputs into an economic model.

**Key points for using the health economic method to specify the target difference**

- Allows a comprehensive approach to the value of a RCT; in particular, the costs of the intervention and its comparator and research can be considered in conjunction with possible benefits and consequences of decision-making. The flexible modelling framework allows any type of outcome to be incorporated.
- The perspective adopted is critical – the viewpoint and values that are used to determine the scope of costs and benefits incorporated into the model structure.
- Uncertainty around inputs can be substantial and extensive sensitivity analyses will likely be needed. Some inputs (e.g. time horizon) will be particularly challenging to specify as well as appropriately representing the statistical relationship of multiple parameters. These could also be based on empirical data and/or expert opinion.
- This is a resource-intensive and complex approach to determining the sample size.
- Unlikely to be accepted as the sole basis for study design at present despite intuitive appeal. Patients and clinicians may be resistant to the formal inclusion of cost into the design and thereby the primary interpretation of studies. Expressing the difference in conventional way is likely to be necessary as it is more intuitive to stakeholders and also furthers the science of interventions. It could provide additional justification for conducting a large and expensive trial (e.g. when there is a small effect and/or events are rare).

Opinion-seeking method
This includes all formal approaches for specifying the target difference based on eliciting expert opinion (often clinicians’ although it can be patients’ or others’). Possible approaches include forming a panel of experts,26 surveying the membership of a professional or patient body254 or interviewing individuals.26 The elicitation can also seek to take into account the trade-off between treatments in terms of positive impacts (e.g. reducing the risk of a heart attack) and negative impacts (e.g. risk of stroke) to determine the value that is viewed as important.462 This process has been carried out explicitly for determining the target difference for a RCT in both a Bayesian and a conventional framework and can be extended to incorporate treatment decision-making.26,45,47,242,243,443

Key points for using the opinion-seeking method to specify the target difference
- Allows for varying degrees of complexity of the scenario (e.g. consideration of related effects or impact on practice) and any outcome type (binary,47,451 continuous242,451 and survival30).
- The perspective is critical – whose opinions are being sought.
- A realistic and/or important target difference can be sought.
- A target difference that takes into account other outcomes and/or consequences (e.g. a target difference that would lead to a health professional changing practice) or focuses exclusively on a single outcome can be sought.

The presentation of the scenario is important. Ideally a mechanism to confirm/probe the initial response will be used.463

Pilot study method
Data from a pilot (preliminary) study can be used to estimate a realistic difference in the outcome of interest in order to determine the target difference.256 A common approach is to undertake a pilot study before conducting the main RCT; in the drug regulatory setting the observed difference in a Phase II study can be used to determine the target difference for a Phase III study. The distinction between this method and the RoEB method is that, with this method, a study is conducted for the purpose of informing a future definitive study as opposed to using one or more existing studies. A pilot study is most useful in situations in which it can be conducted readily and quickly (e.g. rapid recruitment and short outcome follow-up).

There is general acceptance that the estimates from the pilot study should not be used without allowance for the imprecision in these estimates.42,44 A pilot study is best suited to estimating the SD and another method should be used to specify the mean difference for a continuous primary outcome. Similarly, for a binary outcome, its value is in estimating the control group event proportion. It should be noted that the observed SD (or control group proportion for a binary outcome) is itself an estimate of the true value of the variation and may be inaccurate.42,464 Although the imprecision of the SES estimate can be calculated for a continuous outcome, given the likely small size this is likely to be uninformative.257
Key points for using the pilot study method to specify the target difference

- There is a need to assess the relevance of the pilot study to the design of a new RCT study. Some down-weighting (whether formally or informally) may be needed according to the relevance of the study and methodology used. For example, a Phase II study should be used to directly specify a (realistic) target difference for a Phase III study only if the population and outcome measurement are judged to be sufficiently similar.
- Helpful for estimating outcome components such as variability of a continuous outcome (or control group rate for a binary outcome) although the estimation of the target difference is typically imprecise because of a small sample size.
- This approach can be used in conjunction with another method (e.g. using an opinion-seeking method to determine an important difference) to allow full specification of the target difference.

Review of evidence base method

The target difference can be derived by undertaking a RoEB. There are two main approaches that could be adopted. Current studies that measured the outcome for a specific research question can be collated to assess the likely observed difference. Ideally, this would be based on a systematic review of RCTs, and possibly meta-analysis of the outcome of interest, that directly address the research question at hand. The scope of the studies considered can be enlarged. For example, in the absence of randomised evidence, evidence from observational studies could be used in a similar manner. Additionally, studies within a disease area might be considered as opposed to restricting to a specific research question. The focus may be restricted to one component of the target difference (e.g. control group proportion). In a similar manner, the value for other parameters (e.g. CoV in equivalence trial sample size formula) could be determined. The second main approach is to review studies that sought to determine an important difference (e.g. reviewing multiple anchors, distribution or SES methods).

This method can be used for any type of outcome. Consideration of the imprecision and the potential bias in the current evidence is needed, as is the degree to which it is relevant [i.e. consideration of population, intervention, control, outcome and time frame (PICOT)] to the research question. This approach can be formalised by carrying out a simulation study that uses the current meta-analysis data and examines the impact of a new RCT in a hypothetical updated meta-analysis. Doing so obliges the RCT to be analysed as a meta-analysis in conjunction with the rest of the current evidence, given that this is how the study’s sample size was justified. In such a set-up, the publication of new studies during the time when a trial is being conducted would warrant recalculation of the sample size. When no direct evidence exists, the use of evidence from a ‘similar’ comparison could be used, although judgement is needed about the relevance of such data to the decision question and the risk of bias. However, even if existing evidence is formally incorporated into the sample size, consideration of an important difference is still needed. Use of statistical evidence of a (non-zero) difference would imply that any difference in this outcome is ‘important’ and therefore further specification (e.g. using another method) may be needed.

Key points for using the review of evidence base method to specify the target difference

- It should be based on a systematic search of available evidence.
- It can be used for any outcome type (including continuous, binary, ordinal and time-to-event outcomes).
- A choice must be made whether an important and/or a realistic difference is sought.
- A number of issues need to be considered when assessing an observed difference:
  - Is the evidence available directly relevant to the research question at hand (PICOT assessment)?
  - Is the existing evidence of a robust nature? Are there multiple studies available and were they conducted in a methodologically robust manner? What was the risk of bias?
  - Is the outcome of interest fully reported? Individual patient data are seldom available and reporting of outcome is often selective.
Determination of a realistic (target) difference can, and when possible should, be based on a systematic review and associated meta-analysis of RCTs, although imprecision in the estimate needs to be considered.

The use of prior evidence can be formalised through simulation of the impact of a new study on the meta-analysis result, although this implies that a particular analysis will be conducted and the new study will be analysed alongside the current evidence.

**Standardised effect size**

Under this method, the magnitude of the proposed effect size on a standardised scale is calculated and the value of observing such a difference is inferred by reference to the universe of possible standardised effects. Some authors categorise this method as a subtype of the distribution method. These methods have been separated out because of the widespread use of ‘effect size’ approach to determine the target difference of a RCT. Additionally, the term ‘standardised effect size’ is used to clarify that under this approach the effect size is determined according to its magnitude on the standardised scale. Confusingly, the term ‘effect size’ is sometimes used to refer to the target difference on the original scale as well as the standardised scale.

Different types of SES metrics exist. For a continuous outcome, the standardised difference (typically expressed as Cohen’s ‘effect size’, i.e. mean difference divided by the SD) is overwhelmingly the most commonly used although other formulae exist (e.g. Hedges’ g). Cut-offs of 0.2, 0.5 and 0.8 for small, medium and large effects are often used following Cohen’s suggestions, which were based on his experience. The corresponding values of the original scale depend on the SD. For example, if the SD is 5, then small, medium and large effects would be mean differences of 1, 2.5 and 4 respectively.

The SD used to calculate the SES is typically the pooled SD across the two groups. A common alternative, the second type of SES, is to calculate the effect size of the change in a before-and-after treatment study in which the treatment received is widely accepted to be effective. Such an effect size is likely to be larger than would be observed between active treatments (the first type of effect). Minor variations in how this type of effect size is calculated exist, for example the SRM, which uses the SD of the change score as opposed to the before-treatment SD. However, an effect size using the SD of the change score will be larger than that observed between treatment groups, as the within-person variance is removed; this leads to a smaller denominator and hence a larger effect. More complex effect sizes for a repeated measures situation have also been proposed. A third type of effect size is to use a reference population and the impact of ‘diseased’ compared with ‘non-diseased’ populations.

The value of 0.2 SD has been proposed as the MCID and therefore could be used to define the target difference. Some other support for the usefulness of Cohen’s criterion exists (0.5 SD has been suggested as being a meaningful value); however, the current empirical evidence is insufficient to justify its widespread use for a range of outcomes, types of interventions and comparisons, and in disparate populations and disease areas. It seems reasonable to expect different sizes of standardised effect depending on whether an active control is used or not or whether a pragmatic or an explanatory study is planned (see Chapter 3). Cocks and colleagues undertook a novel hybrid approach using expert opinion to categorise studies (without reference to the actual results) as having a trivial, small, medium or large effect for the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30 (EORTC QLQ-C30) quality-of-life tool. A meta-analysis of the observed effects within each category was then performed to calculate the magnitude of a trivial, small, moderate and large effect size (and mean differences). The results were broadly consistent with Cohen’s values with, typically, effect sizes of <0.2 being classed as trivial, 0.5 often being classed as medium and 0.8 as large; there was variation in the cut-offs, ranging from 0.1 to 0.2, 0.4 to 0.7 and 0.6 to 1.1 for small, medium and large effects, respectively, depending on the subscale.
For a binary or survival (time-to-event) outcome, an odds or risk ratio (binary) or hazard ratio (survival) could be utilised, with the spectrum of values interpreted in a similar manner, with a doubling or halving of the ratio (odds, risk or hazard) sometimes taken to imply a large effect (see Chapter 3). The outcome’s definition of an event is key to ascertaining importance. However, halving of the rate (risk ratio of 0.5) of an event from 50% to 25% (absolute difference of 25%) is a very large absolute target difference but a halving of the rate from 1% to 0.5% (absolute difference of 0.5%) will often be unimportant. As a consequence, both the relative and absolute difference needs to be taken into consideration. Hence, for a binary or survival outcome, unlike a continuous outcome, the two components cannot be readily incorporated into a single value; the typical effect size measures (e.g. odds, risk and hazard ratios) are relative measures and do not take into account the absolute level, which is also important. Correspondingly, an absolute target difference (e.g. 25%) can be used but it does not uniquely identify the relative effect. As a consequence, the target difference for such outcomes does not uniquely specify the sample size (given the statistical parameters and analysis) and the control group proportion or equivalent needs to be considered and reported in conjunction with the target difference. Other SES metrics exist for a binary outcome (e.g. Cohen’s h) although they are rarely used. Approximate values for the odds ratio can be calculated using Cohen’s cut-offs, giving 1.44, 2.48 and 4.27, respectively, for a small, medium and large effect.

**Key points for using the standardised effect size to specify the target difference**

- The SES for a continuous outcome should be calculated as the difference between groups divided by the appropriate SD. For a parallel group trial, the SD will typically be an estimate of the (common) group SD, which corresponds to an unadjusted analysis of the final scores; the SD of the within-person change score could be used when an analysis of change scores is planned. The benefit of removing within-person variance, such as through an analysis which adjusted for the baseline value, can also be incorporated when the correlation can be estimated.

- A SES from a before-and-after treatment study is unlikely to be representative of that achievable in a treatment study, particularly when two active treatments are compared.

- Use of Cohen’s criteria of interpretation is difficult to justify, although widespread. Modifications to this effect size scale have been suggested (see Chapters 2 and 3). For example, pragmatic trials are generally accepted to have smaller effects than more efficacy-focused studies. The SES may differ in magnitude between clinical areas and outcomes and when the standard treatment is very effective.

- Changes in the variability (e.g. population spectrum) for a continuous outcome can result in a different standardised effect even though the mean difference remains the same. It is important that an estimate of the variability is also specified and that the sample is similar to the anticipated RCT population. For a binary outcome, the target difference (whether a relative or an absolute difference) should be considered in conjunction with the control group event proportion.

- It is most appropriate as a fall-back option if other more context-relevant methods for specifying the target difference cannot be used.

**Reporting of the sample size calculation**

The assumptions made in the sample size calculation should be clearly specified. All inputs should be clearly stated so that the calculation can be replicated. When the calculation deviates from the conventional approach (shown in Box 5), whether by research question or statistical framework, this should be clearly specified. Formal adjustment of the significance level for multiple comparisons or interim analyses should be specified. We recommend that trial protocols clearly and fully state the sample size calculation, including when the approach taken differs from the conventional approach (e.g. Bayesian framework), statistical parameters and the target difference, with justification of the choice of values. Because of space restrictions, in many publications the main trial paper is likely to contain less detail than is desirable. Nevertheless, we recommend a minimum set of items for the main trial results paper along with
full specification in the trial protocol. The recommended list of items given below for the paper (as well as for the protocol) is more extensive than that in the CONSORT statement (including the 2010 version).

Specification of the target difference in the sample size calculation section of a RCT protocol or results paper varies according to the type of primary outcome. Illustrative examples of a protocol section for a RCT with a binary, continuous and survival primary outcome are given in Boxes 7–9 respectively.

**BOX 7 Protocol sample size calculation example: binary primary outcome [Men After Prostate Surgery (MAPS) trial]**

The primary outcome is urinary continence. The sample size was based on a target difference of 15% absolute difference (85% vs. 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. The control group proportion 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.

**Reporting items for the randomised controlled trial protocol**

- State any divergence from the conventional approach.
- State the primary outcome (and any other outcome that the study sample size calculation is based on), or state why there is not one.
- Reference the formula/simulation approach if standard binary, continuous or survival outcome formulae are not used. The primary analysis should be stated in the statistical analysis section.
- State the values used for statistical parameters (e.g. significance level and power).
- State the underlying basis used for specifying the target difference:
  - an important difference as judged by a stakeholder
  - a realistic difference based on current knowledge or
  - both an important and a realistic difference.
- Express the target difference according to the outcome type:
  - Binary – state the target difference as an absolute and/or a 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.
  - Continuous – state the target mean difference on the natural scale, the common SD and the SES (mean difference divided by the SD). It is preferable to also provide the anticipated control group mean even though it is not required for the sample size calculation.
  - Survival (time-to-event) – state the target difference as an absolute and/or relative difference; provide the control group event proportion, and the intervention and control group survival distributions; additionally, the planned length of follow-up should be stated. If both an absolute and a relative difference are provided, clarify if either takes primacy in terms of the sample size calculation.
- Explain the choice of target difference – specify and reference any formal method used or relevant previous research.
- State the sample size based on the assumptions specified above (for a survival outcome, the number of events required should also be stated). If any factors (e.g. allowance for loss to follow-up) that alter the required sample size are incorporated they should also be specified along with the final sample size.
Reporting items for the randomised controlled trial results paper

- State any divergence from the conventional approach.
- State the primary outcome (and any other outcome that the study sample size calculation is based on), or state why there is not one.
- State the values used for statistical parameters (e.g. significance level and power).
- Express the target difference according to the outcome type:
  - Binary – state the target difference as an absolute and/or a 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.
  - Continuous – state the target mean difference on the natural scale, the common SD and the SES (mean difference divided by the SD). It is preferable to also provide the anticipated control group mean even though it is not required for the sample size calculation.
  - Survival (time-to-event) – state the target difference as an absolute and/or relative difference; provide either the intervention and control group event proportions or the intervention and control group survival distributions; additionally, the planned length of follow-up should be stated. If both an absolute and a relative difference are provided, clarify if either takes primacy in terms of the sample size calculation.
- State the sample size based on the assumptions specified above (for a survival outcome, the number of events required should also be stated). If any factors (e.g. allowance for loss to follow-up) that alter the required sample size are incorporated they should also be specified along with the final sample size.
- Reference the trial protocol for further details.

**BOX 8** Protocol sample size calculation example: continuous primary outcome (FILMS)

The primary outcome is Early Treatment Diabetic Retinopathy Study (ETDRS) distance visual acuity. A target difference of a mean difference of five letters with a common SD of 12 was 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 was based on two previous studies – one observational comparative study and one RCT. This target difference is equivalent to a 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 a total of 246 participants.

**BOX 9** Protocol sample size calculation example: survival primary outcome [Arterial Revascularisation Trial (ART)]

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 difference and control group mortality proportions are realistic based on a systematic review of observational (cohort) studies. 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).

Summary

The specification of the target difference is a key element of RCT design. There is a clear need for an increased use of formal methods for its specification. Although no single method provides a perfect solution to a difficult question, raising the standard of RCT sample size calculations and the corresponding reporting of them would be a step forward. This would aid health professionals, patients, researchers...
and funders in judging the strength of the available evidence and ensure better use of scarce resources. Guidance for researchers on the sample size calculation with particular reference to specifying the target difference and how this should be reported in trial protocols and reports was produced. Although our examples and framing are from a medical context, the issue is relevant to non-medical areas as well.

A few points are worth particular emphasis. There is a place for conducting RCT sample size calculations on the premise that they will be analysed with current evidence as opposed to an exclusive stand-alone basis. The use of a method that focuses on a realistic difference is generally an insufficient basis for specifying the target difference unless any difference in the primary outcome is clearly ‘important’, for example mortality. The distribution method is suboptimal and other methods should be given preference. The pilot study method can only be reasonably used in conjunction with another method as the uncertainty around the target difference will be too large for it to be useful on its own. Further research into the implementation and practicality of alternative methods (e.g. health economic and opinion-seeking), and exploration of the justification of another (SES), is needed. Specific research priorities are listed in the next section.

Further research priorities

1. A comprehensive review of observed effects in different clinical areas, populations and outcomes is needed to assess the generalisability of Cohen’s interpretation for continuous outcomes, and to provide guidance for binary and time-to-event measures. To achieve this, an accessible database of SESs should be set up and maintained. This would aid the prioritisation of research and help researchers, funders, patients and health-care professional assess the impact of interventions.
2. Prospective comparison of different formal methods for specifying the target difference is needed in the design of RCTs to assess the relative impact of different methods.
3. Practical use of the health economic approach is needed; the possibility of developing a decision model structure that reflects the view of a particular funder (e.g. HTA programme) and incorporates all relevant aspects should be explored.
4. Further exploration of the implementation of the opinion-seeking approach in particular is needed. The reliability of a suggested target difference that would lead to changes in practice should be explored, as well as the opinion of different stakeholders.
5. The value of the pilot study for estimating parameters (e.g. control group event proportion) for a definitive study should be further explored by comparing pilot study estimates with the resultant definitive trial results.
6. Qualitative research on the process of determining a target difference in the context of developing a RCT should be carried out to explore the determining factors and interplay of influences.
Acknowledgements

We would like to thank the respondents to the surveys; Lara Kemp and Janice Cruden for secretarial support; Tara Gurung, Mark Forrest and Fiona Stewart for helping with abstract screening, the online survey and retrieving references respectively, and Marion Campbell and Adrian Grant for serving on the project advisory group, which provided guidance on the project’s conduct and interpretation of findings.

The HSRU, Institute of Applied Health Sciences, University of Aberdeen, is core-funded by the Chief Scientist Office of the Scottish Government Health Directorates. Jonathan Cook held MRC UK training (reference no. G0601938) and methodology (reference no. G1002292) fellowships while this research was undertaken.

Contribution of authors

Jonathan A Cook led the writing of Chapters 1, 3 and 4. Jennifer Hislop was the lead systematic reviewer, led the writing of Chapter 2 and was the project research fellow. Temitope E Adewuyi and Kirsten Harrild contributed to the systematic review and commented on the report. Jonathan A Cook, Jennifer Hislop, Douglas G Altman, Craig R Ramsay, Cynthia Fraser, Brian Buckley, Peter Fayers, Andrew H Briggs, John D Norrie, Ian Harvey and Luke D Vale were members of the project steering group and commented on and edited the final report. Ian Ford and Dean Fergusson were members of the project advisory group and commented on the project’s management and commented on the report. Cynthia Fraser developed and executed the searches. Jonathan A Cook and Luke D Vale provided oversight for the whole project.
References


44. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry* 2006;**63**:484–9. http://dx.doi.org/10.1001/archpsyc.63.5.484


49. Landorf KB, Radford JA. Minimal important difference: values for the Foot Health Status Questionnaire, Foot Function Index and Visual Analogue Scale. *Foot* 2008;**18**:15–19. http://dx.doi.org/10.1016/j.foot.2007.06.006


132. Wheaton L, Pope J. The minimally important difference for patient-reported outcomes in spondyloarthropathies including pain, fatigue, sleep, and Health Assessment Questionnaire. *J Rheumatol* 2010;37:816–22. [http://dx.doi.org/10.3899/jrheum.090086](http://dx.doi.org/10.3899/jrheum.090086)


137. Wytrich KW. Minimal important difference thresholds and the standard error of measurement: is there a connection? *J Biopharm Stat* 2004;14:97–110. [http://dx.doi.org/10.1081/BIP-120028508](http://dx.doi.org/10.1081/BIP-120028508)


188. Schmitz N, Hartkamp N, Franke GH. Assessing clinically significant change: application to the SCL-90-R. *Psychol Rep* 2000;**86**:263–74. [http://dx.doi.org/10.2466/pr0.2000.86.1.263](http://dx.doi.org/10.2466/pr0.2000.86.1.263)


191. Matthey S. Calculating clinically significant change in postnatal depression studies using the Edinburgh Postnatal Depression Scale. *J Affect Disord* 2004;**78**:269–72. [http://dx.doi.org/10.1016/S0165-0327(02)00313-0](http://dx.doi.org/10.1016/S0165-0327(02)00313-0)


255. Burbach D, Molnar FJ, St John P, Man-Son-Hing M. Key methodological features of randomized controlled trials of Alzheimer’s disease therapy. Minimal clinically important difference,


273. Klassen AF. Quality of life of children with attention deficit hyperactivity disorder. *Expert Rev Pharmacoecon Outcomes Res* 2005;5:95–103. [http://dx.doi.org/10.1586/14737167.5.1.95](http://dx.doi.org/10.1586/14737167.5.1.95)


275. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003;41:582–92. [http://dx.doi.org/10.1097/01.MLR.0000062554.74615.4C](http://dx.doi.org/10.1097/01.MLR.0000062554.74615.4C)


319. de Morton NA, Davidson M, Keating JL. Validity, responsiveness and the minimal clinically important difference for the de Morton Mobility Index (DEMMI) in an older acute medical population. *BMC Geriatrics* 2010;10:72. [http://dx.doi.org/10.1186/1471-2318-10-72](http://dx.doi.org/10.1186/1471-2318-10-72)


393. Merkies IS, van Nes SI, Hanna K, Hughes RA, Deng C. Confirming the efficacy of intravenous immunoglobulin in CIDP through minimum clinically important differences: shifting from statistical significance to clinical relevance. *J Neurol Neurosurg Psychiatry* 2010;**81**:1194–9. [http://dx.doi.org/10.1136/jnnp.2009.194324](http://dx.doi.org/10.1136/jnnp.2009.194324)

394. Copay AG, Glassman SD, Subach BR, Berven S, Schuler TC, Carreon LY. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. *Spine J* 2008;**8**:968–74. [http://dx.doi.org/10.1016/j.spinee.2007.11.006](http://dx.doi.org/10.1016/j.spinee.2007.11.006)


REFERENCES

421. UKCRC Registered Clinical Trials Units. URL: www.ukcrc-ctu.org.uk/ (accessed December 2011).


Appendix 1 Protocol

Title
Assessing methods to specify the target difference for a randomised controlled trial (Difference ELicitation in TriAls – DELTA review)

Background on target differences

Calculation of sample size
The randomised controlled trial (RCT) is widely considered to be the gold standard for the comparison of the effectiveness of health interventions. Central to its validity is an a priori sample size calculation which sets the recruitment target for a particular study, which (assuming the target recruitment is reached) provides reassurance that the trial result will be informative because it is likely to detect a difference with the appropriate level of statistical precision.

To calculate the sample size for a superiority trial, a compromise is required between the possibility of being mislead by chance and the risk of not identifying a genuine difference. Rejecting the null hypothesis when it is true (Type I error) would lead to a trial concluding that one treatment is superior than another when in reality there is no significant difference in treatment effect. The significance level of the test ($\alpha$) is the probability of the occurrence of Type I error. Failing to reject the null hypothesis when it is false (Type II error) would lead to a trial concluding that one treatment is not superior to another, when in reality that treatment is superior. The probability of the occurrence of Type II error is 1 minus the power of the test, or put another way, the power of the test is denoted by $1 - \beta$, where $\beta$ is the probability of Type II error. Commonly used values are 0.01 or 0.05 for a type I error (the statistical significance level) and 0.1 or 0.2 for a type II error (which would give 90% or 80% power to detect a difference of the size specified) under the conventional (Neyman–Pearson) statistical approach. Once these two criteria are set, and the statistical tests to be conducted during the analysis stage are chosen, the sample size is determined depending on the magnitude of difference to be detected. This ‘targeted’ difference, is the magnitude of difference which the RCT is designed to reliably investigate.

For equivalence (or non-inferiority) trials, as opposed to superiority trials, a range of values around zero will be required within which the interventions are deemed to be effectively equivalent (or not inferior), in order to establish the magnitude of difference that the RCT is designed to investigate. The limits of this range are points at which the differences between treatments are believed to become important and one of the treatments is considered superior: the smallest difference between one of these points and zero is the minimum important difference between treatments.

Once the target difference (or in the case of an equivalence/inferiority trial limits) is determined then the method of estimating the sample size will depend upon the proposed statistical analysis, trial design (e.g. cluster randomised or individual randomisation trial) and statistical properties specified (e.g. agreement for paired data). The general approach is similar across studies under the standard Neyman–Pearson. Other statistical approaches for defining the required sample size are Fisherian, Bayesian and decision-theoretic Bayesian approaches, along with a hybrid of both the Bayesian and Neyman–Pearson approaches. Economic-based methods tend to follow a Bayesian approach. However, a recent review of RCT sample size calculations identified only the Neyman–Pearson approach in widespread usage. Regardless of the statistical method used the key issue is what magnitude of a difference (given it is statistically detected) is of practical interest.
The target difference

From both a scientific and ethical standpoint, selecting an appropriate target difference is of crucial importance. For example, two drugs for treating hypertension may differ in how well they reduce blood pressure. A small difference based upon blood pressure may have limited clinical, patient or economic significance. As a consequence, specifying too small a target difference would be a wasteful (and unethical) use of data and resources. Conversely, too large a target difference could lead to an important difference being easily overlooked, which could also be a wasteful (and unethical) use of data and resources. Furthermore, an undersized study may not usefully contribute to the knowledge base and could detrimentally impact upon decision making.5

An important development has been the concept of the minimum clinically important difference (MCID) as a rationale to define the target difference. Originally, this was defined as the ‘the smallest difference … which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient’s management’6, but has also been referred to as the ‘minimum difference that is important to a patient’.7 The concept applies equally well to the minimum difference that is important to a clinician, or to society, though obviously there could be variations between these groups (patient, clinician and society) in terms of what magnitude of a difference each considers to be important. A clinician might consider it to be the difference that would result in a change of treatment strategy for the patient, society might consider it to be the difference that would result in a return to productive employment for the patient, whereas a patient might consider it any difference that they are personally aware of.8

Variations have been suggested such as the ‘minimally clinically important improvement’ and the ‘sufficiently important difference’, which seeks to adopt a wider perspective by taking into account cost, risk and harms.9,10,11 In fact, a variety of economic approaches have been suggested from both a conventional and Bayesian perspective.11,12 All seek to ascertain a cut-point for a scale (whether directly measureable or latent) upon which an ‘important’ difference or change can be separated from an ‘unimportant’ one. Most work has been carried out on patient reported outcomes, (reflecting the belief that patients find it more difficult than clinicians to specify an important difference) and also the challenge of interpreting quality of life measures.6,13 In addition, there are pragmatic challenges in the interpretation of a mean value for the minimum important difference, where this value is the average change in score derived from all patients who have experienced what they would classify as an important change. Using a mean value will classify those with a change score below the mean as not having experienced an important change. In reality however all patients had experienced what they would classify as an important change. Selecting cut-offs for individuals has been suggested as a possible solution, but more generally it has been argued that the interpretation of important change needs to be considered differently when considering individuals or considering groups.8

In practice, the target difference is often not formally based upon these concepts and in many cases (at least from trial reports) appears to be determined upon convenience or some other informal basis.14 A variety of methods have been proposed to formally determine a target difference (including those for the MCID and its variants).7,11

Existing methods for specifying target differences

There are at least six main formal approaches to identify a targeted difference upon which to base the sample size calculation:

1. Opinion seeking methods: Formal methods for determining the target difference on the basis of eliciting expert opinion (usually clinicians) have been proposed through organisation of either a conference of experts, surveys of members of professional bodies or regulatory committees, or individual interviews.15 A formal Bayesian elicitation enabling both expectation of the difference and also a range of values (for which the trial result could be categorised as one or the other of the treatments being superior or equivalent) has been used.16 Clinical judgement may identify interventions
which are expected to be similar and for which empirical evidence is available (see method 5 below).
Informally, clinical opinion will always be one of the aspects considered.

2. **Distribution methods**: Such methods typically determine a value that is larger than the inherent imprecision in the measurement and therefore likely to represent a meaningful difference (this includes determining SDC/MDC). Other methods are based upon the nature of the outcome (e.g. a fraction of the response range for a visual analogue scale).

3. **Standardised effect size approach**: Under such an approach, the statistical characteristic of the outcome measure is used to define the target difference. For a continuous outcome, the standardised difference (most commonly expressed as Cohen’s ‘effect size’) can be used. Interpretation of the ‘effect size’ approach is heavily reliant on the work by Cohen\(^17\) in giving values of 0.2, 0.5 and 0.8 for small, medium and large effects. Alternatively, the measurement error associated with the outcome can be accounted for (e.g. based upon test–retest reliability) to provide a value that can be characterised as a non-spurious difference. For a binary or time-to-event outcome, risk or hazard ratios respectively, could be utilised. As for the standardised mean difference, an interpretation can be applied to the spectrum of values.

4. **Anchor-based methods**: Under such methods a difference in an outcome measure can be defined as signifying a ‘change’ in status by asking an assessor to judge whether a change (beneficial or otherwise) has occurred. Commonly, patients assess their own change (e.g. before versus after treatment). The values associated with those experiencing a change would then be used to determine the magnitude of difference in the outcome which signifies an important difference. From this the target difference is determined. Variations exist in terms of who assesses change (e.g. patient or clinician), how they assess change (e.g. a change and/or an important changes), what they compare against (e.g. before and after treatment or another patient with the same condition) and how the responses are summarised (e.g. the mean value or receiver operating curve cut-point determination). Rarely have the methods been used and set up with a RCT specifically in mind.\(^{11}\) A Bayesian application could use prior information to determine the size needed for the posterior distribution to rule in or rule out an important difference.

5. **Commissioned research**: A preliminary or pilot study may be carried out where there is little evidence, or even experience, to guide expectations. A pilot study provides support for estimates where relatively small misspecifications could have substantial impact upon corresponding precision and the power to detect a difference (e.g. screening trial).\(^{18}\)

6. **Review of evidence base**: The target difference can be derived using current evidence. Ideally this would be based upon a systematic review of RCTs, and possibly meta-analysis, of the outcome(s) of interest directly addressing the research question at hand. In the absence of randomised evidence, observational evidence has been used in a similar manner. Trials are based upon what the current evidence base suggests is plausible for the parameters of interest. Conventionally, studies are powered in isolation from any current evidence but a formal meta-analysis sample size approach could allow previous evidence to be incorporated in a power calculation akin to the Bayesian Neyman–Pearson hybrid approach.\(^{19}\)

7. **Health Economic approaches**: Recent approaches have used the net monetary benefit (NMB) statistic to define a target difference.\(^{12,20,21}\) An intervention is considered efficient if the NMB for an intervention compared with a comparator is greater than zero (or that the likelihood that the NMB is greater than zero is acceptable). The Bayesian expected value of sampling information approach weighs the expected benefits provided by new research against the expected costs of this research for the key parameters, determining the cost-effectiveness. A distribution for the target difference in a parameter or group of parameters can be inferred.

**Aims and Objectives:**

The aim of this project is to consider all potentially relevant methods of defining a target difference in order to develop clear guidance for researchers on appropriate methods to use under varying circumstances.
To achieve this aim, the following are key objectives for the project:

1. To conduct a systematic review of the methods for identifying a target difference (that has been developed either within or outside the health field), critically appraising the usefulness of each method for different types of RCTs.
2. To identify the methods currently considered as ‘best’ practice using a survey of UK- and Ireland-based Clinical Trial Units, MRC Trial Hubs and the membership of the Society of Clinical Trials.
3. To develop draft guidance to be discussed in a workshop and symposium, and incorporate feedback from these events into finalised guidance for researchers.
4. To identify future research needs

An expert Advisory Group will monitor and review the progress of the project along with the Project Steering Group (grant holders and lead project research fellow).

**Methods**

**Objective 1: Systematic review of the methods to identify a target difference**

**Search Strategy**

The search strategy will involve conducting an electronic literature search of both biomedical and some non-biomedical databases (e.g. Econlit), building on preliminary work already undertaken as part of the original grant application process (see Appendix 2).

It is proposed that both biomedical and some non-biomedical databases are searched, as studies using methods that may be of relevance to RCTs can often be performed in other fields (e.g. behavioural sciences), and restriction to the biomedical field may miss ‘novel’ methods of relevance to the biomedical field. Relevant literature from biomedicine, the social sciences and science and technology fields will therefore be searched, and the databases to be considered include:

- MEDLINE
- EMBASE
- CENTRAL
- Cochrane Methodology Register
- Science Citation index (SCI)
- Econlit
- PsycINFO
- Education Resources Information Centre (ERIC)

There will be no language restriction and searches will be undertaken on dates from 1966 onwards (or from the start of database coverage). A limit on the number of databases searched may be required if the number of records identified is particularly high.

The search of electronic databases will also be augmented in other ways including:

1. Cited reference search
   - The Web of Science and Scopus databases will be used to identify studies referencing any of the key methodological papers that have already been identified through the main electronic database search.
2. Reference lists of included studies
   - These will be checked to identify new methods or variants of a known method.
3. Hand searching
The relevant literature is likely to be distributed across many fields and journals. If a particular journal is the source of many papers, hand-searching of this journal may be considered. However, this method is resource-intensive and the extent to which it will be undertaken depends on the resources available following completion of higher priority tasks. In addition to hand searching journals, standard clinical trial text books will be reviewed to ascertain if they reference or describe a method for eliciting a target difference. General clinical trials books or books on calculating the sample size for clinical trial published in the last 5 years will be reviewed. Additionally, older textbook that are viewed as influential in the trial community will be also reviewed.

4. Contacting those with an interest in the field
   - A number of key figures are involved in this project as either applicants or named collaborators. In addition, authors of key studies identified may be contacted for information on additional available evidence.

5. Grey literature searching
   - Guidance documents from regulatory authorities (e.g. FDA) and known international standards organisations (e.g. International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use) will be searched.

6. Methods being used by UK Clinical researchers
   - Information on methods currently in use will be identified from the surveys (see Objective 2 for more details).

An exploration of best combinations of subject headings and text word searching (searching in titles, abstracts and author specified keywords) will be undertaken before full searches are carried out. It is anticipated that few indexed terms will be suitable and therefore the search strategies are expected to mainly consist of text words and phrases using appropriate synonyms, truncation symbols and adjacency operators.

Inclusion and Exclusion Criteria
This review will concentrate on papers identifying new methods (or a significant variant of established methods) for determining the target difference, although established methods will be referenced and are likely to be identified by the search strategy. Papers published, in any language, will be included which specify methods for determining the target difference (either explicitly or implicitly).

It is likely that a variety of methods will meet our inclusion criteria, and although they should be relevant to RCTs, they may not be reported in RCTs themselves or necessarily used in this context. As the focus of the review is to identify methods for establishing target differences, no restriction regarding the study design (e.g. RCT, quasi-experimental, etc.) in which it may have been applied will be made. Where a method has been identified from a report of a primary study, details of the type of primary study from which it came will be noted at data extraction stage. In terms of outcomes, all types of outcome (e.g. dichotomous, continuous) relevant to clinical trials, including efficacy, effectiveness and cost-effectiveness, will be eligible. Included papers will have to report a real or hypothetical example which seeks to determine a difference (explicitly or implicitly via providing a basis for study size), based on at least one outcome of relevance to clinical trials or which could be used for this purpose.

Inclusion criteria are:

- Reporting a method which could be used to specify a target difference. A method may implicitly specify the target difference by determining the optimal study sample size. The assessment must be based on at least one outcome of relevance to clinical trials or could be used for this purpose. The use of a method in a hypothetical scenario is eligible for inclusion.

Exclusion criteria are:

- Studies failing to report a method for specifying a target difference.
Appendix 1

- Systematic reviews of methods for specifying the target difference or of outcome scales. These reports will be retained and their reference lists reviewed for potentially eligible studies. Such papers will only be included if it is the primary reference for a relevant method.
- Studies reporting only on the statistical considerations for sample size (e.g. a new formula for the sample size calculation) will not be considered sufficient to meet the inclusion criteria.
- Studies which discuss a metric (e.g. risk ratio or number needed to treat) without reference to how a specific difference could be determined will not be considered relevant for inclusion.

It is anticipated that it will not be possible during the initial abstract screening phase, to exclude or include with absolute certainty, potentially relevant papers reporting target difference methods. This is expected because abstracts are brief summaries and relevant information on the method used to elicit a target difference may not be sufficient within an abstract to allow a final decision on inclusion to be made. In addition, it is likely that many papers reporting methods for eliciting target difference may report the use of existing methods rather than new methods (or substantial variation of existing methods) not previously identified. As a result, many papers may report the same method. It will therefore be useful to provisionally categorise papers by their abstracts at the review screening stages, in order to split included papers depending on the method reported, prior to extracting more detailed information (e.g. outcome measures) at the full-text data extraction stage.

Titles and abstracts of the search strategy results will be screened by one reviewer in the first instance, but where there is uncertainty the opinion of a second reviewer will be used, and if necessary a third member of the team will act as an arbiter where there is disagreement. Full text papers will be obtained where, on initial screening of the abstract, the work is considered to be potentially relevant and these papers will be assessed to confirm inclusion or exclusion in the review.

A register of studies meeting the inclusion criteria will be organised using Reference Manager bibliographic software using the key word facility to classify articles by type, reference source and methodology.

Data Extraction Strategy
At the screening stages of the review, included (or potentially relevant) papers will be categorised by the method they report. Following the categorisation of papers by the method reported, data will be extracted from papers to help summarise the variation and range of applicability of each method. Data extracted at this stage will include (where reported):

1. What is measured, e.g. single measure of clinical effectiveness or safety, composite measure of clinical effectiveness and/or safety, a measure of overall (or disease specific) health, or cost/cost-effectiveness measure.
2. Type of outcome measure, e.g. binary, ordinal, continuous or time-to-event.
3. Relevant summary measure reported, e.g. risk ratio, absolute risk difference, mean difference.
4. Size of the sample used to elicit value for important difference (where reported).
5. Perspective used to define target difference, e.g. patients’ and clinicians’

Data will also be extracted on the following details (where they have been reported):

- The context in which the difference was elicited (e.g. real or hypothetical RCT).
- Terminology used to describe the important difference.
- Methodological details and noteworthy features (e.g. unique variations).

It may be necessary to extract different information depending on the method used. For example, details of any formulas used for distribution methods are unlikely to apply to expert opinion methods. As a result, the common factors listed above will be extracted along with specific information relevant to each particular method, and therefore no generic data extraction form will be used.
Method of Analysis
A summary description of each method found will be produced by reporting extracted details (and also any slight variants of the method). The key characteristics of each method will be categorised. A cube classification system for studies of responsiveness has been proposed. However, some modification/or simplifications to the original method may be necessary when applied to the available evidence, particularly in considering the implications of applying markedly different methods to RCTs.

Once the methods have been classified and the evidence available summarised, an assessment of the strengths and weaknesses of each method will be undertaken. This will link the review of the available evidence to the development of methodological guidance for eliciting targeted differences (see Objective 3). The criteria used to critique all methods will be developed in relation to the key focus of this review (i.e. applicability of methods to a clinical trial setting) and may include aspects such as the practical feasibility of using the method, the appropriateness of the method, its reliability and validity etc.). The critique will also evaluate whether particular methods are better suited to particular stages of development of an intervention. Uncertainty will be considered with regard to both the calculation of precision around the proposed target difference found using each method, and the extent to which the target difference value can account for, or might be modified by, other potential outcome measures for the same study population being considered.

The criteria itself will be developed and finalised in consultation and agreement with the project steering and advisory groups. Existing checklists that have been developed for other methodological reviews may also be used to help develop the criteria to be used to assess the strengths and weaknesses of each method being considered by this review, although it is likely that no existing checklist will be fully relevant or applicable to this review.

Objective 2: Surveys
In order to assess awareness and usage of different methods, two surveys will be sent; one to the international Society for Clinical Trials (SCT) and one to UK-based Trialists. The aim of the surveys is to evaluate the methods used in practice for establishing target differences, and it is hoped this will provide practical information on the possible strengths and weaknesses of different approaches for eliciting a target difference. While the two surveys will essentially be the same, the second will be slightly more extensive (see below for details). The methodology of the two surveys are outlined below.

Methodology of the Surveys
Survey 1: SCT membership
All members of the Society for Clinical Trials will be surveyed through the email distribution list. The survey will ask generic questions about the individual responding (position, affiliation, location and whether they are involved in the design of RCTs). They will be asked about their awareness and usage of methods for determining the target difference with the opportunity to suggest an additional method provided. A brief summary of each of the seven identified methods will be provided in the online form. Additionally, the respondent will be asked whether they would be willing to recommend the use of any of the methods. Finally, an opportunity to comment on the issue will be provided. Members will receive an email invitation sent via the society’s email distribution list, inviting members to complete an online survey. The invitation will include a brief introduction to this issue and the aim of the survey. The online survey will be designed bespoke for this purpose by the Health Services Research Unit (HSRU) Programming team. Once potential participants receive the email, they will then be able to access the survey by clicking on, or typing the URL hyperlink provided. Participants may then complete and submit their response. An email reminder will also be sent out one week after the initial email invitation. As it is not possible to tailor reminders to individuals who have not completed the survey, only a general reminder to the entire study sample will be sent.
Survey 2: UK- and Ireland-based trialists

For the Survey of UK- and Ireland-based trialists, one named individual per unit or hub (e.g. Director) for the clinical trial units and MRC Hubs and Regional NIHR Research Design Services will be sampled. Where the same individual holds a position at more than one entity only one survey will be sent. They will be requested to forward on to the appropriate member of their group if they are not personally able to complete.

In addition to the information collected in the SCT survey, this survey (see Appendix 4) includes questions on the approach used for the most recent trial developed including the underlying basis adopted for the target difference (e.g. realistic difference or important difference) any methods for determining the target difference used. Additionally, they will be asked if there is anything that would aid them in the design of RCTs and if they would be happy to be contacted. If appropriate, a structured telephone discussion will be used to elicit further details on their response to the survey.

The initial request will be personalised and send by post, and will include an invitation letter, paper survey and description of the methods available for determining the difference. A paper reminder will be sent after a period of two weeks from the initial notification of the survey. An additional (final) email reminder will be sent after an additional week with an electronic invitation, version of the survey and description of the methods.

The survey invitations and formats will be piloted with members of the project team and local researchers in the first instance.

Ethical Review

The surveys will be submitted for review and approval by the appropriate University of Aberdeen’s College of Life Sciences and Medicine Ethics Review Board (CERB). This project will abide by the MRC’s guidance on Good Research Practice and conform to the University of Aberdeen’s Research Governance Guidelines.

Data Management and Consent

The responses to the online survey and submitted survey data will be stored within a secure database on a secure server within HSRU. HSRU follows the University of Aberdeen policies on I.T. Security and Data Protection and all staff have signed, and are required to adhere to, a ‘Protecting Information Policy’, and are also expected to adhere to the principles of Good Clinical Practice.

Data Analysis

Each survey will be analysed separately. The response rate will be defined as the respective number of responding participants divided by the number of potential participants. Data will be summarised across responses to identified methods for eliciting target difference are used in practice, the opinions of those working in trial design, on the methods identified by the systematic review and in current use. No statistical analysis is planned. Survey results will be discussed with the steering and advisory groups, and will be used to develop the guidance on methods for eliciting the target difference. In addition, the survey findings will be disseminated to each of the surveyed groups.

Objectives 3 and 4: Developing Guidance documentation and identification of future research needs

The draft guidance will be developed from the results of the systematic review and questionnaire survey (Objectives 1 and 2). It will detail the strengths and weaknesses of each approach identified, and will be divided into separate guidance sections on each method. The usefulness for particularly study designs (e.g. phase II, III or IV) and types of outcome and summary measure) will be considered. It is felt that this is crucial to enabling researchers and funders to implement the guidance. Draft guidance will be developed by a subset of the steering group and will be presented to the combined project (steering committee and advisory group) team who will provide critical feedback with particular reference to its clarity, relevance and practicality. This will take place at a workshop of the project members towards the end of the project.
The guidance will be further refined and finalised following this meeting in order to incorporate findings from the day, and then disseminated. The process has been developed to provide substantial input from outwith the project steering group to ensure the guidance produced meet the needs of key stakeholders.

It is hoped that the identification of future research needs will occur from both the systematic review and survey components of the project. Additionally, it is anticipated that the process of developing the guidance will highlight areas research needs and that those involved in this project will be well placed to identify where there are current gaps in the knowledge on target differences, and possible solutions that might best improve the existing evidence base on this subject.

**Project Members**

*Steering Group Members*
- Dr Jonathan Cook
- Professor Luke Vale
- Ms Jenni Hislop
- Ms Cynthia Fraser
- Dr Craig Ramsay
- Professor Peter Fayers
- Professor Andrew Briggs
- Professor John Norrie
- Professor Doug Altman
- Professor Ian Harvey
- Dr Brian Buckley

*Advisory Committee Members*
- Professor Marion Campbell
- Professor Adrian Grant
- Professor Ian Ford
- Professor Dean Fergusson

**Competing Interests of Members**
None stated.

**References**


2. Pocock SJ. *Clinical Trials – A practical approach*. 1996 John Wiley & Sons Ltd.


Appendix 2  Literature search strategies

MEDLINE (1966 to November Week 2 2010), EMBASE (1980 to 2010 Week 45), Medline In Process & Other Non-Indexed Citations (17 November 2010)

Ovid Multifile Search URL: https://shibboleth.ovid.com/

1. mcid.tw.
2. (target$ adj1 difference?).tw.
3. change score.tw.
4. change point.tw.
5. (clinical$ importan$ adj2 (difference? or change? or improvement? or effect?)).tw.
6. (minim$ importan$ adj2 (difference? or change? or improvement? or effect?)).tw.
7. (clinical$ meaningful$ adj2 (difference? or change? or improvement? or effect?)).tw.
8. (minim$ meaningful$ adj2 (difference? or change? or improvement? or effect?)).tw.
9. (smallest meaningful$ adj2 (difference? or change? or improvement? or effect?)).tw.
10. (minim$ significant$ adj2 (difference? or change? or improvement? or effect?)).tw.
11. (smallest significant$ adj2 (difference? or change? or improvement? or effect?)).tw.
12. (minim$ detect$ adj2 (difference? or change? or improvement? or effect?)).tw.
13. (smallest detect$ adj2 (difference? or change? or improvement? or effect?)).tw.
14. (sufficient$ importan$ adj2 (difference? or change? or improvement? or effect?)).tw.
15. (sufficient$ meaningful$ adj2 (difference? or change? or improvement? or effect?)).tw.
16. (minim$ clinical$ adj2 (important or detectable or meaningful)).tw.
17. ((calculat$ or determin$ or comput$) adj2 meaningful).tw.
18. ((calculat$ or determin$ or comput$) adj2 detectable).tw.
19. ((calculat$ or determin$ or comput$) adj2 important adj2 (difference? or change? or improvement? or effect?)).tw.
20. ((calculat$ or determin$ or comput$) adj2 meaningful adj2 (difference? or change? or improvement? or effect?)).tw.
21. ((calculat$ or determin$ or comput$) adj2 detectable adj2 (difference? or change? or improvement? or effect?)).tw.
22. (definition$ adj2 (difference? or change? or improvement?)).tw.
23. ((responsiveness adj2 (calculat$ or determin$ or comput$)) and (measure$ or scale$ or score$ or rating$)).tw.
24. *sample size/
25. or/1-24

PsycINFO (1967 to January Week 2 2011)

Ovid Search URL: https://shibboleth.ovid.com/

1. mcid.tw.
2. (target$ adj1 difference?).tw.
3. change score.tw.
4. change point.tw.
5. (clinical$ importan$ adj2 (difference? or change? or improvement? or effect?)).tw.
6. (minim$ importan$ adj2 (difference? or change? or improvement? or effect?)).tw.
7. (clinical$ meaningful$ adj2 (difference? or change? or improvement? or effect?)).tw.
APPENDIX 2

8. (minim$ meaningful$ adj2 (difference? or change? or improvement? or effect?)).tw.
9. (smallest meaningful$ adj2 (difference? or change? or improvement? or effect?)).tw.
10. (minim$ significant$ adj2 (difference? or change? or improvement? or effect?)).tw.
11. (smallest significant$ adj2 (difference? or change? or improvement? or effect?)).tw.
12. (minim$ detect$ adj2 (difference? or change? or improvement? or effect?)).tw.
13. (sufficient$ importan$ adj2 (difference? or change? or improvement? or effect?)).tw.
14. (sufficient$ meaningful$ adj2 (difference? or change? or improvement? or effect?)).tw.
15. (minim$ clinical$ adj2 (important or detectable or meaningful)).tw.
16. ((calculat$ or determin$ or comput$) adj2 meaningful).tw.
17. ((calculat$ or determin$ or comput$) adj2 detectable).tw.
18. (smallest detect$ adj2 (difference? or change? or improvement? or effect?)).tw.
19. ((calculat$ or determin$ or comput$) adj2 important adj2 (difference? or change? or improvement? or effect?)).tw.
20. ((calculat$ or determin$ or comput$) adj2 meaningful adj2 (difference? or change? or improvement? or effect?)).tw.
21. ((calculat$ or determin$ or comput$) adj2 detectable adj2 (difference? or change? or improvement? or effect?)).tw.
22. *sample size/
23. (definition$ adj2 (difference? or change? or improvement?)).tw.
24. ((responsiveness adj2 (calculat$ or determine$ or comput$)) and (measure$ or scale$ or score$ or rating$ or metric$)).tw.
25. clinical$ importan$.id.
26. clinical$ significan$.id.
27. clinical$ meaningful$.id.
28. minim$ importan$.id.
29. minim$ significan$.id.
30. minim$ meaningful$.id.
31. smallest meaningful$.id.
32. smallest significan$.id.
33. smallest importan$.id.
34. sufficient$ importan$.id.
35. sufficient$ meaningful$.id.
36. minim$ clinical$.id.
37. minim$ detect$.id.
38. importan$ difference$.id.
39. meaningful$ difference$.id.
40. minim$ difference$.id.
41. or/1-40

The Cochrane Library (Cochrane Central Register of Controlled Trials and Cochrane Methodology Register Issue 1 2011)

URL: www3.interscience.wiley.com/

#1 (mcid) or (target* NEXT difference*) or (change NEXT score) or (Change NEXT point)
#2 (clinical* NEXT importan*) NEAR/2 (difference* or change* or improvement* or effect*)
#3 (minim* NEXT importan*) NEAR/2 (difference* or change* or improvement* or effect*)
#4 (clinical* NEXT meaningful*) NEAR/2 (difference* or change* or improvement* or effect*)
#5 (minim* NEXT meaningful*) NEAR/2 (difference* or change* or improvement* or effect*)
#6 (smallest NEXT meaningful*) NEAR/2 (difference* or change* or improvement* or effect*)
#7 (minim* NEXT significant*) NEAR/2 (difference* or change* or improvement* or effect*)
#8 (smallest NEXT significant*) NEAR/2 (difference* or change* or improvement* or effect*)
#9 (minim* NEXT detect*) NEAR/2 (difference* or change* or improvement* or effect*)
#10 (smallest NEXT detect*) NEAR/2 (difference* or change* or improvement* or effect*)
#11 (sufficent* NEXT importan*) NEAR/2 (difference* or change* or improvement* or effect*)
#12 (sufficent* NEXT meaningful*) NEAR/2 (difference* or change* or improvement* or effect*)
#13 (minim* NEXT clinical*) NEAR/2 (difference* or change* or improvement* or effect*)
#14 (calculat* or determin* or comput*) NEAR/2 (meaningful)
#15 (calculat* or determin* or comput*) NEAR/2 (detectable)
#16 (calculat* or determin* or comput*) NEAR/2 (important)
#17 (calculat* or determin* or comput*) NEAR/2 (detectable)
#18 (definition*) NEAR/2 (difference* or change* or improvement*)
#19 (responsiveness) NEAR/2 (calculat* or determine* or comput*)
#20 MeSH descriptor Sample Size
#21 (#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #20)

Science Citation Index (1970 to 22 January 2011)

ISI Web of Knowledge URL: http://wok.mimas.ac.uk/

#1 TS=mcid
#2 TS="targeted difference"
#3 TS="change score"
#4 TS="change score"
#5 TS="change point"
#6 TS="clinical* important" SAME TS=(change* or difference* or improvement* or effect*)
#7 TS="minimal* important" SAME TS=(change* or difference* or improvement* or effect)
#8 TS="clinical* meaningful" SAME TS=(change* or difference* or improvement* or effect*)
#9 TS="minimal* meaningful" SAME TS=(change* or difference* or improvement* or effect*)
#10 TS="smallest meaningful" SAME TS=(change* or difference* or improvement* or effect*)
#11 TS="minimal* significant" SAME TS=(change* or difference* or improvement* or effect*)
#12 TS="smallest significant" SAME TS=(change* or difference* or improvement* or effect*)
#13 TS="minimal* detect*" SAME TS=(change* or difference* or improvement* or effect*)
#14 TS="smallest detect*" SAME TS=(change* or difference* or improvement* or effect*)
#15 TS="sufficient* important" SAME TS=(change* or difference* or improvement* or effect*)
#16 TS="sufficient* meaningful" SAME TS=(change* or difference* or improvement* or effect*)
#17 TS="minim* clinical* importan"
#18 TS="minim* clinical* detectable"
#19 TS="minim* clinical* meaningful"
#20 #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17 or #18 or #19 AND Document Type=(Article)

EconLit (1984 to 31 January 2011)

CSA Illumina URL: www.csa1.co.uk/

S1 TX mcid or TX target* w2 difference*
S2 TX “change score” or TX “change point”
S3 TX minim* w2 importan or TX sufficient* w2 importan* or TX smallest w2 importan*
S4 TX minim* w2 meaningful* or TX smallest w2 meaningful* or TX sufficient* w2 meaningful*
S5 TX minim* w2 significant* or TX smallest w2 significant* or TX smallest w2 detect*
S6 TX minim* w2 detect* or TX minim* w2 difference or TX meaningful w2 difference
Education Resources Information Centre (1960 to 28 January 2011)

Proquest URL: http://search.proquest.com/

(((mcid or (target* within 2 difference*)) or ("change score") or ("change point")) or (minim* within 2 importan*) or (sufficient* within 2 importan*) or (smallest within 2 importan*) or (minim* within 2 meaningful*) or (smallest within 2 meaningful*) or (sufficient* within 2 meaningful*) or (minim* within 2 significant*) or (smallest within 2 significant*) or (sufficient* within 2 detect*) or (minim* within 2 detect*) or (minim* within 2 difference) or (minim* within 2 difference) or (meaningful within 2 difference) or (meaningful within 2 change) or (smallest within 2 change) or (important within 2 change) or (minim* within 2 change) or (smallest within 2 change))

Scopus (28 January 2011)

URL: www.scopus.com/

((TITLE-ABS-KEY(mcid) AND DOCTYPE(ip)) or (TITLE-ABS-KEY("target* difference**") AND DOCTYPE(ip)) or (TITLE-ABS-KEY("change point") AND DOCTYPE(ip)) or (TITLE-ABS-KEY("minim* important") AND DOCTYPE(ip)) or (TITLE-ABS-KEY("minim* clinical") AND DOCTYPE(ip)) or (TITLE-ABS-KEY("clinical* meaningful") AND DOCTYPE(ip)) or (TITLE-ABS-KEY("clinical* significant") AND DOCTYPE(ip)) or (TITLE-ABS-KEY("minim* clinical*") AND DOCTYPE(ip)) or (TITLE-ABS-KEY("clinical* meaningful*") AND DOCTYPE(ip)))

Clinical trial books/guidelines consulted


## Appendix 3  List of included studies

### Anchor method ($n = 253$)


47. Coeytaux RR, Kaufman JS, Chao R, Mann JD, Devellis RF. Four methods of estimating the minimal important difference score were compared to establish a clinically significant change in Headache Impact Test. *J Clin Epidemiol* 2006;59:374–80.


126. Kelly AM. Does the clinically significant difference in visual analog scale pain scores vary with gender, age, or cause of pain? Acad Emerg Med 1998;5:1086–90.


144. Landorf KB, Radford JA. Minimal important difference: values for the Foot Health Status Questionnaire, Foot Function Index and visual analogue scale. *Foot* 2008;18:15–19.


237. Weisscher N, Vermeulen M, Roos YB, de Haan RJ. What should be defined as good outcome in stroke trials; a modified Rankin score of 0–1 or 0–2? J Neurol 2008;255:867–74.


240. Wheaton L, Pope J. The minimally important difference for patient-reported outcomes in spondyloarthropathies including pain, fatigue, sleep, and Health Assessment Questionnaire. *J Rheumatol* 2010;37:816–22.


**Distribution method (n = 171)**


26. Dunngalvin A, Cullinane C, Daly DA, Flokstra-De Blok BMJ, Dubois AEJ, Hourihane JB. Longitudinal validity and responsiveness of the Food Allergy Quality of Life Questionnaire – Parent Form in children 0–12 years following positive and negative food challenges. *Clin Exp Allergy* 2010;**40**:476–85.


139. Spratt KF. Patient-level minimal clinically important difference based on clinical judgment and minimally detectable measurement difference: a rationale for the SF-36 physical function scale in the SPORT intervertebral disc herniation cohort. *Spine* 2009;**34**:1722–31.


**Health economic method (n = 13)**


**Opinion-seeking method (n = 60)**


**Pilot study method (n = 5)**


**Review of evidence base method (n = 22)**


**Standardised effect size (n = 37)**


20. Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. *Arch Gen Psychiatry* 1995;52:867–71.


**Multiple methods (n = 216)**


33. Copay AG, Glassman SD, Subach BR, Berven S, Schuler TC, Carreon LY. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. *Spine J* 2008;8:968–74.


39. de Morton NA, Davidson M, Keating JL. Validity, responsiveness and the minimal clinically important difference for the de Morton Mobility Index (DEMMI) in an older acute medical population. *BMC Geriatrics* 2010;10:72.


44. Demoulin C, Ostelo R, Krottnener JA, Smeets RJEM. Quebec back pain disability scale was responsive and showed reasonable interpretability after a multidisciplinary treatment. *J Clin Epidemiol* 2010;63:1249–55.


130. Patrick DL, Burns T, Morosini P, Gagnon DD, Rothman M, Adriaenssen I. Measuring social functioning with the personal and social performance scale in patients with acute symptoms of


155. Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol* 2004;57:1008–18.


Appendix 4 Survey form sent to UK- and Ireland-based triallists
A survey of current practice and usage of formal methods regarding determination of the target difference in randomised controlled trial sample size calculations

DELTA is funded by the Medical Research Council UK’s Methodology Research Programme. The project group includes collaborators from the UK, Ireland and Canada.
Introduction:

Deciding how many participants are needed for a randomised controlled trial (RCT) is a key issue. Most methods for determining the trial size incorporate a difference (e.g. 10% difference in success rate) which the trial is designed to detect. From both a scientific and ethical standpoint, selecting an appropriate (target) difference is of crucial importance. The choice of target difference has received relatively little attention until recently, though a variety of formal methods have been proposed. This short survey seeks to find out the extent to which leading experts on RCT design are aware of these methods, and to determine which of these methods are in current use. It has been sent to the Directors UKCRC registered Clinical Trial Units (CTUs), MRC UK Hubs for Trial Methodology and NIHR Research Design Services (RDS).

We would be extremely grateful if you could take the time (approximately 10 minutes) to complete this short survey.

1. What is your position? (tick all that apply if you hold a position with more than one group)
   - CTU Director/Co-Director
   - MRC Trial Hub Director/Co-Director
   - RDS Director
   - Other, please state.................................................................................................................

2. Which group are you responding on behalf of? (tick all that apply if are responding on behalf of more than one group. Please treat them as a single entity for subsequent questions.)
   - CTU
   - MRC Trial Hub
   - RDS

3. Which types of interventions does your group’s trial portfolio cover? (please tick all that apply.
   - Pharmacological
   - Non-pharmacological
4. Which phases of trials are represented in your group’s RCT portfolio? (please tick all that apply.)

- Phase II
- Phase III
- Phase IV
- Other, please state

5. Which clinical areas does your group’s RCT portfolio cover? (please tick all that apply)

- Blood
- Cancer
- Cardiovascular
- Dementias and Neurodegenerative Diseases
- Diabetes
- Ear
- Eye
- Genetics & Congenital Disease
- Infection
- Inflammatory and Immune
- Injuries & Emergencies
- Medicines for Children
- Mental Health
- Metabolic and Endocrine
- Musculoskeletal
- Neurological
Formal methods for determining the target difference:

A variety of methods have been proposed to determine the target difference. One approach is to use a method to identify the minimal clinically important difference; the smallest value that would be viewed as clinically important. If the treatments differ by this amount it would likely alter clinical practice. Other approaches are more data-driven i.e. What is a realistic value given what we already know? Some of the available methods can be used to determine a realistic value or a value which is viewed as clinically important.

6. Which of the following methods are you aware of? (Please tick all that apply)

- Anchor methods
- Distribution methods
- Health economic methods
- Opinion seeking methods
- Pilot study
- Review of evidence base
- Standardised effect size approach

Other, please state……………………………………………………………………………………………………………………………………………….

I am not aware of any of these methods
7. Which of the following methods has your group used when designing an RCT? (Please tick all that apply)

- Anchor methods
- Distribution methods
- Health economic methods
- Opinion seeking methods
- Pilot study
- Review of evidence base
- Standardised effect size approach
- Other, please state
- We have not used any of these methods

8. Which of the following methods would you be happy to recommend? (Please tick all that apply)

- Anchor methods
- Distribution methods
- Health economic methods
- Opinion seeking methods
- Pilot study
- Review of evidence base
- Standardised effect size approach
- Other, please state
- I would not recommend any method
Thinking about the trial most recently developed by your group:

9. What was the primary outcome(s) in your most recent trial? (Please tick all that apply)
   - Generic quality of life (e.g. EQ-5D)
   - Disease-specific quality of life (e.g. Oxford Knee Score)
   - Other patient reported outcome (non quality of life measure)
   - Mortality
   - Clinical functional measure (e.g. forced expiratory volume - FEV)
   - Economic outcome (e.g. incremental cost per QALY)
   - Other, please state name and type of measure
     ...
   - There was no primary outcome, please state why

10. Which methods did you use to determine the target difference for the primary outcomes? (Please tick all that apply)
   - Anchor methods
   - Distribution methods
   - Health economic methods
   - Opinion seeking methods
   - Pilot study
   - Review of evidence base
   - Standardised effect size approach
   - Other, please state
   - No formal method was used
11. What was the underlying principle(s) adopted in determining the difference? (Please tick all that apply)

<table>
<thead>
<tr>
<th>Option</th>
<th>Tick</th>
</tr>
</thead>
<tbody>
<tr>
<td>A realistic difference given the interventions under evaluation</td>
<td></td>
</tr>
<tr>
<td>A difference which would lead to an achievable sample size</td>
<td></td>
</tr>
<tr>
<td>A difference that would be viewed as important by a relevant stakeholder group (e.g. clinicians)</td>
<td></td>
</tr>
<tr>
<td>Other, please state</td>
<td></td>
</tr>
</tbody>
</table>

Final questions:

12. Is there anything related to determining the target difference which would help your group in its role in designing/conducting RCT’s?

13. Please feel free to use this space to comment on any aspect related to this topic or survey that you feel is important.

14. Would you be happy for us to contact you if we have further questions?

<table>
<thead>
<tr>
<th>Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
</tr>
</tbody>
</table>

If ‘Yes’, please provide contact details:
THANK YOU FOR TAKING THE TIME TO COMPLETE THIS SURVEY!

YOUR HELP IS GREATLY APPRECIATED.

If you have queries about this survey or about the DELTA project please contact: Jonathan Cook; Tel 01224 438166; email: j.a.cook@abdn.ac.uk.

www.abdn.ac.uk/hsru/research/assessment/methodological/delta/
This report presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.