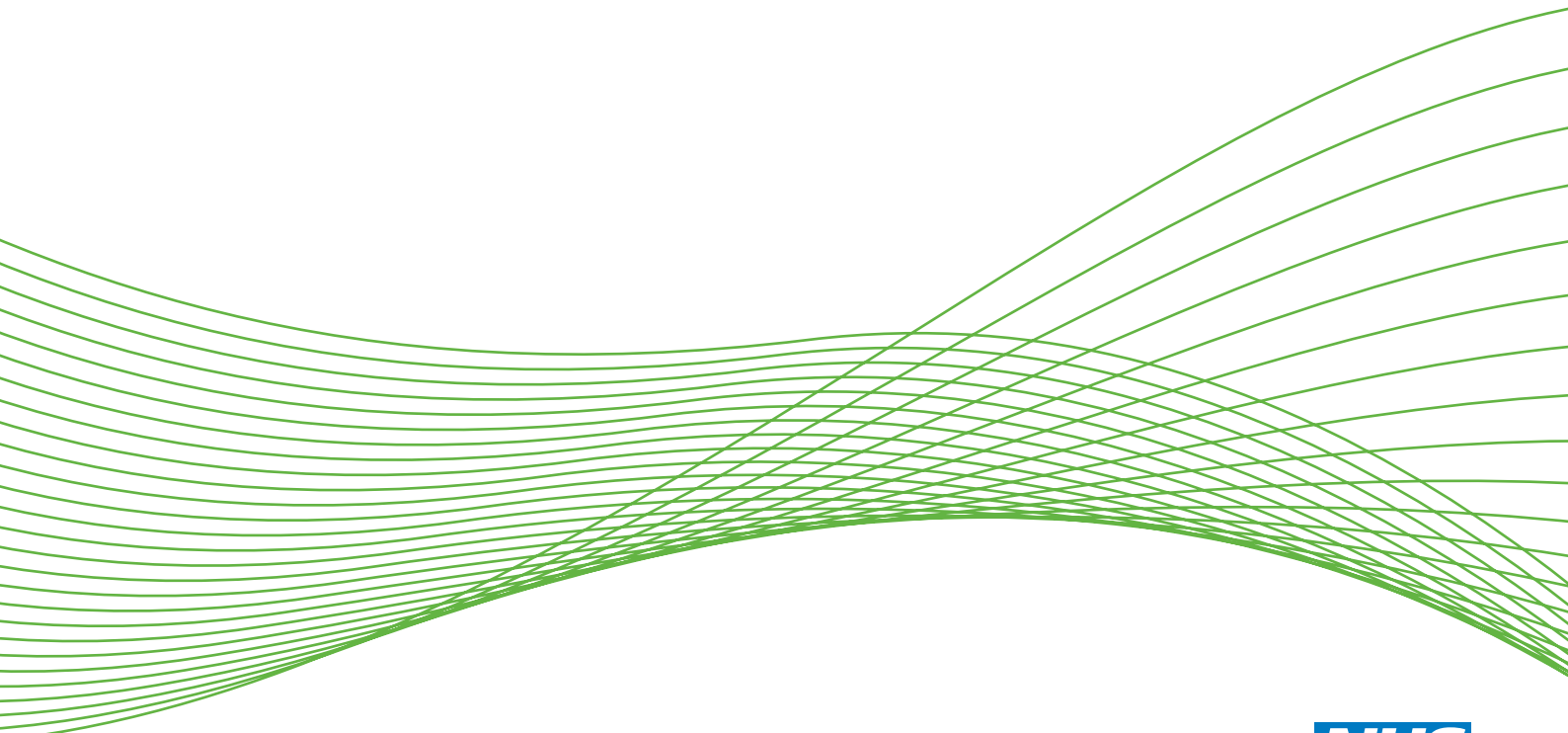


Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review

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



**National Institute for
Health Research**

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:

Cook JA, Hislop J, Adewuyi TE, Harrild K, Altman DG, Ramsay C, *et al.* Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technol Assess* 2014;**18**(28).

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

ISSN 1366-5278 (Print)

ISSN 2046-4924 (Online)

Five-year impact factor: 5.596

Health Technology Assessment is indexed in MEDLINE, CINAHL, EMBASE, The Cochrane Library and the ISI Science Citation Index and is assessed for inclusion in the Database of Abstracts of Reviews of Effects.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (<http://www.publicationethics.org/>).

Editorial contact: nihredit@southampton.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta. Print-on-demand copies can be purchased from the report pages of the NIHR Journals Library website: www.journalslibrary.nihr.ac.uk

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 trialists) 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 but 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 programme.

Contents

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

<i>Health economic method</i>	24
<i>Opinion-seeking method</i>	28
<i>Pilot study method</i>	31
<i>Review of evidence base method</i>	32
<i>Standardised effect size</i>	33
<i>Combination of methods</i>	38
Discussion	42
<i>Key findings</i>	42
<i>Strengths and limitations</i>	44
Chapter 3 Surveys of trialists' current practice	45
Introduction	45
Methodology of the surveys	45
<i>Survey 1: Society of Clinical Trials membership</i>	45
<i>Survey 2: UK- and Ireland-based trialists</i>	45
<i>Ethical review</i>	46
<i>Data analysis</i>	46
Results	46
<i>Survey 1: Society of Clinical Trials membership</i>	46
<i>Survey 2: UK- and Ireland-based trialists</i>	48
Discussion	52
<i>Key findings</i>	52
<i>Strengths and limitations</i>	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
<i>Background</i>	55
<i>Research question</i>	56
<i>Sample size calculation</i>	57
<i>The role of the primary outcome</i>	58
Specifying the target difference	59
Description and guidance on the use of individual methods for specifying the target difference	62
<i>Anchor method</i>	62
<i>Distribution method</i>	63
<i>Health economic method</i>	63
<i>Opinion-seeking method</i>	64
<i>Pilot study method</i>	64
<i>Review of evidence base method</i>	65
<i>Standardised effect size</i>	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

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

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.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.² The importance and impact of the target difference chosen can be demonstrated using two motivating examples (*Table 1*).

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

Trial	Target difference ^a	Result	Triallists' interpretation
Norwegian Spine Study ³	Mean difference of 10 points in the Oswestry Disability Index (SD 18 points)	−8.4 points, 95% CI −13.2 to −6.6 points	'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'
Full-thickness macular hole and Internal Limiting Membrane peeling study (FILMS) ⁴	Mean difference of six letters in distance visual acuity (SD 12 letters)	4.8 letters, 95% CI −0.3 to 9.8 letters	'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'

CI, confidence interval; SD, standard deviation.

^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.³ 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.⁴ 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.⁵ 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.⁶ 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 triallists [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/.

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,¹⁰ 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.¹¹ 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

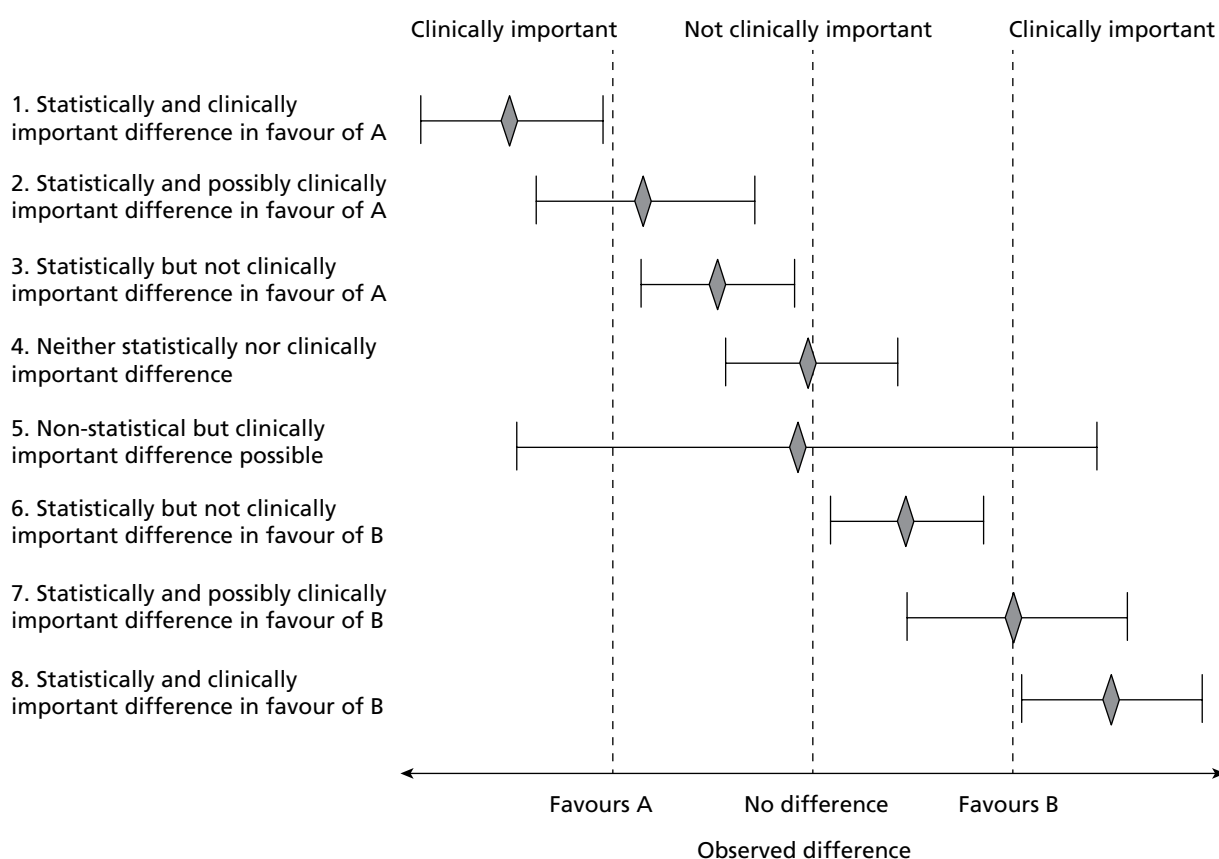


FIGURE 1 Statistically and clinically important difference.

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.^{12–14}

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.¹⁴ 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

beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management',¹⁵ but has also been referred to as the 'minimum difference that is important to a patient'.¹⁶ Variations have been suggested such as the 'minimally clinically important improvement', the 'patient acceptable symptom state' and the 'sufficiently important difference',¹⁶ which seeks to adopt a wider perspective by taking into account cost, risk and harms.^{17–19} 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.^{18,20} 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.^{15,21} 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.^{22–24} 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.^{23,25} 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 difference²⁶ 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

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

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

be potentially answered in a single study.²³ 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.²² 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.²² 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.^{27,28}

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.²⁴ 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.²⁴ 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.²⁹ The CHART trial³⁰ 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)³¹ or stopping a trial if benefit/futility has been shown (sometimes called group sequential trial).^{29,32} 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.²⁹ Trend analysis models have been proposed that adjust the sample size according to the trend in the current data.³³ They are not commonly used because of the risk of bias (similar to play-the-winner designs)³⁴ 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.³⁵ 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.³⁶ 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.²⁴ 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.³⁷ 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.²⁶ A variety of methods have been proposed to formally determine a target difference (including those for the MCID and its variants).^{16,18} 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.^{27,28,37}

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/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 Number of search results obtained from each included database

Database	Number of titles and abstracts identified by the search strategy ^a
MEDLINE/MEDLINE In-Process & Other Non-Indexed Citations/EMBASE	7189
CENTRAL	487
Science Citation Index	1898
EconLit	255
PsycINFO	1367
Cochrane Methodology Register	10
ERIC	158
Scopus	121
Total	11,485

a Number of titles and abstracts after deduplication.

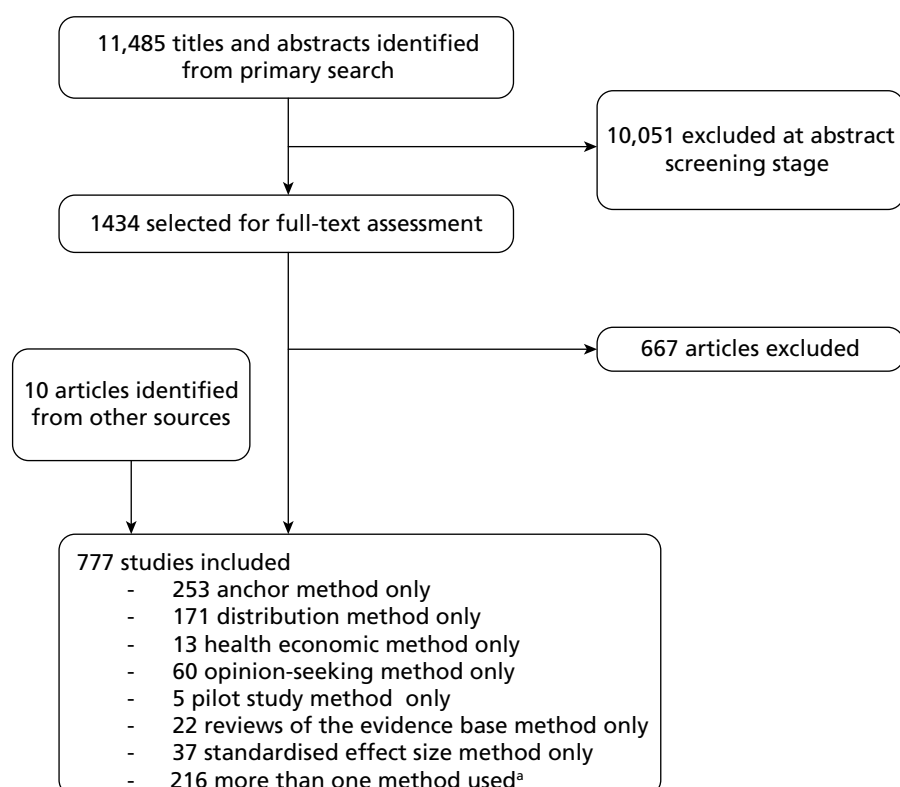


FIGURE 2 The screening process. a, For a breakdown of studies in which more than one method was used in combination, please see *Combination of methods*.

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.^{37,38} 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.³⁹ 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.⁴⁰ In addition, a newly published study was forwarded to the principal investigators by a colleague who thought that it would be relevant for inclusion.⁴¹ Eight further studies were identified from citations of full-text studies screened or were sent for assessment by members of the steering group.^{30,42–48} 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.⁴⁹ 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 colleagues⁵⁰ 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'.^{51,52} Cousens and colleagues⁵³ 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.

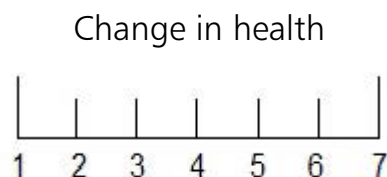
Variations on the anchor method

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.⁵⁴

Please rate the change in the patient’s health from before they received treatment to now by circling one of the following options:



Options:

1. marked improvement
2. moderate improvement
3. minimal improvement
4. unchanged
5. minimal worsening
6. moderate worsening
7. marked worsening.

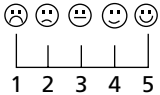

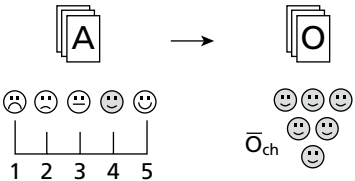
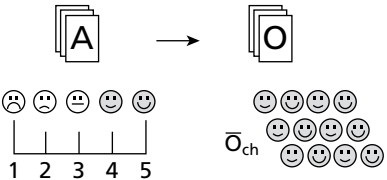
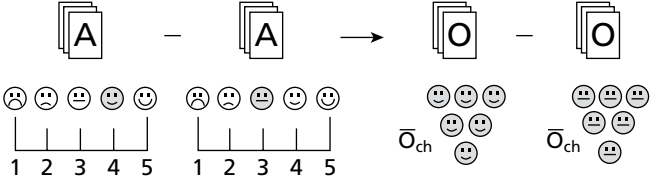
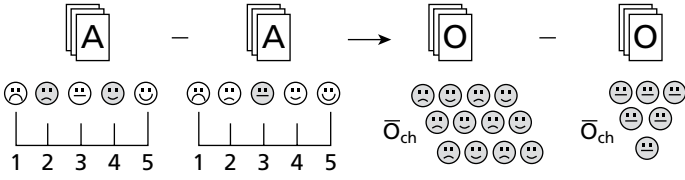
Study design	Estimating the MCID – select minor variations (i–iv) in approach
<p>The study sample (30 patients) complete the instrument of interest (O) both before and after the treatment is received. The change score is calculated for each patient:</p> <p style="text-align: center;">O_{ch}</p> <p>The anchor (A) is a five-point ordinal scale:</p>  <p>It is only completed after treatment for each patient:</p> 	<p>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.</p>  <p>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 (\bar{O}_{ch}) for all those who selected points four or five on A, is the MCID for O.</p>  <p>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 (\bar{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.</p>  <p>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 (\bar{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.</p> 

FIGURE 3 Anchor method illustrative example.

and linear discriminant analysis].^{73–77} Harman and colleagues⁷⁸ 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^{79–83} 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,^{84–86} based on the method established by Jaeschke and colleagues,⁸⁷ although five-point^{69,88,89} and seven-point^{54,90–92} 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).⁹³ 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.^{77,94–97} For example, Khanna and colleagues⁹⁵ 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.^{74,98–100} For example, Colwell and colleagues⁷⁴ 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).^{101,102} 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).¹⁰³ 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.⁷⁵

The anchor question was most often posed to patients alone^{63,104} 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,¹⁰⁵ 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.^{69,71} Parents' views were also occasionally used, especially for studies of paediatric outcomes.^{106–109} 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).¹¹⁰ 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.^{78,111,112} 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.¹¹³ It is worth noting that the anchor method was not always successful in deriving values for interpreting an important difference.^{114,115}

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).^{76,116} In most instances, a rationale for calculating the MCID was not explicitly given. As many studies looked at instrument development, calculation of

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,⁶⁹ 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.^{59,80,88,110,116} 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.^{85,88,122,126,133} 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.^{69,127} Information on a 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 suitably relevant.^{63,110} 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.^{70,75,85,126,129} 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).^{76,116,129,132} There are examples of methods used to counteract this bias. For example, Wyrwich and colleagues,^{134–137} 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.¹³⁸

The method proposed by Redelmeier and colleagues,⁸⁰ 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 complex⁸³ 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.^{70,120,128} 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.⁶⁹ 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,⁶⁸ also minimise the potential for misunderstanding the meaning of a question being used.¹¹⁸ 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.^{63,122} 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.^{63,105} 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.^{76,85,95,122,132,133} Furthermore, the MCID estimate may vary between studies. The agreement or otherwise of the estimate with different studies should be explored.^{54,64,85,121,126}

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).¹³⁹

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 $1.96\sqrt{2}\text{SEM}$ (i.e. 2.77SEM) originates from the measurement error associated with the first and second (repeat) measurement from a test-retest scenario.¹⁴⁰ 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 alpha¹⁴¹ 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).¹⁴²⁻¹⁴⁴ 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 $2.77\sqrt{\text{Sw}}$. A simplified formula, the SD of the differences, SDw, can be used in place of $\sqrt{\text{Sw}}$ if there is no real source of a difference between the first and the second measurements.¹⁴⁵⁻¹⁴⁷ 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',¹⁴⁸ the 'intra-rater standard deviation',¹⁴⁹ or the square root of the 'mean square error'.¹⁵⁰ 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.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 minimised¹⁵² or if test–retest reliability can be improved by, for example, using a larger number of repetitions.^{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 $1.96\sqrt{2}$. 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.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,^{156–160} 1.65 SEM,¹⁶¹ 2.0 SEM^{151,162,163} or even 2.58 SEM.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.^{156,157,159,165–167} Sometimes the value $\sqrt{2}$ SEM is used (approximately a 68% CI),¹⁶⁸ or a 90% CI is used in conjunction with it 1.65 $\sqrt{2}$ SEM,^{169–174} or the t-distribution value is used instead of the z-value.^{145,146} Another study calculated the standard error (SE) of the difference from the SEM using the formula:

$$\sqrt{SEM_{baseline}^2 + SEM_{follow-up}^2} \quad 175 \quad (1)$$

Other variations in implementing the general approach are possible.

A similar approach was proposed by Jacobson and colleagues.^{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 $\sqrt{2(SE)^2}$ 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.^{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.^{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.^{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 SD¹⁷⁹ or 1.5 SDs¹⁸⁰ have also been proposed, as has a weighted cut-point when data are not normally distributed.^{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 sample¹⁸¹ 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.¹⁸⁵ 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.^{186,187} Nevertheless, with overlap, there may still be difficulties in identifying the separate distributions of each group.¹⁸⁸ Minor variants include using Cohen's *d* to calculate the percentage of non-overlap.¹⁸⁹ 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,¹⁹⁰ whereas those experiencing comparably smaller changes in magnitude may cross the cut-off point.¹⁹¹ It is worth noting that these approaches have been used outside clinical settings, for example in marriage counselling,¹⁹² 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.¹⁹³

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).¹⁹⁴ 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 Hoeven¹⁹⁵ defined the 'minimum significant difference' as 'for a given significance level α is the smallest shift δ such that $p(\Delta = \delta) \leq \alpha$ '. The delta symbols used represent the sample (δ) and population (Δ) 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 colleagues¹⁹⁶ (and similarly by Pijls and colleagues¹⁹⁷) as:

$$\sqrt{(z_{\alpha} + z_{\beta})^2 \times \sigma^2 / n} \quad (2)$$

where α is the significance level, $1 - \beta$ is the statistical power, σ^2 is the variance of the within-person differences, z_{α} 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 colleagues¹⁹⁸ and Bridges and Farrar¹⁹⁹ for two independent groups (equal group size and variance), and allowing for unknown variance, as:

$$\sqrt{2}(t_{\alpha, v} + t_{\beta, v})SD / \sqrt{n} \quad (3)$$

where $t_{\alpha, v}$ (and $t_{\beta, v}$ 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 α and β are as above. Anderson and colleagues²⁰⁰ used a similar formula to define the 'minimum significant difference', which allowed for unequal groups sizes although without regard to statistical power:

$$t_{\alpha, v} \sqrt{S_w \frac{1}{n_1} + \frac{1}{n_2}} \quad (4)$$

where n_1 and n_2 are the number of observations in the respective groups and t_{α_v} and S_w are as before.

In the context of samples of soil, Hoss and colleagues²⁰¹ 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):

$$\frac{100t_{\alpha, n_c+n_r-2}\sqrt{\frac{(SD_c)^2}{n_c} + \frac{(SD_r)^2}{n_r}}}{\bar{X}_r} \quad (5)$$

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 \bar{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, Fuchsman and colleagues,²⁰² Kropmans and colleagues²⁰³ and Warren-Hicks and colleagues²⁰⁴). 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 procedure²⁰⁵ and Dunnett's multiple comparison test.²⁰⁶ In a similar manner, though without the formal statistical test framework, Ndlovu and colleagues²⁰⁷ 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:

$2.8\text{CoV},^{208}$

$2\frac{\sum \text{CoV}}{n_{\text{replications}}},^{209}$ and

$2\sqrt{2}\text{precision error}^{210}$

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 colleagues²¹¹ 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 clearly 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.²¹² 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.^{23,213}

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^{214,215} 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's²¹⁴ 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.²¹⁵ 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.²¹⁶ 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).

BOX 3 Summary of Detsky's²¹⁴ 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.

Following the same general principle as Torgerson and colleagues,²¹⁵ Samsa and Matchar²¹⁷ 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 Gray¹⁷ took an approach that in its basic form was conceptually similar to that of Torgerson and colleagues.²¹⁵ 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 (δC) 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:

$$\delta C = R_c \Delta \tilde{E} \quad (6)$$

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 (ΔC) should be less than the critical difference in cost (δC). 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,¹⁷ O'Hagan and Stevens¹⁹ 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 ΔE and ΔC 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

TABLE 3 Two worked examples based on the Torgerson and colleagues²¹⁵ approach

	Worked example 1	Worked example 2
Research question	For a given difference in cost what difference in effectiveness could give the (maximum) acceptable incremental cost-effectiveness ratio (ICER)	For a given difference in cost what difference in effectiveness could give an 'efficient' (as defined by economic theory) allocation of resources
Definitions	C_o = Cost of open hernia repair = £1009 C_L = Cost of laparoscopic hernia repair = £1190 ΔC = Difference in cost = $C_L - C_o = £1190 - £1009 = £181$ ΔE = Difference in effectiveness (QALYs) = $E_L - E_o$ = unknown Incremental cost per unit of effectiveness (ICER) = $\Delta C / \Delta E$	E_o = Effectiveness of open hernia repair = 4.42 QALYs E_L = Effectiveness of laparoscopic hernia repair = unknown E_N = Effectiveness of no intervention = 2 QALYs C_o = Cost of open hernia repair = £1009 C_L = Cost of laparoscopic hernia repair = £1190 C_N = Cost of no intervention = £500 The technical efficiency for an intervention can be defined as the ratio of costs to effects for that intervention. If $C_o/E_o = C_L/E_L$ holds then the two interventions are judged to be economically equivalent. If, however, the use of no (neither) intervention had a cost of C_N (in terms of the costs of management) and an effectiveness of E_N then the formula can be revised to $\Delta C_o / \Delta E_o = \Delta C_L / \Delta E_L$, where Δ symbolises the difference from the 'no intervention' situation
Working	Assuming an ICER of £20,000: $\Delta E = \Delta C / \text{ICER} = 181 / 20,000 = 0.01$ QALYs can be used as the target difference	Ignoring the 'no intervention' option: $E_L = C_L / (C_o / E_o) = 1190 / (1009 / 4.42) = 5.21$ QALYs if equivalent Therefore, $5.21 - 4.42 = 0.79$ QALYs can be used as the target difference Incorporating the 'no intervention' option: $\Delta E_L = \Delta C_L / (\Delta C_o / \Delta E_o) = (1190 - 500) / [(1009 - 500) / (4.42 - 2)] = 3.28$ QALYs Therefore, $3.28 - 2.42 = 0.96$ QALYs can be used as the target difference

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.^{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 Pezeshk²¹⁸ 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 colleagues²²⁰ and Willan²¹⁹ 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 Eckermann³⁸ 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's¹⁹ 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, Bacchetti and colleagues²²¹ argue that looking at the value of information is difficult and risky for investigators planning a trial because 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, $\text{cost-efficiency} = v_n/c_n$. 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 Eckermann³⁸ 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.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, triallists); 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,^{124,222–225} the views of both clinicians and patients,^{135,226–229} or the views of multidisciplinary experts (e.g. including non-clinicians such as triallists).^{41,230–236} The methods used to recruit experts also varied. In most studies, opinions were sought from a sample of

study investigators,^{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.^{134,136} Experts who were listed as members of relevant specialty organisations were recruited either through random selection²²² or by contacting all members of the list (e.g. members of the Royal College of Psychiatrists).²³⁷ Convenience sampling (e.g. from patients attending an outpatient clinic²²⁷ or those who noticed an advertisement requesting participants)^{124,224} was also used. For face-to-face meetings, participants could be recruited from among conference delegates.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 5²³⁹ to 1584,²⁴⁰ although these data were not always reported.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.²⁶ In one case the method involved a face-to-face meeting and additional elicitation although the method was not reported.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 specified⁴³ or was derived using established techniques for such processes (e.g. nominal group technique).²³⁸ Specified processes commonly involved some kind of quantitative mechanism for eliciting opinions to summarise agreement, for example through voting,²³⁰ or the compilation of individual attendees' scoring results for a set of paper patients.²⁴¹ Splitting the number of attendees into groups was used in some face-to-face meetings.^{229,238} If survey methods were used alongside face-to-face meetings, survey responses could be gathered before the meeting,²³⁸ during the meeting²³⁴ or after the meeting.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 using survey and face-to-face methods for eliciting an important difference used a Delphi process (normally two 'rounds'),^{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).^{242–245} Several studies used an 80% consensus threshold,^{246–248} or a value very close to this level ('all but two' out of nine experts).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.^{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.^{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,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).^{16,222,224,225,227,228,250} Another study using interview techniques²²⁶ 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 colleagues²⁵¹ 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.^{26,30,47} Fayers and colleagues²⁶ asked surgeons about their expectations regarding a particular treatment. Similarly, the

TABLE 4 Opinion-seeking method example: Delphi method approach for an antirheumatic drug trial²⁴²

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

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).³⁰ Latthe and colleagues⁴⁵ developed a distribution for doctors' prior beliefs about the effectiveness of a particular treatment, using a survey method for seeking opinion. Oremus and colleagues⁴⁷ 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 colleagues²⁴⁷ 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^{248,252} and Giannini and colleagues.²⁴⁶

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).²⁵³

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,^{26,30} 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 envisaged²⁴⁴ and the sample may not be representative of the general population.²²³ Face-to-face meetings may be influenced by group processes.¹³⁴ 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.²⁴² 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.⁴¹ 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).²⁶ Indeed, the specialty may influence values^{26,254} and the views of participants may change over time.²³⁹ 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.²⁵⁰ The wording of the question is important

as it needs to be clear and specific enough to prevent different participants making varying assumptions about the situation they have been presented with.²⁵⁵ Interview-based methods (telephone or face-to-face) might affect estimates¹²⁴ and questions may need to be rephrased,²³¹ 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,¹²⁴ may be considered too artificial²²⁶ and may be difficult to comprehend,²²³ 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.²²⁸

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.²⁵⁶ Similarly, a Phase II study in the regulatory drug setting could be viewed as a pilot for the Phase III study.²⁵⁷ 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 trialists 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.⁴⁴ One paper²⁵⁸ 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 colleagues²⁵⁹ 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,⁴² 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 colleagues²⁵⁷ 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^{42,44,257–259} and one used it in combination with another method²⁶⁰ (see *Combination of methods* for details). Two studies^{258,259} had conducted pilot interventions designed to reduce pain, whereas in the study by Browne,⁴² a hypothetical

example involving renal patients was provided to illustrate the method used. The sample sizes of the simulated pilot studies ranged from 5⁴² to 200²⁵⁷ and for the real pilot studies ranged from 12²⁵⁸ to 160.²⁶⁰

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.²⁵⁹ 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);²⁵⁸ 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.^{257,260} 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^{261–268} 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 colleagues²⁶⁴ 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 colleagues²⁶¹ 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 colleagues²⁶⁴). Of note, Revicki and colleagues²⁶⁷ 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,²⁶⁹ Thomas and colleagues²⁷⁰ and Woods and colleagues,²⁷¹ 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 colleagues²⁷⁰ 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.²⁷¹ The study by Johnston and colleagues²⁶⁹ 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.^{272,273} One study by Schwartz and colleagues²⁷⁴ found some evidence to dispute Norman

and colleagues¹²⁷⁵ concept of the universality of half a SD. The study by Nietzel and colleagues²⁷⁶ calculated an effect size with reference to a normative population. RCI calculations comparing treated groups with a normative population were also found.^{273,276–278}

Julious²³ 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 colleagues²⁸ 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 Lammers²⁷⁹ 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^{262,264,268–270,275} 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.^{263,264,280} 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.²⁶⁷ The formal use of meta-analysis results to determine the sample size²⁸ 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 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.²⁸¹ 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:

$$\frac{X_{group1} - X_{group2}}{\sqrt{\frac{SD_{group1}^2 + SD_{group2}^2}{2}}} \quad (7)$$

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:^{282–290}

$$\frac{X_{follow-up} - X_{baseline}}{SD_{baseline}} \quad (8)$$

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.^{291–294} The resulting measure is sometimes called the standardised response mean (SRM):⁹

$$\frac{X_{follow-up} - X_{baseline}}{SD_{change}} \quad (9)$$

A number of studies used both approaches.^{295–298} Howard and colleagues⁴³ 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 scaling²⁹³

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.^{289,299} In the simplest case (i.e. equal group sizes) it is calculated as:

$$\frac{X_{follow-up} - X_{baseline}}{\sqrt{\frac{SD_{follow-up}^2 + SD_{baseline}^2}{2}}} \quad (10)$$

In some studies, the SES was of the mean change between groups rather than within one group over time.^{289,290,300–302} A further variation on the choice of SD was described by Horton,³⁰³ who compared the difference in means between two groups, divided by the largest SD value.

The study by Andrew and Rockwood²⁹⁹ 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:

$$\frac{X_{follow-up} - X_{baseline}}{SD_{pooled} \sqrt{1-r}} \quad (11)$$

Fredrickson and colleagues³⁰⁴ 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:

$$d = 2t_c \left[\frac{1-r}{n} \right]^{1/2} \quad (12)$$

where t_c 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.^{273,305} Harris and colleagues³⁰⁵ 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

$$x_1 = \frac{\mu_{pre} - \mu_{norm}}{\sigma_{norm}} \text{ and } x_2 = \frac{\mu_{post} - \mu_{norm}}{\sigma_{norm}} \quad (13)$$

A lower value for the score used in this study indicated a better outcome. A similar method was used by Rajagopalan and colleagues.³⁰⁶ 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.²⁸⁹ 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 research³⁰⁷ 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.³⁰⁸

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,^{282,287,290,293,301,303} whereas others used the subjects' self-assessment^{292,297,302,309–311} or both^{296,299,312,313} (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.^{314–316} Similarly, Rockwood and colleagues³¹² used the combined judgement of the patient and the caregiver as well as the clinician's assessment. Basoglu and colleagues²⁹¹ 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 norms²⁷³ or people with a condition of interest with those without the condition.³⁰⁶ Some papers were interested in assessing the effect of a treatment on health or comparing the effects of different treatments.^{289,301,304,305} The study by Fredrickson and colleagues³⁰⁴ 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.^{283,300,302,317} The larger a standardised effect, the fewer the number of subjects required to detect a difference between groups or time points.²⁸³ Pyne and colleagues³⁰² 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 colleagues³⁰⁴ 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 colleagues²⁸⁷ 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 colleagues³¹² 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.^{290,299,318–322} The a priori value for clinically meaningful change varied between studies from 0.2 SDs^{290,318} to 0.5 SDs,^{299,319,322} 0.8 SDs³²¹ and both 0.4 SDs and 0.8 SDs.³²⁰ Alternative values for small, medium and large SESs (0.1, 0.5 and 1.0) have been suggested although without any identifiable usage.³²³ 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 SDs),^{82,114,324} a third of a SD (~ 0.33 SDs)^{325,326} or even 0.3 SDs.³²⁷ Occasionally, studies considered multiple levels as being of interest for interpretation of an important difference, for example both 0.5 SDs and 0.2 SDs.^{328–331} Sometimes larger values were used, and two studies^{332,333} used 1 SD as an effect size cut-off, of which the study by Hurst and Bolton³³² had also used 0.6 SDs. When studies did not cite Cohen specifically, many still interpreted their SES estimates according to his guidelines.³³⁴ 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',³³⁵ 'moderate',^{336,337} or 'small to moderate'.³³⁸

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.³⁰⁴ In addition, it is relatively easy to calculate effect sizes for published studies, provided that they have reported sufficient information.^{283,304} 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 colleagues²⁸³ 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;^{271,307,339} 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 metric²⁷¹ 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)³⁰⁸ 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.³³⁹ 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.³⁴⁰ Corresponding risk ratio values vary according to the control group event proportion.³⁰⁷ 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.²⁹⁰ Some attempts to assess the generalisability of the guidelines have been undertaken^{285,341,342} and provide some reassurance. Gordon and colleagues²⁹³ 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 colleagues³⁰⁰ and Pyne and colleagues³⁰² made use of an anchor to categorise the health or changes in health of their study subjects. Dumas and colleagues²⁸² called for further research using an external criterion to investigate clinically meaningful changes. Matza and colleagues²⁸⁴ 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 colleagues³⁰¹ 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

TABLE 5 Use of additional method(s)

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

TABLE 6 Studies utilising more than one method: method combinations

Methods							No. of studies
Anchor	Distribution	Health economic	Opinion-seeking	Pilot study	RoEB	SES	
✓	✓						70
✓						✓	46
✓			✓				8
✓					✓		1
✓	✓					✓	63
✓	✓		✓				2
✓			✓			✓	2
✓	✓		✓			✓	1
✓			✓		✓	✓	1
	✓					✓	13
	✓		✓				3
	✓		✓			✓	1
			✓			✓	1
					✓	✓	1
			✓		✓		2
				✓	✓		1

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 colleagues³⁵⁶ 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.³⁴⁶ Swigris and colleagues³⁵⁷ determined their final value from the mean of all MCID point estimates found, and de Morton and

colleagues³⁵⁸ 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.^{319,329,359} Walters and Brazier³⁶⁰ 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 colleagues³⁶¹ 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.³⁶² In another case the anchor method estimate was already known but the study verified that the value was in agreement with the distribution method.³⁶³ McLeod and colleagues³⁶⁴ 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 Wyrwich¹³⁷ and Wyrwich and colleagues^{365,366} 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.^{362,363,367-369} Some studies that had used 2.77SEM (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.³⁷⁰ 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.¹⁶⁶ Quinn and Wells³⁷¹ 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.^{161,372-374} Fewer studies used both methods to identify a point estimate of important change^{172,367,375,376} and in at least one instance this was because there was agreement in the values found using both methods.^{172,174} More often, rather than reporting point estimates, studies reported a range of values for important differences, based on all methods used.^{211,314,377,378} Other studies used anchor-based values rather than distribution-based values to infer the level of an important difference.^{175,379,380} 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.³⁸¹ Other studies also used the largest estimate, which in the case of Kocks and colleagues¹¹² was an average of two anchor values (as this method yielded higher results than the SEM), whereas Bols and colleagues³⁸² specified that the point estimate had to be larger than the smallest 'real' change value found using a distribution method. Hsieh and colleagues¹¹⁸ used the larger of two anchor estimates, which exceeded the measurement error approach estimate.

The study by Lauridsen and colleagues³⁸³ 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.¹⁸³ Two studies examined the agreement between methods. As mentioned above, Wyrwich and colleagues³⁶³ considered the similarity between SEM values and anchor values, whereas Rejas and colleagues¹⁶⁴ used formal methods (Cohen's kappa and tau-b) to assess agreement.¹⁶⁴ Both studies identified the closest agreement as being found with the anchor and 1.0SEM methods. Polson and colleagues³⁸⁴ 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.^{173,385} 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.^{386,387} A study by Bagó and colleagues³⁸⁸ 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,^{389–392} 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 colleagues²⁷⁵ 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.³⁸⁹ Kemmler and colleagues³⁸⁹ 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 colleagues³⁹⁰ 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 change³⁹³ 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.³⁹⁴

There were some notable variations across studies in the summarisation of estimates. A study by Broom and colleagues³²⁵ 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 SDs³⁹⁵ and the other excluding anchor estimates for which the SES was not within the 0.2–0.5 SD range.³⁹⁶ 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).^{326,397–399} Yount and colleagues⁴⁰⁰ reported an overall range, but had also used point estimates for the anchor results and a range for the other methods. Cole and colleagues⁴⁰¹ 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 colleagues⁴⁰² 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 colleagues⁴⁰³ calculated 95% CIs for each of the methods used and, finding overlap, used an anchor estimate arguing that the distribution method estimate supported it. Dubois and colleagues⁴⁰⁴ found a variable range of estimates across methods and suggested the use of cumulative distribution curves rather than single values. Brouwer and colleagues³²⁷ 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's²¹⁴ 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^{405–407} used Delphi techniques followed by typical anchor methods for determining an

important difference. The study by Binkley and colleagues⁴⁰⁸ 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 colleagues⁴⁰⁹ 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 colleagues⁴¹⁰ 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.⁴¹¹ Liang⁴¹² 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.⁴¹³ 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 colleagues⁴¹⁴ 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)²⁷² although determination of a final estimate of an important (and/or a realistic) difference may still be challenging. Mills and colleagues⁴¹⁵ 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.^{43,416} One study by Samsa and colleagues²⁶⁰ 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.²⁶⁰ For two studies that used four methods, the opinion-seeking method was used to confirm or synthesise the findings of the other methods.^{417,418}

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.^{323,341}

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 trialists (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 trialists 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.¹³⁸

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.³³ The use of the 'CI approach' was highlighted by one respondent⁴²³ 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.

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

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

^a Based on $n = 179$ as one respondent did not complete.

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

Method	Aware of, <i>n</i> (%)	Used, <i>n</i> (%)	Recommend, <i>n</i> (%)	Recommend if used, <i>n</i> (%)
Anchor	77 (43)	59 (33)	54 (30)	42 (71)
Distribution	104 (58)	72 (40)	60 (33)	49 (68)
Health economic	69 (38)	16 (9)	28 (16)	11 (69)
Opinion-seeking	106 (59)	72 (40)	58 (32)	40 (56)
Pilot study	162 (90)	133 (74)	117 (65)	103 (77)
RoEB	156 (87)	132 (73)	132 (73)	118 (89)
SES	138 (77)	104 (58)	73 (41)	65 (63)
Other	0 (0)	0 (0)	0 (0)	NA
None	3 (2)	6 (3)	6 (3)	NA

NA, non-applicable.

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)

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

continued

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

Characteristics	n (% of respondents)
Reproductive health	14 (42)
Respiratory	14 (42)
Skin	10 (30)
Stroke	17 (52)

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

b Some respondents represented more than one group.

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 lowest levels were for the distribution method (26%) closely followed by the health economic method (32%). 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 but 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.

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

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

NA, non-applicable.

^a Based on *n* = 33 as one CTU had only recently started.**TABLE 11** Survey 2: most recent trial (*n* = 33)^a

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

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

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

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 triallists 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^{22,424,425} or, at least, the level of use is higher for the type of RCTs that the UK and Ireland triallist 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 triallists. 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

triallists. 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.²² 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. Bayesian²⁴ or update of a meta-analysis⁴⁸).

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 groups⁴²⁶ 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 triallists regarding determination of the target difference.

The survey of UK- and Ireland-based triallists 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 triallists 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 triallist 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.⁴²⁷ 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.⁴²⁸ 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.^{20,21}

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.^{22,23,428–430} 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).²³ Alternative statistical approaches have been proposed and are briefly discussed below.^{24,37,431,432} 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.^{29,433} 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).²⁹

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

- Stand-alone definitive study.
- Superiority question evaluating evidence of a difference (in either direction).
- Neyman–Pearson framework used to calculate the sample size. This requires specification of:
 - primary outcome
 - statistical parameters (significance level and power)
 - target difference^a
 - other component(s) of the sample size calculation (e.g. common SD).

a See *Specifying the target difference* for details on specification of the target difference and other components of the sample size calculation, which vary depending on the outcome type (e.g. binary).

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,⁴⁸ 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.^{28,428} 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.²⁷ 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.^{434,435}

Very large studies designed as a single ‘final’ study are sometimes referred to as ‘mega’ or ‘large simple’ trials,^{435,436} for example the CRASH trials, which investigated the impact of treating trauma patients with corticosteroids⁴³⁷ and tranexamic acid.⁴³⁸ 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.⁴⁸ 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).⁴³⁹ 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).²⁸

Sample size calculation

General considerations

Statistical sample size calculation is not an exact science.⁴³³ 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)⁴⁴⁰ 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 level⁴⁴¹ 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.⁴²⁸ 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.^{24,428} 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 trials⁴⁴²) 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.^{24,443,444} 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.^{24,29} Health economic-based methods developed since the late 1990s tend to follow a Bayesian approach¹⁹ and sometimes also utilise a decision-theoretic framework³⁸ (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).²⁴ 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.²² 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).²³ 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.⁴²⁸ 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.⁴⁴⁵ 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;⁴⁴⁶ 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.⁴⁴⁷

In some circumstances more than one primary outcome may be used.⁴²⁸ For example, more than one outcome may be used to cover the different aspects of the purported difference.⁶ 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.⁴²⁸ 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.⁴⁴⁸ 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.⁴²⁸

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.⁴⁴⁹ 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.⁴⁵⁰

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.²⁶ A large amount of literature exists on defining a (clinically) important difference.^{2,9,15} The most

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’.¹⁵ Variants exist on this basic approach.^{16,18} 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.^{38,220} 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^{23,33,430,453} or ‘at least as large as the MCID’.⁴⁵⁴ 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.^{23,453} 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) trial⁴⁵¹

Realistic difference: based on a systematic review of RCTs,⁴⁵² 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.⁴²³ 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.^{213,455} 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 a 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.⁸⁷ 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)⁶³ or by another (e.g. health professional or parent).^{93,105} 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.⁴⁰ Using between-patient contrasts (sometimes referred to as the social comparison method),² in which patients whose disease is at varying stages are used and a judgement is made for each pair of patients, is also possible.^{80,83} 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.⁵⁹ 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).^{69,87,90} Additionally, how the judgement is used to determine the MID varies. Positive (improvement) and negative (deterioration) change may be considered together or separately.¹⁰³ 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 difference⁵⁹ although the optimal cut-point may be determined using ROC curve methods.⁴⁵⁶ Finally, an outcome (e.g. quality-of-life measure) that has been ‘anchored’ can itself be used to anchor another outcome.⁴⁵⁶ 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.⁴⁵⁷

A distinction between what is ‘important’ at an individual level and what is ‘important’ at a group level has been made by some authors.^{12,456,458} 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.⁴⁵⁹ 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.⁴⁵⁶ 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.¹⁷⁶ 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.^{197,201} 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 score¹³⁹ 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).⁴⁵⁸

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.¹⁷ 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.^{214,215} 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.^{19,38,220,460,461}

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} continuous^{242,451} 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.^{27,28,48,465} 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).^{262,265,461}

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.²⁸ 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)?⁴⁶⁶
 - 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?⁴⁶⁷
 - Is the outcome of interest fully reported? Individual patient data are seldom available and reporting of outcome is often selective.⁴⁶⁸

- 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,²⁸ 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.^{2,9} These methods have been separated out because of the widespread use of 'effect size' approach to determine the target difference of a RCT.^{23,453,469} 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.^{281,304} For a continuous outcome, the standardised difference (typically expressed as Cohen's *d* 'effect size', i.e. mean difference divided by the SD) is overwhelmingly the most commonly used although other formulae exist (e.g. Hedges' *g*).⁴⁷⁰ 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.²⁹⁵⁻²⁹⁸ 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.²⁹⁸ 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.^{304,471} A third type of effect size is to use a reference population and the impact of 'diseased' compared with 'non-diseased' populations.²⁸¹

The value of 0.2 SD has been proposed as the MCID and therefore could be used to define the target difference.² Some other support for the usefulness of Cohen's criterion exists (0.5 SD has been suggested as being a meaningful value);^{260,275,342,424} 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.⁷ 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*).⁴⁷² Cocks and colleagues³⁴¹ 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*).³⁰⁸ 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.²⁷¹ 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.³⁴⁰

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.⁴⁷³
- 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*).⁴⁷⁴ For example, pragmatic trials⁴⁷⁵ 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.^{455,476}

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).^{1,477} 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) trial⁴⁵¹]

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.⁴⁵² 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.^{23,430} 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.

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

The primary outcome is Early Treatment Diabetic Retinopathy Study (ETDRS) distance visual acuity.⁴⁷⁸ 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 study⁴⁷⁹ and one RCT.⁴⁸⁰ 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.

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 9 Protocol sample size calculation example: survival primary outcome [Arterial Revascularisation Trial (ART)]⁴⁸¹

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.⁴⁸² 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 contributed to 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

1. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, *et al.* The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;**134**:663–94.
2. Copay AG, Subach BR, Glassman SD, Polly J, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J* 2007;**7**:541–6. <http://dx.doi.org/10.1016/j.spinee.2007.01.008>
3. Hellum C, Johnsen LG, Storheim K, Nygaard I, Brox JI, Rossvoll I, *et al.* Surgery with disc prosthesis versus rehabilitation in patients with low back pain and degenerative disc: two year follow-up of randomised study. *BMJ* 2011;**342**:d2786. <http://dx.doi.org/10.1136/bmj.d2786>
4. Lois N, Burr J, Norrie J, Vale L, Cook J, McDonald A, *et al.* Internal limiting membrane peeling versus no peeling for idiopathic full-thickness macular hole: a pragmatic randomized controlled trial. *Invest Ophthalmol Vis Sci* 2011;**52**:1586–92. <http://dx.doi.org/10.1167/iops.10-6287>
5. National Institute for Health and Care Excellence. *Guide to the methods of technology*. London: NICE; 2008. URL: www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf (accessed March 2012).
6. Andrews PJ, Avenell A, Noble DW, Campbell MK, Croal BL, Simpson WG, *et al.* Randomised trial of glutamine, selenium, or both, to supplement parenteral nutrition for critically ill patients. *BMJ* 2011;**342**:d1542.
7. Lenth RV. Some practical guidelines for effective sample size determination. *Am Stat* 2001;**55**: 187–93. <http://dx.doi.org/10.1198/000313001317098149>
8. Lenth RV. 'A first course in the design of experiments: a linear models approach' by Weber & Skillins: book review. *Am Stat* 2001;**55**:370. <http://dx.doi.org/10.1198/000313001753272367>
9. Wells G, Beaton D, Shea B, Boers M, Simon L, Strand V, *et al.* Minimal clinically important differences: review of methods. *J Rheumatol* 2001;**28**:406–12.
10. Zilak ST. Matrix v. Siracusano and Student v. Fisher. *Significance* 2011;**8**:131–4.
11. Blanton H, Jaccard J. Arbitrary metrics in psychology. *Am Psychol* 2006;**61**:27–41. <http://dx.doi.org/10.1037/0003-066X.61.1.27>
12. Cella D, Bullinger M, Scott C, Barofsky I, Clinical Significance Consensus Meeting Group. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. *Mayo Clin Proc* 2002;**77**:384–92. <http://dx.doi.org/10.4065/77.4.384>
13. Fayers PM, Machin D. *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes*. Chichester: Wiley; 2007.
14. Walters SJ. *Quality of life outcomes in clinical trials and health-care evaluation: a practical guide to analysis and interpretation*. Chichester: Wiley; 2009.
15. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MICD): a literature review and directions for future research. *Curr Opin Rheumatol* 2002;**14**:109–14. <http://dx.doi.org/10.1097/00002281-200203000-00006>
16. Barrett B, Brown D, Mundt M, Brown R. Sufficiently important difference: expanding the framework of clinical significance. *Med Decis Making* 2005;**25**:250–61. <http://dx.doi.org/10.1177/0272989X05276863>

17. Briggs AH, Gray AM. Power and sample size calculations for stochastic cost-effectiveness analysis. *Med Decis Making* 1998;**18**:S81–92. <http://dx.doi.org/10.1177/0272989X9801800210>
18. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? *Pharmacoeconomics* 2000;**18**:419–23. <http://dx.doi.org/10.2165/00019053-200018050-00001>
19. O'Hagan A, Stevens JW. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Med Decis Making* 2001;**21**:219–30.
20. Chan KBY, Man-Son-Hing M, Molnar FJ, Laupacis A. How well is the clinical importance of study results reported? An assessment of randomized controlled trials. *Can Med Assoc J* 2001;**165**:1197–202.
21. Molnar FJ, Man-Son-Hing M, Fergusson D. Systematic review of measures of clinical significance employed in randomized controlled trials of drugs for dementia. *J Am Geriatr Soc* 2009;**57**:536–46. <http://dx.doi.org/10.1111/j.1532-5415.2008.02122.x>
22. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. *BMJ* 2009;**338**:b1732.
23. Julious SA. *Sample sizes for clinical trials*. Boca Raton, FL: CRC Press; 2010.
24. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. Chichester: John Wiley; 2003. <http://dx.doi.org/10.1002/0470092602>
25. Pocock SJ. *Clinical trials: a practical approach*. Chichester: Wiley; 1983.
26. Fayers PM, Cuschieri A, Fielding J, Craven J, Uscinska B, Freedman LS. Sample size calculation for clinical trials: the impact of clinician beliefs. *Br J Cancer* 2000;**82**:213–19. <http://dx.doi.org/10.1054/bjoc.1999.0902>
27. Clarke M, Hopewell S, Chalmers I. Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting. *Lancet* 2010;**376**:20–1. [http://dx.doi.org/10.1016/S0140-6736\(10\)61045-8](http://dx.doi.org/10.1016/S0140-6736(10)61045-8)
28. Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. *Stat Med* 2007;**26**:2479–500. <http://dx.doi.org/10.1002/sim.2704>
29. Berry SM, Carlin BE, Lee JJ, Muller P. *Bayesian adaptive methods for clinical trials*. London: Taylor & Francis; 2011.
30. Parmar MK, Griffiths GO, Spiegelhalter DJ, Souhami RL, Altman DG, van der Scheuren E, et al. Monitoring of large randomised clinical trials: a new approach with Bayesian methods. *Lancet* 2001;**358**:375–81. [http://dx.doi.org/10.1016/S0140-6736\(01\)05558-1](http://dx.doi.org/10.1016/S0140-6736(01)05558-1)
31. Lee SM, Ying KC. Model calibration in the continual reassessment method. *Clin Trials* 2009;**6**:227–38. <http://dx.doi.org/10.1177/1740774509105076>
32. Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 1994;**50**:337–49. <http://dx.doi.org/10.2307/2533377>
33. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of clinical trials*. New York, NY: Springer; 2010. <http://dx.doi.org/10.1007/978-1-4419-1586-3>
34. Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. *Cancer Treat Rep* 1985;**69**:1375–81.
35. Parmiginani G. *Modeling in medical decision making: a Bayesian approach*. Chichester: Wiley; 2002.
36. Lindley DV. Discussion of the paper by Spiegelhalter, Freedman & Parmar [Bayesian approaches to randomized trials]. *J R Stat Soc Ser A* 1994;**157**:393.

37. Willan AR. Power function arguments in support of an alternative approach for analyzing management trials. *Control Clin Trials* 1994;**15**:211–19. [http://dx.doi.org/10.1016/0197-2456\(94\)90058-2](http://dx.doi.org/10.1016/0197-2456(94)90058-2)
38. Willan AR, Eckermann S. Optimal clinical trial design using value of information methods with imperfect implementation. *Health Econ* 2010;**19**:549–61.
39. Cook JA, Ramsay CR, Vale LD, DELTA group. Guidance on minimally important clinical difference and trial size is needed. *Br Med J* 2012;**383**:D4375.
40. Gatchel RJ, Mayer TG. Testing minimal clinically important difference: consensus or conundrum? *Spine J* 2010;**10**:321–7. <http://dx.doi.org/10.1016/j.spinee.2009.10.015>
41. Kirkby HM, Wilson S, Calvert M, Draper H. Using e-mail recruitment and an online questionnaire to establish effect size: a worked example. *BMC Med Res Methodol* 2011;**11**(89).
42. Browne RH. On the use of a pilot sample for sample size determination. *Stat Med* 1995;**14**: 1933–40. <http://dx.doi.org/10.1002/sim.4780141709>
43. Howard R, Phillips P, Johnson T, O'Brien J, Sheehan B, Lindsay J, et al. Determining the minimum clinically important differences for outcomes in the DOMINO trial. *Int J Geriatr Psychiatry* 2011;**26**: 812–17. <http://dx.doi.org/10.1002/gps.2607>
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>
45. Latthe PM, Brauholtz DA, Hills RK, Khan KS, Lilford R. Measurement of beliefs about effectiveness of laparoscopic uterosacral nerve ablation. *BJOG* 2005;**112**:243–6. <http://dx.doi.org/10.1111/j.1471-0528.2004.00304.x>
46. Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. *Arch Gen Psychiatry* 1995;**52**:867–71. <http://dx.doi.org/10.1001/archpsyc.1995.03950220077014>
47. Oremus M, Collet JP, Corcos J, Shapiro SH. A survey of physician efficacy requirements to plan clinical trials. *Pharmacoepidemiol Drug Saf* 2002;**11**:677–85. <http://dx.doi.org/10.1002/pds.750>
48. Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. *BMC Med Res Methodol* 2009;**9**:29. <http://dx.doi.org/10.1186/1471-2288-9-29>
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>
50. Piva SR, Fitzgerald GK, Irrgang JJ, Bouzubar F, Starz TW. Get up and go test in patients with knee osteoarthritis. *Arch Phys Med Rehabil* 2004;**85**:284–9. <http://dx.doi.org/10.1016/j.apmr.2003.05.001>
51. Kropmans TJ, Dijkstra PU, Stegenga B, Stewart R, de Bont LG. Smallest detectable difference in outcome variables related to painful restriction of the temporomandibular joint. *J Dent Res* 1999;**78**:784–9. <http://dx.doi.org/10.1177/00220345990780031101>
52. Kropmans TJ, Dijkstra PU, van Veen A, Stegenga B, de Bont LG. The smallest detectable difference of mandibular function impairment in patients with a painfully restricted temporomandibular joint. *J Dent Res* 1999;**78**:1445–9. <http://dx.doi.org/10.1177/00220345990780081001>
53. Cousens SN, Rosser DA, Murdoch IE, Laidlaw DA. A simple model to predict the sensitivity to change of visual acuity measurements. *Optom Vis Sci* 2004;**81**:673–7. <http://dx.doi.org/10.1097/01.opx.0000144745.42600.76>

54. Bastyr EJ III, Price KL, Bril V, MBBQ Study Group. Development and validity testing of the neuropathy total symptom score-6: questionnaire for the study of sensory symptoms of diabetic peripheral neuropathy. *Clin Ther* 2005;**27**:1278–94. <http://dx.doi.org/10.1016/j.clinthera.2005.08.002>
55. Browne JP, van der Meulen JH, Lewsey JD, Lamping DL, Black N. Mathematical coupling may account for the association between baseline severity and minimally important difference values. *J Clin Epidemiol* 2010;**63**:865–74. <http://dx.doi.org/10.1016/j.jclinepi.2009.10.004>
56. Colangelo KJ, Pope JE, Peschken C. The minimally important difference for patient reported outcomes in systemic lupus erythematosus including the HAQ-DI, pain, fatigue, and SF-36. *J Rheumatol* 2009;**36**:2231–7. <http://dx.doi.org/10.3899/jrheum.090193>
57. Hayran O, Mumcu G, Inanc N, Ergun T, Direskeneli H. Assessment of minimal clinically important improvement by using Oral Health Impact Profile-14. *Clin Exp Rheumatol* 2009;**27**(Suppl.):84.
58. Landorf KB, Radford JA, Hudson S. Minimal important difference (MID) of two commonly used outcome measures for foot problems. *J Foot Ankle Res* 2010;**3**:7. <http://dx.doi.org/10.1186/1757-1146-3-7>
59. Yalcin I, Patrick DL, Summers K, Kinchen K, Bump RC. Minimal clinically important differences in incontinence quality-of-life scores in stress urinary incontinence. *Urology* 2006;**67**:1304–8. <http://dx.doi.org/10.1016/j.urology.2005.12.006>
60. Spiegel B, Bolus R, Harris LA, Lucak S, Naliboff B, Esrailian E, et al. Measuring irritable bowel syndrome patient-reported outcomes with an abdominal pain numeric rating scale. *Aliment Pharmacol Ther* 2009;**30**:1159–70. <http://dx.doi.org/10.1111/j.1365-2036.2009.04144.x>
61. Dommasch ED, Shin DB, Troxel AB, Margolis DJ, Gelfand JM. Reliability, validity and responsiveness to change of the Patient Report of Extent of Psoriasis Involvement (PREPI) for measuring body surface area affected by psoriasis. *Br J Dermatol* 2010;**162**:835–42. <http://dx.doi.org/10.1111/j.1365-2133.2009.09589.x>
62. John MT, Reissmann DR, Szentpétery A, Steele J. An approach to define clinical significance in prosthodontics. *J Prosthodont* 2009;**18**:455–60. <http://dx.doi.org/10.1111/j.1532-849X.2009.00457.x>
63. Tafazal SI, Sell PJ. Outcome scores in spinal surgery quantified: excellent, good, fair and poor in terms of patient-completed tools. *Eur Spine J* 2006;**15**:1653–60. <http://dx.doi.org/10.1007/s00586-005-0028-1>
64. DeRogatis LR, Graziottin A, Bitzer J, Schmitt S, Koochaki PE, Rodenberg C. Clinically relevant changes in sexual desire, satisfying sexual activity and personal distress as measured by the profile of female sexual function, sexual activity log, and personal distress scale in postmenopausal women with hypoactive sexual desire disorder. *J Sex Med* 2009;**6**:175–83. <http://dx.doi.org/10.1111/j.1743-6109.2008.01058.x>
65. Aletaha D, Funovits J, Ward MM, Smolen JS, Kvien TK. Perception of improvement in patients with rheumatoid arthritis varies with disease activity levels at baseline. *Arthritis Rheum* 2009;**61**:313–20. <http://dx.doi.org/10.1002/art.24282>
66. Kvamme MK, Kristiansen IS, Lie E, Kvien TK. Identification of cutpoints for acceptable health status and important improvement in patient-reported outcomes, in rheumatoid arthritis, psoriatic arthritis, and ankylosing spondylitis. *J Rheumatol* 2010;**37**:26–31. <http://dx.doi.org/10.3899/jrheum.090449>
67. Yamaguchi N, Poudel KC, Poudel-Tandukar K, Shakya D, Ravens-Sieberer U, Jimba M. Reliability and validity of a Nepalese version of the Kiddo-KINDL in adolescents. *Bioscience Trends* 2010;**4**:178–85.
68. Stratford PW, Binkley J, Solomon P, Gill C, Finch E. Assessing change over time in patients with low back pain. *Phys Ther* 1994;**74**:528–33.

69. Deyo RA, Inui TS. Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Serv Res* 1984;**19**:275–89.
70. Eberle E, Ottillinger B. Clinically relevant change and clinically relevant difference in knee osteoarthritis. *Osteoarthritis Cartilage* 1999;**7**:502–3. <http://dx.doi.org/10.1053/joca.1999.0246>
71. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland–Morris Back Pain Questionnaire: part 1. *Phys Ther* 1998;**78**:1186–96.
72. Vela LI, Denegar CR. The Disablement in the Physically Active Scale, part II: the psychometric properties of an outcomes scale for musculoskeletal injuries. *J Athlet Train* 2010;**45**:630–41. <http://dx.doi.org/10.4085/1062-6050-45.6.630>
73. Cappelleri JC, Bushmakina AG, McDermott AM, Dukes E, Sadosky A, Petrie CD, et al. Measurement properties of the Medical Outcomes Study Sleep Scale in patients with fibromyalgia. *Sleep Med* 2009;**10**:766–70. <http://dx.doi.org/10.1016/j.sleep.2008.09.004>
74. Colwell HH, Hunt BJ, Pasta DJ, Palo WA, Mathias SD, Joseph-Ridge N. Gout Assessment Questionnaire: initial results of reliability, validity and responsiveness. *Int J Clin Pract* 2006;**60**:1210–17. <http://dx.doi.org/10.1111/j.1742-1241.2006.01104.x>
75. Metz SM, Wyrwich KW, Babu AN, Kroenke K, Tierney WM, Wolinsky FD. A comparison of traditional and Rasch cut points for assessing clinically important change in health-related quality of life among patients with asthma. *Qual Life Res* 2006;**15**:1639–49. <http://dx.doi.org/10.1007/s11136-006-0036-6>
76. Russell IJ, Crofford LJ, Leon T, Cappelleri JC, Bushmakina AG, Whalen E, et al. The effects of pregabalin on sleep disturbance symptoms among individuals with fibromyalgia syndrome. *Sleep Med* 2009;**10**:604–10. <http://dx.doi.org/10.1016/j.sleep.2009.01.009>
77. Tilson JK, Sullivan KJ, Cen SY, Rose DK, Koradia CH, Azen SP, et al. Meaningful gait speed improvement during the first 60 days poststroke: minimal clinically important difference. *Phys Ther* 2010;**90**:196–208. <http://dx.doi.org/10.2522/ptj.20090079>
78. Harman JS, Manning WG, Lurie N, Liu CF. Interpreting results in mental health research. *Mental Health Serv Res* 2001;**3**:91–7. <http://dx.doi.org/10.1023/A:1011564918552>
79. Pouchot J, Kherani RB, Brant R, Lacaille D, Lehman AJ, Ensworth S, et al. Determination of the minimal clinically important difference for seven fatigue measures in rheumatoid arthritis. *J Clin Epidemiol* 2008;**61**:705–13. <http://dx.doi.org/10.1016/j.jclinepi.2007.08.016>
80. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. *J Clin Epidemiol* 1996;**49**:1215–19. [http://dx.doi.org/10.1016/S0895-4356\(96\)00206-5](http://dx.doi.org/10.1016/S0895-4356(96)00206-5)
81. Ringash J, Bezjak A, O'Sullivan B, Redelmeier DA. Interpreting differences in quality of life: the FACT-H&N in laryngeal cancer patients. *Qual Life Res* 2004;**13**:725–33. <http://dx.doi.org/10.1023/B:QURE.0000021703.47079.46>
82. Ringash J, O'Sullivan B, Bezjak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. *Cancer* 2007;**110**:196–202. <http://dx.doi.org/10.1002/cncr.22799>
83. Brant R, Sutherland L, Hilsden R. Examining the minimum important difference. *Stat Med* 1999;**18**:2593–603. [http://dx.doi.org/10.1002/\(SICI\)1097-0258\(19991015\)18:19<2593::AID-SIM392>3.0.CO;2-T](http://dx.doi.org/10.1002/(SICI)1097-0258(19991015)18:19<2593::AID-SIM392>3.0.CO;2-T)
84. Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. *Arch Phys Med Rehabil* 2006;**87**:32–9. <http://dx.doi.org/10.1016/j.apmr.2005.08.130>

85. Fritz JM, Hebert J, Koppenhaver S, Parent E. Beyond minimally important change: defining a successful outcome of physical therapy for patients with low back pain. *Spine* 2009;**34**:2803–9. <http://dx.doi.org/10.1097/BRS.0b013e3181ae2bd4>
86. Mintken PE, Glynn P, Cleland JA. Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain. *J Shoulder Elbow Surg* 2009;**18**:920–6. <http://dx.doi.org/10.1016/j.jse.2008.12.015>
87. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;**10**:407–15. [http://dx.doi.org/10.1016/0197-2456\(89\)90005-6](http://dx.doi.org/10.1016/0197-2456(89)90005-6)
88. Barber BL, Santanello NC, Epstein RS. Impact of the global on patient perceivable change in an asthma specific QOL questionnaire. *Qual Life Res* 1996;**5**:117–22. <http://dx.doi.org/10.1007/BF00435976>
89. Emshoff R, Emshoff I, Bertram S. Estimation of clinically important change for visual analog scales measuring chronic temporomandibular disorder pain. *J Orofac Pain* 2010;**24**:262–9.
90. Bennett RM, Bushmakina AG, Cappelleri JC, Zlateva G, Sadosky AB. Minimal clinically important difference in the fibromyalgia impact questionnaire. *J Rheumatol* 2009;**36**:1304–11. <http://dx.doi.org/10.3899/jrheum.081090>
91. Irrgang JJ, Anderson AF, Boland AL, Harner CD, Neyret P, Richmond JC, et al. Responsiveness of the International Knee Documentation Committee Subjective Knee Form. *Am J Sports Med* 2006;**34**:1567–73. <http://dx.doi.org/10.1177/0363546506288855>
92. Thomas K, Ruby J, Peter JV, Cherian AM. Comparison of disease-specific and a generic quality of life measure in patients with bronchial asthma. *Natl Med J India* 1995;**8**:258–60.
93. Brunner HI, Klein-Gitelman MS, Miller MJ, Barron A, Baldwin N, Trombley M, et al. Minimal clinically important differences of the childhood health assessment questionnaire. *J Rheumatol* 2005;**32**:150–61.
94. Fisher K. Assessing clinically meaningful change following a programme for managing chronic pain. *Clin Rehabil* 2008;**22**:252–9. <http://dx.doi.org/10.1177/0269215507081928>
95. Khanna D, Tseng CH, Furst DE, Clements PJ, Elashoff R, Roth M, et al. Minimally important differences in the Mahler's Transition Dyspnoea Index in a large randomized controlled trial – results from the Scleroderma Lung Study. *Rheumatology* 2009;**48**:1537–40. <http://dx.doi.org/10.1093/rheumatology/kep284>
96. Chiou CF, Sherbourne CD, Cornelio I, Lubeck DP, Paulus HE, Dylan M, et al. Development and validation of the revised Cedars-Sinai health-related quality of life for rheumatoid arthritis instrument. *Arthritis Rheum* 2006;**55**:856–63. <http://dx.doi.org/10.1002/art.22090>
97. Roy JS, Macdermid JC, Faber KJ, Drosdowech DS, Athwal GS. The simple shoulder test is responsive in assessing change following shoulder arthroplasty. *J Orthop Sports Phys Ther* 2010;**40**:413–21.
98. Brod M, Hammer M, Kragh N, Lessard S, Bushnell DM. Development and validation of the Treatment Related Impact Measure of Weight (TRIM-Weight). *Health Qual Life Outcomes* 2010;**8**:19. <http://dx.doi.org/10.1186/1477-7525-8-19>
99. Picado C, Badiola C, Perulero N, Sastre J, Bel JM, Pez V, et al. Validation of the Spanish version of the Asthma Control Questionnaire. *Clin Ther* 2008;**30**:1918–31. <http://dx.doi.org/10.1016/j.clinthera.2008.10.005>
100. Thienthong S, Pratheepawanit N, Limwattananon C, Maoleekoonpairaj S, Lertsanguansinchai P, Chanvej L. Pain and quality of life of cancer patients: a multi-center study in Thailand. *J Med Assoc Thai* 2006;**89**:1120–6.

101. Escobar A, Quintana JM, Bilbao A, Aróstegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. *Osteoarthritis Cartilage* 2007;**15**:273–80.
102. Singh SJ, Jones PW, Evans R, Morgan MD. Minimum clinically important improvement for the incremental shuttle walking test. *Thorax* 2008;**63**:775–7. <http://dx.doi.org/10.1136/thx.2007.081208>
103. Lee BB, King MT, Simpson JM, Haran MJ, Stockler MR, Marial O, *et al.* Validity, responsiveness, and minimal important difference for the SF-6D health utility scale in a spinal cord injured population. *Value Health* 2008;**11**:680–8. <http://dx.doi.org/10.1111/j.1524-4733.2007.00311.x>
104. Kragt JJ, Nielsen IM, van der Linden FA, Uitdehaag BM, Polman CH. How similar are commonly combined criteria for EDSS progression in multiple sclerosis? *Mult Scler* 2006;**12**:782–6. <http://dx.doi.org/10.1177/1352458506070931>
105. Riddle DL, Stratford PW, Binkley JM. Sensitivity to change of the Roland–Morris back pain questionnaire: part 2. *Phys Ther* 1998;**78**:1197–207.
106. Dempster H, Porepa M, Young N, Feldman BM. The clinical meaning of functional outcome scores in children with juvenile arthritis. *Arthritis Rheum* 2001;**44**:1768–74. [http://dx.doi.org/10.1002/1529-0131\(200108\)44:8<1768::AID-ART312>3.0.CO;2-Q](http://dx.doi.org/10.1002/1529-0131(200108)44:8<1768::AID-ART312>3.0.CO;2-Q)
107. Gong GW, Young NL, Dempster H, Porepa M, Feldman BM. The Quality of My Life questionnaire: the minimal clinically important difference for pediatric rheumatology patients. *J Rheumatol* 2007;**34**:581–7.
108. Kingsberg S, Shifren J, Wekselman K, Rodenberg C, Koochaki P, Derogatis L. Evaluation of the clinical relevance of benefits associated with transdermal testosterone treatment in postmenopausal women with hypoactive sexual desire disorder. *J Sex Med* 2007;**4**:1001–8. <http://dx.doi.org/10.1111/j.1743-6109.2007.00526.x>
109. Filocamo G, Schiappapietra B, Bertamino M, Pistorio A, Ruperto N, Magni-Manzoni S, *et al.* A new short and simple health-related quality of life measurement for paediatric rheumatic diseases: initial validation in juvenile idiopathic arthritis. *Rheumatology* 2010;**49**:1272–80. <http://dx.doi.org/10.1093/rheumatology/keq065>
110. Potter LP, Mathias SD, Raut M, Kianifard F, Tavakkol A. The OnyCOE-t questionnaire: responsiveness and clinical meaningfulness of a patient-reported outcomes questionnaire for toenail onychomycosis. *Health Qual Life Outcomes* 2006;**4**:50. <http://dx.doi.org/10.1186/1477-7525-4-50>
111. Dixon T, Lim LL, Oldridge NB. The MacNew heart disease health-related quality of life instrument: reference data for users. *Qual Life Res* 2002;**11**:173–83. <http://dx.doi.org/10.1023/A:1015005109731>
112. Kocks JW, Tuinenga MG, Uil SM, van den Berg JW, Ståhl E, van der Molen T. Health status measurement in COPD: the minimal clinically important difference of the clinical COPD questionnaire. *Respir Res* 2006;**7**:62. <http://dx.doi.org/10.1186/1465-9921-7-62>
113. van Grootel RJ, van der Glas HW. Statistically and clinically important change of pain scores in patients with myogenous temporomandibular disorders. *Eur J Pain* 2009;**13**:506–10. <http://dx.doi.org/10.1016/j.ejpain.2008.06.002>
114. Kawata AK, Revicki DA, Thakkar R, Jiang P, Krause S, Davidson MH, *et al.* Flushing ASessment Tool (FAST): psychometric properties of a new measure assessing flushing symptoms and clinical impact of niacin therapy. *Clin Drug Investig* 2009;**29**:215–29. <http://dx.doi.org/10.2165/00044011-200929040-00001>

115. Pepin V, Laviolette L, Brouillard C, Sewell L, Singh SJ, Revill SM, *et al*. Significance of changes in endurance shuttle walking performance. *Thorax* 2011;**66**:115–20. <http://dx.doi.org/10.1136/thx.2010.146159>
116. Sekhon S, Pope J, Canadian Scleroderma Research Group, Baron M. The minimally important difference in clinical practice for patient-centered outcomes including health assessment questionnaire, fatigue, pain, sleep, global visual analog scale, and SF-36 in scleroderma. *J Rheumatol* 2010;**37**:591–8. <http://dx.doi.org/10.3899/jrheum.090375>
117. Yamashita K, Ohzono K, Hiroshima K. Patient satisfaction as an outcome measure after surgical treatment for lumbar spinal stenosis: testing the validity and discriminative ability in terms of symptoms and functional status. *Spine* 2006;**31**:2602–8. <http://dx.doi.org/10.1097/01.brs.0000240717.25787.7d>
118. Hsieh YW, Wang CH, Sheu CF, Hsueh IP, Hsieh CL. Estimating the minimal clinically important difference of the Stroke Rehabilitation Assessment of Movement measure. *Neurorehabil Neural Repair* 2008;**22**:723–7. <http://dx.doi.org/10.1177/1545968308316385>
119. Cella DF, Bonomi AE, Lloyd SR, Tulsky DS, Kaplan E, Bonomi P. Reliability and validity of the Functional Assessment of Cancer Therapy-Lung (FACT-L) quality of life instrument. *Lung Cancer* 1995;**12**:199–220. [http://dx.doi.org/10.1016/0169-5002\(95\)00450-F](http://dx.doi.org/10.1016/0169-5002(95)00450-F)
120. Tannenbaum C, Brouillette J, Michaud J, Korner-Bitensky N, Dumoulin C, Corcos J, *et al*. Responsiveness and clinical utility of the geriatric self-efficacy index for urinary incontinence. *J Am Geriatr Soc* 2009;**57**:470–5. <http://dx.doi.org/10.1111/j.1532-5415.2008.02146.x>
121. 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. <http://dx.doi.org/10.1111/j.1553-2712.1998.tb02667.x>
122. Santanello NC, Zhang J, Seidenberg B, Reiss TF, Barber BL. What are minimal important changes for asthma measures in a clinical trial? *Eur Respir J* 1999;**14**:23–7. <http://dx.doi.org/10.1034/j.1399-3003.1999.14a06.x>
123. Shauver MJ, Chung KC. The minimal clinically important difference of the Michigan hand outcomes questionnaire. *J Hand Surg Am* 2009;**34**:509–14. <http://dx.doi.org/10.1016/j.jhsa.2008.11.001>
124. Barrett B, Harahan B, Brown D, Zhang Z, Brown R. Sufficiently important difference for common cold: severity reduction. *Ann Family Med* 2007;**5**:216–23. <http://dx.doi.org/10.1370/afm.698>
125. van Stel HF, Maille AR, Colland VT, Everaerd W. Interpretation of change and longitudinal validity of the quality of life for respiratory illness questionnaire (QoLRIQ) in inpatient pulmonary rehabilitation. *Qual Life Res* 2003;**12**:133–45. <http://dx.doi.org/10.1023/A:1022213223673>
126. Mannion AF, Porchet F, Lattig F, Jeszenszky D, Bartanusz V, *et al*. The quality of spine surgery from the patient's perspective: part 2. Minimal clinically important difference for improvement and deterioration as measured with the Core Outcome Measures Index. *Eur Spine J* 2009;**18**(Suppl. 3): 374–9. <http://dx.doi.org/10.1007/s00586-009-0931-y>
127. Glassman SD, Copay AG, Berven SH, Polly DW, Subach BR, Carreon LY. Defining substantial clinical benefit following lumbar spine arthrodesis. *J Bone Joint Surg Am* 2008;**90**:1839–47. <http://dx.doi.org/10.2106/JBJS.G.01095>
128. Piva SR, Gil AB, Moore CG, Fitzgerald GK. Responsiveness of the activities of daily living scale of the knee outcome survey and numeric pain rating scale in patients with patellofemoral pain. *J Rehabil Med* 2009;**41**:129–35. <http://dx.doi.org/10.2340/16501977-0295>
129. Pope JE, Khanna D, Norrie D, Ouimet JM. The minimally important difference for the health assessment questionnaire in rheumatoid arthritis clinical practice is smaller than in randomized controlled trials. *J Rheumatol* 2009;**36**:254–9. <http://dx.doi.org/10.3899/jrheum.080479>

130. Tashjian RZ, Deloach J, Green A, Porucznik CA, Powell AP. Minimal clinically important differences in ASES and simple shoulder test scores after nonoperative treatment of rotator cuff disease. *J Bone Joint Surg Am* 2010;**92**:296–303. <http://dx.doi.org/10.2106/JBJS.H.01296>
131. Puente-Maestu L, Villar F, de Miguel J, Stringer WW, Sanz P, Sanz ML, et al. Clinical relevance of constant power exercise duration changes in COPD. *Eur Respir J* 2009;**34**:340–5. <http://dx.doi.org/10.1183/09031936.00078308>
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>
133. Farrar JT, Troxel AB, Stott C, Duncombe P, Jensen MP. Validity, reliability, and clinical importance of change in a 0–10 numeric rating scale measure of spasticity: a post hoc analysis of a randomized, double-blind, placebo-controlled trial. *Clin Ther* 2008;**30**:974–85. <http://dx.doi.org/10.1016/j.clinthera.2008.05.011>
134. Wyrwich KW, Nelson HS, Tierney WM, Babu AN, Kroenke K, Wolinsky FD. Clinically important differences in health-related quality of life for patients with asthma: an expert consensus panel report. *Ann Allergy Asthma Immunol* 2003;**91**:148–53. [http://dx.doi.org/10.1016/S1081-1206\(10\)62169-2](http://dx.doi.org/10.1016/S1081-1206(10)62169-2)
135. Wyrwich KW, Fihn SD, Tierney WM, Kroenke K, Babu AN, Wolinsky FD. Clinically important changes in health-related quality of life for patients with chronic obstructive pulmonary disease: an expert consensus panel report. *J Gen Intern Med* 2003;**18**:196–202. <http://dx.doi.org/10.1046/j.1525-1497.2003.20203.x>
136. Wyrwich KW, Spertus JA, Kroenke K, Tierney WM, Babu AN, Wolinsky FD, et al. Clinically important differences in health status for patients with heart disease: an expert consensus panel report. *Am Heart J* 2004;**147**:615–22. <http://dx.doi.org/10.1016/j.ahj.2003.10.039>
137. Wyrwich 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>
138. ten Klooster PM, Drossaers-Bakker KW, Taal E, van de Laar MA. Patient-perceived satisfactory improvement (PPSI): interpreting meaningful change in pain from the patient's perspective. *Pain* 2006;**121**:151–7. <http://dx.doi.org/10.1016/j.pain.2005.12.021>
139. Sarna L, Cooley ME, Brown JK, Chernecky C, Elashoff D, Kotlerman J. Symptom severity 1 to 4 months after thoracotomy for lung cancer. *Am J Crit Care* 2008;**17**:455–67.
140. Grotle M, Brox JI, Iltstad NK. Reliability, validity and responsiveness of the fear-avoidance beliefs questionnaire: methodological aspects of the Norwegian version. *J Rehabil Med* 2006;**38**:346–53. <http://dx.doi.org/10.1080/16501970600722403>
141. Wolfe F, Michaud K, Li T. Sleep disturbance in patients with rheumatoid arthritis: evaluation by medical outcomes study and visual analog sleep scales. *J Rheumatol* 2006;**33**:1942–51.
142. Stratford PW, Binkley JM. A comparison study of the back pain functional scale and Roland Morris Questionnaire. North American Orthopaedic Rehabilitation Research Network. *J Rheumatol* 2000;**27**:1928–36.
143. Prushansky T, Handelszalts S, Pevzner E. Reproducibility of pressure pain threshold and visual analog scale findings in chronic whiplash patients. *Clin J Pain* 2007;**23**:339–45. <http://dx.doi.org/10.1097/AJP.0b013e31803157ff>
144. Weir JP. Quantifying test–retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Condition Res* 2005;**19**:231–40.

145. Ijzerman MJ, Baardman G, van Hof MA, Boom HB, Hermens HJ, Veltink PH. Validity and reproducibility of crutch force and heart rate measurements to assess energy expenditure of paraplegic gait. *Arch Phys Med Rehabil* 1999;**80**:1017–23. [http://dx.doi.org/10.1016/S0003-9993\(99\)90054-0](http://dx.doi.org/10.1016/S0003-9993(99)90054-0)
146. Ijzerman MJ, Nene AV. Feasibility of the physiological cost index as an outcome measure for the assessment of energy expenditure during walking. *Arch Phys Med Rehabil* 2002;**83**:1777–82. <http://dx.doi.org/10.1053/apmr.2002.35655>
147. Wassenberg S, Fischer-Kahle V, Herborn G, Rau R. A method to score radiographic change in psoriatic arthritis. *Z Rheumatol* 2001;**60**:156–66. <http://dx.doi.org/10.1007/s003930170064>
148. Edgar DW, Briffa NK, Cole J, Tan MH, Khoo B, Goh J, et al. Measurement of acute edema shifts in human burn survivors – the reliability and sensitivity of bioimpedance spectroscopy as an objective clinical measure. *J Burn Care Res* 2009;**30**:818–23. <http://dx.doi.org/10.1097/BCR.0b013e3181b487bc>
149. Rau R, Wassenberg S, Herborn G, Stucki G, Gebler A. A new method of scoring radiographic change in rheumatoid arthritis. *J Rheumatol* 1998;**25**:2094–107.
150. Knols RH, Stappaerts KH, Fransen J, Uebelhart D, Aufdemkampe G. Isometric strength measurement for muscle weakness in cancer patients: reproducibility of isometric muscle strength measurements with a hand-held pull-gauge dynamometer in cancer patients. *Support Care Cancer* 2002;**10**:430–8. <http://dx.doi.org/10.1007/s00520-002-0343-6>
151. Kolotkin RL, Crosby RD, Williams GR, Hartley GG, Nicol S. The relationship between health-related quality of life and weight loss. *Obes Res* 2001;**9**:564–71. <http://dx.doi.org/10.1038/oby.2001.73>
152. Geertzen JH, Dijkstra PU, Stewart RE, Groothoff JW, Ten Duis HJ, Eisma WH. Variation in measurements of range of motion: a study in reflex sympathetic dystrophy patients. *Clin Rehabil* 1998;**12**:254–64. <http://dx.doi.org/10.1191/026921598675343181>
153. Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. *Phys Ther* 1993;**73**:386–95.
154. Van Meeteren J, Roebroeck ME, Stam HJ. Test–retest reliability in isokinetic muscle strength measurements of the shoulder. *J Rehabil Med* 2002;**34**:91–5. <http://dx.doi.org/10.1080/165019702753557890>
155. van Baalen B, Odding E, van Woensel MP, Roebroeck ME. Reliability and sensitivity to change of measurement instruments used in a traumatic brain injury population. *Clin Rehabil* 2006;**20**:686–700. <http://dx.doi.org/10.1191/0269215506cre982oa>
156. Krebs EE, Bair MJ, Damush TM, Tu W, Wu J, Kroenke K. Comparative responsiveness of pain outcome measures among primary care patients with musculoskeletal pain. *Med Care* 2010;**48**:1007–14.
157. Modi AC, Zeller MH. Validation of a parent-proxy, obesity-specific quality-of-life measure: sizing them up. *Obesity* 2008;**16**:2624–33. <http://dx.doi.org/10.1038/oby.2008.416>
158. Duru G, Fantino B. The clinical relevance of changes in the Montgomery–Asberg Depression Rating Scale using the minimum clinically important difference approach. *Curr Med Res Opin* 2008;**24**:1329–35. <http://dx.doi.org/10.1185/030079908X291958>
159. Movsas B, Scott C, Watkins-Bruner D. Pretreatment factors significantly influence quality of life in cancer patients: a Radiation Therapy Oncology Group (RTOG) analysis. *Int J Radiat Oncol Biol Phys* 2006;**65**:830–5. <http://dx.doi.org/10.1016/j.ijrobp.2006.01.004>
160. Gnat R, Kuszewski M, Koczar R, Dziewonska A. Reliability of the passive knee flexion and extension tests in healthy subjects. *J Manipulative Physiol Ther* 2010;**33**:659–65. <http://dx.doi.org/10.1016/j.jmpt.2010.09.001>

161. Brunner HI, Higgins GC, Klein-Gitelman MS, Lapidus SK, Olson JC, Onel K, *et al.* Minimal clinically important differences of disease activity indices in childhood-onset systemic lupus erythematosus. *Arthritis Care Res* 2010;**62**:950–9. <http://dx.doi.org/10.1002/acr.20154>
162. Mannion AF, Junge A, Fairbank JC, Dvorak J, Grob D. Development of a German version of the Oswestry Disability Index. Part 1: cross-cultural adaptation, reliability, and validity. *Eur Spine J* 2006;**15**:55–65. <http://dx.doi.org/10.1007/s00586-004-0815-0>
163. Fritz JM, Piva SR. Physical impairment index: reliability, validity, and responsiveness in patients with acute low back pain. *Spine* 2003;**28**:1189–94. <http://dx.doi.org/10.1097/01.BRS.0000067270.50897.DB>
164. Rejas J, Pardo A, Ruiz MA. Standard error of measurement as a valid alternative to minimally important difference for evaluating the magnitude of changes in patient-reported outcomes measures. *J Clin Epidemiol* 2008;**61**:350–6. <http://dx.doi.org/10.1016/j.jclinepi.2007.05.011>
165. Fitzpatrick R, Norquist JM, Jenkinson C. Distribution-based criteria for change in health-related quality of life in Parkinson's disease. *J Clin Epidemiol* 2004;**57**:40–4. <http://dx.doi.org/10.1016/j.jclinepi.2003.07.003>
166. Gabel CP, Michener LA, Burkett B, Neller A. The Upper Limb Functional Index: development and determination of reliability, validity, and responsiveness. *J Hand Ther* 2006;**19**:328–48. <http://dx.doi.org/10.1197/j.jht.2006.04.001>
167. Lowe B, Unutzer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. *Med Care* 2004;**42**:1194–201. <http://dx.doi.org/10.1097/00005650-200412000-00006>
168. Wang SS, Normile SO, Lawshe BT. Reliability and smallest detectable change determination for serratus anterior muscle strength and endurance tests. *Physiother Theory Pract* 2006;**22**:33–42. <http://dx.doi.org/10.1080/09593980500422461>
169. Taylor R, Jayasinghe UW, Koelmeyer L, Ung O, Boyages J. Reliability and validity of arm volume measurements for assessment of lymphedema. *Phys Ther* 2006;**86**:205–14.
170. Dawson J, Doll H, Coffey J, Jenkinson C, Oxford and Birmingham Foot and Ankle Clinical Research Group. Responsiveness and minimally important change for the Manchester-Oxford foot questionnaire (MOXFQ) compared with AOFAS and SF-36 assessments following surgery for hallux valgus. *Osteoarthritis Cartilage* 2007;**15**:918–31. <http://dx.doi.org/10.1016/j.joca.2007.02.003>
171. Dawson J, Doll H, Boller I, Fitzpatrick R, Little C, Rees J, *et al.* Comparative responsiveness and minimal change for the Oxford Elbow Score following surgery. *Qual Life Res* 2008;**17**:1257–67. <http://dx.doi.org/10.1007/s11136-008-9409-3>
172. Wang YC, Hart DL, Stratford PW, Mioduski JE. Clinical interpretation of a lower-extremity functional scale-derived computerized adaptive test. *Phys Ther* 2009;**89**:957–68. <http://dx.doi.org/10.2522/ptj.20080359>
173. Las Hayas C, Quintana JM, Padierna JA, Bilbao A, Munoz P, Francis CE. Health-Related Quality of Life for Eating Disorders questionnaire version-2 was responsive 1-year after initial assessment. *J Clin Epidemiol* 2007;**60**:825–33. <http://dx.doi.org/10.1016/j.jclinepi.2006.10.004>
174. Wang YC, Hart DL, Stratford PW, Mioduski JE. Clinical interpretation of computerized adaptive test outcome measures in patients with foot/ankle impairments. *J Orthop Sports Phys Ther* 2009;**39**:753–64.
175. Hvidsten K, Carlsson M, Stecher VJ, Symonds T, Levinson I. Clinically meaningful improvement on the quality of erection questionnaire in men with erectile dysfunction. *Int J Impot Res* 2010;**22**:45–50. <http://dx.doi.org/10.1038/ijir.2009.47>

176. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991;**59**:12–19. <http://dx.doi.org/10.1037/0022-006X.59.1.12>
177. Asenlof P, Denison E, Lindberg P. Idiographic outcome analyses of the clinical significance of two interventions for patients with musculoskeletal pain. *Behav Res Ther* 2006;**44**:947–65. <http://dx.doi.org/10.1016/j.brat.2005.07.005>
178. Bowersox NW, Saunders SM, Wojcik JV. An evaluation of the utility of statistical versus clinical significance in determining improvement in alcohol and other drug (AOD) treatment in correctional settings. *Alcohol Treat Q* 2009;**27**:113–29. <http://dx.doi.org/10.1080/07347320802591700>
179. Kendall PC, Marrs-Garcia A, Nath SR, Sheldrick RC. Normative comparisons for the evaluation of clinical significance. *J Consult Clin Psychol* 1999;**67**:285–99. <http://dx.doi.org/10.1037/0022-006X.67.3.285>
180. Pekarik G, Wolff CB. Relationship of satisfaction to symptom change, follow-up adjustment, and clinical significance. *Prof Psychol* 1996;**27**:202–8. <http://dx.doi.org/10.1037/0735-7028.27.2.202>
181. Atkins DC, Bedics JD, McGlinchey JB, Beauchaine TP. Assessing clinical significance: does it matter which method we use? *J Consult Clin Psychol* 2005;**73**:982–9. <http://dx.doi.org/10.1037/0022-006X.73.5.982>
182. Choi KH, Buskey W, Johnson B. Evaluation of counseling outcomes at a university counseling center: the impact of clinically significant change on problem resolution and academic functioning. *J Counsel Psychol* 2010;**57**:297–303. <http://dx.doi.org/10.1037/a0020029>
183. Crosby RD, Kolotkin RL, Williams GR. An integrated method to determine meaningful changes in health-related quality of life. *J Clin Epidemiol* 2004;**57**:1153–60. <http://dx.doi.org/10.1016/j.jclinepi.2004.04.004>
184. Iverson GL, Sawyer DC, McCracken LM, Kozora E. Assessing depression in systemic lupus erythematosus: determining reliable change. *Lupus* 2001;**10**:266–71. <http://dx.doi.org/10.1191/096120301680416959>
185. Moleiro C, Beutler LE. Clinically significant change in psychotherapy for depressive disorders. *J Affect Disord* 2009;**115**:220–4. <http://dx.doi.org/10.1016/j.jad.2008.09.009>
186. Grundy CT, Lambert MJ, Grundy EM. Assessing clinical significance: application to the Hamilton Rating Scale for Depression. *J Ment Health* 1996;**5**:25–33. <http://dx.doi.org/10.1080/09638239650037162>
187. Tingey RC. Assessing clinical significance: extensions in method and application to the SCL-90-R. *Diss Abstract Int* 1989;**50**:1659.
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>
189. Seggar LB, Lambert MJ, Hansen NB. Assessing clinical significance: application to the Beck Depression Inventory. *Behav Ther* 2002;**33**:253–69. [http://dx.doi.org/10.1016/S0005-7894\(02\)80028-4](http://dx.doi.org/10.1016/S0005-7894(02)80028-4)
190. Ankuta GY, Abeles N. Client satisfaction, clinical significance, and meaningful change in psychotherapy. *Prof Psychol* 1993;**24**:70–4. <http://dx.doi.org/10.1037/0735-7028.24.1.70>
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)
192. Hawley DR. Assessing change with preventive interventions: the reliable change index. *Fam Relat* 1995;**44**:278–84. <http://dx.doi.org/10.2307/585526>

193. Newnham EA, Harwood KE, Page AC. Evaluating the clinical significance of responses by psychiatric inpatients to the mental health subscales of the SF-36. *J Affect Disord* 2007;**98**:91–7. <http://dx.doi.org/10.1016/j.jad.2006.07.001>
194. Mavissakalian M. Clinically significant improvement in agoraphobia research. *Behav Res Ther* 1986;**24**:369–70. [http://dx.doi.org/10.1016/0005-7967\(86\)90198-1](http://dx.doi.org/10.1016/0005-7967(86)90198-1)
195. van der Hoeven N. Calculation of the minimum significant difference at the NOEC using a non-parametric test. *Ecotoxicol Environ Saf* 2008;**70**:61–6. <http://dx.doi.org/10.1016/j.ecoenv.2007.06.010>
196. Valk GD, Grootenhuys PA, van Eijk JT, Bouter LM, Bertelsmann FW. Methods for assessing diabetic polyneuropathy: validity and reproducibility of the measurement of sensory symptom severity and nerve function tests. *Diabetes Res Clin Pract* 2000;**47**:87–95. [http://dx.doi.org/10.1016/S0168-8227\(99\)00111-4](http://dx.doi.org/10.1016/S0168-8227(99)00111-4)
197. Pijls LT, de Vries H, Donker AJ, van Eijk JT. Reproducibility and biomarker-based validity and responsiveness of a food frequency questionnaire to estimate protein intake. *Am J Epidemiol* 1999;**150**:987–95. <http://dx.doi.org/10.1093/oxfordjournals.aje.a010108>
198. Hanson ML, Sanderson H, Solomon KR. Variation, replication, and power analysis of *Myriophyllum* spp. microcosm toxicity data. *Environ Toxicol Chem* 2003;**22**:1318–29.
199. Bridges TS, Farrar JD. The influence of worm age, duration of exposure and endpoint selection on bioassay sensitivity for *Neanthes arenaceodentata* (Annelida: Polychaeta). *Environ Toxicol Chem* 1997;**16**:1650–8.
200. Anderson BS, Hunt JW, Phillips BM, Tudor S, Fairey R, Newman J, et al. Comparison of marine sediment toxicity test protocols for the amphipod *Rhepoxynius abronius* and the polychaete worm *Nereis (Neanthes) arenaceodentata*. *Environ Toxicol Chem* 1998;**17**:859–66.
201. Hoss S, Jansch S, Moser T, Junker T, Rombke J. Assessing the toxicity of contaminated soils using the nematode *Caenorhabditis elegans* as test organism. *Ecotoxicol Environ Saf* 2009;**72**:1811–18. <http://dx.doi.org/10.1016/j.ecoenv.2009.07.003>
202. Fuchsman PC, Barber TR, Sheehan PJ. Sediment toxicity evaluation for hexachlorobenzene: spiked sediment tests with *Leptocheirus plumulosus*, *Hyalella azteca*, and *Chironomus tentans*. *Arch Environ Contam Toxicol* 1998;**35**:573–9. <http://dx.doi.org/10.1007/s002449900418>
203. Kropmans T, Dijkstra P, Stegenga B, Stewart R, de Bont L. Smallest detectable difference of maximal mouth opening in patients with painfully restricted temporomandibular joint function. *Eur J Oral Sci* 2000;**108**:9–13. <http://dx.doi.org/10.1034/j.1600-0722.2000.00747.x>
204. Warren-Hicks WJ, Parkhurst BR, Moore DRJ, Teed RS, Baird RB, Berger R, et al. Assessment of whole effluent toxicity test variability: partitioning sources of variability. *Environ Toxicol Chem* 2000;**19**:94–104.
205. Burgoyne CF, Mercante DE, Thompson HW. Change detection in regional and volumetric disc parameters using longitudinal confocal scanning laser tomography. *Ophthalmology* 2002;**109**:455–66. [http://dx.doi.org/10.1016/S0161-6420\(01\)01005-3](http://dx.doi.org/10.1016/S0161-6420(01)01005-3)
206. Gully JR, Bottomley JP, Baird RB. Effects of sporophyll storage on giant kelp *Macrocystis pyrifera* (Agardh) bioassay. *Environment Toxicol Chem* 1999;**18**:1474–81.
207. Ndlovu AM, Farrell TJ, Webber CE. Coherent scattering and bone mineral measurement: the dependence of sensitivity on angle and energy. *Med Phys* 1991;**18**:985–9. <http://dx.doi.org/10.1118/1.596614>
208. Gonnelli S, Cepollaro C, Montagnani A, Martini S, Gennari L, Mangeri M, et al. Heel ultrasonography in monitoring alendronate therapy: a four-year longitudinal study. *Osteoporos Int* 2002;**13**:415–21. <http://dx.doi.org/10.1007/s001980200048>

209. Rosen HN, Moses AC, Garber J, Ross DS, Lee SL, Greenspan SL. Utility of biochemical markers of bone turnover in the follow-up of patients treated with bisphosphonates. *Calcif Tissue Int* 1998;**63**:363–8. <http://dx.doi.org/10.1007/s002239900541>
210. Lodder MC, Lems WF, Ader HJ, Marthinsen AE, van Coeverden SC, Lips P, et al. Reproducibility of bone mineral density measurement in daily practice. *Ann Rheum Dis* 2004;**63**:285–9. <http://dx.doi.org/10.1136/ard.2002.005678>
211. Abrams P, Kelleher C, Huels J, Quebe-Fehling E, Omar MA, Steel M. Clinical relevance of health-related quality of life outcomes with darifenacin. *BJU Int* 2008;**102**:208–13. <http://dx.doi.org/10.1111/j.1464-410X.2008.07523.x>
212. Patten C, Kothari D, Whitney J, Lexell J, Lum PS. Reliability and responsiveness of elbow trajectory tracking in chronic poststroke hemiparesis. *J Rehabil Res Dev* 2003;**40**:487–500. <http://dx.doi.org/10.1682/JRRD.2003.11.0487>
213. Wang D, Bakhai A. *Clinical trials: a practical guide to design, analysis, and reporting*. London: Remedica; 2006.
214. Detsky AS. Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials. *Stat Med* 1990;**9**:173–84. <http://dx.doi.org/10.1002/sim.4780090124>
215. Torgerson DJ, Ryan M, Ratcliffe J. Economics in sample size determination for clinical trials. *QJM* 1995;**88**:517–21.
216. McCormack K, Wake B, Perez J, Fraser C, Cook JA, Vale L, et al. Systematic review of the clinical effectiveness and cost-effectiveness of laparoscopic surgery for inguinal hernia repair. *Health Technol Assess* 2005;**9**(14).
217. Samsa GP, Matchar DB. Have randomized controlled trials of neuroprotective drugs been underpowered? An illustration of three statistical principles. *Stroke* 2001;**32**:669–74. <http://dx.doi.org/10.1161/01.STR.32.3.669>
218. Gittins JC, Pezeshk H. A decision theoretic approach to sample size determination in clinical trials. *J Biopharm Stat* 2002;**12**:535–51. <http://dx.doi.org/10.1081/BIP-120016234>
219. Willan AR. Optimal sample size determinations from an industry perspective based on the expected value of information. *Clin Trials* 2008;**5**:587–94. <http://dx.doi.org/10.1177/1740774508098413>
220. Kikuchi T, Pezeshk H, Gittins J. A Bayesian cost–benefit approach to the determination of sample size in clinical trials. *Stat Med* 2008;**27**:68–82. <http://dx.doi.org/10.1002/sim.2965>
221. Bacchetti P, McCulloch CE, Segal MR. Simple, defensible sample sizes based on cost efficiency. *Biometrics* 2008;**64**:577–85. http://dx.doi.org/10.1111/j.1541-0420.2008.01004_1.x
222. Aarabi M, Skinner J, Price CE, Jackson PR. Patients' acceptance of antihypertensive therapy to prevent cardiovascular disease: a comparison between South Asians and Caucasians in the United Kingdom. *Eur J Cardiovasc Prevent Rehabil* 2008;**15**:59–66. <http://dx.doi.org/10.1097/HJR.0b013e3282f07973>
223. Allison DB, Elobeid MA, Cope MB, Brock DW, Faith MS, Vander VS, et al. Sample size in obesity trials: patient perspective versus current practice. *Med Decis Making* 2010;**30**:68–75. <http://dx.doi.org/10.1177/0272989X09340583>
224. Barrett B, Brown R, Mundt M, Dye L, Alt J, Safdar N, et al. Using benefit harm tradeoffs to estimate sufficiently important difference: the case of the common cold. *Med Decis Making* 2005;**25**:47–55. <http://dx.doi.org/10.1177/0272989X04273147>

225. Wong RK, Gafni A, Whelan T, Franssen E, Fung K. Defining patient-based minimal clinically important effect sizes: a study in palliative radiotherapy for painful unresectable pelvic recurrences from rectal cancer. *Int J Radiat Oncol Biol Phys* 2002;**54**:661–9. [http://dx.doi.org/10.1016/S0360-3016\(02\)02995-4](http://dx.doi.org/10.1016/S0360-3016(02)02995-4)
226. Bloom LF, Lapierre NM, Wilson KG, Curran D, DeForge DA, Blackmer J. Concordance in goal setting between patients with multiple sclerosis and their rehabilitation team. *Am J Physical Med Rehabil* 2006;**85**:807–13. <http://dx.doi.org/10.1097/01.phm.0000237871.91829.30>
227. Bryce RL, Bradley MT, McCormick SM. To what extent would women prefer chorionic villus sampling to amniocentesis for prenatal diagnosis? *Paediatr Perinat Epidemiol* 1989;**3**:137–45. <http://dx.doi.org/10.1111/j.1365-3016.1989.tb00507.x>
228. McAlister FA, O'Connor AM, Wells G, Grover SA, Laupacis A. When should hypertension be treated? The different perspectives of Canadian family physicians and patients. *CMAJ* 2000;**163**:403–8.
229. Stone MA, Inman RD, Wright JG, Maetzel A. Validation exercise of the Ankylosing Spondylitis Assessment Study (ASAS) group response criteria in ankylosing spondylitis patients treated with biologics. *Arthritis Rheum* 2004;**51**:316–20. <http://dx.doi.org/10.1002/art.20414>
230. Bellm LA, Cunningham G, Durnell L, Eilers J, Epstein JB, Fleming T, et al. Defining clinically meaningful outcomes in the evaluation of new treatments for oral mucositis: oral mucositis patient provider advisory board. *Cancer Invest* 2002;**20**:793–800. <http://dx.doi.org/10.1081/CNV-120002497>
231. Boers M, Tugwell P. OMERACT conference questionnaire results. OMERACT Committee. *J Rheumatol* 1993;**20**:552–4.
232. Burgess P, Trauer T, Coombs T, McKay R, Pirkis J. What does 'clinical significance' mean in the context of the Health of the Nation Outcome Scales? *Aust Psychiatry* 2009;**17**:141–8. <http://dx.doi.org/10.1080/10398560802460453>
233. Mosca M, Lockshin M, Schneider M, Liang MH, Albrecht J, Aringer M, et al. Response criteria for cutaneous SLE in clinical trials. *Clin Exp Rheumatol* 2007;**25**:666–71.
234. Rider LG, Giannini EH, Harris-Love M, Joe G, Isenberg D, Pilkington C, et al. Defining clinical improvement in adult and juvenile myositis. *J Rheumatol* 2003;**30**:603–17.
235. Tubach F, Ravaud P, Beaton D, Boers M, Bombardier C, Felson DT, et al. Minimal clinically important improvement and patient acceptable symptom state for subjective outcome measures in rheumatic disorders. *J Rheumatol* 2007;**34**:1188–93.
236. Wells G, Anderson J, Boers M, Felson D, Heiberg T, Hewlett S, et al. MCID/Low Disease Activity State Workshop: summary, recommendations, and research agenda. *J Rheumatol* 2003;**30**:1115–18.
237. Brown KA. Unilateral and bilateral electroconvulsive therapy: what informs Scottish psychiatrists' choices? *Psychiatric Bull* 2009;**33**:95–98. <http://dx.doi.org/10.1192/pb.bp.107.018853>
238. Fried BJ, Boers M, Baker PR. A method for achieving consensus on rheumatoid arthritis outcome measures: the OMERACT conference process. *J Rheumatol* 1993;**20**:548–51.
239. Freedman LS, Lowe D, Macaskill P. Stopping rules for clinical trials. *Stat Med* 1983;**2**:167–74. <http://dx.doi.org/10.1002/sim.4780020210>
240. Bayle FJ, Misdrahi D, Llorca PM, Lancon C, Olivier V, Quintin P, et al. [Acute schizophrenia concept and definition: investigation of a French psychiatrist population]. *Encephale* 2005;**31**:10–17. [http://dx.doi.org/10.1016/S0013-7006\(05\)82367-X](http://dx.doi.org/10.1016/S0013-7006(05)82367-X)

241. Rantz MJ, Petroski GF, Madsen RW, Scott J, Mehr DR, Popejoy L, *et al.* Setting thresholds for MDS (minimum data set) quality indicators for nursing home quality improvement reports. *Jt Comm J Qual Improv* 1997;**23**:602–11.
242. Bellamy N, Anastassiades TP, Buchanan WW, Davis P, Lee P, McCain GA, *et al.* Rheumatoid arthritis antirheumatic drug trials. III. Setting the delta for clinical trials of antirheumatic drugs – results of a consensus development (Delphi) exercise. *J Rheumatol* 1991;**18**:1908–15.
243. Bellamy N, Buchanan WW, Esdaile JM, Fam AG, Kean WF, Thompson JM, *et al.* Ankylosing spondylitis antirheumatic drug trials. III. Setting the delta for clinical trials of antirheumatic drugs – results of a consensus development (Delphi) exercise. *J Rheumatol* 1991;**18**:1716–22.
244. Harding G, Leidy NK, Meddis D, Kleinman L, Wagner S, O'Brien CD. Interpreting clinical trial results of patient-perceived onset of effect in asthma: methods and results of a Delphi panel. *Curr Med Res Opin* 2009;**25**:1563–71. <http://dx.doi.org/10.1185/03007990902914403>
245. Bellamy N, Carrette S, Ford PM, Kean WF, le Riche NG, Lussier A, *et al.* Osteoarthritis antirheumatic drug trials. III. Setting the delta for clinical trials – results of a consensus development (Delphi) exercise. *J Rheumatol* 1992;**19**:451–7.
246. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. *Arthritis Rheum* 1997;**40**:1202–9.
247. Rider LG, Giannini EH, Brunner HI, Ruperto N, James-Newton L, Reed AM, *et al.* International consensus on preliminary definitions of improvement in adult and juvenile myositis. *Arthritis Rheum* 2004;**50**:2281–90. <http://dx.doi.org/10.1002/art.20349>
248. Ruperto N, Ravelli A, Oliveira S, Alessio M, Mihaylova D, Pasic S, *et al.* The Pediatric Rheumatology International Trials Organization/American College of Rheumatology provisional criteria for the evaluation of response to therapy in juvenile systemic lupus erythematosus: prospective validation of the definition of improvement. *Arthritis Rheum* 2006;**55**:355–63. <http://dx.doi.org/10.1002/art.22002>
249. Oliveira VC, Ferreira PH, Ferreira ML, Tiburcio L, Pinto RZ, Oliveira W, *et al.* People with low back pain who have externalised beliefs need to see greater improvements in symptoms to consider exercises worthwhile: an observational study. *Aust J Physiother* 2009;**55**:271–5. [http://dx.doi.org/10.1016/S0004-9514\(09\)70007-8](http://dx.doi.org/10.1016/S0004-9514(09)70007-8)
250. Massel D, Cruickshank M. Greater expectations in a cancer trial: absolute more than relative survival increases, community more than academic clinicians. *Cancer Invest* 2000;**18**:798–803. <http://dx.doi.org/10.3109/07357900009012213>
251. Man-Son-Hing M, Laupacis A, O'Connor A, Wells G, Lemelin J, Wood W, *et al.* Warfarin for atrial fibrillation. The patient's perspective. *Arch Intern Med* 1996;**156**:1841–8. <http://dx.doi.org/10.1001/archinte.1996.00440150095011>
252. Ruperto N, Pistorio A, Ravelli A, Rider LG, Pilkington C, Oliveira S, *et al.* The Paediatric Rheumatology International Trials Organisation provisional criteria for the evaluation of response to therapy in juvenile dermatomyositis. *Arthritis Care Res* 2010;**62**:1533–41. <http://dx.doi.org/10.1002/acr.20280>
253. Kirby S, Chuang-Stein C, Morris M. Determining a minimum clinically important difference between treatments for a patient-reported outcome. *J Biopharm Stat* 2010;**20**:1043–54. <http://dx.doi.org/10.1080/10543400903315757>
254. van Walraven C, Mahon JL, Moher D, Bohm C, Laupacis A. Surveying physicians to determine the minimal important difference: implications for sample-size calculation. *J Clin Epidemiol* 1999;**52**:717–23. [http://dx.doi.org/10.1016/S0895-4356\(99\)00050-5](http://dx.doi.org/10.1016/S0895-4356(99)00050-5)
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,

- sample size and trial duration. *Dement Geriatr Cogn Disord* 1999;**10**:534–40. <http://dx.doi.org/10.1159/000017201>
256. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. *BMC Med Res Methodol* 2010;**10**:1. <http://dx.doi.org/10.1186/1471-2288-10-1>
257. Wang SJ, Hung HM, O'Neill RT. Adapting the sample size planning of a phase III trial based on phase II data. *Pharm Stat* 2006;**5**:85–97. <http://dx.doi.org/10.1002/pst.217>
258. Johnstone R, Donaghy M, Martin D. A pilot study of a cognitive-behavioural therapy approach to physiotherapy, for acute low back pain patients, who show signs of developing chronic pain. *Adv Physiother* 2002;**4**:182–8. <http://dx.doi.org/10.1080/14038190260501622>
259. Salter GC, Roman M, Bland MJ, MacPherson H. Acupuncture for chronic neck pain: a pilot for a randomised controlled trial. *BMC Musculoskelet Disord* 2006;**7**:99. <http://dx.doi.org/10.1186/1471-2474-7-99>
260. Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J, Matchar D. Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. *Pharmacoeconomics* 1999;**15**:141–55. <http://dx.doi.org/10.2165/00019053-199915020-00003>
261. Blumenauer B. Quality of life in patients with rheumatoid arthritis: which drugs might make a difference? *Pharmacoeconomics* 2003;**21**:927–40. <http://dx.doi.org/10.2165/00019053-200321130-00002>
262. Bombardier C, Hayden J, Beaton DE. Minimal clinically important difference. Low back pain: outcome measures. *J Rheumatol* 2001;**28**:431–8.
263. Campbell JD, Gries KS, Watanabe JH, Ravelo A, Dmochowski RR, Sullivan SD. Treatment success for overactive bladder with urinary urge incontinence refractory to oral antimuscarinics: a review of published evidence. *BMC Urol* 2009;**9**:18. <http://dx.doi.org/10.1186/1471-2490-9-18>
264. Cranney A, Welch V, Wells G, Adachi J, Shea B, Simon L, et al. Discrimination of changes in osteoporosis outcomes. *J Rheumatol* 2001;**28**:413–21.
265. Feise RJ, Menke JM. Functional Rating Index: literature review. *Med Sci Monit* 2010;**16**:RA25–36.
266. Muller U, Duetz MS, Roeder C, Greenough CG. Condition-specific outcome measures for low back pain: part I: validation. *Eur Spine J* 2004;**13**:301–13. <http://dx.doi.org/10.1007/s00586-003-0665-1>
267. Revicki DA, Feeny D, Hunt TL, Cole BF. Analyzing oncology clinical trial data using the Q-TWiST method: clinical importance and sources for health state preference data. *Qual Life Res* 2006;**15**:411–23. <http://dx.doi.org/10.1007/s11136-005-1579-7>
268. Schünemann HJ, Goldstein R, Mador MJ, McKim D, Stahl E, Puhan M, et al. A randomised trial to evaluate the self-administered standardised chronic respiratory questionnaire. *Eur Respir J* 2005;**25**:31–40. <http://dx.doi.org/10.1183/09031936.04.00029704>
269. Johnston MF, Hays RD, Hui KK. Evidence-based effect size estimation: an illustration using the case of acupuncture for cancer-related fatigue. *BMC Complement Altern Med* 2009;**9**:1. <http://dx.doi.org/10.1186/1472-6882-9-1>
270. Thomas JR, Lochbaum MR, Landers DM, He C. Planning significant and meaningful research in exercise science: estimating sample size. *Res Q Exerc Sport* 1997;**68**:33–43.
271. Woods SW, Stolar M, Sernyak MJ, Charney DS. Consistency of atypical antipsychotic superiority to placebo in recent clinical trials. *Biol Psychiatry* 2001;**49**:64–70. [http://dx.doi.org/10.1016/S0006-3223\(00\)00973-2](http://dx.doi.org/10.1016/S0006-3223(00)00973-2)

272. Smith M, Wells J, Borrie M. Treatment effect size of memantine therapy in Alzheimer disease and vascular dementia. *Alzheimer Dis Assoc Disord* 2006;**20**:133–7. <http://dx.doi.org/10.1097/00002093-200607000-00002>
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>
274. Schwartz CE, Bode R, Repucci N, Becker J, Sprangers MAG, Fayers PM. The clinical significance of adaptation to changing health: a meta-analysis of response shift. *Qual Life Res* 2006;**15**:1533–50. <http://dx.doi.org/10.1007/s11136-006-0025-9>
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>
276. Nietzel MT, Russell RL, Hemmings KA, Gretter ML. Clinical significance of psychotherapy for unipolar depression: a meta-analytic approach to social comparison. *J Consult Clin Psychol* 1987;**55**:156–61. <http://dx.doi.org/10.1037/0022-006X.55.2.156>
277. Fisher PL. The efficacy of psychological treatments for generalised anxiety disorder? In Davey GCL, Wells A, editors. *Worry and its psychological disorders: theory, assessment and treatment*. Chichester: John Wiley; 2006. pp. 359–77. <http://dx.doi.org/10.1002/9780470713143.ch20>
278. Sheldrick RC, Kendall PC, Heimberg RG. The clinical significance of treatments: a comparison of three treatments for conduct disordered children. *Clin Psychol* 2001;**8**:418–30.
279. Zanen P, Lammers JW. Sample sizes for comparative inhaled corticosteroid trials with emphasis on showing therapeutic equivalence. *Eur J Clin Pharmacol* 1995;**48**:179–84. <http://dx.doi.org/10.1007/BF00198295>
280. Barbui C, Violante A, Garattini S. Does placebo help establish equivalence in trials of new antidepressants? *Eur Psychiatry* 2000;**15**:268–73. [http://dx.doi.org/10.1016/S0924-9338\(00\)00233-9](http://dx.doi.org/10.1016/S0924-9338(00)00233-9)
281. Huberty CJ. A history of effect size indices. *Educ Psychol Measure* 2002;**62**:227. <http://dx.doi.org/10.1177/0013164402062002002>
282. Dumas HM, Haley SM, Bedell GM, Hull EM. Social function changes in children and adolescents with acquired brain injury during inpatient rehabilitation. *Pediatr Rehabil* 2001;**4**:177–85.
283. Haymes SA, Johnston AW, Heyes AD. Preliminary investigation of the responsiveness of the Melbourne Low Vision ADL index to low-vision rehabilitation. *Optom Vis Sci* 2001;**78**:373–80. <http://dx.doi.org/10.1097/00006324-200106000-00008>
284. Matza LS, Johnston JA, Faries DE, Malley KG, Brod M. Responsiveness of the Adult Attention-Deficit/Hyperactivity Disorder Quality of Life Scale (AAQoL). *Qual Life Res* 2007;**16**:1511–20. <http://dx.doi.org/10.1007/s11136-007-9254-9>
285. Norman GR. The relation between the minimally important difference and patient benefit. *COPD* 2005;**2**:69–73. <http://dx.doi.org/10.1081/COPD-200051249>
286. van de Port IG, Ketelaar M, Schepers VP, Van den Bos GA, Lindeman E. Monitoring the functional health status of stroke patients: the value of the Stroke-Adapted Sickness Impact Profile-30. *Disabil Rehabil* 2004;**26**:635–40. <http://dx.doi.org/10.1080/09638280410001672481>
287. van der Putten JJ, Hobart JC, Freeman JA, Thompson AJ. Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel Index and the Functional Independence Measure. *J Neurol Neurosurg Psychiatry* 1999;**66**:480–4. <http://dx.doi.org/10.1136/jnnp.66.4.480>

288. Dawson J, Fitzpatrick R, Carr A. The assessment of shoulder instability. The development and validation of a questionnaire. *J Bone Joint Surg Br* 1999;**81**:420–6.
289. Krakow B, Melendrez D, Sisley B, Warner TD, Krakow J, Leahigh L, et al. Nasal dilator strip therapy for chronic sleep-maintenance insomnia and symptoms of sleep-disordered breathing: a randomized controlled trial. *Sleep Breath* 2006;**10**:16–28. <http://dx.doi.org/10.1007/s11325-005-0037-7>
290. Rockwood K, Stolee P. Responsiveness of outcome measures used in an antidementia drug trial. *Alzheimer Dis Assoc Disord* 2000;**14**:182–5. <http://dx.doi.org/10.1097/00002093-200007000-00010>
291. Basoglu M, Livanou M, Salcioglu E, Kalender D. A brief behavioural treatment of chronic post-traumatic stress disorder in earthquake survivors: results from an open clinical trial. *Psychol Med* 2003;**33**:647–54. <http://dx.doi.org/10.1017/S0033291703007360>
292. Cramer JA, Cuffel BJ, Divan V, Al-Sabbagh A, Glassman M. Patient satisfaction with an injection device for multiple sclerosis treatment. *Acta Neurol Scand* 2006;**113**:156–62. <http://dx.doi.org/10.1111/j.1600-0404.2005.00568.x>
293. Gordon JE, Powell C, Rockwood K. Goal attainment scaling as a measure of clinically important change in nursing-home patients. *Age Ageing* 1999;**28**:275–81. <http://dx.doi.org/10.1093/ageing/28.3.275>
294. Myles PS, Hunt JO, Fletcher H, Solly R, Woodward D, Kelly S. Relation between quality of recovery in hospital and quality of life at 3 months after cardiac surgery. *Anesthesiology* 2001;**95**:862–7. <http://dx.doi.org/10.1097/00000542-200110000-00013>
295. Merkies IS, Schmitz PI, van der Meché, Samijn JP, van Doorn PA, Inflammatory Neuropathy Cause and Treatment (INCAT) group. Quality of life complements traditional outcome measures in immune-mediated polyneuropathies. *Neurology* 2002;**59**:84–91. <http://dx.doi.org/10.1212/WNL.59.1.84>
296. Nilsson AK, Roos EM, Westerlund JP, Roos HP, Lohmander LS. Comparative responsiveness of measures of pain and function after total hip replacement. *Arthritis Rheum* 2001;**45**:258–62. [http://dx.doi.org/10.1002/1529-0131\(200106\)45:3<258::AID-ART258>3.0.CO;2-L](http://dx.doi.org/10.1002/1529-0131(200106)45:3<258::AID-ART258>3.0.CO;2-L)
297. Tuzun EH, Eker L, Aytar A, Daskapan A, Bayramoglu M. Acceptability, reliability, validity and responsiveness of the Turkish version of WOMAC osteoarthritis index. *Osteoarthritis Cartilage* 2005;**13**:28–33. <http://dx.doi.org/10.1016/j.joca.2004.10.010>
298. van Tubergen A, Landewe R, Heuft-Dorenbosch L, Spoorenberg A, Van Der Heijde D, van der Tempel H, et al. Assessment of disability with the World Health Organization Disability Assessment Schedule II in patients with ankylosing spondylitis. *Ann Rheum Dis* 2003;**62**:140–5. <http://dx.doi.org/10.1136/ard.62.2.140>
299. Andrew MK, Rockwood K. A five-point change in Modified Mini-Mental State Examination was clinically meaningful in community-dwelling elderly people. *J Clin Epidemiol* 2008;**61**:827–31. <http://dx.doi.org/10.1016/j.jclinepi.2007.10.022>
300. Cheung YB, Goh C, Thumboo J, Khoo KS, Wee J. Variability and sample size requirements of quality-of-life measures: a randomized study of three major questionnaires. *J Clin Oncol* 2005;**23**:4936–44. <http://dx.doi.org/10.1200/JCO.2005.07.141>
301. Konst EM, Pahl C, Weersink-Braks H, De Boo T, Pahl-Andersen B, Kuipers-Jagtman AM, et al. Cost-effectiveness of infant orthopedic treatment regarding speech in patients with complete unilateral cleft lip and palate: a randomized three-center trial in the Netherlands (Dutchcleft). *Cleft Palate Craniofac J* 2004;**41**:71–7. <http://dx.doi.org/10.1597/02-069>

302. Pyne JM, Sullivan G, Kaplan R, Williams DK. Comparing the sensitivity of generic effectiveness measures with symptom improvement in persons with schizophrenia. *Med Care* 2003;**41**:208–17. <http://dx.doi.org/10.1097/01.MLR.0000044900.72470.D4>
303. Horton AM. Estimation of clinical significance: a brief note. *Psychol Rep* 1980;**47**:141–2. <http://dx.doi.org/10.2466/pr0.1980.47.1.141>
304. Fredrickson A, Snyder PJ, Cromer J, Thomas E, Lewis M, Maruff P. The use of effect sizes to characterize the nature of cognitive change in psychopharmacological studies: an example with scopolamine. *Hum Psychopharmacol* 2008;**23**:425–36. <http://dx.doi.org/10.1002/hup.942>
305. Harris MA, Greco P, Wysocki T, White NH. Family therapy with adolescents with diabetes: a litmus test for clinically meaningful change. *Fam Syst Health* 2001;**19**:159–68. <http://dx.doi.org/10.1037/h0089445>
306. Rajagopalan R, Laitinen D, Dietz B. Impact of lipoatrophy on quality of life in HIV patients receiving anti-retroviral therapy. *AIDS Care* 2008;**20**:1197–201. <http://dx.doi.org/10.1080/09540120801926993>
307. Higgins JPT, Greene S. *Cochrane handbook for systematic reviews of interventions version 5.1.0*. The Cochrane Collaboration; 2011. URL: www.cochrane-handbook.org/ (accessed June 2012).
308. Hackshaw AK. *A concise guide to clinical trials*. Oxford: Wiley-Blackwell; 2009. <http://dx.doi.org/10.1002/9781444311723>
309. Burton HJ, Kline SA, Cooper BS, Rabinowitz A, Dodek A. Assessing risk for major depression on patients selected for percutaneous transluminal coronary angioplasty: is it a worthwhile venture? *Gen Hosp Psychiatry* 2003;**25**:200–8. [http://dx.doi.org/10.1016/S0163-8343\(03\)00016-1](http://dx.doi.org/10.1016/S0163-8343(03)00016-1)
310. McKee G. Are there meaningful longitudinal changes in health related quality of life–SF36, in cardiac rehabilitation patients? *Eur J Cardiovasc Nursing* 2009;**8**:40–7. <http://dx.doi.org/10.1016/j.ejcnurse.2008.04.004>
311. Lai JS, Cella D, Kupst MJ, Holm S, Kelly ME, Bode RK, et al. Measuring fatigue for children with cancer: development and validation of the pediatric Functional Assessment of Chronic Illness Therapy–Fatigue (pedsFACIT-F). *J Pediatr Hematol Oncol* 2007;**29**:471–9. <http://dx.doi.org/10.1097/MPH.0b013e318095057a>
312. Rockwood K, Fay S, Song X, MacKnight C, Gorman M, Video-Imaging Synthesis of Treating Alzheimer’s Disease (VISTA) investigators. Attainment of treatment goals by people with Alzheimer’s disease receiving galantamine: a randomized controlled trial. *CMAJ* 2006;**174**:1099–105. <http://dx.doi.org/10.1503/cmaj.051432>
313. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg Br* 1996;**78**:185–90.
314. Rentz AM, Matza LS, Secnik K, Swensen A, Revicki DA. Psychometric validation of the child health questionnaire (CHQ) in a sample of children and adolescents with attention-deficit/hyperactivity disorder. *Qual Life Res* 2005;**14**:719–34. <http://dx.doi.org/10.1007/s11136-004-0832-9>
315. Morris C, Doll H, Davies N, Wainwright A, Theologis T, Willett K, et al. The Oxford Ankle Foot Questionnaire for children: responsiveness and longitudinal validity. *Qual Life Res* 2009;**18**:1367–76. <http://dx.doi.org/10.1007/s11136-009-9550-7>
316. Oeffinger D, Bagley A, Rogers S, Gorton G, Kryscio R, Abel M, et al. Outcome tools used for ambulatory children with cerebral palsy: responsiveness and minimum clinically important differences. *Dev Med Child Neurol* 2008;**50**:918–25. <http://dx.doi.org/10.1111/j.1469-8749.2008.03150.x>

317. Bolton JE, Breen AC. The Bournemouth Questionnaire: a short-form comprehensive outcome measure. I. Psychometric properties in back pain patients. *J Manipulative Physiol Ther* 1999;**22**: 503–10. [http://dx.doi.org/10.1016/S0161-4754\(99\)70001-1](http://dx.doi.org/10.1016/S0161-4754(99)70001-1)
318. Drinkwater EJ, Pritchett EJ, Behm DG. Effect of instability and resistance on unintentional squat-lifting kinetics. *Int J Sports Physiol Perform* 2007;**2**:400–13.
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>
320. Gompertz P, Pound P, Ebrahim S. Validity of the extended activities of daily living scale. *Clin Rehabil* 1994;**8**:275–80. <http://dx.doi.org/10.1177/026921559400800401>
321. Harrill WC, Pillsbury HC III, McGuirt WF, Stewart MG. Radiofrequency turbinate reduction: a NOSE evaluation. *Laryngoscope* 2007;**117**:1912–19. <http://dx.doi.org/10.1097/MLG.0b013e3181271414>
322. Spiegel B, Camilleri M, Bolus R, Andresen V, Chey WD, Fehnel S, et al. Psychometric evaluation of patient-reported outcomes in irritable bowel syndrome randomized controlled trials: a Rome Foundation report. *Gastroenterology* 2009;**137**:1944–53. <http://dx.doi.org/10.1053/j.gastro.2009.08.047>
323. Machin D, Day S, Greene S, editors. *Textbook of clinical trials*. Chichester: John Wiley; 2006. <http://dx.doi.org/10.1002/9780470010167>
324. Ries AL. Minimally clinically important difference for the UCSD Shortness of Breath Questionnaire, Borg Scale, and Visual Analog Scale. *COPD* 2005;**2**:105–10. <http://dx.doi.org/10.1081/COPD-200050655>
325. Broom R, Du H, Clemons M, Eton D, Dranitsaris G, Simmons C, et al. Switching breast cancer patients with progressive bone metastases to third-generation bisphosphonates: measuring impact using the Functional Assessment of Cancer Therapy-Bone Pain. *J Pain Symptom Manage* 2009;**38**:244–57. <http://dx.doi.org/10.1016/j.jpainsymman.2008.08.005>
326. McNair PJ, Prapavessis H, Collier J, Bassett S, Bryant A, Larmer P. The lower-limb tasks questionnaire: an assessment of validity, reliability, responsiveness, and minimal important differences. *Arch Phys Med Rehabil* 2007;**88**:993–1001. <http://dx.doi.org/10.1016/j.apmr.2007.05.008>
327. Brouwer CN, Schilder AG, van Stel HF, Rovers MM, Veenhoven RH, Grobbee DE, et al. Reliability and validity of functional health status and health-related quality of life questionnaires in children with recurrent acute otitis media. *Qual Life Res* 2007;**16**:1357–73. <http://dx.doi.org/10.1007/s11136-007-9242-0>
328. Brach JS, Perera S, Studenski S, Katz M, Hall C, Verghese J. Meaningful change in measures of gait variability in older adults. *Gait Posture* 2010;**31**:175–9. <http://dx.doi.org/10.1016/j.gaitpost.2009.10.002>
329. Kelleher CJ, Pleil AM, Reese PR, Burgess SM, Brodish PH. How much is enough and who says so? *BJOG* 2004;**111**:605–12. <http://dx.doi.org/10.1111/j.1471-0528.2004.00129.x>
330. Kvam AK, Wisløff F, Fayers PM. Minimal important differences and response shift in health-related quality of life; a longitudinal study in patients with multiple myeloma. *Health Qual Life Outcomes* 2010;**8**:79. <http://dx.doi.org/10.1186/1477-7525-8-79>
331. Nieves JW, Li T, Zion M, Gussekloo J, Pahor M, Bernabei R, et al. The clinically meaningful change in physical performance scores in an elderly cohort. *Aging Clin Exp Res* 2007;**19**:484–91.
332. Hurst H, Bolton J. Assessing the clinical significance of change scores recorded on subjective outcome measures. *J Manipulative Physiol Ther* 2004;**27**:26–35. <http://dx.doi.org/10.1016/j.jmpt.2003.11.003>

333. Wright P, Marshall L, Smith AB, Velikova G, Selby P. Measurement and interpretation of social distress using the social difficulties inventory (SDI). *Eur J Cancer* 2008;**44**:1529–35. <http://dx.doi.org/10.1016/j.ejca.2008.04.011>
334. Fairchild CJ, Chalmers RL, Begley CG. Clinically important difference in dry eye: change in IDEEL-symptom bother. *Optom Vis Sci* 2008;**85**:699–707. <http://dx.doi.org/10.1097/OPX.0b013e3181824e0d>
335. Karsten J, Hartman CA, Ormel J, Nolen WA, Penninx BWJH. Subthreshold depression based on functional impairment better defined by symptom severity than by number of DSM-IV symptoms. *J Affect Disord* 2010;**123**:230–7. <http://dx.doi.org/10.1016/j.jad.2009.10.013>
336. Locker D, Jokovic A, Clarke M. Assessing the responsiveness of measures of oral health-related quality of life. *Community Dent Oral Epidemiol* 2004;**32**:10–18. <http://dx.doi.org/10.1111/j.1600-0528.2004.00114.x>
337. Malden PE, Thomson WM, Jokovic A, Locker D. Changes in parent-assessed oral health-related quality of life among young children following dental treatment under general anaesthetic. *Community Dent Oral Epidemiol* 2008;**36**:108–17. <http://dx.doi.org/10.1111/j.1600-0528.2007.00374.x>
338. Wiebe S, Matijevic S, Eliasziw M, Derry PA. Clinically important change in quality of life in epilepsy. *J Neurol Neurosurg Psychiatry* 2002;**73**:116–20. <http://dx.doi.org/10.1136/jnnp.73.2.116>
339. Cohen J. *Statistical power: analysis of behavioural sciences*. New York, NY: Academic Press; 1977.
340. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat Med* 2000;**19**:3127–31. [http://dx.doi.org/10.1002/1097-0258\(20001130\)19:22<3127::AID-SIM784>3.0.CO;2-M](http://dx.doi.org/10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M)
341. Cocks K, King MT, Velikova G, Martyn St-James M, Fayers PM, Brown JM. Evidence-based guidelines for determination of sample size and interpretation of the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. *J Clin Oncol* 2011;**29**:89–96. <http://dx.doi.org/10.1200/JCO.2010.28.0107>
342. King MT, Stockler MR, Cella DF, Osoba D, Eton DT, Thompson J, et al. Meta-analysis provides evidence-based effect sizes for a cancer-specific quality-of-life questionnaire, the FACT-G. *J Clin Epidemiol* 2010;**63**:270–81. <http://dx.doi.org/10.1016/j.jclinepi.2009.05.001>
343. Rossi MD, Eberle T, Roche M, Waggoner M, Blake R, Burwell B, et al. Delaying knee replacement and implications on early postoperative outcomes: a pilot study. *Orthopedics* 2009;**32**:885–93. <http://dx.doi.org/10.3928/01477447-20091020-06>
344. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. *J Clin Epidemiol* 1995;**48**:1369–78. [http://dx.doi.org/10.1016/0895-4356\(95\)00054-2](http://dx.doi.org/10.1016/0895-4356(95)00054-2)
345. Wyrwich K, Harnam N, Revicki DA, Locklear JC, Svedsäter H, Endicott J. Assessing health-related quality of life in generalized anxiety disorder using the Quality Of Life Enjoyment and Satisfaction Questionnaire. *Int Clin Psychopharmacol* 2009;**24**:289–95. <http://dx.doi.org/10.1097/YIC.0b013e32832d6bf4>
346. Funk GF, Karnell LH, Smith RB, Christensen AJ. Clinical significance of health status assessment measures in head and neck cancer: what do quality-of-life scores mean? *Arch Otolaryngol Head Neck Surg* 2004;**130**:825–9. <http://dx.doi.org/10.1001/archotol.130.7.825>
347. Robinson D Jr, Zhao N, Gathany T, Kim LL, Cella D, Revicki D. Health perceptions and clinical characteristics of relapsing-remitting multiple sclerosis patients: baseline data from an international clinical trial. *Curr Med Res Opin* 2009;**25**:1121–30. <http://dx.doi.org/10.1185/03007990902797675>

348. Drossman D, Morris CB, Hu Y, Toner BB, Diamant N, Whitehead WE, *et al.* Characterization of health related quality of life (HRQOL) for patients with functional bowel disorder (FBD) and its response to treatment. *Am J Gastroenterol* 2007;**102**:1442–53. <http://dx.doi.org/10.1111/j.1572-0241.2007.01283.x>
349. Gold SM, Schulz H, Stein H, Solf K, Schulz KH, Heesen C. Responsiveness of patient-based and external rating scales in multiple sclerosis: head-to-head comparison in three clinical settings. *J Neurol Sci* 2010;**290**:102–6. <http://dx.doi.org/10.1016/j.jns.2009.10.020>
350. Puhan MA, Frey M, Büchi S, Schünemann HJ. The minimal important difference of the hospital anxiety and depression scale in patients with chronic obstructive pulmonary disease. *Health Qual Life Outcomes* 2008;**6**:46. <http://dx.doi.org/10.1186/1477-7525-6-46>
351. Terwee CB, Dekker FW, Mourits MP, Gerding MN, Baldeschi L, Kalmann R, *et al.* Interpretation and validity of changes in scores on the Graves' ophthalmopathy quality of life questionnaire (GO-QOL) after different treatments. *Clin Endocrinol (Oxf)* 2001;**54**:391–8. <http://dx.doi.org/10.1046/j.1365-2265.2001.01241.x>
352. Lasch K, Joish VN, Zhu Y, Rosa K, Qiu C, Crawford B. Validation of the sleep impact scale in patients with major depressive disorder and insomnia. *Curr Med Res Opin* 2009;**25**:1699–710. <http://dx.doi.org/10.1185/03007990902973201>
353. Arbuckle RA, Humphrey L, Vardeva K, Arondekar B, Danten-Viala M, Scott JA, *et al.* Psychometric evaluation of the Diabetes Symptom Checklist-Revised (DSC-R) – a measure of symptom distress. *Value Health* 2009;**12**:1168–75. <http://dx.doi.org/10.1111/j.1524-4733.2009.00571.x>
354. Barnes ML, Vaidyanathan S, Williamson PA, Lipworth BJ. The minimal clinically important difference in allergic rhinitis. *Clin Exp Allergy* 2010;**40**:242–50. <http://dx.doi.org/10.1111/j.1365-2222.2009.03381.x>
355. Miskala PH, Hawkins BS, Mangione CM, Bass EB, Bressler NM, Dong LM, *et al.* Responsiveness of the National Eye Institute Visual Function Questionnaire to changes in visual acuity: findings in patients with subfoveal choroidal neovascularization – SST Report No. 1. *Arch Ophthalmol* 2003;**121**:531–9. <http://dx.doi.org/10.1001/archophth.121.4.531>
356. Middel B, Stewart R, Bouma J, van Sonderen E, van den Heuvel WJ. How to validate clinically important change in health-related functional status. Is the magnitude of the effect size consistently related to magnitude of change as indicated by a global question rating? *J Eval Clin Pract* 2001;**7**:399–410. <http://dx.doi.org/10.1046/j.1365-2753.2001.00298.x>
357. Swigris JJ, Wamboldt FS, Behr J, Du Bois RM, King TE, Raghu G, *et al.* The 6 minute walk in idiopathic pulmonary fibrosis: longitudinal changes and minimum important difference. *Thorax* 2010;**65**:173–7. <http://dx.doi.org/10.1136/thx.2009.113498>
358. de Morton NA, Davidson M, Keating JL. The de Morton Mobility Index (DEMMI): an essential health index for an ageing world. *Health Qual Life Outcomes* 2008;**6**:63. <http://dx.doi.org/10.1186/1477-7525-6-63>
359. Laviolette L, Bourbeau J, Bernard S, Lacasse Y, Pepin V, Breton MJ, *et al.* Assessing the impact of pulmonary rehabilitation on functional status in COPD. *Thorax* 2008;**63**:115–21. <http://dx.doi.org/10.1136/thx.2006.076844>
360. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health Qual Life Outcomes* 2003;**1**:4. <http://dx.doi.org/10.1186/1477-7525-1-4>
361. Shulman LM, Gruber-Baldini AL, Anderson KE, Fishman PS, Reich SG, Weiner WJ. The clinically important difference on the unified Parkinson's disease rating scale. *Arch Neurol* 2010;**67**:64–70. <http://dx.doi.org/10.1001/archneurol.2009.295>

362. Wolfe F, Michaud K. Assessment of pain in rheumatoid arthritis: minimal clinically significant difference, predictors, and the effect of anti-tumor necrosis factor therapy. *J Rheumatol* 2007;**34**:1674–83.
363. Wyrwich KW, Tierney WM, Wolinsky FD. Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. *Qual Life Res* 2002;**11**:1–7. <http://dx.doi.org/10.1023/A:1014485627744>
364. McLeod LD, Fehnel SE, Brandman J, Symonds T. Evaluating minimal clinically important differences for the acne-specific quality of life questionnaire. *Pharmacoeconomics* 2003;**21**:1069–79. <http://dx.doi.org/10.2165/00019053-200321150-00001>
365. Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Med Care* 1999;**37**:469–78. <http://dx.doi.org/10.1097/00005650-199905000-00006>
366. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol* 1999;**52**:861–73. [http://dx.doi.org/10.1016/S0895-4356\(99\)00071-2](http://dx.doi.org/10.1016/S0895-4356(99)00071-2)
367. Kupferberg DH, Kaplan RM, Slymen DJ, Ries AL. Minimal clinically important difference for the UCSD Shortness of Breath Questionnaire. *J Cardiopulm Rehabil* 2005;**25**:370–7. <http://dx.doi.org/10.1097/00008483-200511000-00011>
368. Stull DE, Vernon MK, Canonica GW, Crespi S, Sandor D. Using the Congestion Quantifier Seven-Item Test to assess change in patient symptoms and their impact. *Allergy Asthma Proc* 2008;**29**:295–303. <http://dx.doi.org/10.2500/aap.2008.29.3119>
369. Tsai CL, Hodder RV, Page JH, Cydulka RK, Rowe BH, Camargo CA Jr. The short-form chronic respiratory disease questionnaire was a valid, reliable, and responsive quality-of-life instrument in acute exacerbations of chronic obstructive pulmonary disease. *J Clin Epidemiol* 2008;**61**:489–97. <http://dx.doi.org/10.1016/j.jclinepi.2007.07.003>
370. Vos CJ, Verhagen AP, Koes BW. Reliability and responsiveness of the Dutch version of the Neck Disability Index in patients with acute neck pain in general practice. *Eur Spine J* 2006;**15**:1729–36. <http://dx.doi.org/10.1007/s00586-006-0119-7>
371. Quinn JV, Wells GA. An assessment of clinical wound evaluation scales. *Acad Emerg Med* 1998;**5**:583–6. <http://dx.doi.org/10.1111/j.1553-2712.1998.tb02465.x>
372. Terwee CB, Roorda LD, Knol DL, De Boer MR, de Vet HC. Linking measurement error to minimal important change of patient-reported outcomes. *J Clin Epidemiol* 2009;**62**:1062–7. <http://dx.doi.org/10.1016/j.jclinepi.2008.10.011>
373. Moser JS, Barker KL, Doll HA, Carr AJ. Comparison of two patient-based outcome measures for shoulder instability after nonoperative treatment. *J Shoulder Elbow Surg* 2008;**17**:886–92. <http://dx.doi.org/10.1016/j.jse.2008.05.040>
374. Martin RL, Irrgang JJ, Burdett RG, Conti SF, Van Swearingen JM. Evidence of validity for the Foot and Ankle Ability Measure (FAAM). *Foot Ankle Int* 2005;**26**:968–83.
375. Childs JD, Piva SR. Psychometric properties of the functional rating index in patients with low back pain. *Eur Spine J* 2005;**14**:1008–12. <http://dx.doi.org/10.1007/s00586-005-0900-z>
376. Ekeberg OM, Bautz-Holter E, Keller A, Tveitå EK, Juel NG, et al. A questionnaire found disease-specific WORC index is not more responsive than SPADI and OSS in rotator cuff disease. *J Clin Epidemiol* 2010;**63**:575–84. <http://dx.doi.org/10.1016/j.jclinepi.2009.07.012>
377. Holland AE, Hill CJ, Conron M, Munro P, McDonald CF. Small changes in six-minute walk distance are important in diffuse parenchymal lung disease. *Respir Med* 2009;**103**:1430–5. <http://dx.doi.org/10.1016/j.rmed.2009.04.024>

378. Hendriks EJ, Bernards AT, de Bie RA, de Vet HC. The minimal important change of the PRAFAB questionnaire in women with stress urinary incontinence: results from a prospective cohort study. *Neurourol Urodyn* 2008;**27**:379–87. <http://dx.doi.org/10.1002/nau.20554>
379. Wang YC, Hart DL, Werneke M, Stratford PW, Mioduski JE. Clinical interpretation of outcome measures generated from a lumbar computerized adaptive test. *Phys Ther* 2010;**90**:1323–35. <http://dx.doi.org/10.2522/ptj.20090371>
380. Carreon LY, Glassman SD, Campbell MJ, Anderson PA. Neck Disability Index, short form-36 physical component summary, and pain scales for neck and arm pain: the minimum clinically important difference and substantial clinical benefit after cervical spine fusion. *Spine J* 2010;**10**:469–74. <http://dx.doi.org/10.1016/j.spinee.2010.02.007>
381. Young BA, Walker MJ, Strunce JB, Boyles RE, Whitman JM, Childs JD. Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. *Spine J* 2009;**9**:802–8. <http://dx.doi.org/10.1016/j.spinee.2009.06.002>
382. Bols EM, Hendriks EJ, Deutekom M, Berghmans BC, Baeten CG, de Bie RA. Inconclusive psychometric properties of the Vaizey score in fecally incontinent patients: a prospective cohort study. *Neurourol Urodyn* 2010;**29**:370–7.
383. Lauridsen HH, Manniche C, Korsholm L, Grunnet-Nilsson N, Hartvigsen J. What is an acceptable outcome of treatment before it begins? Methodological considerations and implications for patients with chronic low back pain. *Eur Spine J* 2009;**18**:1858–66. <http://dx.doi.org/10.1007/s00586-009-1070-1>
384. Polson K, Reid D, McNair PJ, Larmer P. Responsiveness, minimal importance difference and minimal detectable change scores of the shortened disability arm shoulder hand (QuickDASH) questionnaire. *Manual Ther* 2010;**15**:404–7. <http://dx.doi.org/10.1016/j.math.2010.03.008>
385. Bilbao A, Quintana JM, Escobar A, Garcia S, Andradas E, Bare M, et al. Responsiveness and clinically important differences for the VF-14 index, SF-36, and visual acuity in patients undergoing cataract surgery. *Ophthalmology* 2009;**116**:418–24. <http://dx.doi.org/10.1016/j.ophtha.2008.11.020>
386. Demoulin C, Ostelo R, Knottnerus 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. <http://dx.doi.org/10.1016/j.jclinepi.2009.08.029>
387. Wuang YP, Su CY. Reliability and responsiveness of the Bruininks-Oseretsky Test of Motor Proficiency-Second Edition in children with intellectual disability. *Res Dev Disabil* 2009;**30**:847–55. <http://dx.doi.org/10.1016/j.ridd.2008.12.002>
388. Bagó J, Pérez-Gruoso FJ, Les E, Hernández P, Pellisé F. Minimal important differences of the SRS-22 Patient Questionnaire following surgical treatment of idiopathic scoliosis. *Eur Spine J* 2009;**18**:1898–904. <http://dx.doi.org/10.1007/s00586-009-1066-x>
389. Kemmler G, Zabernigg A, Gattringer K, Rumpold G, Giesinger J, Sperner-Unterweger B, et al. A new approach to combining clinical relevance and statistical significance for evaluation of quality of life changes in the individual patient. *J Clin Epidemiol* 2010;**63**:171–9. <http://dx.doi.org/10.1016/j.jclinepi.2009.03.016>
390. Lemieux J, Beaton DE, Hogg-Johnson S, Bordeleau LJ, Goodwin PJ. Three methods for minimally important difference: no relationship was found with the net proportion of patients improving. *J Clin Epidemiol* 2007;**60**:448–55. <http://dx.doi.org/10.1016/j.jclinepi.2006.08.006>
391. Rentz AM, Yu R, Iler-Lissner S, Leyendecker P. Validation of the Bowel Function Index to detect clinically meaningful changes in opioid-induced constipation. *J Med Econ* 2009;**12**:371–83. <http://dx.doi.org/10.3111/13696990903430481>

392. Schurch B, Denys P, Kozma CM, Reese PR, Slaton T, Barron R. Reliability and validity of the Incontinence Quality of Life questionnaire in patients with neurogenic urinary incontinence. *Arch Phys Med Rehabil* 2007;**88**:646–52. <http://dx.doi.org/10.1016/j.apmr.2007.02.009>
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>
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>
395. Yost KJ. Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the Functional Assessment of Cancer Therapy. *Value Health* 2005;**8**:117–27. <http://dx.doi.org/10.1111/j.1524-4733.2005.08202.x>
396. Yost KJ, Cella D, Chawla A, Holmgren E, Eton DT, Ayanian JZ, et al. Minimally important differences were estimated for the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) instrument using a combination of distribution- and anchor-based approaches. *J Clin Epidemiol* 2005;**58**:1241–51. <http://dx.doi.org/10.1016/j.jclinepi.2005.07.008>
397. Kozma CM, Slaton TL, Monz BU, Hodder R, Reese PR. Development and validation of a patient satisfaction and preference questionnaire for inhalation devices. *Treat Resp Med* 2005;**4**:41–52. <http://dx.doi.org/10.2165/00151829-200504010-00005>
398. Lin KC, Hsieh YW, Wu CY, Chen CL, Jang Y, Liu JS. Minimal detectable change and clinically important difference of the Wolf Motor Function Test in stroke patients. *Neurorehabil Neural Repair* 2009;**23**:429–34. <http://dx.doi.org/10.1177/1545968308331144>
399. Yang M, Morin CM, Schaefer K, Wallenstein GV. Interpreting score differences in the Insomnia Severity Index: using health-related outcomes to define the minimally important difference. *Curr Med Res Opin* 2009;**25**:2487–94. <http://dx.doi.org/10.1185/03007990903167415>
400. Yount S, List M, Du H, Yost K, Bode R, Brockstein B, et al. A randomized validation study comparing embedded versus extracted FACT Head and Neck Symptom Index scores. *Qual Life Res* 2007;**16**:1615–26. <http://dx.doi.org/10.1007/s11136-007-9270-9>
401. Cole JC, Lin P, Rupnow MF. Minimal important differences in the Migraine-Specific Quality of Life Questionnaire (MSQ) version. *Cephalalgia* 2009;**29**:1180–7. <http://dx.doi.org/10.1111/j.1468-2982.2009.01852.x>
402. Williams VS, Morlock RJ, Feltner D. Psychometric evaluation of a visual analog scale for the assessment of anxiety. *Health Qual Life Outcomes* 2010;**8**:57. <http://dx.doi.org/10.1186/1477-7525-8-57>
403. Barber MD, Spino C, Janz NK, Brubaker L, Nygaard I, Nager CW, et al. The minimum important differences for the urinary scales of the Pelvic Floor Distress Inventory and Pelvic Floor Impact Questionnaire. *Am J Obstet Gynecol* 2009;**200**:580–7. <http://dx.doi.org/10.1016/j.ajog.2009.02.007>
404. Dubois D, Gilet H, Viala-Danten M, Tack J. Psychometric performance and clinical meaningfulness of the Patient Assessment of Constipation-Quality of Life questionnaire in prucalopride (RESOLOR) trials for chronic constipation. *Neurogastroenterol Motil* 2010;**22**:e54–63. <http://dx.doi.org/10.1111/j.1365-2982.2009.01408.x>
405. Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T, et al. Estimating clinically significant differences in quality of life outcomes. *Qual Life Res* 2005;**14**:285–95. <http://dx.doi.org/10.1007/s11136-004-0705-2>

406. Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Measuring patient and clinician perspectives to evaluate change in health-related quality of life among patients with chronic obstructive pulmonary disease. *J Gen Intern Med* 2007;**22**:161–70. <http://dx.doi.org/10.1007/s11606-006-0063-6>
407. Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Triangulating patient and clinician perspectives on clinically important differences in health-related quality of life among patients with heart disease. *Health Serv Res* 2007;**42**:2257–74. <http://dx.doi.org/10.1111/j.1475-6773.2007.00733.x>
408. Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. North American Orthopaedic Rehabilitation Research Network. *Phys Ther* 1999;**79**:371–83.
409. Wells G, Li T, Maxwell L, MacLean R, Tugwell P. Determining the minimal clinically important differences in activity, fatigue, and sleep quality in patients with rheumatoid arthritis. *J Rheumatol* 2007;**34**:280–9.
410. Raj AA, Pavord DI, Birring SS. Clinical cough IV: what is the minimal important difference for the Leicester Cough Questionnaire? *Handbook Exp Pharmacol* 2009;**187**:311–20. http://dx.doi.org/10.1007/978-3-540-79842-2_16
411. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. OMERACT Committee. *J Rheumatol* 1993;**20**:561–5.
412. Liang MH. The American College of Rheumatology response criteria for systemic lupus erythematosus clinical trials – measures of overall disease activity. *Arthritis Rheum* 2004;**50**:3418–26. <http://dx.doi.org/10.1002/art.20628>
413. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain* 2008;**9**:105–21. <http://dx.doi.org/10.1016/j.jpain.2007.09.005>
414. Ornetti P, Brandt K, Hellio-Le Graverand MP, Hochberg M, Hunter DJ, Kloppenburg M, et al. OARSI-OMERACT definition of relevant radiological progression in hip/knee osteoarthritis. *Osteoarthritis Cartilage* 2009;**17**:856–63. <http://dx.doi.org/10.1016/j.joca.2009.01.007>
415. Mills K, Blanch P, Vicenzino B. Identifying clinically meaningful tools for measuring comfort perception of footwear. *Med Sci Sports Exerc* 2010;**42**:1966–71. <http://dx.doi.org/10.1249/MSS.0b013e3181dbacc8>
416. Symonds T, Spino C, Sisson M, Soni P, Martin M, Gunter L, et al. Methods to determine the minimum important difference for a sexual event diary used by postmenopausal women with hypoactive sexual desire disorder. *J Sex Med* 2007;**4**:1328–35. <http://dx.doi.org/10.1111/j.1743-6109.2007.00562.x>
417. Spiegel BM, Younossi ZM, Hays RD, Revicki D, Robbins S, Kanwal F. Impact of hepatitis C on health related quality of life: a systematic review and quantitative assessment. *Hepatology* 2005;**41**:790–800. <http://dx.doi.org/10.1002/hep.20659>
418. Vernon MK, Revicki DA, Awad AG, Dirani R, Panish J, Canuso CM, et al. Psychometric evaluation of the Medication Satisfaction Questionnaire (MSQ) to assess satisfaction with antipsychotic medication among schizophrenia patients. *Schizophr Res* 2010;**118**:271–8. <http://dx.doi.org/10.1016/j.schres.2010.01.021>
419. Society for Clinical Trials. URL: www.sctweb.org/ (accessed December 2011).
420. MRC Network of Hubs for Trials Methodology Research. URL: www.methodologyhubs.mrc.ac.uk/ (accessed December 2011).

421. UKCRC Registered Clinical Trials Units. URL: www.ukcrc-ctu.org.uk/ (accessed December 2011).
422. NIHR Research Design Services. URL: www.nihr.ac.uk/research/Pages/ResearchDesignService.aspx (accessed December 2011).
423. Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. *Control Clin Trials* 2002;**23**:2–14. [http://dx.doi.org/10.1016/S0197-2456\(01\)00196-9](http://dx.doi.org/10.1016/S0197-2456(01)00196-9)
424. Gayet-Ageron A, Agoritsas T, Combescure C, Bagamery K, Courvoisier DS, Perneger TV. What differences are detected by superiority trials or ruled out by noninferiority trials? A cross-sectional study on a random sample of two-hundred two-arms parallel group randomized clinical trials. *BMC Med Res Methodol* 2010;**10**:93. <http://dx.doi.org/10.1186/1471-2288-10-93>
425. Goudie AC, Sutton AJ, Jones DR, Donald A. Empirical assessment suggests that existing evidence could be used more fully in designing randomized controlled trials. *J Clin Epidemiol* 2010;**63**: 983–91. <http://dx.doi.org/10.1016/j.jclinepi.2010.01.022>
426. Cook JV, Dickinson HO, Eccles MP. Response rates in postal surveys of healthcare professionals between 1996 and 2005: an observational study. *BMC Health Serv Res* 2009;**9**:160. <http://dx.doi.org/10.1186/1472-6963-9-160>
427. McDonald A, Knight R, Campbell MK, Entwistle VA, Cook JA, Grant A, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. *Trials* 2006;**7**:7. <http://dx.doi.org/10.1186/1745-6215-7-9>
428. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005;**365**:1348–53. [http://dx.doi.org/10.1016/S0140-6736\(05\)61034-3](http://dx.doi.org/10.1016/S0140-6736(05)61034-3)
429. Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet* 2005;**365**:1159–62. [http://dx.doi.org/10.1016/S0140-6736\(05\)71879-1](http://dx.doi.org/10.1016/S0140-6736(05)71879-1)
430. Matthews JN. *Introduction to randomized controlled clinical trials*. London: Taylor & Francis; 2006. <http://dx.doi.org/10.1201/9781420011302>
431. Bland JM. The tyranny of power: is there a better way to calculate sample size? *BMJ* 2009;**339**: b3985. <http://dx.doi.org/10.1136/bmj.b3985>
432. Borm GF, Bloem BR, Munneke M, Teerenstra S. A simple method for calculating power based on a prior trial. *J Clin Epidemiol* 2010;**63**:992–7. <http://dx.doi.org/10.1016/j.jclinepi.2009.10.011>
433. Piantadosi S. *Clinical trials – a methodologic perspective*. Chichester: Wiley Interscience; 2005. <http://dx.doi.org/10.1002/0471740136>
434. Guyatt GH, Mills EJ, Elbourne D. In the era of systematic reviews, does the size of an individual trial still matter. *PLoS Med* 2008;**5**:e4. <http://dx.doi.org/10.1371/journal.pmed.0050004>
435. Shrier I, Platt RW, Steele RJ. Mega-trials vs. meta-analysis: precision vs. heterogeneity? *Contemp Clin Trials* 2007;**28**:324–8. <http://dx.doi.org/10.1016/j.cct.2006.11.007>
436. Peto R, Baigent C. Trials: the next 50 years. Large scale randomised evidence of moderate benefits. *BMJ* 1998;**317**:1170–1. <http://dx.doi.org/10.1136/bmj.317.7167.1170>
437. Roberts I, Yates D, Sandercock P, Farrell B, Wasserberg J, Lomas G, et al. Effect of intravenous corticosteroids on death within 14 days in 10008 adults with clinically significant head injury (MRC CRASH trial): randomised placebo-controlled trial. *Lancet* 2004;**364**:1321–8. [http://dx.doi.org/10.1016/S0140-6736\(04\)17188-2](http://dx.doi.org/10.1016/S0140-6736(04)17188-2)
438. CRASH-2 Trial Collaborators, Shakur H, Roberts I, Bautista R, Caballero J, Coats T, et al. Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): a randomised, placebo-controlled trial. *Lancet* 2010;**376**: 23–32. [http://dx.doi.org/10.1016/S0140-6736\(10\)60835-5](http://dx.doi.org/10.1016/S0140-6736(10)60835-5)

439. Bonjer HJ, Hop WC, Nelson H, Sargent DJ, Lacy AM, Castells A, *et al.* Laparoscopically assisted vs open colectomy for colon cancer: a meta-analysis. *Arch Surg* 2007;**142**:298–303. <http://dx.doi.org/10.1001/archsurg.142.3.298>
440. Senn S. Controversies concerning randomization and additivity in clinical trials. *Stat Med* 2004;**23**:3729–53. <http://dx.doi.org/10.1002/sim.2074>
441. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics* 1980;**36**:343–6. <http://dx.doi.org/10.2307/2529990>
442. Campbell MK, Mollison J, Grimshaw JM. Cluster trials in implementation research: estimation of intracluster correlation coefficients and sample size. *Stat Med* 2001;**20**:391–9. [http://dx.doi.org/10.1002/1097-0258\(20010215\)20:3<391::AID-SIM800>3.0.CO;2-Z](http://dx.doi.org/10.1002/1097-0258(20010215)20:3<391::AID-SIM800>3.0.CO;2-Z)
443. Girling AJ, Lilford RJ, Braunholtz DA, Gillett WR. Sample-size calculations for trials that inform individual treatment decisions: a ‘true-choice’ approach. *Clin Trials* 2007;**4**:15–24. <http://dx.doi.org/10.1177/1740774506075872>
444. Kadane JB. An application of robust Bayesian-analysis to a medical experiment. *J Stat Plan Infer* 1994;**40**:221–8. [http://dx.doi.org/10.1016/0378-3758\(94\)90122-8](http://dx.doi.org/10.1016/0378-3758(94)90122-8)
445. Kass MA, Heuer DK, Higginbotham EJ, Johnson CA, Keltner JL, Miller JP, *et al.* The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch Ophthalmol* 2009;**120**:701–13.
446. Burr JM, Botello Pinzon P, Takwoingi Y, Hernandez R, Vazquez-Montes M, Elders A, *et al.* Surveillance for ocular hypertension: evidence synthesis and economic evaluation. *Health Technol Assess* 2012;**16**(29).
447. Williamson PA, Altman DG, Blazeby J, Clarke M. COMET (Core Outcome Measures in Effectiveness Trials) Initiative. 2012. URL: www.comet-initiative.org/ (accessed June 2012).
448. Goodacre S, Bradburn M, Fitzgerald P, Cross E, Collinson P, Gray A, *et al.* The RATPAC (Randomised Assessment of Treatment using Panel Assay of Cardiac markers) trial: a randomised controlled trial of point-of-care cardiac markers in the emergency department. *Health Technol Assess* 2011;**15**(23).
449. Senn S, Julious S. Measurement in clinical trials: a neglected issue for statisticians? *Stat Med* 2009;**28**:3189–209. <http://dx.doi.org/10.1002/sim.3603>
450. Wittes J. Commentary on ‘Measurement in clinical trials: a neglected issue for statisticians?’. *Stat Med* 2009;**28**:3220–2.
451. Glazener C, Boachie C, Buckley B, Cochran C, Dorey G, Grant A, *et al.* Urinary incontinence in men after formal one-to-one pelvic-floor muscle training following radical prostatectomy or transurethral resection of the prostate (MAPS): two parallel randomised controlled trials. *Lancet* 2011;**378**:328–37. [http://dx.doi.org/10.1016/S0140-6736\(11\)60751-4](http://dx.doi.org/10.1016/S0140-6736(11)60751-4)
452. Hunter KF, Glazener CM, Moore KN. Conservative management for postprostatectomy urinary incontinence. *Cochrane Database Syst Rev* 2007;**2**:CD001843.
453. Peace KE, Chen DG. *Clinical trial methodology*. London: Chapman & Hall; 2010. <http://dx.doi.org/10.1201/EBK1584889175>
454. van Tulder M, Malmivaara A, Hayden J, Koes B. Statistical significance versus clinical importance: trials on exercise therapy for chronic low back pain as example. *Spine* 2007;**32**:1785–90. <http://dx.doi.org/10.1097/BRS.0b013e3180b9ef49>

455. International Conference on Harmonisation. *Statistical principles for clinical trials (E9): ICH tripartite guideline*. Geneva: International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH); 1998. URL: www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html (accessed March 2012).
456. de Vet HC, Terwee CB. The minimal detectable change should not replace the minimal important difference. *J Clin Epidemiol* 2010;**63**:804–5. <http://dx.doi.org/10.1016/j.jclinepi.2009.12.015>
457. Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *Eur J Health Econ* 2010;**11**:215–25. <http://dx.doi.org/10.1007/s10198-009-0168-z>
458. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;**77**:371–83. <http://dx.doi.org/10.4065/77.4.371>
459. Wosje KS, Knipstein BL, Kalkwarf HJ. Measurement error of DXA: interpretation of fat and lean mass changes in obese and non-obese children. *J Clin Densitom* 2006;**9**:335–40. <http://dx.doi.org/10.1016/j.jocd.2006.03.016>
460. Eckermann S, Karnon J, Willan AR. The value of value of information: best informing research design and prioritization using current methods. *Pharmacoeconomics* 2010;**28**:699–709. <http://dx.doi.org/10.2165/11537370-000000000-00000>
461. Schünemann HJ, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. Measurement properties and interpretability of the Chronic respiratory disease questionnaire (CRQ). *COPD* 2005;**2**:81–9. <http://dx.doi.org/10.1081/COPD-200050651>
462. Naylor CD, Llewellyn-Thomas HA. Can there be a more patient-centred approach to determining clinically important effect sizes for randomized treatment trials? *J Clin Epidemiol* 1994;**47**:787–95. [http://dx.doi.org/10.1016/0895-4356\(94\)90176-7](http://dx.doi.org/10.1016/0895-4356(94)90176-7)
463. O'Hagan A, Buck CE, Daneshkhan A, Eiser JR, Garthwaite PH, Jenkinson DJ, *et al.* *Uncertain judgements: eliciting experts' probabilities*. Chichester: John Wiley; 2006. <http://dx.doi.org/10.1002/0470033312>
464. Vickers AJ. Underpowering in randomized trials reporting a sample size calculation. *J Clin Epidemiol* 2003;**56**:717–20. [http://dx.doi.org/10.1016/S0895-4356\(03\)00141-0](http://dx.doi.org/10.1016/S0895-4356(03)00141-0)
465. Herbison P, Hay-Smith J, Gillespie WJ. Meta-analyses of small numbers of trials often agree with longer-term results. *J Clin Epidemiol* 2011;**64**:145–53. <http://dx.doi.org/10.1016/j.jclinepi.2010.02.017>
466. Rios LP, Ye C, Thabane L. Association between framing of the research question using the PICOT format and reporting quality of randomized controlled trials. *BMC Med Res Methodol* 2010;**10**:11. <http://dx.doi.org/10.1186/1471-2288-10-11>
467. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, *et al.* The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;**343**:d5928. <http://dx.doi.org/10.1136/bmj.d5928>
468. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;**336**:924–6. <http://dx.doi.org/10.1136/bmj.39489.470347.AD>
469. Cook TD, de Mets DL. *Introduction to statistical methods for clinical trials*. London: Chapman & Hall; 2008.
470. Rosenthal R, Rubin DB. Meta-analytic procedures for combining studies with multiple effect sizes. *Psychol Bull* 1986;**99**:400–6. <http://dx.doi.org/10.1037/0033-2909.99.3.400>

471. Dunlop WP, Cortina JM, Vaslow JB, Burke JM. Meta-analysis of experiments with matched groups or repeated measures design. *Psychol Methods* 1996;**1**:170–7. <http://dx.doi.org/10.1037/1082-989X.1.2.170>
472. Zwarenstein M, Treweek S. What kind of randomized trials do we need? *J Clin Epidemiol* 2009;**62**:461–3. <http://dx.doi.org/10.1016/j.jclinepi.2009.01.011>
473. Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. *BMJ* 2001;**323**:1123–4. <http://dx.doi.org/10.1136/bmj.323.7321.1123>
474. Day S. *Dictionary for clinical trials*. Chichester: John Wiley; 2007.
475. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, *et al*. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *BMJ* 2008;**337**:a2390. <http://dx.doi.org/10.1136/bmj.a2390>
476. Flather M, Aston H, Stables R. *Handbook of clinical trials*. London: Remedica; 2001.
477. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;**340**:c332. <http://dx.doi.org/10.1136/bmj.c332>
478. Early Treatment Diabetic Retinopathy Study design and baseline patient characteristics. ETDRS report number 7. *Ophthalmology* 1991;**98**(Suppl. 5):56.
479. Brooks HL Jr. Macular hole surgery with and without internal limiting membrane peeling. *Ophthalmology* 1948;**107**:1939–48. [http://dx.doi.org/10.1016/S0161-6420\(00\)00331-6](http://dx.doi.org/10.1016/S0161-6420(00)00331-6)
480. Paques M, Chastang C, Mathis A, Sahel J, Massin P, Dosquet C, *et al*. Effect of autologous platelet concentrate in surgery for idiopathic macular hole: results of a multicenter, double-masked, randomized trial. Platelets in Macular Hole Surgery Group. *Ophthalmology* 1999;**106**:932–8. [http://dx.doi.org/10.1016/S0161-6420\(99\)00512-6](http://dx.doi.org/10.1016/S0161-6420(99)00512-6)
481. Taggart DP, Lees B, Gray A, Altman DG, Flather M, Channon K, *et al*. Protocol for the Arterial Revascularisation Trial (ART). A randomised trial to compare survival following bilateral versus single internal mammary grafting in coronary revascularisation. *Trials* 2006;**7**:7. <http://dx.doi.org/10.1186/1745-6215-7-7>
482. Taggart DP, D'Amico R, Altman DG. Effect of arterial revascularisation on survival: a systematic review of studies comparing bilateral and single internal mammary arteries. *Lancet* 2001;**358**:870–5. [http://dx.doi.org/10.1016/S0140-6736\(01\)06069-X](http://dx.doi.org/10.1016/S0140-6736(01)06069-X)

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.¹ 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 misled 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 (α) 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 β 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.² 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.³ 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.⁴ 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 Cohen¹⁷ 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 a 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 (grantholders 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.

- 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.¹¹ 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 sent 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

1. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D *et al.* The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;**134**:663–94.
2. Pocock SJ. *Clinical Trials – A practical approach*. 1996 John Wiley & Sons Ltd.
3. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. Chichester: John Wiley & Sons; 2003.
4. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. *BMJ* **338**:b1732, 2009.
5. Fayers PM and Machin D. Sample size: how many patients are necessary? *British Journal of Cancer* 1995;**72**:1–9.

6. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;**10**:407–15.
7. Copay AG, Subach BR, Glassman SD, Polly J, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine Journal* 2007;**7**:541–6.
8. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MICD): A literature review and directions for future research. *Curr Opin Rheumatol* 2002;**14**:109–14.
9. Kvien TK, Heiberg T, Hagen KB. Minimum clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean? *Ann Rheum Dis* 2007;**66**(Suppl III):iii40–iii41.
10. Barrett B, Brown D, Mundt M, Brown R. Sufficiently important difference: expanding the framework of clinical significance. *Med Decis Making* 2005;**25**:250–61.
11. Wells G, Beaton D, Shea B, Boers M, Simon L, Strand V et al. Minimal clinically important differences: Review of methods. *J Rheumatol* 2001;**28**:406–12.
12. Briggs AH, Gray AM. Power and sample size calculations for stochastic cost-effectiveness analysis. *Med Decis Making* 1998;**18**(Suppl.):S81–92.
13. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? *Pharmacoeconomics* 2000;**18**:419–23.
14. Chan KBY, Man-Son-Hing M, Molnar FJ, Laupacis A. How well is the clinical importance of study results reported? An assessment of randomized controlled trials. *Can Med Assoc J* 2001;**165**:1197–202.
15. Molnar FJ, Man-Son-Hing M, Fergusson D. Systematic review of measures of clinical significance employed in randomized controlled trials of drugs for dementia. *J Am Geriatr Soc* 2009;**57**:536–46.
16. Fayers PM, Cuschieri A, Fielding J, Craven J, Uscinska B, Freedman LS. Sample size calculation for clinical trials: the impact of clinician beliefs. *Br J Cancer* 2000;**82**:213–9.
17. Cohen J. *Statistical power: analysis of behavioural sciences*. New York: Academic Press; 1977.
18. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: Recommendations for good practice. *J Eval Clin Pract* 2004;**10**:307–12.
19. Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. *Stat Med* 2007;**26**:2479–500.
20. O'Hagan A, Stevens JW. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Med Decis Making* 2001;**21**:219–30.
21. Laska EM, Meisner M, Siegel C. Power and sample size in cost-effectiveness analysis. *Med Decis Making* 1999;**19**:339–43.

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 determine\$ 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.

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 (sufficient* NEXT importan*) NEAR/2 (difference* or change* or improvement* or effect*)
 #12 (sufficient* 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* important"
 #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

S7 TX minim* w2 change or TX smallest w2 difference
 S8 TX meaningful w2 change
 S9 TX smallest w2 change
 S10 S1 or S2 or S3 or S4 or S5 or S6 or S7 or S8 or S9

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 (smallest 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 difference)

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* meaningful") AND DOCTYPE(ip)) or (TITLE-ABS-KEY("minim* 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

1. Berry SM, Carlin BP, Lee JJ, Muller P. *Bayesian adaptive methods for clinical trials*. London: Taylor & Francis; 2011.
2. Chin RY. *Principles and practice of clinical trial medicine*. Amsterdam: Elsevier; 2008.
3. Cleophas TJ, Zwinderman AH, Cleophas, TF, Cleophas EP. *Statistics applied to clinical trials*. London: Springer; 2009.
4. Cook TD, DeMets DL. *Introduction to statistical methods for clinical trials*. London: Chapman & Hall; 2008.
5. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of clinical trials*. New York, NY: Springer; 2010.
6. Hackshaw AK. *A concise guide to clinical trials*. Oxford: Wiley-Blackwell; 2009.
7. Julious SA. *Sample sizes for clinical trials*. Boca Raton, FL: CRC Press; 2010.
8. Machin D, Day S, Green S, editors. *Textbook of clinical trials*. Chichester: John Wiley; 2006.
9. Matthews J. *Introduction to randomized controlled trials*. London: Chapman & Hall; 2006.
10. Peace KE, Chen D. *Clinical trial methodology*. London: Chapman & Hall; 2011.
11. Piantadosi S. *Clinical trials: a methodologic perspective*. Hoboken, NJ: Wiley; 2005.

12. Pocock SJ. *Clinical trials: a practical approach*. Chichester: Wiley; 1983.
13. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. Hoboken, NJ: Wiley; 2003.
14. Walters SJ. *Quality of life outcomes in clinical trials and health-care evaluation: a practical guide to analysis and interpretation*. Chichester: Wiley; 2009.
15. Wang D, Bakhai A. *Clinical trials: a practical guide to design, analysis, and reporting*. London: Remedica; 2006.
16. International Conference on Harmonisation. *Statistical principles for clinical trials (E9): ICH tripartite guidelines*. Geneva: International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use; 1998. URL: www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html (accessed March 2012).

Appendix 3 List of included studies

Anchor method (n = 253)

1. Aletaha D, Funovits J, Ward MM, Smolen JS, Kvien TK. Perception of improvement in patients with rheumatoid arthritis varies with disease activity levels at baseline. *Arthritis Rheum* 2009;**61**:313–20.
2. Allen CJ, Parameswaran K, Belda J, Anvari M. Reproducibility, validity, and responsiveness of a disease-specific symptom questionnaire for gastroesophageal reflux disease. *Dis Esophagus* 2000;**13**:265–70.
3. Allen PF, O'Sullivan M, Locker D. Determining the minimally important difference for the Oral Health Impact Profile-20. *Eur J Oral Sci* 2009;**117**:129–34.
4. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis Rheum* 2001;**45**:384–91.
5. Angst F, Aeschlimann A, Michel BA, Stucki G. Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities. *J Rheumatol* 2002;**29**:131–8.
6. Badia X, Díez-Pérez A, Lahoz R, Lizán L, Nogués X, Iborra J. The ECOS-16 questionnaire for the evaluation of health related quality of life in post-menopausal women with osteoporosis. *Health Qual Life Outcomes* 2004;**2**:41.
7. Barber BL, Santanello NC, Epstein RS. Impact of the global on patient perceivable change in an asthma specific QOL questionnaire. *Qual Life Res* 1996;**5**:117–22.
8. Barrett B, Brown RL, Mundt MP, Thomas GR, Barlow SK, Highstrom AD, *et al.* Validation of a short form Wisconsin Upper Respiratory Symptom Survey (WURSS-21). *Health Qual Life Outcomes* 2009;**7**:76.
9. Bastyr EJ III, Price KL, Bril V, MBBQ Study Group. Development and validity testing of the neuropathy total symptom score-6: questionnaire for the study of sensory symptoms of diabetic peripheral neuropathy. *Clin Ther* 2005;**27**:1278–94.
10. Bellamy N, Bell MJ, Goldsmith CH, Pericak D, Walker V, Raynauld JP, *et al.* Evaluation of WOMAC 20, 50, 70 response criteria in patients treated with hylan G-F 20 for knee osteoarthritis. *Ann Rheum Dis* 2005;**64**:881–5.
11. Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. *Arch Phys Med Rehabil* 2006;**87**:32–9.
12. Bennett RM, Bushmakini AG, Cappelleri JC, Zlateva G, Sadosky AB. Minimal clinically important difference in the fibromyalgia impact questionnaire. *J Rheumatol* 2009;**36**:1304–11.
13. Bennett SJ, Oldridge NB, Eckert GJ, Embree JL, Browning S, Hou N, *et al.* Comparison of quality of life measures in heart failure. *Nurs Res* 2003;**52**:207–16.
14. Bessette L, Sangha O, Kuntz KM, Keller RB, Lew RA, Fossel AH, *et al.* Comparative responsiveness of generic versus disease-specific and weighted versus unweighted health status measures in carpal tunnel syndrome. *Med Care* 1998;**36**:491–502.
15. Bijur PE, Chang AK, Esses D, Gallagher EJ. Identifying the minimum clinically significant difference in acute pain in the elderly. *Ann Emerg Med* 2010;**56**:517–21.

16. Black N, Browne J, van der Meulen J, Jamieson L, Copley L, Lewsey J. Is there overutilisation of cataract surgery in England? *Br J Ophthalmol* 2009;**93**:13–17.
17. Bohannon RW. Responsiveness of measurements of knee extension force obtained by hand-held dynamometry: a preliminary analysis. *Isokinet Exerc Sci* 2009;**17**:169–72.
18. Bolton JE. Sensitivity and specificity of outcome measures in patients with neck pain: detecting clinically significant improvement. *Spine* 2004;**29**:2410–17.
19. Bonniaud V, Bryant D, Parratte B, Guyatt G. Qualiveen, a urinary-disorder specific instrument: 0.5 corresponds to the minimal important difference. *J Clin Epidemiol* 2008;**61**:505–10.
20. Brant R, Sutherland L, Hilsden R. Examining the minimum important difference. *Stat Med* 1999;**18**:2593–603.
21. Brod M, Christensen T, Kongs JH, Bushnell DM. Examining and interpreting responsiveness of the Diabetes Medication Satisfaction measure. *J Med Econ* 2009;**12**:309–16.
22. Brod M, Hammer M, Kragh N, Lessard S, Bushnell DM. Development and validation of the Treatment Related Impact Measure of Weight (TRIM-Weight). *Health Qual Life Outcomes* 2010;**8**:19.
23. Bronfort G, Bouter LM. Responsiveness of general health status in chronic low back pain: a comparison of the COOP charts and the SF-36. *Pain* 1999;**83**:201–9.
24. Browne JP, van der Meulen JH, Lewsey JD, Lamping DL, Black N. Mathematical coupling may account for the association between baseline severity and minimally important difference values. *J Clin Epidemiol* 2010;**63**:865–74.
25. Bruynesteyn K, Van Der Linden S, Landewé R, Gubler F, Weijers R, Van Der Heijde D. Progression of rheumatoid arthritis on plain radiographs judged differently by expert radiologists and rheumatologists. *J Rheumatol* 2004;**31**:1088–94.
26. Bushnell DM, Martin ML, Moore KA, Richter HE, Rubin A, Patrick DL. Menorrhagia Impact Questionnaire: assessing the influence of heavy menstrual bleeding on quality of life. *Curr Med Res Opin* 2010;**26**:2745–55.
27. Cappelleri JC, Althof SE, O’Leary MP, Glina S, King R, Stecher VJ, *et al.* Clinically meaningful improvement on the Self-Esteem And Relationship questionnaire in men with erectile dysfunction. *Qual Life Res* 2007;**16**:1203–10.
28. Cappelleri JC, Bushmakina AG, McDermott AM, Dukes E, Sadosky A, Petrie CD, *et al.* Measurement properties of the Medical Outcomes Study Sleep Scale in patients with fibromyalgia. *Sleep Med* 2009;**10**:766–70.
29. Carragee EJ, Cheng I. Minimum acceptable outcomes after lumbar spinal fusion. *Spine J* 2010;**10**:313–20.
30. Carreon LY, Sanders JO, Diab M, Sucato DJ, Sturm PF, Glassman SD, *et al.* The minimum clinically important difference in Scoliosis Research Society-22 Appearance, Activity, And Pain domains after surgical correction of adolescent idiopathic scoliosis. *Spine* 2010;**35**:2079–83.
31. Cella DF, Bonomi AE, Lloyd SR, Tulskey DS, Kaplan E, Bonomi P. Reliability and validity of the Functional Assessment of Cancer Therapy-Lung (FACT-L) quality of life instrument. *Lung Cancer* 1995;**12**:199–220.
32. Chan KS, Chen WH, Gan TJ, Hsieh R, Chen C, Lakshminarayanan M, *et al.* Development and validation of a composite score based on clinically meaningful events for the opioid-related symptom distress scale. *Qual Life Res* 2009;**18**:1331–40.

33. Chan L, Mulgaonkar S, Walker R, Arns W, Ambühl P, Schiavelli R. Patient-reported gastrointestinal symptom burden and health-related quality of life following conversion from mycophenolate mofetil to enteric-coated mycophenolate sodium. *Transplantation* 2006;**81**:1290–7.
34. Chaput de Saintonge DM, Kirwan JR, Evans SJ, Crane GJ. How can we design trials to detect clinically important changes in disease severity? *Br J Clin Pharmacol* 1988;**26**:355–62.
35. Chen H-Y, Hung Y-C, Yang T-C, Yeh L-S, Chang W-C. A scored storage symptoms questionnaire to screen urodynamic stress incontinence in women with overactive bladder. *Mid-Taiwan J Med* 2006;**11**:222–9.
36. Chesworth BM, Mahomed NN, Bourne RB, Davis AM, OJRR Study Group. Willingness to go through surgery again validated the WOMAC clinically important difference from THR/TKR surgery. *J Clin Epidemiol* 2008;**61**:907–18.
37. Chhabra SK. Acute bronchodilator response has limited value in differentiating bronchial asthma from COPD. *J Asthma* 2005;**42**:367–72.
38. Childs JD, Piva SR. Psychometric properties of the functional rating index in patients with low back pain. *Eur Spine J* 2005;**14**:1008–12.
39. Chiou CF, Sherbourne CD, Cornelio I, Lubeck DP, Paulus HE, Dylan M, *et al.* Development and validation of the revised Cedars-Sinai health-related quality of life for rheumatoid arthritis instrument. *Arthritis Rheum* 2006;**55**:856–63.
40. Chiou CF, Sherbourne CD, Cornelio I, Lubeck DP, Paulus HE, Dylan M, *et al.* Revalidation of the original Cedars-Sinai health-related quality of life in rheumatoid arthritis questionnaire. *J Rheumatol* 2006;**33**:256–62.
41. Chung VC, Wong VC, Lau CH, Hui H, Lam TH, Zhong LX, *et al.* Using Chinese version of MYMOP in Chinese medicine evaluation: validity, responsiveness and minimally important change. *Health Qual Life Outcomes* 2010;**8**:111.
42. Clayson D, Wild D, Doll H, Keating K, Gondek K. Validation of a patient-administered questionnaire to measure the severity and bothersomeness of lower urinary tract symptoms in uncomplicated urinary tract infection (UTI): the UTI Symptom Assessment questionnaire. *BJU Int* 2005;**96**:350–9.
43. Cleland JA, Fritz JM, Whitman JM, Palmer JA. The reliability and construct validity of the Neck Disability Index and patient specific functional scale in patients with cervical radiculopathy. *Spine* 2006;**31**:598–602.
44. Cleland JA, Childs JD, Whitman JM. Psychometric properties of the Neck Disability Index and Numeric Pain Rating Scale in patients with mechanical neck pain. *Arch Phys Med Rehabil* 2008;**89**:69–74.
45. Clifton JC, Finley RJ, Gelfand G, Graham AJ, Inculet R, Malthaner R, *et al.* Development and validation of a disease-specific quality of life questionnaire (EQOL) for potentially curable patients with carcinoma of the esophagus. *Dis Esophagus* 2007;**20**:191–201.
46. Coelho RA, Siqueira FB, Ferreira PH, Ferreira ML. Responsiveness of the Brazilian-Portuguese version of the Oswestry Disability Index in subjects with low back pain. *Eur Spine J* 2008;**17**:1101–6.
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.
48. Colangelo KJ, Pope JE, Peschken C. The minimally important difference for patient reported outcomes in systemic lupus erythematosus including the HAQ-DI, pain, fatigue, and SF-36. *J Rheumatol* 2009;**36**:2231–7.

49. Colwell HH, Hunt BJ, Pasta DJ, Palo WA, Mathias SD, Joseph-Ridge N. Gout Assessment Questionnaire: initial results of reliability, validity and responsiveness. *Int J Clin Pract* 2006;**60**:1210–17.
50. Colwell HH, Mathias SD, Turner MP, Lu J, Wright N, Peeters M, *et al.* Psychometric evaluation of the FACT Colorectal Cancer Symptom Index (FCSI-9): reliability, validity, responsiveness, and clinical meaningfulness. *Oncologist* 2010;**15**:308–16.
51. Costelloe L, O'Rourke K, Kearney H, McGuigan C, Gribbin L, Duggan M, *et al.* The patient knows best: significant change in the physical component of the Multiple Sclerosis Impact Scale (MSIS-29 physical). *J Neurol Neurosurg Psychiatr* 2007;**78**:841–4.
52. Cramer J, Rosenheck R, Xu WC, Henderson W, Thomas J, Charney D. Detecting improvement in quality of life and symptomatology in schizophrenia. *Schizophr Bull* 2001;**27**:227–34.
53. Cramer JA, Hammer AE, Kustra RP. Improved mood states with lamotrigine in patients with epilepsy. *Epilepsy Behav* 2004;**5**:702–7.
54. Cramer JA, Hammer AE, Kustra RP. Quality of life improvement with conversion to lamotrigine monotherapy. *Epilepsy Behav* 2004;**5**:224–30.
55. Crossley KM, Bennell KL, Cowan SM, Green S. Analysis of outcome measures for persons with patellofemoral pain: which are reliable and valid? *Arch Phys Med Rehabil* 2004;**85**:815–22.
56. Dawson J, Fitzpatrick R, Carr A. The assessment of shoulder instability. The development and validation of a questionnaire. *J Bone Joint Surg Br* 1999;**81**:420–6.
57. De Boer MR, de Vet HC, Terwee CB, Moll AC, Iker-Dieben HJ, van Rens GH. Changes to the subscales of two vision-related quality of life questionnaires are proposed. *J Clin Epidemiol* 2005;**58**:1260–8.
58. de Boer YA, Hazes JM, Winia PC, Brand R, Rozing PM. Comparative responsiveness of four elbow scoring instruments in patients with rheumatoid arthritis. *J Rheumatol* 2001;**28**:2616–23.
59. De La Loge C, Trudeau E, Marquis P, Revicki DA, Rentz AM, Stanghellini V, *et al.* Responsiveness and interpretation of a quality of life questionnaire specific to upper gastrointestinal disorders. *Clin Gastroenterol Hepatol* 2004;**2**:778–86.
60. de Lemos J, Tweeddale M, Chittock D. Measuring quality of sedation in adult mechanically ventilated critically ill patients. The Vancouver Interaction and Calmness Scale. Sedation Focus Group. *J Clin Epidemiol* 2000;**53**:908–19.
61. Dempster H, Porepa M, Young N, Feldman BM. The clinical meaning of functional outcome scores in children with juvenile arthritis. *Arthritis Rheum* 2001;**44**:1768–74.
62. DeRogatis LR, Graziottin A, Bitzer J, Schmitt S, Koochaki PE, Rodenberg C. Clinically relevant changes in sexual desire, satisfying sexual activity and personal distress as measured by the profile of female sexual function, sexual activity log, and personal distress scale in postmenopausal women with hypoactive sexual desire disorder. *J Sex Med* 2009;**6**:175–83.
63. Desmond DW, Tatemichi TK, Stern Y, Sano M. The determination of clinically meaningful cognitive decline: development and use of an alternative method. *Arch Clin Neuropsychol* 1995;**10**:535–42.
64. de Vet HC, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RW, *et al.* Three ways to quantify uncertainty in individually applied 'minimally important change' values. *J Clin Epidemiol* 2010;**63**:37–45.
65. Deyo RA, Inui TS. Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Serv Res* 1984;**19**:275–89.
66. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Control Clin Trials* 1991;**12**(Suppl.):158S.

67. Dommasch ED, Shin DB, Troxel AB, Margolis DJ, Gelfand JM. Reliability, validity and responsiveness to change of the Patient Report of Extent of Psoriasis Involvement (PREPI) for measuring body surface area affected by psoriasis. *Br J Dermatol* 2010;**162**:835–42.
68. Donovan KA, Jacobsen PB, Small BJ, Munster PN, Andrykowski MA. Identifying clinically meaningful fatigue with the Fatigue Symptom Inventory. *J Pain Symptom Manage* 2008;**36**:480–7.
69. Doyle C, Crump M, Pintilie M, Oza AM. Does palliative chemotherapy palliate? Evaluation of expectations, outcomes, and costs in women receiving chemotherapy for advanced ovarian cancer. *J Clin Oncol* 2001;**19**:1266–74.
70. Drossman DA, Patrick DL, Whitehead WE, Toner BB, Diamant NE, Hu Y, *et al.* Further validation of the IBS-QOL: a disease-specific quality-of-life questionnaire. *Am J Gastroenterol* 2000;**95**:999–1007.
71. Drulovic J, Riise T, Nortvedt M, Pekmezovic T, Manigoda M. Self-rated physical health predicts change in disability in multiple sclerosis. *Mult Scler* 2008;**14**:999–1002.
72. Eberle E, Ottillinger B. Clinically relevant change and clinically relevant difference in knee osteoarthritis. *Osteoarthritis Cartilage* 1999;**7**:502–3.
73. Emshoff R, Emshoff I, Bertram S. Estimation of clinically important change for visual analog scales measuring chronic temporomandibular disorder pain. *J Orofac Pain* 2010;**24**:262–9.
74. Escobar A, Quintana JM, Bilbao A, Aróstegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after total knee replacement. *Osteoarthritis Cartilage* 2007;**15**:273–80.
75. Falissard B, Lukasiewicz M, Corruble E. The MDP75: a new approach in the determination of the minimal clinically meaningful difference in a scale or a questionnaire. *J Clin Epidemiol* 2003;**56**:618–21.
76. Faraone SV, Pliszka SR, Olvera RL, Skolnik R, Biederman J. Efficacy of Adderall and methylphenidate in attention deficit hyperactivity disorder: a reanalysis using drug–placebo and drug–drug response curve methodology. *J Child Adolesc Psychopharmacol* 2001;**11**:171–80.
77. Farrar JT, Portenoy RK, Berlin JA, Kinman JL, Strom BL. Defining the clinically important difference in pain outcome measures. *Pain* 2000;**88**:287–94.
78. Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain* 2001;**94**:149–58.
79. Farrar JT, Berlin JA, Strom BL. Clinically important changes in acute pain outcome measures: a validation study. *J Pain Symptom Manage* 2003;**25**:406–11.
80. Farrar JT, Troxel AB, Stott C, Duncombe P, Jensen MP. Validity, reliability, and clinical importance of change in a 0–10 numeric rating scale measure of spasticity: a post hoc analysis of a randomized, double-blind, placebo-controlled trial. *Clin Ther* 2008;**30**:974–85.
81. Farrar JT, Pritchett YL, Robinson M, Prakash A, Chappell A. The clinical importance of changes in the 0 to 10 numeric rating scale for worst, least, and average pain intensity: analyses of data from clinical trials of duloxetine in pain disorders. *J Pain* 2010;**11**:109–18.
82. Filocamo G, Davis, S, Pistorio A, Bertamino M, Ruperto N, *et al.* Evaluation of 21-numbered circle and 10-centimeter horizontal line visual analog scales for physician and parent subjective ratings in juvenile idiopathic arthritis. *J Rheumatol* 2010;**37**:1534–41.
83. Filocamo G, Schiappapietra B, Bertamino M, Pistorio A, Ruperto N, Magni-Manzoni S, *et al.* A new short and simple health-related quality of life measurement for paediatric rheumatic diseases: initial validation in juvenile idiopathic arthritis. *Rheumatology (Oxford)* 2010;**49**:1272–80.
84. Fisher K. Assessing clinically meaningful change following a programme for managing chronic pain. *Clin Rehabil* 2008;**22**:252–9.

85. Fletcher KE, French CT, Irwin RS, Corapi KM, Norman GR. A prospective global measure, the Punum Ladder, provides more valid assessments of quality of life than a retrospective transition measure. *J Clin Epidemiol* 2010;**63**:1123–31.
86. Forouzanfar T, Weber WEJ, Kemler M, van Kleef M. What is a meaningful pain reduction in patients with complex regional pain syndrome type 1? *Clin J Pain* 2003;**19**:281–5.
87. Fortin PR, Abrahamowicz M, Clarke AE, Neville C, Du Berger R, Fraenkel L, *et al.* Do lupus disease activity measures detect clinically important change? *J Rheumatol* 2000;**27**:1421–8.
88. Fritz JM, Hebert J, Koppenhaver S, Parent E. Beyond minimally important change: defining a successful outcome of physical therapy for patients with low back pain. *Spine* 2009;**34**:2803–9.
89. Gallagher EJ, Liebman M, Bijur PE. Prospective validation of clinically important changes in pain severity measured on a visual analog scale. *Ann Emerg Med* 2001;**38**:633–8.
90. Gallagher EJ, Bijur PE, Latimer C, Silver W. Reliability and validity of a visual analog scale for acute abdominal pain in the ED. *Am J Emerg Med* 2002;**20**:287–90.
91. Gatchel RJ, Mayer TG. Testing minimal clinically important difference: consensus or conundrum? *Spine J* 2010;**10**:321–7.
92. Gill M, Windemuth R, Steele R, Green SM. A comparison of the Glasgow Coma Scale score to simplified alternative scores for the prediction of traumatic brain injury outcomes. *Ann Emerg Med* 2005;**45**:37–42.
93. Glassman SD, Copay AG, Berven SH, Polly DW, Subach BR, Carreon LY. Defining substantial clinical benefit following lumbar spine arthrodesis. *J Bone Joint Surg Am* 2008;**90**:1839–47.
94. Goldsmith CH, Boers M, Bombardier C, Tugwell P. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. OMERACT Committee. *J Rheumatol* 1993;**20**:561–5.
95. Goldsmith KA, Dyer MT, Schofield PM, Buxton MJ, Sharples LD. Relationship between the EQ-5D index and measures of clinical outcomes in selected studies of cardiovascular interventions. *Health Qual Life Outcomes* 2009;**7**:96.
96. Goligher EC, Pouchot J, Brant R, Kherani RB, Zubieta JA, Lacaille D, *et al.* Minimal clinically important difference for 7 measures of fatigue in patients with systemic lupus erythematosus. *J Rheumatol* 2008;**35**:635–42.
97. Gong GW, Young NL, Dempster H, Porepa M, Feldman BM. The Quality of My Life questionnaire: the minimal clinically important difference for pediatric rheumatology patients. *J Rheumatol* 2007;**34**:581–7.
98. Gowland C, Huijbregts M, McClung A, McNern A. Measuring clinically important change with the Chedoke-McMaster Stroke Assessment. *Can J Rehabil* 1993;**7**:14–16.
99. Greco NJ, Anderson AF, Mann BJ, Cole BJ, Farr J, Nissen CW, *et al.* Responsiveness of the International Knee Documentation Committee Subjective Knee Form in comparison to the Western Ontario and McMaster Universities Osteoarthritis Index, modified Cincinnati Knee Rating System, and Short Form 36 in patients with focal articular cartilage defects. *Am J Sports Med* 2010;**38**:891–902.
100. Gummeson C, Atroshi I, Ekdahl C. The disabilities of the arm, shoulder and hand (DASH) outcome questionnaire: longitudinal construct validity and measuring self-rated health change after surgery. *BMC Musculoskelet Disord* 2003;**4**:11.
101. Hagen KB, Smedstad LM, Uhlig T, Kvien TK. The responsiveness of health status measures in patients with rheumatoid arthritis: comparison of disease-specific and generic instruments. *J Rheumatol* 1999;**26**:1474–80.

102. Hart DL, Wang YC, Stratford PW, Mioduski JE. A computerized adaptive test for patients with hip impairments produced valid and responsive measures of function. *Arch Phys Med Rehabil* 2008;**89**:2129–39.
103. Hart DL, Wang YC, Stratford PW, Mioduski JE. Computerized adaptive test for patients with foot or ankle impairments produced valid and responsive measures of function. *Qual Life Res* 2008;**17**:1081–91.
104. Hart DL, Wang YC, Stratford PW, Mioduski JE. Computerized adaptive test for patients with knee impairments produced valid and responsive measures of function. *J Clin Epidemiol* 2008;**61**:1113–24.
105. Hawthorne G, Osborne R. Population norms and meaningful differences for the Assessment of Quality of Life (AQoL) measure. *Aust N Z J Public Health* 2005;**29**:136–42.
106. Hayran O, Mumcu G, Inanc N, Ergun T, Direskeneli H. Assessment of minimal clinically important improvement by using Oral Health Impact Profile-14 in Behcet's disease. *Clin Exp Rheumatol* 2009;**27**(2 Suppl. 3):S79–84.
107. Hazell P, Lewin T, Sly K. What is a clinically important level of improvement in symptoms of attention-deficit/hyperactivity disorder? *Aust N Z J Psychiatry* 2005;**39**:354–8.
108. Higgins PDR, Schwartz M, Mapili J, Krokos I, Leung J, Zimmermann EM. Patient defined dichotomous end points for remission and clinical improvement in ulcerative colitis. *Gut* 2005;**54**:782–8.
109. Hiroe T, Kojima M, Yamamoto I, Nojima S, Kinoshita Y, Hashimoto N, *et al.* Gradations of clinical severity and sensitivity to change assessed with the Beck Depression Inventory-II in Japanese patients with depression. *Psychiatry Res* 2005;**135**:229–35.
110. Hoffman DL, Sadosky A, Dukes EM, Alvir J. How do changes in pain severity levels correspond to changes in health status and function in patients with painful diabetic peripheral neuropathy? *Pain* 2010;**149**:194–201.
111. Homma Y, Koyama N. Minimal clinically important change in urinary incontinence detected by a quality of life assessment tool in overactive bladder syndrome with urge incontinence. *Neurourol Urodyn* 2006;**25**:228–35.
112. Hsieh YW, Wang CH, Sheu CF, Hsueh IP, Hsieh CL. Estimating the minimal clinically important difference of the Stroke Rehabilitation Assessment of Movement measure. *Neurorehabil Neural Repair* 2008;**22**:723–7.
113. Huang IC, Thompson LA, Chi YY, Knapp CA, Revicki DA, Seid M, *et al.* The linkage between pediatric quality of life and health conditions: establishing clinically meaningful cutoff scores for the PedsQL. *Value Health* 2009;**12**:773–81.
114. Huber AM, Feldman BM, Rennebohm RM, Hicks JE, Lindsley CB, Perez MD, *et al.* Validation and clinical significance of the Childhood Myositis Assessment Scale for assessment of muscle function in the juvenile idiopathic inflammatory myopathies. *Arthritis Rheum* 2004;**50**:1595–603.
115. Irrgang JJ, Anderson AF, Boland AL, Harner CD, Neyret P, Richmond JC, *et al.* Responsiveness of the International Knee Documentation Committee Subjective Knee Form. *Am J Sports Med* 2006;**34**:1567–73.
116. Iyer LV, Haley SM, Watkins MP, Dumas HM. Establishing minimal clinically important differences for scores on the pediatric evaluation of disability inventory for inpatient rehabilitation. *Phys Ther* 2003;**83**:888–98.
117. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1989;**10**:407–15.

118. Jaeschke R, Guyatt GH, Keller J, Singer J. Interpreting changes in quality-of-life score in N of 1 randomized trials. *Control Clin Trials* 1991;**12**(Suppl.):233S.
119. Jensen MP, Chen C, Brugger AM. Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain. *J Pain* 2003;**4**:407–14.
120. John MT, Reissmann DR, Steele J. An approach to define clinical significance in prosthodontics. *J Prosthodont* 2009;**18**:455–60.
121. Jones G, Jenkinson C, Kennedy S. Evaluating the responsiveness of the Endometriosis Health Profile Questionnaire: the EHP-30. *Qual Life Res* 2004;**13**:705–13.
122. Jowett SL, Seal CJ, Barton JR, Welfare MR. The short inflammatory bowel disease questionnaire is reliable and responsive to clinically important change in ulcerative colitis. *Am J Gastroenterol* 2001;**96**:2921–8.
123. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. *J Clin Epidemiol* 1994;**47**:81–7.
124. Juniper EF, Chauhan A, Neville E, Chatterjee A, Svensson K, rk AC *et al*. Clinicians tend to overestimate improvements in asthma control: an unexpected observation. *Primary Care Respirat J* 2004;**13**:181–4.
125. Karras DJ, Sammon ME, Terregino CA, Lopez BL, Griswold SK, Arnold GK. Clinically meaningful changes in quantitative measures of asthma severity. *Acad Emerg Med* 2000;**7**:327–34.
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.
127. Khan NA, Yazici Y, Calvo-Alen J, Dadonienė J, Gossec L, Hansen TM, *et al*. Reevaluation of the role of duration of morning stiffness in the assessment of rheumatoid arthritis activity. *J Rheumatol* 2009;**36**:2435–42.
128. Khanna D, Furst DE, Hays RD, Park GS, Wong WK, Seibold JR, *et al*. Minimally important difference in diffuse systemic sclerosis: results from the D-penicillamine study. *Ann Rheum Dis* 2006;**65**:1325–9.
129. Khanna D, Furst DE, Wong WK, Tsevat J, Clements PJ, Park GS, *et al*. Reliability, validity, and minimally important differences of the SF-6D in systemic sclerosis. *Qual Life Res* 2007;**16**:1083–92.
130. Khanna D, Tseng CH, Furst DE, Clements PJ, Elashoff R, Roth M, *et al*. Minimally important differences in the Mahler's Transition Dyspnoea Index in a large randomized controlled trial – results from the Scleroderma Lung Study. *Rheumatology (Oxford)* 2009;**48**:1537–40.
131. Khanna PP, Maranian P, Gregory J, Khanna D. The minimally important difference and patient acceptable symptom state for the Raynaud's condition score in patients with Raynaud's phenomenon in a large randomised controlled clinical trial. *Ann Rheum Dis* 2010;**69**:588–91.
132. Kingsberg S, Shifren J, Wekselman K, Rodenberg C, Koochaki P, Derogatis L. Evaluation of the clinical relevance of benefits associated with transdermal testosterone treatment in postmenopausal women with hypoactive sexual desire disorder. *J Sex Med* 2007;**4**:1001–8.
133. Kirby S, Chuang-Stein C, Morris M. Determining a minimum clinically important difference between treatments for a patient-reported outcome. *J Biopharm Stat* 2010;**20**:1043–54.
134. Knox SA, King MT. Validation and calibration of the SF-36 health transition question against an external criterion of clinical change in health status. *Qual Life Res* 2009;**18**:637–45.
135. Kosinski M, Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE Jr. Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. *Arthritis Rheum* 2000;**43**:1478–87.

136. Kovacs FM, Abaira V, Royuela A, Corcoll J, Alegre L, Tomas M, *et al.* Minimum detectable and minimal clinically important changes for pain in patients with nonspecific neck pain. *BMC Musculoskelet Disord* 2008;**9**:43.
137. Kovacs FM, Bago J, Royuela A, Seco J, Gimenez S, Muriel A, *et al.* Psychometric characteristics of the Spanish version of instruments to measure neck pain disability. *BMC Musculoskelet Disord* 2008;**9**:42.
138. Kragt JJ, Nielsen IM, van der Linden FA, Uitdehaag BM, Polman CH. How similar are commonly combined criteria for EDSS progression in multiple sclerosis? *Mult Scler* 2006;**12**:782–6.
139. Kundhal PS, Critch JN, Zachos M, Otley AR, Stephens D, Griffiths AM. Pediatric Crohn Disease Activity Index: responsive to short-term change. *J Pediatr Gastroenterol Nutr* 2003;**36**:83–9.
140. Kuzniar T, Patkowski J, Liebhart J, Wytrychowski K, Dobek R, Slusarz R, *et al.* [Validation of the Polish version of St. George's respiratory questionnaire in patients with bronchial asthma.] *Pneumonol Alergol Pol* 1999;**67**:497–503.
141. Kvam AK, Wisløff F, Fayers PM. Minimal important differences and response shift in health-related quality of life; a longitudinal study in patients with multiple myeloma. *Health Qual Life Outcomes* 2010;**8**:79.
142. Kvamme MK, Kristiansen IS, Lie E, Kvien TK. Identification of cutpoints for acceptable health status and important improvement in patient-reported outcomes, in rheumatoid arthritis, psoriatic arthritis, and ankylosing spondylitis. *J Rheumatol* 2010;**37**:26–31.
143. Kwok T, Pope JE. Minimally important difference for patient-reported outcomes in psoriatic arthritis: Health Assessment Questionnaire and pain, fatigue, and global visual analog scales. *J Rheumatol* 2010;**37**:1024–8.
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.
145. Landorf KB, Radford JA, Hudson S. Minimal important difference (MID) of two commonly used outcome measures for foot problems. *J Foot Ankle Res* 2010;**3**:7.
146. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. *BMC Musculoskelet Disord* 2006;**7**:82.
147. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Danish version of the Oswestry disability index for patients with low back pain. Part 2: sensitivity, specificity and clinically significant improvement in two low back pain populations. *Eur Spine J* 2006;**15**:1717–28.
148. Lee BB, King MT, Simpson JM, Haran MJ, Stockler MR, Marial O, *et al.* Validity, responsiveness, and minimal important difference for the SF-6D health utility scale in a spinal cord injured population. *Value Health* 2008;**11**:680–8.
149. Lee JS, Hobden E, Stiell IG, Wells GA. Clinically important change in the visual analog scale after adequate pain control. *Acad Emerg Med* 2003;**10**:1128–30.
150. Leggin BG, Michener LA, Shaffer MA, Brenneman SK, Iannotti JP, Williams GR Jr. The Penn shoulder score: reliability and validity. *J Orthop Sports Phys Ther* 2006;**36**:138–51.
151. Lehman LA, Sindhu BS, Shechtman O, Romero S, Velozo CA. A comparison of the ability of two upper extremity assessments to measure change in function. *J Hand Ther* 2010;**23**:31–9.
152. Levine SZ, Rabinowitz J, Engel R, Etschel E, Leucht S. Extrapolation between measures of symptom severity and change: an examination of the PANSS and CGI. *Schizophr Res* 2008;**98**:318–22.

153. Lewis JD, Chuai S, Nessel L, Lichtenstein GR, Aberra FN, Ellenberg JH. Use of the noninvasive components of the Mayo score to assess clinical response in ulcerative colitis. *Inflamm Bowel Dis* 2008;**14**:1660–6.
154. Lubeck DP, Whitmore K, Sant GR, Alvarez-Horine S, Lai C. Psychometric validation of the O'leary-Sant interstitial cystitis symptom index in a clinical trial of pentosan polysulfate sodium. *Urology* 2001;**57**(6 Suppl. 1):62–6.
155. Luo N, Tan LC, Zhao Y, Lau PN, Au WL, Li SC. Determination of the longitudinal validity and minimally important difference of the 8-item Parkinson's Disease Questionnaire (PDQ-8). *Mov Disord* 2009;**24**:183–7.
156. Luo N, Johnson J, Coons SJ. Using instrument-defined health state transitions to estimate minimally important differences for four preference-based health-related quality of life instruments. *Med Care* 2010;**48**:365–71.
157. Maillefert JF, Nguyen M, Gueguen A, Berdah L, Lequesne M, Mazières B, et al. Relevant change in radiological progression in patients with hip osteoarthritis. II. Determination using an expert opinion approach. *Rheumatology (Oxford)* 2002;**41**:148–52.
158. Maksymowych WP, Richardson R, Mallon C, Van Der Heijde D, Boonen A. Evaluation and validation of the patient acceptable symptom state (PASS) in patients with ankylosing-spondylitis. *Arthritis Rheum* 2007;**57**:133–9.
159. Mancuso CA, Peterson MG. Different methods to assess quality of life from multiple follow-ups in a longitudinal asthma study. *J Clin Epidemiol* 2004;**57**:45–54.
160. Mannion AF, Junge A, Grob D, Dvorak J, Fairbank JC. Development of a German version of the Oswestry Disability Index. Part 2: sensitivity to change after spinal surgery. *Eur Spine J* 2006;**15**:66–73.
161. Mannion AF, Porchet F, Kleinstück FS, Lattig F, Jeszenszky D, Bartanusz V, et al. The quality of spine surgery from the patient's perspective: part 2. Minimal clinically important difference for improvement and deterioration as measured with the Core Outcome Measures Index. *Eur Spine J* 2009;**18**(Suppl. 3):374–9.
162. Martin RL, Philippon MJ. Evidence of reliability and responsiveness for the hip outcome score. *Arthroscopy* 2008;**24**:676–82.
163. Martinez-Martin P, Prieto L, Forjaz MJ. Longitudinal metric properties of disability rating scales for Parkinson's disease. *Value Health* 2006;**9**:386–93.
164. Metz SM, Wyrwich KW, Babu AN, Kroenke K, Tierney WM, Wolinsky FD. A comparison of traditional and Rasch cut points for assessing clinically important change in health-related quality of life among patients with asthma. *Qual Life Res* 2006;**15**:1639–49.
165. Mintken PE, Glynn P, Cleland JA. Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain. *J Shoulder Elbow Surg* 2009;**18**:920–6.
166. Morrow SA, Drake A, Zivadinov R, Munschauer F, Weinstock-Guttman B, Benedict RHB. Predicting loss of employment over three years in multiple sclerosis: clinically meaningful cognitive decline. *Clin Neuropsychol* 2010;**24**:1131–45.
167. Mulhall JP, Goldstein I, Bushmakina AG, Cappelleri JC, Hvidsten K. Validation of the erection hardness score. *J Sex Med* 2007;**4**:1626–34.
168. Mulhall JP, King R, Kirby M, Hvidsten K, Symonds T, Bushmakina AG, et al. Evaluating the sexual experience in men: validation of the sexual experience questionnaire. *J Sex Med* 2008;**5**:365–76.

169. Naylor CD, Llewellyn-Thomas HA. Can there be a more patient-centred approach to determining clinically important effect sizes for randomized treatment trials? *J Clin Epidemiol* 1994;**47**:787–95.
170. Nemann HJ, Griffith L, Jaeschke R, Goldstein R, Stubbing D, Guyatt GH. Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. *J Clin Epidemiol* 2003;**56**:1170–6.
171. Newell D, Bolton JE. Responsiveness of the Bournemouth questionnaire in determining minimal clinically important change in subgroups of low back pain patients. *Spine* 2010;**35**:1801–6.
172. Nicholl D, Nasrallah H, Nuamah I, Akhras KS, Gagnon DD, Gopal S. Personal and social functioning in schizophrenia: defining a clinically meaningful measure of maintenance in relapse prevention. *Curr Med Res Opin* 2010;**26**:1471–84.
173. Özyürekoğlu T, McCabe SJ, Goldsmith LJ, LaJoie AS. The minimal clinically important difference of the Carpal Tunnel Syndrome Symptom Severity Scale. *J Hand Surg Am* 2006;**31**:733–8.
174. Pandina GJ, Garibaldi GM, Revicki DA, Kleinman L, Turkoz I, Kujawa MJ, et al. Psychometric evaluation of a Patient-Rated Most Troubling Symptom Scale for Depression: findings from a secondary analysis of a clinical trial. *Int Clin Psychopharmacol* 2010;**25**:51–9.
175. Patrick D, Gagnon DD, Zagari MJ. Improvements in quality of life associated with epoetin alfa treatment are clinically, as well as statistically, significant. *Curr Med Res Opin Suppl* 2005;**21**:S3–5.
176. Patrick DL, Martin ML, Bushnell DM, Yalcin I, Wagner TH, Buesching DP. Quality of life of women with urinary incontinence: further development of the incontinence quality of life instrument (I-QOL). *Urology* 1999;**53**:71–6.
177. Patrick DL, Gagnon DD, Zagari MJ, Mathijs R, Sweetenham J, Epoetin Alfa Study Group. Assessing the clinical significance of health-related quality of life (HrQOL) improvements in anaemic cancer patients receiving epoetin alfa. *Eur J Cancer* 2003;**39**:335–45.
178. Pavy S, Brophy S, Calin A. Establishment of the minimum clinically important difference for the bath ankylosing spondylitis indices: a prospective study. *J Rheumatol* 2005;**32**:80–5.
179. Perez T, Arnould B, Grosbois JM, Bosch V, Guillemin I, Bravo ML, et al. Validity, reliability, and responsiveness of a new short Visual Simplified Respiratory Questionnaire (VSRQ) for health-related quality of life assessment in chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis* 2009;**4**:9–18.
180. Peto V, Jenkinson C, Fitzpatrick R. Determining minimally important differences for the PDQ-39 Parkinson's disease questionnaire. *Age Ageing* 2001;**30**:299–302.
181. Petrou S, Morrell J, Spiiby H. Assessing the empirical validity of alternative multi-attribute utility measures in the maternity context. *Health Qual Life Outcomes* 2009;**7**:40.
182. Picado C, Badiola C, Perulero N, Sastre J, Olaguibel JM, López Viña A, et al. Validation of the Spanish version of the Asthma Control Questionnaire. *Clin Ther* 2008;**30**:1918–31.
183. Piva SR, Gil AB, Moore CG, Fitzgerald GK. Responsiveness of the activities of daily living scale of the knee outcome survey and numeric pain rating scale in patients with patellofemoral pain. *J Rehabil Med* 2009;**41**:129–35.
184. Pope JE, Khanna D, Norrie D, Ouimet JM. The minimally important difference for the health assessment questionnaire in rheumatoid arthritis clinical practice is smaller than in randomized controlled trials. *J Rheumatol* 2009;**36**:254–9.
185. Potter LP, Mathias SD, Raut M, Kianifard F, Tavakkol A. The OnyCOE-t questionnaire: responsiveness and clinical meaningfulness of a patient-reported outcomes questionnaire for toenail onychomycosis. *Health Qual Life Outcomes* 2006;**4**:50.

186. Pouchot J, Kherani RB, Brant R, Lacaille D, Lehman AJ, Ensworth S, *et al.* Determination of the minimal clinically important difference for seven fatigue measures in rheumatoid arthritis. *J Clin Epidemiol* 2008;**61**:705–13.
187. Puente-Maestu L, Villar F, de Miguel J, Stringer WW, Sanz P, Sanz ML, *et al.* Clinical relevance of constant power exercise duration changes in COPD. *Eur Respir J* 2009;**34**:340–5.
188. Purcell A, Fleming J, Bennett S, Burmeister B, Haines T. Determining the minimal clinically important difference criteria for the Multidimensional Fatigue Inventory in a radiotherapy population. *Support Care Cancer* 2010;**18**:307–15.
189. Rambo WW, Pinto JN. Employees' perception of pay increases. *J Occup Psychol* 1989;**62**:135–45.
190. Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. *J Clin Epidemiol* 1996;**49**:1215–19.
191. Riddle DL, Stratford PW, Binkley JM. Sensitivity to change of the Roland–Morris back pain questionnaire: part 2. *Phys Ther* 1998;**78**:1197–207.
192. Ringash J, Bezjak A, O'Sullivan B, Redelmeier DA. Interpreting differences in quality of life: the FACT-H&N in laryngeal cancer patients. *Qual Life Res* 2004;**13**:725–33.
193. Ringash J, O'Sullivan B, Bezjak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. *Cancer* 2007;**110**:196–202.
194. Roberts G, Hurley C, Lack G. Development of a quality-of-life assessment for the allergic child or teenager with multisystem allergic disease. *J Allergy Clin Immunol* 2003;**111**:491–7.
195. Rockwood K, Howlett S, Stadnyk K, Carver D, Powell C, Stolee P. Responsiveness of goal attainment scaling in a randomized controlled trial of comprehensive geriatric assessment. *J Clin Epidemiol* 2003;**56**:736–43.
196. Rockwood K, Fay S, Gorman M. The ADAS-cog and clinically meaningful change in the VISTA clinical trial of galantamine for Alzheimer's disease. *Int J Geriatr Psychiatry* 2010;**25**:191–201.
197. Rodrigues G, Bezjak A, Osoba D, Catton P, Tsuji D, Taylor D, *et al.* The relationship of changes in EORTC QLQ-C30 scores to ratings on the Subjective Significance Questionnaire in men with localized prostate cancer. *Qual Life Res* 2004;**13**:1235–46.
198. Roy JS, Macdermid JC, Faber KJ, Drosdowech DS, Athwal GS. The simple shoulder test is responsive in assessing change following shoulder arthroplasty. *J Orthop Sports Phys Ther* 2010;**40**:413–21.
199. Russell IJ, Crofford LJ, Leon T, Cappelleri JC, Bushmakina AG, Whalen E, *et al.* The effects of pregabalin on sleep disturbance symptoms among individuals with fibromyalgia syndrome. *Sleep Med* 2009;**10**:604–10.
200. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *Eur J Pain* 2004;**8**:283–91.
201. Santana MJ, Au HJ, Dharma-Wardene M, Hewitt JD, Dupere D, Hanson J, *et al.* Health-related quality of life measures in routine clinical care: can FACT-Fatigue help to assess the management of fatigue in cancer patients? *Int J Technol Assess Health Care* 2009;**25**:90–6.
202. Santanello NC, Zhang J, Seidenberg B, Reiss TF, Barber BL. What are minimal important changes for asthma measures in a clinical trial? *Eur Respir J* 1999;**14**:23–7.
203. Schrag A, Sampaio C, Counsell N, Poewe W. Minimal clinically important change on the unified Parkinson's disease rating scale. *Mov Disord* 2006;**21**:1200–7.
204. Schwartz AL, Meek PM, Nail LM, Fargo J, Lundquist M, Donofrio M, *et al.* Measurement of fatigue: determining minimally important clinical differences. *J Clin Epidemiol* 2002;**55**:239–44.

205. Sekhon S, Pope J, Canadian Scleroderma Research Group, Baron M. The minimally important difference in clinical practice for patient-centered outcomes including health assessment questionnaire, fatigue, pain, sleep, global visual analog scale, and SF-36 in scleroderma. *J Rheumatol* 2010;**37**:591–8.
206. Shauver MJ, Chung KC. The minimal clinically important difference of the Michigan hand outcomes questionnaire. *J Hand Surg Am* 2009;**34**:509–14.
207. Sheldon EA, Bird SR, Smugar SS, Tershakovec AM. Correlation of measures of pain, function, and overall response – results pooled from two identical studies of etoricoxib in chronic low back pain. *Spine* 2008;**33**:533–8.
208. Shirai K, Iso H, Fukuda H, Toyoda Y, Takatorige T, Tatara K. Factors associated with ‘Ikigai’ among members of a public temporary employment agency for seniors (Silver Human Resources Centre) in Japan; gender differences. *Health Qual Life Outcomes* 2006;**4**:12.
209. Singh SJ, Jones PW, Evans R, Morgan MD. Minimum clinically important improvement for the incremental shuttle walking test. *Thorax* 2008;**63**:775–7.
210. Sloman R, Wruble AW, Rosen G, Rom M. Determination of clinically meaningful levels of pain reduction in patients experiencing acute postoperative pain. *Pain Manage Nurs* 2006;**7**:153–8.
211. Spiegel B, Bolus R, Harris LA, Lucak S, Naliboff B, Esrailian E, et al. Measuring irritable bowel syndrome patient-reported outcomes with an abdominal pain numeric rating scale. *Aliment Pharmacol Ther* 2009;**30**:1159–70.
212. Stratford PW, Binkley J, Solomon P, Gill C, Finch E. Assessing change over time in patients with low back pain. *Phys Ther* 1994;**74**:528–33.
213. Stratford PW, Levy DR. Assessing valid change over time in patients with lateral epicondylitis at the elbow. *Clin J Sport Med* 1994;**4**:88–91.
214. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland–Morris Back Pain Questionnaire: part 1. *Phys Ther* 1998;**78**:1186–96.
215. Tafazal SI, Sell PJ. Outcome scores in spinal surgery quantified: excellent, good, fair and poor in terms of patient-completed tools. *Eur Spine J* 2006;**15**:1653–60.
216. Tannenbaum C, Brouillette J, Michaud J, Korner-Bitensky N, Dumoulin C, Corcos J, et al. Responsiveness and clinical utility of the geriatric self-efficacy index for urinary incontinence. *J Am Geriatr Soc* 2009;**57**:470–5.
217. Tashjian RZ, Deloach J, Porucznik CA, Powell AP. Minimal clinically important differences (MCID) and patient acceptable symptomatic state (PASS) for visual analog scales (VAS) measuring pain in patients treated for rotator cuff disease. *J Shoulder Elbow Surg* 2009;**18**:927–32.
218. Tashjian RZ, Deloach J, Green A, Porucznik CA, Powell AP. Minimal clinically important differences in ASES and simple shoulder test scores after nonoperative treatment of rotator cuff disease. *J Bone Joint Surg Am* 2010;**92**:296–303.
219. ten Klooster PM, Drossaers-Bakker KW, Taal E, van de Laar MA. Patient-perceived satisfactory improvement (PPSI): interpreting meaningful change in pain from the patient’s perspective. *Pain* 2006;**121**:151–7.
220. Thienthong S, Pratheepawanit N, Limwattananon C, Maoleekoonpairaj S, Lertsanguansinchai P, Chanvej L. Pain and quality of life of cancer patients: a multi-center study in Thailand. *J Med Assoc Thai* 2006;**89**:1120–6.
221. Thomas K, Ruby J, Peter JV, Cherian AM. Comparison of disease-specific and a generic quality of life measure in patients with bronchial asthma. *Natl Med J India* 1995;**8**:258–60.

222. Thomson WM. Measuring change in dry-mouth symptoms over time using the Xerostomia Inventory. *Gerodontology* 2007;**24**:30–5.
223. Tilson JK, Sullivan KJ, Cen SY, Rose DK, Behrman AL, Wu SS, *et al.* Determination of the minimal clinically important difference in the lower extremity Fugl–Meyer Motor Score in the first 60 days post-stroke. *Stroke* 2008;**39**:693.
224. Tilson JK, Sullivan KJ, Cen SY, Rose DK, Koradia CH, Azen SP, *et al.* Meaningful gait speed improvement during the first 60 days poststroke: minimal clinically important difference. *Phys Ther* 2010;**90**:196–208.
225. Todd KH, Funk JP. The minimum clinically important difference in physician-assigned visual analog pain scores. *Acad Emerg Med* 1996;**3**:142–6.
226. Tractenberg RE, Jin S, Patterson M, Schneider LS, Gamst A, Thomas RG, *et al.* Qualifying change: a method for defining clinically meaningful outcomes of change score computation. *J Am Geriatr Soc* 2000;**48**:1478–82.
227. Tubach F, Ravaud P, Baron G, Falissard B, Logeart I, Bellamy N, *et al.* Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. *Ann Rheum Dis* 2005;**64**:29–33.
228. Tugwell P, Bombardier C, Buchanan WW, Goldsmith CH, Grace E, Hanna B. The MACTAR Patient Preference Disability Questionnaire – an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis. *J Rheumatol* 1987;**14**:446–51.
229. Tuli SK, Yerby SA, Katz JN. Methodological approaches to developing criteria for improvement in lumbar spinal stenosis surgery. *Spine* 2006;**31**:1276–80.
230. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, *et al.* Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. *J Clin Epidemiol* 2009;**62**:374–9.
231. Turner JA, Franklin G, Heagerty PJ, Wu R, Egan K, Fulton-Kehoe D, *et al.* The association between pain and disability. *Pain* 2004;**112**:307–14.
232. van Bennekom CA, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the rehabilitation activities profile and the Barthel index. *J Clin Epidemiol* 1996;**49**:39–44.
233. Vela LI, Denegar CR. The Disablement in the Physically Active Scale, part II: the psychometric properties of an outcomes scale for musculoskeletal injuries. *J Athl Train* 2010;**45**:630–41.
234. Ward MM, Marx AS, Barry NN. Identification of clinically important changes in health status using receiver operating characteristic curves. *J Clin Epidemiol* 2000;**53**:279–84.
235. Weatherall M, Marsh S, Shirtcliffe P, Williams M, Travers J, Beasley R. Quality of life measured by the St George's Respiratory Questionnaire and spirometry. *Eur Respir J* 2009;**33**:1025–30.
236. Weinreb NJ, Cappellini MD, Cox TM, Giannini EH, Grabowski GA, Hwu WL, *et al.* A validated disease severity scoring system for adults with type 1 Gaucher disease. *Genetics Med* 2010;**12**:44–51.
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.
238. Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. *J Rheumatol* 1993;**20**:557–60.
239. Welsing PM, Borm GF, van RP. Minimal clinically important difference in radiological progression of joint damage. A definition based on patient perspective. *J Rheumatol* 2006;**33**:501–7.

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.
241. Wilson HD, Mayer TG, Gatchel RJ. The lack of association between changes in functional outcomes and work retention in a chronic disabling occupational spinal disorder population: implications for the minimum clinical important difference. *Spine* 2010;**36**:474–80.
242. Witek TJ Jr, Mahler DA. Meaningful effect size and patterns of response of the transition dyspnea index. *J Clin Epidemiol* 2003;**56**:248–55.
243. Witek TJ Jr, Mahler DA. Minimal important difference of the transition dyspnoea index in a multinational clinical trial. *Eur Respir J* 2003;**21**:267–72.
244. Wolfe F, Michaud K, Strand V. Expanding the definition of clinical differences: from minimally clinically important differences to really important differences. Analyses in 8931 patients with rheumatoid arthritis. *J Rheumatol* 2005;**32**:583–9.
245. Wyrwich KW, Spratt DI, Gass M, Yu H, Bobula JD. Identifying meaningful differences in vasomotor symptoms among menopausal women. *Menopause* 2008;**15**:698–705.
246. Xu W, Collet JP, Shapiro S, Lin Y, Yang T, Wang C, *et al.* Validation and clinical interpretation of the St George's Respiratory Questionnaire among COPD patients, China. *Int J Tuberc Lung Dis* 2009;**13**:181–9.
247. Yalcin I, Patrick DL, Summers K, Kinchen K, Bump RC. Minimal clinically important differences in Incontinence Quality-of-Life scores in stress urinary incontinence. *Urology* 2006;**67**:1304–8.
248. Yalcin I, Peng G, Viktrup L, Bump RC. Reductions in stress urinary incontinence episodes: what is clinically important for women? *Neurourol Urodyn* 2010;**29**:344–7.
249. Yamaguchi N, Poudel KC, Poudel-Tandukar K, Shakya D, Ravens-Sieberer U, Jimba M. Reliability and validity of a Nepalese version of the Kiddo-KINDL in adolescents. *Biosci Trends* 2010;**4**:178–85.
250. Yamashita K, Ohzono K, Hiroshima K. Patient satisfaction as an outcome measure after surgical treatment for lumbar spinal stenosis: testing the validity and discriminative ability in terms of symptoms and functional status. *Spine* 2006;**31**:2602–8.
251. Young IA, Cleland JA, Michener LA, Brown C. Reliability, construct validity, and responsiveness of the neck disability index, patient-specific functional scale, and numeric pain rating scale in patients with cervical radiculopathy. *Am J Phys Med Rehabil* 2010;**89**:831–9.
252. Zangger P, Kachura JR, Bombardier C, Redelmeier DA, Badley EM, Bogoch ER. Assessing damage in individual joints in rheumatoid arthritis: a new method based on the Larsen system. *Joint Bone Spine* 2004;**71**:389–96.
253. Zanolli G. Outcome assessment in lumbar spine surgery. *Acta Orthop* 2005;**76**:2–47.

Distribution method (n = 171)

1. Anderson BS, Hunt JW, Phillips BM, Tudor S, Fairey R, Newman J, *et al.* Comparison of marine sediment toxicity test protocols for the amphipod *Rhepoxynius abronius* and the polychaete worm *Nereis (Neanthes) arenaceodentata*. *Environ Toxicol Chem* 1998;**17**:859–66.
2. Ankuta GY, Abeles N. Client satisfaction, clinical significance, and meaningful change in psychotherapy. *Prof Psychol* 1993;**24**:70–4.
3. Arveschoug AK, Revsbech P, Brøchner-Mortensen J. Sources of variation in the determination of distal blood pressure measured using the strain gauge technique. *Clin Physiol* 1998;**18**:361–8.

4. Asenlof P, Denison E, Lindberg P. Idiographic outcome analyses of the clinical significance of two interventions for patients with musculoskeletal pain. *Behav Res Ther* 2006;**44**:947–65.
5. Atkins DC, Bedics JD, McGlinchey JB, Beauchaine TP. Assessing clinical significance: does it matter which method we use? *J Consult Clin Psychol* 2005;**73**:982–9.
6. Auleley GR, Duche A, Drape JL, Dougados M, Ravaud P. Measurement of joint space width in hip osteoarthritis: influence of joint positioning and radiographic procedure. *Rheumatology (Oxford)* 2001;**40**:414–19.
7. Auleley GR, Benbouazza K, Spoorenberg A, Collantes E, Hajjaj-Hassouni N, Van Der Heijde D, *et al.* Evaluation of the smallest detectable difference in outcome or process variables in ankylosing spondylitis. *Arthritis Rheum* 2002;**47**:582–7.
8. Bauer S, Lambert MJ, Nielsen SL. Clinical significance methods: a comparison of statistical techniques. *J Pers Assess* 2004;**82**:60–70.
9. Borstad JD, Mathlowetz KM, Minday LE, Prabhu B, Christopherson DE, Ludewig PM. Clinical measurement of posterior shoulder flexibility. *Manual Ther* 2007;**12**:386–9.
10. Bowersox NW, Saunders SM, Wojcik JV. An evaluation of the utility of statistical versus clinical significance in determining improvement in alcohol and other drug (AOD) treatment in correctional settings. *Alcohol Treat Q* 2009;**27**:113–29.
11. Brehm MA, Nollet F, Harlaar J. Energy demands of walking in persons with postpoliomyelitis syndrome: relationship with muscle strength and reproducibility. *Arch Phys Med Rehabil* 2006;**87**:136–40.
12. Bridges TS, Farrar JD. The influence of worm age, duration of exposure and endpoint selection on bioassay sensitivity for *Neanthes arenaceodentata* (Annelida: Polychaeta). *Environ Toxicol Chem* 1997;**16**:1650–8.
13. Bruynesteyn K, Landew R, van der Linden S, Van Der Heijde D. Radiography as primary outcome in rheumatoid arthritis: acceptable sample sizes for trials with 3 months' follow up. *Ann Rheum Dis* 2004;**63**:1413–18.
14. Burgoyne CF, Mercante DE, Thompson HW. Change detection in regional and volumetric disc parameters using longitudinal confocal scanning laser tomography. *Ophthalmology* 2002;**109**:455–66.
15. Castañeda S, González-Alvaro I, Rodríguez-Salvanés F, Quintana ML, Laffon A, Vadillo JA. Reproducibility of metacarpophalangeal bone mass measurements obtained by dual-energy X-ray absorptiometry in healthy volunteers and patients with early arthritis. *J Clin Densitom* 2007;**10**:298–305.
16. Cella D, Yount S, Sorensen M, Chartash E, Sengupta N, Grober J. Validation of the Functional Assessment of Chronic Illness Therapy Fatigue Scale relative to other instrumentation in patients with rheumatoid arthritis. *J Rheumatol* 2005;**32**:811–19.
17. Chan EY, Bridge PD, Dundas I, Pao CS, Healy MJ, McKenzie SA. Repeatability of airway resistance measurements made using the interrupter technique. *Thorax* 2003;**58**:344–7.
18. Choi KH, Buskey W, Johnson B. Evaluation of counseling outcomes at a university counseling center: the impact of clinically significant change on problem resolution and academic functioning. *J Counsel Psychol* 2010;**57**:297–303.
19. Cousens SN, Rosser DA, Murdoch IE, Laidlaw DA. A simple model to predict the sensitivity to change of visual acuity measurements. *Optomet Vis Sci* 2004;**81**:673–7.
20. Crosby RD, Kolotkin RL, Williams GR. An integrated method to determine meaningful changes in health-related quality of life. *J Clin Epidemiol* 2004;**57**:1153–60.

21. Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. *Phys Ther* 2002;**82**:8–24.
22. De Beurs E, Van Dyck R, Van Balkom AJLM, Lange A, Koele P. Assessing the clinical significance of outcome in agoraphobia research: A comparison of two approaches. *Behav Ther* 1994;**25**:147–58.
23. de Leon MCE, Diaz JMM, Ruiz EJC. A pilot study of the clinical and statistical significance of a program to reduce eating disorder risk factors in children. *Eat Weight Disord* 2008;**13**:111–18.
24. de Morton NA, Lane K. Validity and reliability of the de Morton Mobility Index in the subacute hospital setting in a geriatric evaluation and management population. *J Rehabil Med* 2010;**42**:956–61.
25. Denton DL, Fox JF, Fulk FA. Enhancing toxicity test performance by using a statistical criterion. *Environ Toxicol Chem* 2003;**22**:2323–8.
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.
27. Duru G, Fantino B. The clinical relevance of changes in the Montgomery–Asberg Depression Rating Scale using the minimum clinically important difference approach. *Curr Med Res Opin* 2008;**24**:1329–35.
28. Edgar DW, Briffa NK, Cole J, Tan MH, Khoo B, Goh J, *et al.* Measurement of acute edema shifts in human burn survivors – the reliability and sensitivity of bioimpedance spectroscopy as an objective clinical measure. *J Burn Care Res* 2009;**30**:818–23.
29. Ferguson RJ, Robinson AB, Splaine M. Use of the reliable change index to evaluate clinical significance in SF-36 outcomes. *Qual Life Res* 2002;**11**:509–16.
30. Ferguson SA, Marras WS, Burr DL, Woods S, Mendel E, Gupta P. Quantification of a meaningful change in low back functional impairment. *Spine* 2009;**34**:2060–5.
31. Fitzgerald MP, Ayuste D, Brubaker L. How do urinary diaries of women with an overactive bladder differ from those of asymptomatic controls? *BJU Int* 2005;**96**:365–7.
32. Fitzpatrick C, Simpson JM, Valentine JD, Ryder S, Peacock-Edwards T, Sidnell P, *et al.* The measurement properties and performance characteristics among older people of TURN180, a test of dynamic postural stability. *Clin Rehabil* 2005;**19**:412–18.
33. Fitzpatrick R, Norquist JM, Jenkinson C. Distribution-based criteria for change in health-related quality of life in Parkinson's disease. *J Clin Epidemiol* 2004;**57**:40–4.
34. Fritz JM, Piva SR. Physical impairment index: reliability, validity, and responsiveness in patients with acute low back pain. *Spine* 2003;**28**:1189–94.
35. Gabel CP, Michener LA, Burkett B, Neller A. The Upper Limb Functional Index: development and determination of reliability, validity, and responsiveness. *J Hand Ther* 2006;**19**:328–48.
36. Gagnon M, Ladouceur R. Defining clinically significant changes in the treatment of child stutterers. *Percept Mot Skills* 1991;**73**:375–8.
37. Geertzen JH, Dijkstra PU, Stewart RE, Groothoff JW, Ten Duis HJ, Eisma WH. Variation in measurements of range of motion: a study in reflex sympathetic dystrophy patients. *Clin Rehabil* 1998;**12**:254–64.
38. Glick ID, Clarkin JF, Haas GL, Spencer JH. Clinical significance of inpatient family intervention: conclusions from a clinical trial. *Hosp Community Psychiatry* 1993;**44**:869–73.

39. Glickstein J, Buyon J, Kim M, Friedman D, PRIDE investigators. The fetal Doppler mechanical PR interval: a validation study. *Fetal Diagn Ther* 2004;**19**:31–4.
40. Gnat R, Kuszewski M, Koczar R, Dziewonska A. Reliability of the passive knee flexion and extension tests in healthy subjects. *J Manipulative Physiol Ther* 2010;**33**:659–65.
41. Gonnelli S, Cepollaro C, Montagnani A, Martini S, Gennari L, Mangeri M, et al. Heel ultrasonography in monitoring alendronate therapy: a four-year longitudinal study. *Osteoporos Int* 2002;**13**:415–21.
42. Grooten WJ, Puttemans V, Larsson RJ. Reliability of isokinetic supine bench press in healthy women using the Ariel Computerized Exercise System. *Scand J Med Sci Sport* 2002;**12**:218–22.
43. Grotle M, Brox JI, Iltstad NK. Reliability, validity and responsiveness of the fear-avoidance beliefs questionnaire: methodological aspects of the Norwegian version. *J Rehabil Med* 2006;**38**:346–53.
44. Grundy CT, Lambert MJ, Grundy EM. Assessing clinical significance: application to the Hamilton Rating Scale for Depression. *J Ment Health* 1996;**5**:25–33.
45. Gully JR, Bottomley JP, Baird RB. Effects of sporophyll storage on giant kelp *Macrocystis pyrifera* (Agardh) bioassay. *Environ Toxicol Chem* 1999;**18**:1474–81.
46. Hafkenscheid A, Kuipers A, Marinkelle A. [Self-rating symptom inventories in single-case treatment efficacy studies: are the indices of 'clinical significance' and 'reliable change' applicable to clinical practice?] *Gedragstherapie* 1998;**31**:221–39.
47. Hanson ML, Sanderson H, Solomon KR. Variation, replication, and power analysis of *Myriophyllum* spp. microcosm toxicity data. *Environ Toxicol Chem* 2003;**22**:1318–29.
48. Harrill WC, Pillsbury HC III, McGuirt WF, Stewart MG. Radiofrequency turbinate reduction: a NOSE evaluation. *Laryngoscope* 2007;**117**:1912–19.
49. Hart DL. Test–retest reliability of an abbreviated self-report overall health status measure. *J Orthop Sport Phys Ther* 2003;**33**:734–44.
50. Hawley DR. Assessing change with preventive interventions: the reliable change index. *Fam Relat* 1995;**44**:278–84.
51. Hebert R, Spiegelhalter DJ, Brayne C. Setting the minimal metrically detectable change on disability rating scales. *Arch Phys Med Rehabil* 1997;**78**:1305–8.
52. Herpel LB, Kanner RE, Lee SM, Fessler HE, Sciurba FC, Connett JE, et al. Variability of spirometry in chronic obstructive pulmonary disease: results from two clinical trials. *Am J Respir Crit Care Med* 2006;**173**:1106–13.
53. Hicks GE, George SZ, Nevitt MA, Cauley JA, Vogt MT. Measurement of lumbar lordosis: inter-rater reliability, minimum detectable change and longitudinal variation. *J Spinal Disord Tech* 2006;**19**:501–6.
54. Hoiland-Carlsen PF, Lauritzen SL, Marving J, Rasmussen S, Hesse B, Folke K, et al. The reliability of measuring left ventricular ejection fraction by radionuclide cardiography: evaluation by the method of variance components. *Br Heart J* 1988;**59**:653–62.
55. Holmback AM, Lexell J. Reproducibility of isokinetic ankle dorsiflexor strength and fatigue measurements in healthy older subjects. *Isokinet Exerc Sci* 2007;**15**:263–70.
56. Holz O, Jörres R, Krause T, Magnussen H. Reproducibility of basal and induced DNA single-strand breaks detected by the single-cell gel electrophoresis assay in human peripheral mononuclear leukocytes. *Int Arch Occup Environ Health* 1995;**67**:305–10.
57. Horton AM. Estimation of clinical significance: a brief note. *Psychol Rep* 1980;**47**:141–2.

58. Hoss S, Jansch S, Moser T, Junker T, Rombke J. Assessing the toxicity of contaminated soils using the nematode *Caenorhabditis elegans* as test organism. *Ecotoxicol Environ Saf* 2009;**72**:1811–18.
59. Ijzerman MJ, Baardman G, van't Hof MA, Boom HB, Hermens HJ, Veltink PH. Validity and reproducibility of crutch force and heart rate measurements to assess energy expenditure of paraplegic gait. *Arch Phys Med Rehabil* 1999;**80**:1017–23.
60. Ijzerman MJ, Nene AV. Feasibility of the physiological cost index as an outcome measure for the assessment of energy expenditure during walking. *Arch Phys Med Rehabil* 2002;**83**:1777–82.
61. Isella V, Atzeni L, Iurlaro S, Villa ML, Russo A, Forapani E, *et al.* Assessing clinically relevant cognitive decline: preliminary data on a new method. *Neurol Sci* 2003;**24**:236–41.
62. Iverson GL, Sawyer DC, McCracken LM, Kozora E. Assessing depression in systemic lupus erythematosus: determining reliable change. *Lupus* 2001;**10**:266–71.
63. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991;**59**:12–19.
64. Kendall PC, Marrs-Garcia A, Nath SR, Sheldrick RC. Normative comparisons for the evaluation of clinical significance. *J Consult Clin Psychol* 1999;**67**:285–99.
65. Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. *BMC Musculoskelet Disord* 2005;**6**:3.
66. Kjeklen I, Dagfinrud H, Uhlig T, Mowinckel P, Kvien TK, Finset A. Reliability of the Canadian Occupational Performance Measure in patients with ankylosing spondylitis. *J Rheumatol* 2005;**32**:1503–9.
67. Klopčič M, Jakovljević M. Repeatability and reliability of a new thermotactile quantitative sensory testing algorithm. *Zdrav Vestn* 2009;**78**:365–70.
68. Knols RH, Stappaerts KH, Fransen J, Uebelhart D, Aufdemkampe G. Isometric strength measurement for muscle weakness in cancer patients: reproducibility of isometric muscle strength measurements with a hand-held pull-gauge dynamometer in cancer patients. *Support Care Cancer* 2002;**10**:430–8.
69. Kolotkin RL, Crosby RD, Williams GR, Hartley GG, Nicol S. The relationship between health-related quality of life and weight loss. *Obes Res* 2001;**9**:564–71.
70. Kolta S, Ravaud P, Fechtenbaum J, Dougados M, Roux C. Follow-up of individual patients on two DXA scanners of the same manufacturer. *Osteoporos Int* 2000;**11**:709–13.
71. Krebs EE, Bair MJ, Damush TM, Tu W, Wu J, Kroenke K. Comparative responsiveness of pain outcome measures among primary care patients with musculoskeletal pain. *Med Care* 2010;**48**:1007–14.
72. Kropmans T, Dijkstra P, Stegenga B, Stewart R, de Bont L. Smallest detectable difference of maximal mouth opening in patients with painfully restricted temporomandibular joint function. *Eur J Oral Sci* 2000;**108**:9–13.
73. Kropmans TJ, Dijkstra PU, Stegenga B, Stewart R, de Bont LG. Smallest detectable difference in outcome variables related to painful restriction of the temporomandibular joint. *J Dent Res* 1999;**78**:784–9.
74. Kropmans TJ, Dijkstra PU, van Veen A, Stegenga B, de Bont LG. The smallest detectable difference of mandibular function impairment in patients with a painfully restricted temporomandibular joint. *J Dent Res* 1999;**78**:1445–9.

75. Kropmans TJ, Dijkstra PU, Stegenga B, Stewart R, de Bont LG. Repeated assessment of temporomandibular joint pain: reasoned decision-making with use of unidimensional and multidimensional pain scales. *Clin J Pain* 2002;**18**:107–15.
76. Ladak HM, Thomas JB, Mitchell JR, Rutt BK, Steinman DA. A semi-automatic technique for measurement of arterial wall from black blood MRI. *Med Phys* 2001;**28**:1098–107.
77. Lassere M, Boers M, Van Der Heijde D, Boonen A, Edmonds J, Saudan A, *et al.* Smallest detectable difference in radiological progression. *J Rheumatol* 1999;**26**:731–9.
78. Lodder MC, Lems WF, Ader HJ, Marthinsen AE, van Coeverden SC, Lips P, *et al.* Reproducibility of bone mineral density measurement in daily practice. *Ann Rheum Dis* 2004;**63**:285–9.
79. Lowe B, Unutzer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the Patient Health Questionnaire-9. *Med Care* 2004;**42**:1194–201.
80. Mahony K, Hunt A, Daley D, Sims S, Adams R. Inter-tester reliability and precision of manual muscle testing and hand-held dynamometry in lower limb muscles of children with spina bifida. *Phys Occup Ther Pediatr* 2009;**29**:44–59.
81. Mannion AF, Junge A, Fairbank JC, Dvorak J, Grob D. Development of a German version of the Oswestry Disability Index. Part 1: cross-cultural adaptation, reliability, and validity. *Eur Spine J* 2006;**15**:55–65.
82. Matthey S. Calculating clinically significant change in postnatal depression studies using the Edinburgh Postnatal Depression Scale. *J Affect Disord* 2004;**78**:269–72.
83. Mavissakalian M. Clinically significant improvement in agoraphobia research. *Behav Res Ther* 1986;**24**:369–70.
84. Mayrovitz HN, Soontupe LB. Wound areas by computerized planimetry of digital images: accuracy and reliability. *Adv Skin Wound Care* 2009;**22**:222–9.
85. McCarthy CJ, Oldham JA. The reliability, validity and responsiveness of an aggregated locomotor function (ALF) score in patients with osteoarthritis of the knee. *Rheumatology (Oxford)* 2004;**43**:514–17.
86. McKinnon AD, Bowyer SM, Hubbard CW, Miley HS, Perkins RW, Thompson RC, *et al.* Environmental measurements with a Comprehensive Nuclear Test Ban Treaty radionuclide particulate monitor. *J Radioanal Nucl Chem* 1998;**235**:115–19.
87. McMillan D, Gilbody S, Richards D. Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. *J Affect Disord* 2010;**127**:122–9.
88. Mitchell JR, Karlik SJ, Lee DH, Eliasziw M, Rice GP, Fenster A. The variability of manual and computer assisted quantification of multiple sclerosis lesion volumes. *Med Phys* 1996;**23**:85–97.
89. Modi AC, Zeller MH. Validation of a parent-proxy, obesity-specific quality-of-life measure: sizing them up. *Obesity* 2008;**16**:2624–33.
90. Moleiro C, Beutler LE. Clinically significant change in psychotherapy for depressive disorders. *J Affect Disord* 2009;**115**:220–4.
91. Monbaliu E, Ortibus E, Roelens F, Desloovere K, Deklerck J, Prinzie P, *et al.* Rating scales for dystonia in cerebral palsy: reliability and validity. *Dev Med Child Neurol* 2010;**52**:570–5.
92. Movsas B, Scott C, Watkins-Bruner D. Pretreatment factors significantly influence quality of life in cancer patients: a Radiation Therapy Oncology Group (RTOG) analysis. *Int J Radiat Oncol Biol Phys* 2006;**65**:830–5.

93. Murphy WJ, Franks JR, Berger EH, Behar A, Casali JG, Dixon-Ernst C, *et al.* Development of a new standard laboratory protocol for estimation of the field attenuation of hearing protection devices: sample size necessary to provide acceptable reproducibility. *J Acoust Soc Am* 2004;**115**:311–23.
94. Ndlovu AM, Farrell TJ, Webber CE. Coherent scattering and bone mineral measurement: the dependence of sensitivity on angle and energy. *Med Phys* 1991;**18**:985–9.
95. Negrete RJ, Hanney WJ, Kolber MJ, Davies GJ, Ansley MK, McBride AB, *et al.* Reliability, minimal detectable change, and normative values for tests of upper extremity function and power. *J Strength Cond Res* 2010;**24**:3318–25.
96. Newnham EA, Harwood KE, Page AC. Evaluating the clinical significance of responses by psychiatric inpatients to the mental health subscales of the SF-36. *J Affect Disord* 2007;**98**:91–7.
97. Nitschke JE, McMeeken JM, Burry HC, Matyas TA. When is a change a genuine change? A clinically meaningful interpretation of grip strength measurements in healthy and disabled women. *J Hand Ther* 1999;**12**:25–30.
98. Ogles BM, Lambert MJ, Sawyer JD. Clinical significance of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *J Consult Clin Psychol* 1995;**63**:321–6.
99. Ostelo RW, de Vet HC, Knol DL, Van Den Brandt PA. 24-item Roland–Morris Disability Questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery. *J Clin Epidemiol* 2004;**57**:268–76.
100. Otterstad JE, Froeland G, St John SM, Holme I. Accuracy and reproducibility of biplane two-dimensional echocardiographic measurements of left ventricular dimensions and function. *Eur Heart J* 1997;**18**:507–13.
101. Overend T, Anderson C, Sawant A, Perryman B, Locking-Cusolito H. Relative and absolute reliability of physical function measures in people with end-stage renal disease. *Physiother Can* 2010;**62**:122–8.
102. Parabiaghi A, Barbato A, D’Avanzo B, Erlicher A, Lora A. Assessing reliable and clinically significant change on Health of the Nation Outcome Scales: method for displaying longitudinal data. *Aust N Z J Psychiatry* 2005;**39**:719–25.
103. Patten C, Kothari D, Whitney J, Lexell J, Lum PS. Reliability and responsiveness of elbow trajectory tracking in chronic poststroke hemiparesis. *J Rehabil Res Dev* 2003;**40**:487–500.
104. Pekarik G, Wolff CB. Relationship of satisfaction to symptom change, follow-up adjustment, and clinical significance. *Prof Psychol* 1996;**27**:202–8.
105. Pennell DJ, Mavrogeni SI, Forbat SM, Karwatowski SP, Underwood SR. Adenosine combined with dynamic exercise for myocardial perfusion imaging. *J Am Coll Cardiol* 1995;**25**:1300–9.
106. Pieltain C, De Curtis M, Gérard P, Rigo J. Weight gain composition in preterm infants with dual energy X-ray absorptiometry. *Pediatr Res* 2001;**49**:120–4.
107. Pijls LT, de Vries H, Donker AJ, van Eijk JT. Reproducibility and biomarker-based validity and responsiveness of a food frequency questionnaire to estimate protein intake. *Am J Epidemiol* 1999;**150**:987–95.
108. Piva SR, Fitzgerald GK, Irrgang JJ, Bouzubar F, Starz TW. Get up and go test in patients with knee osteoarthritis. *Arch Phys Med Rehabil* 2004;**85**:284–9.
109. Piva SR, Erhard RE, Childs JD, Browder DA. Inter-tester reliability of passive intervertebral and active movements of the cervical spine. *Man Ther* 2006;**11**:321–30.

110. Potter L, McCarthy C, Oldham J. Algometer reliability in measuring pain pressure threshold over normal spinal muscles to allow quantification of anti-nociceptive treatment effects. *Int J Osteopath Med* 2006;**9**:113–19.
111. Prushansky T, Handelzalts S, Pevzner E. Reproducibility of pressure pain threshold and visual analog scale findings in chronic whiplash patients. *Clin J Pain* 2007;**23**:339–45.
112. Quinn JV, Wells GA. An assessment of clinical wound evaluation scales. *Acad Emerg Med* 1998;**5**:583–6.
113. Rau R, Wassenberg S, Herborn G, Stucki G, Gebler A. A new method of scoring radiographic change in rheumatoid arthritis. *J Rheumatol* 1998;**25**:2094–107.
114. Ravaud P, Giraudeau B, Auleley GR, Edouard N, Dougados M, Chastang C. Assessing smallest detectable change over time in continuous structural outcome measures: application to radiological change in knee osteoarthritis. *J Clin Epidemiol* 1999;**52**:1225–30.
115. Ravaud P, Reny JL, Giraudeau B, Porcher R, Dougados M, Roux C. Individual smallest detectable difference in bone mineral density measurements. *J Bone Mineral Res* 1999;**14**:1449–56.
116. Reeves BC, Wood JM, Hill AR. Vistech VCTS 6500 charts – within- and between-session reliability. *Optom Vis Sci* 1991;**68**:728–37.
117. Rentz AM, Yu R, Iler-Lissner S, Leyendecker P. Validation of the Bowel Function Index to detect clinically meaningful changes in opioid-induced constipation. *J Med Econ* 2009;**12**:371–83.
118. Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. *Phys Ther* 1993;**73**:386–95.
119. Roebroeck ME, Harlaar J, Lankhorst GJ. Reliability assessment of isometric knee extension measurements with a computer-assisted hand-held dynamometer. *Arch Phys Med Rehabil* 1998;**79**:442–8.
120. Rosen HN, Moses AC, Garber J, Ross DS, Lee SL, Greenspan SL. Utility of biochemical markers of bone turnover in the follow-up of patients treated with bisphosphonates. *Calcif Tissue Int* 1998;**63**:363–8.
121. Rosen HN, Moses AC, Garber J, Iloputaife ID, Ross DS, Lee SL, et al. Serum CTX: a new marker of bone resorption that shows treatment effect more often than other markers because of low coefficient of variability and large changes with bisphosphonate therapy. *Calcif Tissue Int* 2000;**66**:100–3.
122. Roy JS, Moffet H, McFadyen BJ, Macdermid JC. The kinematics of upper extremity reaching: a reliability study on people with and without shoulder impingement syndrome. *Sports Med Arthrosc Rehabil Ther Technol* 2010;**2**:8.
123. Rucker TL. Calculation of decision levels and minimum detectable concentrations from method blank and sample uncertainty data – Utopian statistics. *J Radioanal Nucl Chem* 2001;**248**:191–6.
124. Sarna L, Cooley ME, Brown JK, Chernenky C, Elashoff D, Kotlerman J. Symptom severity 1 to 4 months after thoracotomy for lung cancer. *Am J Crit Care* 2008;**17**:455–67.
125. Schauenberg H, Strack M. Measuring psychotherapeutic change with the Symptom Checklist SCL 90 R. *Psychother Psychosom* 1999;**68**:199–206.
126. Schmitz N, Hartkamp N, Franke GH. Assessing clinically significant change: application to the SCL-90-R. *Psychol Rep* 2000;**86**:263–74.
127. Schreuders TA, Roebroeck M, van der Kar TJ, Soeters JN, Hovius SE, Stam HJ. Strength of the intrinsic muscles of the hand measured with a hand-held dynamometer: reliability in patients with ulnar and median nerve paralysis. *J Hand Surg Br* 2000;**25**:560–5.

128. Schurch B, Denys P, Kozma CM, Reese PR, Slaton T, Barron R. Reliability and validity of the Incontinence Quality of Life questionnaire in patients with neurogenic urinary incontinence. *Arch Phys Med Rehabil* 2007;**88**:646–52.
129. Seggar LB, Lambert MJ, Hansen NB. Assessing clinical significance: application to the Beck Depression Inventory. *Behav Ther* 2002;**33**:253–69.
130. Shi HY, Lee HH, Chiu CC, Chiu HC, Uen YH, Lee KT. Responsiveness and minimal clinically important differences after cholecystectomy: GIQLI versus SF-36. *J Gastrointest Surg* 2008;**12**:1275–82.
131. Shim JB, Lee SH, Kim H. A study of minimal change in nocturia affecting quality of life. *Korean J Urol* 2009;**50**:241–5.
132. Shuster JJ. Fixing the number of events in large comparative trials with low event rates: a binomial approach. *Control Clin Trials* 1993;**14**:198–208.
133. Simpson JM, Valentine J, Worsfold C. The Standardized Three-metre Walking Test for elderly people (WALK3m): repeatability and real change. *Clin Rehabil* 2002;**16**:843–50.
134. Smidt N, van der Windt DA, Assendelft WJ, Mourits AJ, Deville WL, *et al.* Interobserver reproducibility of the assessment of severity of complaints, grip strength, and pressure pain threshold in patients with lateral epicondylitis. *Arch Phys Med Rehabil* 2002;**83**:1145–50. [Erratum published in *Arch Phys Med Rehabil* 2003;**84**:938.]
135. Soeters JN, Roebroek ME, Holland WP, Hovius SE, Stam HJ. Non-invasive measurement of tendon excursion with a colour Doppler imaging system: a reliability study in healthy subjects. *Scand J Plast Reconstr Surg Hand Surg* 2004;**38**:356–60.
136. Spadoni GF, Stratford PW, Solomon PE, Wishart LR. The evaluation of change in pain intensity: a comparison of the P4 and single-item numeric pain rating scales. *J Orthop Sports Phys Ther* 2004;**34**:187–93.
137. Speer DC. Clinically significant change: Jacobson and Truax (1991) revisited. *J Consult Clin Psychol* 1992;**60**:402–8.
138. Spiegel B, Camilleri M, Bolus R, Andresen V, Chey WD, Fehnel S, *et al.* Psychometric evaluation of patient-reported outcomes in irritable bowel syndrome randomized controlled trials: a Rome Foundation report. *Gastroenterology* 2009;**137**:1944–53.
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.
140. Stevenson TJ. Detecting change in patients with stroke using the Berg Balance Scale. *Aust J Physiother* 2001;**47**:29–38.
141. Stewart TR, Joyce CR. Increasing the power of clinical trials through judgment analysis. *Med Decis Making* 1988;**8**:33–8.
142. Stoddard MF, Dawkins PR, Prince CR, Ammash NM. Left atrial appendage thrombus is not uncommon in patients with acute atrial fibrillation and a recent embolic event: a transesophageal echocardiographic study. *J Am Coll Cardiol* 1995;**25**:452–9.
143. Stratford PW, Binkley JM. A comparison study of the back pain functional scale and Roland Morris Questionnaire. North American Orthopaedic Rehabilitation Research Network. *J Rheumatol* 2000;**27**:1928–36.
144. Strimpakos N, Sakellari V, Gioftos G, Oldham J. Intratester and intertester reliability of neck isometric dynamometry. *Arch Phys Med Rehabil* 2004;**85**:1309–16.
145. Strimpakos N, Sakellari V, Gioftos G, Kapreli E, Oldham J. Cervical joint position sense: an intra- and inter-examiner reliability study. *Gait Posture* 2006;**23**:22–31.

146. Sumner DR, Turner TM, Galante JO. Symmetry of the canine femur: implications for experimental sample size requirements. *J Orthop Res* 1988;**6**:758–65.
147. Taylor R, Jayasinghe UW, Koelmeyer L, Ung O, Boyages J. Reliability and validity of arm volume measurements for assessment of lymphedema. *Phys Ther* 2006;**86**:205–14.
148. Tsang RCC, Wong E, Au T, Yeung AA, Chung I, Fung L, *et al*. Reference values for 6-minute walk test and hand-grip strength in healthy Hong Kong Chinese adults. *Hong Kong Physiother J* 2005;**23**:6–12.
149. Valk GD, Grootenhuis PA, van Eijk JT, Bouter LM, Bertelsmann FW. Methods for assessing diabetic polyneuropathy: validity and reproducibility of the measurement of sensory symptom severity and nerve function tests. *Diabetes Res Clin Pract* 2000;**47**:87–95.
150. van Baalen B, Odding E, van Woensel MP, Roebroek ME. Reliability and sensitivity to change of measurement instruments used in a traumatic brain injury population. *Clin Rehabil* 2006;**20**:686–700.
151. van der Esch M, Steultjens M, Ostelo RW, Harlaar J, Dekker J. Reproducibility of instrumented knee joint laxity measurement in healthy subjects. *Rheumatology (Oxford)* 2006;**45**:595–9.
152. van Dieën JH, Heijblom P. Reproducibility of isometric trunk extension torque, trunk extensor endurance, and related electromyographic parameters in the context of their clinical applicability. *J Orthop Res* 1996;**14**:139–43.
153. van der Heijde D, Dankert T, Nieman F, Rau R, Boers M. Reliability and sensitivity to change of a simplification of the Sharp/van der Heijde radiological assessment in rheumatoid arthritis. *Rheumatology (Oxford)* 1999;**38**:941–7.
154. van der Hoeven N. Calculation of the minimum significant difference at the NOEC using a non-parametric test. *Ecotoxicol Environ Saf* 2008;**70**:61–6.
155. van der Lee JH, Wagenaar RC, Lankhorst GJ, Vogelaar TW, Deville WL, *et al*. Forced use of the upper extremity in chronic stroke patients: results from a single-blind randomized clinical trial. *Stroke* 1999;**30**:2369–75.
156. van der Lee JH, De Groot, V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. *Arch Phys Med Rehabil* 2001;**82**:14–19.
157. van Meeteren J, Roebroek ME, Stam HJ. Test–retest reliability in isokinetic muscle strength measurements of the shoulder. *J Rehabil Med* 2002;**34**:91–5.
158. Vos CJ, Verhagen AP, Koes BW. Reliability and responsiveness of the Dutch version of the Neck Disability Index in patients with acute neck pain in general practice. *Eur Spine J* 2006;**15**:1729–36.
159. Wang Q, Denton DL, Shukla R. Applications and statistical properties of minimum significant difference-based criterion testing in a toxicity testing program. *Environ Toxicol Chem* 2000;**19**:113–17.
160. Wang SS, Normile SO, Lawshe BT. Reliability and smallest detectable change determination for serratus anterior muscle strength and endurance tests. *Physiother Theory Pract* 2006;**22**:33–42.
161. Warren-Hicks WJ, Parkhurst BR, Moore DRJ, Teed RS, Baird RB, Berger R, *et al*. Assessment of whole effluent toxicity test variability: partitioning sources of variability. *Environ Toxicol Chem* 2000;**19**:94–104.
162. Wassenberg S, Fischer-Kahle V, Herborn G, Rau R. A method to score radiographic change in psoriatic arthritis. *Z Rheumatol* 2001;**60**:156–66.
163. Wiebe S, Eliasziw M, Matijevic S. Changes in quality of life in epilepsy: how large must they be to be real? *Epilepsia* 2001;**42**:113–18.

164. Willis C, Niere KR, Hoving JL, Green S, O'Leary EF, Buchbinder R. Reproducibility and responsiveness of the Whiplash Disability Questionnaire. *Pain* 2004;**110**:681–8.
165. Wolfe F, Michaud K, Li T. Sleep disturbance in patients with rheumatoid arthritis: evaluation by medical outcomes study and visual analog sleep scales. *J Rheumatol* 2006;**33**:1942–51.
166. Wosje KS, Knipstein BL, Kalkwarf HJ. Measurement error of DXA: interpretation of fat and lean mass changes in obese and non-obese children. *J Clin Densitom* 2006;**9**:335–40.
167. Yao RT, Ma TY, Shao YP. Lutetium oxyorthosilicate (LSO) intrinsic activity correction and minimal detectable target activity study for SPECT imaging with a LSO-based animal PET scanner. *Phys Med Biol* 2008;**53**:4399–415.
168. Yoshida EMP. [Clinical significance of change in process of brief psychodynamic therapy.] *Paidéia* 2008;**18**:305–16.
169. Zaina F, Negrini S, Atanasio S. TRACE (Trunk Aesthetic Clinical Evaluation), a routine clinical tool to evaluate aesthetics in scoliosis patients: development from the Aesthetic Index (AI) and repeatability. *Scoliosis* 2009;**4**:3.
170. Zeggelink WF, Deurloo EE, Bartelink H, Rutgers EJ, Gilhuijs KG. Reproducibility of the assessment of tumor extent in the breast using multiple image modalities. *Med Phys* 2003;**30**:2919–26.
171. Ziv E, Patish H, Dvir Z. Grip and pinch strength in healthy subjects and patients with primary osteoarthritis of the hand: a reproducibility study. *Orthop J* 2008;**2**:86–90.

Health economic method ($n = 13$)

1. Al MJ, van Hout BA, Michel BC, Rutten FF. Sample size calculation in economic evaluations. *Health Econ* 1998;**7**:327–35.
2. Bacchetti P, McCulloch CE, Segal MR. Simple, defensible sample sizes based on cost efficiency. *Biometrics* 2008;**64**:577–85.
3. Briggs AH, Gray AM. Power and sample size calculations for stochastic cost-effectiveness analysis. *Med Decis Making* 1998;**18**(Suppl.):S81–92.
4. Detsky AS. Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials. *Stat Med* 1990;**9**:173–84.
5. Gittins JC, Pezeshk H. A decision theoretic approach to sample size determination in clinical trials. *J Biopharm Stat* 2002;**12**:535–51.
6. Kikuchi T, Pezeshk H, Gittins J. A Bayesian cost–benefit approach to the determination of sample size in clinical trials. *Stat Med* 2008;**27**:68–82.
7. O'Hagan A, Stevens JW. Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Med Decis Making* 2001;**21**:219–30.
8. Samsa GP, Matchar DB. Have randomized controlled trials of neuroprotective drugs been underpowered? An illustration of three statistical principles. *Stroke* 2001;**32**:669–74.
9. Tan SB, Smith AF. Exploratory thoughts on clinical trials with utilities. *Stat Med* 1998;**17**:2771–91.
10. Torgerson DJ, Ryan M, Ratcliffe J. Economics in sample size determination for clinical trials. *QJM* 1995;**88**:517–21.
11. Willan AR. Optimal sample size determinations from an industry perspective based on the expected value of information. *Clin Trials* 2008;**5**:587–94.

12. Willan A, Kowgier M. Determining optimal sample sizes for multi-stage randomized clinical trials using value of information methods. *Clin Trials* 2008;**5**:289–300.
13. Willan AR, Eckermann S. Optimal clinical trial design using value of information methods with imperfect implementation. *Health Econ* 2010;**19**:549–61.

Opinion-seeking method ($n = 60$)

1. Aakr KM, Thue G, Subramaniam-Haavik S, Bukve T, Morris H, Iler M, *et al.* Postanalytical external quality assessment of urine albumin in primary health care: an international survey. *Clin Chem* 2008;**54**:1630–6.
2. Aarabi M, Skinner J, Price CE, Jackson PR. Patients' acceptance of antihypertensive therapy to prevent cardiovascular disease: a comparison between South Asians and Caucasians in the United Kingdom. *Eur J Cardiovasc Prevent Rehabil* 2008;**15**:59–66.
3. Ad Hoc Committee on Lupus Response Criteria: Cognition Sub-committee, Mikdashi JA, Esdaile JM, Alarc GS, Crofford L, *et al.* Proposed response criteria for neurocognitive impairment in systemic lupus erythematosus clinical trials. *Lupus* 2007;**16**:418–25.
4. Allison DB, Elobeid MA, Cope MB, Brock DW, Faith MS, Vander VS, *et al.* Sample size in obesity trials: patient perspective versus current practice. *Med Decis Making* 2010;**30**:68–75.
5. Barrett B, Brown R, Mundt M, Dye L, Alt J, Safdar N, *et al.* Using benefit harm tradeoffs to estimate sufficiently important difference: the case of the common cold. *Med Decis Making* 2005;**25**:47–55.
6. Barrett B, Harahan B, Brown D, Zhang Z, Brown R. Sufficiently important difference for common cold: severity reduction. *Ann Fam Med* 2007;**5**:216–23.
7. Bayle FJ, Misdrahi D, Llorca PM, Lancon C, Olivier V, Quintin P, *et al.* [Acute schizophrenia concept and definition: investigation of a French psychiatrist population.] *Encephale* 2005;**31**:10–17.
8. Bellamy N, Anastassiades TP, Buchanan WW, Davis P, Lee P, McCain GA, *et al.* Rheumatoid arthritis antirheumatic drug trials. III. Setting the delta for clinical trials of antirheumatic drugs – results of a consensus development (Delphi) exercise. *J Rheumatol* 1991;**18**:1908–15.
9. Bellamy N, Buchanan WW, Esdaile JM, Fam AG, Kean WF, Thompson JM, *et al.* Ankylosing spondylitis antirheumatic drug trials. III. Setting the delta for clinical trials of antirheumatic drugs – results of a consensus development (Delphi) exercise. *J Rheumatol* 1991;**18**:1716–22.
10. Bellamy N, Carette S, Ford PM, Kean WF, le Riche NG, Lussier A, *et al.* Osteoarthritis antirheumatic drug trials. III. Setting the delta for clinical trials – results of a consensus development (Delphi) exercise. *J Rheumatol* 1992;**19**:451–7.
11. Bellm LA, Cunningham G, Durnell L, Eilers J, Epstein JB, Fleming T, *et al.* Defining clinically meaningful outcomes in the evaluation of new treatments for oral mucositis: oral mucositis patient provider advisory board. *Cancer Invest* 2002;**20**:793–800.
12. Bloom LF, Lapierre NM, Wilson KG, Curran D, DeForge DA, Blackmer J. Concordance in goal setting between patients with multiple sclerosis and their rehabilitation team. *Am J Phys Med Rehabil* 2006;**85**:807–13.
13. Boers M, Tugwell P. OMERACT conference questionnaire results. OMERACT Committee. *J Rheumatol* 1993;**20**:552–4.
14. Brown KA. Unilateral and bilateral electroconvulsive therapy: what informs Scottish psychiatrists' choices? *Psychiatric Bull* 2009;**33**:95–8.
15. Bryce RL, Bradley MT, McCormick SM. To what extent would women prefer chorionic villus sampling to amniocentesis for prenatal diagnosis? *Paediatr Perinat Epidemiol* 1989;**3**:137–45.

16. 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, sample size and trial duration. *Dement Geriatr Cogn Disord* 1999;**10**:534–40.
17. Burgess P, Trauer T, Coombs T, McKay R, Pirkis J. What does 'clinical significance' mean in the context of the Health of the Nation Outcome Scales? *Aust Psychiatr* 2009;**17**:141–8.
18. Castrillo-Viguera C, Grasso DL, Simpson E, Shefner J, Cudkowicz ME. Clinical significance in the change of decline in ALSFRS-R. *Amyotroph Lateral Scler* 2010;**11**:178–80.
19. Fayers PM, Cuschieri A, Fielding J, Craven J, Uscinska B, Freedman LS. Sample size calculation for clinical trials: the impact of clinician beliefs. *Br J Cancer* 2000;**82**:213–19.
20. Ferreira ML, Ferreira PH, Herbert RD, Latimer J. People with low back pain typically need to feel 'much better' to consider intervention worthwhile: an observational study. *Aust J Physiother* 2009;**55**:123–7.
21. Freedman LS, Lowe D, Macaskill P. Stopping rules for clinical trials. *Stat Med* 1983;**2**:167–74.
22. Fried BJ, Boers M, Baker PR. A method for achieving consensus on rheumatoid arthritis outcome measures: the OMERACT conference process. *J Rheumatol* 1993;**20**:548–51.
23. Gajewski BJ, Mayo MS. Bayesian sample size calculations in phase II clinical trials using a mixture of informative priors. *Stat Med* 2006;**25**:2554–66.
24. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. *Arthritis Rheum* 1997;**40**:1202–9.
25. Giannini EH, Mehta AB, Hilz MJ, Beck M, Bichet DG, Brady RO, et al. A validated disease severity scoring system for Fabry disease. *Mol Genet Metab* 2010;**99**:283–90.
26. Girling AJ, Lilford RJ, Brauholtz DA, Gillett WR. Sample-size calculations for trials that inform individual treatment decisions: a 'true-choice' approach. *Clin Trials* 2007;**4**:15–24.
27. Harding G, Leidy NK, Meddis D, Kleinman L, Wagner S, O'Brien CD. Interpreting clinical trial results of patient-perceived onset of effect in asthma: methods and results of a Delphi panel. *Curr Med Res Opin* 2009;**25**:1563–71.
28. Kirkby HM, Wilson S, Calvert M, Draper H. Using e-mail recruitment and an online questionnaire to establish effect size: a worked example. *BMC Med Res Methodol* 2011;**11**:89.
29. Kirwan JR, Chaput de Saintonge DM, Joyce CR, Currey HL. Clinical judgment in rheumatoid arthritis. III. British rheumatologists' judgments of 'change in response to therapy'. *Ann Rheum Dis* 1984;**43**:686–94.
30. Kirwan JR, Currey HL, Brooks PM. Measuring physicians' judgment – the use of clinical data by Australian rheumatologists. *Aust N Z J Med* 1985;**15**:738–44.
31. Latthe PM, Brauholtz DA, Hills RK, Khan KS, Lilford R. Measurement of beliefs about effectiveness of laparoscopic uterosacral nerve ablation. *BJOG* 2005;**112**:243–6.
32. Man-Son-Hing M, Laupacis A, O'Connor A, Wells G, Lemelin J, Wood W, et al. Warfarin for atrial fibrillation. The patient's perspective. *Arch Intern Med* 1996;**156**:1841–8.
33. Massel D, Cruickshank M. Greater expectations in a cancer trial: absolute more than relative survival increases, community more than academic clinicians. *Cancer Invest* 2000;**18**:798–803.
34. McAlister FA, O'Connor AM, Wells G, Grover SA, Laupacis A. When should hypertension be treated? The different perspectives of Canadian family physicians and patients. *CMAJ* 2000;**163**:403–8.
35. Miller WR, Manuel JK. How large must a treatment effect be before it matters to practitioners? An estimation method and demonstration. *Drug Alcohol Rev* 2008;**27**:524–8.

36. Milne J, Dwinnel S, Swaby C, Wood S, Ross S. What is the minimum clinically important difference (MCID) required to introduce a new treatment into obstetrical practice? Survey of Canadian obstetricians *Clin Trials* 2007;**4**:424.
37. Mosca M, Lockshin M, Schneider M, Liang MH, Albrecht J, Aringer M, *et al.* Response criteria for cutaneous SLE in clinical trials. *Clin Exp Rheumatol* 2007;**25**:666–71.
38. Oliveira VC, Ferreira PH, Ferreira ML, Tiburcio L, Pinto RZ, Oliveira W, *et al.* People with low back pain who have externalised beliefs need to see greater improvements in symptoms to consider exercises worthwhile: an observational study. *Aust J Physiother* 2009;**55**:271–5.
39. Oremus M, Collet JP, Corcos J, Shapiro SH. A survey of physician efficacy requirements to plan clinical trials. *Pharmacoepidemiol Drug Saf* 2002;**11**:677–85.
40. Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Von KM, *et al.* Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. *Spine* 2008;**33**:90–4.
41. Ouellet D, Werth J, Parekh N, Feltner D, McCarthy B, Lalonde RL. The use of a clinical utility index to compare insomnia compounds: a quantitative basis for benefit–risk assessment. *Clin Pharmacol Ther* 2009;**85**:277–82.
42. Parmar MK, Griffiths GO, Spiegelhalter DJ, Souhami RL, Altman DG, van der Scheuren E, *et al.* Monitoring of large randomised clinical trials: a new approach with Bayesian methods. *Lancet* 2001;**358**:375–81.
43. Rang LC, Murray HE, Wells GA, Macgougan CK. Can peripheral venous blood gases replace arterial blood gases in emergency department patients? *CJEM* 2002;**4**:7–15.
44. Rantz MJ, Petroski GF, Madsen RW, Scott J, Mehr DR, Popejoy L, *et al.* Setting thresholds for MDS (minimum data set) quality indicators for nursing home quality improvement reports. *Jt Comm J Qual Improv* 1997;**23**:602–11.
45. Rider LG, Giannini EH, Harris-Love M, Joe G, Isenberg D, Pilkington C, *et al.* Defining clinical improvement in adult and juvenile myositis. *J Rheumatol* 2003;**30**:603–17.
46. Rider LG, Giannini EH, Brunner HI, Ruperto N, James-Newton L, Reed AM, *et al.* International consensus on preliminary definitions of improvement in adult and juvenile myositis. *Arthritis Rheum* 2004;**50**:2281–90.
47. Ruperto N, Ravelli A, Oliveira S, Alessio M, Mihaylova D, Pasic S, *et al.* The Pediatric Rheumatology International Trials Organization/American College of Rheumatology provisional criteria for the evaluation of response to therapy in juvenile systemic lupus erythematosus: prospective validation of the definition of improvement. *Arthritis Rheum* 2006;**55**:355–63.
48. Ruperto N, Pistorio A, Ravelli A, Rider LG, Pilkington C, Oliveira S, *et al.* The Paediatric Rheumatology International Trials Organisation provisional criteria for the evaluation of response to therapy in juvenile dermatomyositis. *Arthritis Care Res* 2010;**62**:1533–41.
49. Skendzel LP. How physicians use laboratory tests. *JAMA* 1978;**239**:1077–80.
50. Stone MA, Inman RD, Wright JG, Maetzel A. Validation exercise of the Ankylosing Spondylitis Assessment Study (ASAS) group response criteria in ankylosing spondylitis patients treated with biologics. *Arthritis Rheum* 2004;**51**:316–20.
51. Tubach F, Ravaud P, Beaton D, Boers M, Bombardier C, Felson DT, *et al.* Minimal clinically important improvement and patient acceptable symptom state for subjective outcome measures in rheumatic disorders. *J Rheumatol* 2007;**34**:1188–93.

52. Van Der Heijde D, Lassere M, Edmonds J, Kirwan J, Strand V, Boers M. Minimal clinically important difference in plain films in RA: group discussions, conclusions, and recommendations. OMERACT Imaging Task Force. *J Rheumatol* 2001;**28**:914–17.
53. van Walraven C, Mahon JL, Moher D, Bohm C, Laupacis A. Surveying physicians to determine the minimal important difference: implications for sample-size calculation. *J Clin Epidemiol* 1999;**52**:717–23.
54. Wells G, Anderson J, Boers M, Felson D, Heiberg T, Hewlett S, *et al.* MCID/Low Disease Activity State Workshop: summary, recommendations, and research agenda. *J Rheumatol* 2003;**30**:1115–18.
55. Wells G, Boers M, Shea B, Anderson J, Felson D, Johnson K, *et al.* MCID/Low Disease Activity State Workshop: low disease activity state in rheumatoid arthritis. *J Rheumatol* 2003;**30**:1110–11.
56. Wengritzky R, Mettho T, Myles PS, Burke J, Kakos A. Development and validation of a postoperative nausea and vomiting intensity scale. *Br J Anaesth* 2010;**104**:158–66.
57. Wong RK, Gafni A, Whelan T, Franssen E, Fung K. Defining patient-based minimal clinically important effect sizes: a study in palliative radiotherapy for painful unresectable pelvic recurrences from rectal cancer. *Int J Radiat Oncol Biol Phys* 2002;**54**:661–9.
58. Wyrwich KW, Fihn SD, Tierney WM, Kroenke K, Babu AN, Wolinsky FD. Clinically important changes in health-related quality of life for patients with chronic obstructive pulmonary disease: an expert consensus panel report. *J Gen Intern Med* 2003;**18**:196–202.
59. Wyrwich KW, Nelson HS, Tierney WM, Babu AN, Kroenke K, Wolinsky FD. Clinically important differences in health-related quality of life for patients with asthma: an expert consensus panel report. *Ann Allerg Asthm Immunol* 2003;**91**:148–53.
60. Wyrwich KW, Spertus JA, Kroenke K, Tierney WM, Babu AN, Wolinsky FD, *et al.* Clinically important differences in health status for patients with heart disease: an expert consensus panel report. *Am Heart J* 2004;**147**:615–22.

Pilot study method (*n* = 5)

1. Browne RH. On the use of a pilot sample for sample size determination. *Stat Med* 1995;**14**:1933–40.
2. Johnstone R, Donaghy M, Martin D. A pilot study of a cognitive-behavioural therapy approach to physiotherapy, for acute low back pain patients, who show signs of developing chronic pain. *Adv Physiother* 2002;**4**:182–8.
3. 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.
4. Salter GC, Roman M, Bland MJ, MacPherson H. Acupuncture for chronic neck pain: a pilot for a randomised controlled trial. *BMC Musculoskelet Disord* 2006;**7**:99.
5. Wang SJ, Hung HM, O'Neill RT. Adapting the sample size planning of a phase III trial based on phase II data. *Pharm Stat* 2006;**5**:85–97.

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

1. Blumenauer B. Quality of life in patients with rheumatoid arthritis: which drugs might make a difference? *Pharmacoeconomics* 2003;**21**:927–40.

2. Bombardier C, Hayden J, Beaton DE. Minimal clinically important difference. Low back pain: outcome measures. *J Rheumatol* 2001;**28**:431–8.
3. Cranney A, Welch V, Wells G, Adachi J, Shea B, Simon L, *et al.* Discrimination of changes in osteoporosis outcomes. *J Rheumatol* 2001;**28**:413–21.
4. Even C, Friedman S, Dardennes R, Zuber M, Guelfi JD. Prevalence of depression in multiple sclerosis: a review and meta-analysis. *Rev Neurol (Paris)* 2004;**160**:917–25.
5. Feise RJ, Menke JM. Functional Rating Index: literature review. *Med Sci Monit* 2010;**16**:RA25–36.
6. Fisher PL. The efficacy of psychological treatments for generalised anxiety disorder? In Davey GCL, Wells A, editors. *Worry and its psychological disorders: theory, assessment and treatment*. Chichester: John Wiley; 2006. pp. 359–77.
7. Johnston MF, Hays RD, Hui KK. Evidence-based effect size estimation: an illustration using the case of acupuncture for cancer-related fatigue. *BMC Complement Altern Med* 2009;**9**:1.
8. Klassen AF. Quality of life of children with attention deficit hyperactivity disorder. *Exp Rev Pharm Outcome Res* 2005;**5**:95–103.
9. Muller U, Duetz MS, Roeder C, Greenough CG. Condition-specific outcome measures for low back pain: part I: validation. *Eur Spine J* 2004;**13**:301–13.
10. Nemann HJ, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. Measurement properties and interpretability of the Chronic respiratory disease questionnaire (CRQ). *COPD* 2005;**2**:81–9.
11. Nietzel MT, Russell RL, Hemmings KA, Gretter ML. Clinical significance of psychotherapy for unipolar depression: a meta-analytic approach to social comparison. *J Consult Clin Psychol* 1987;**55**:156–61.
12. 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.
13. Revicki DA, Feeny D, Hunt TL, Cole BF. Analyzing oncology clinical trial data using the Q-TWiST method: clinical importance and sources for health state preference data. *Qual Life Res* 2006;**15**:411–23.
14. Schwartz CE, Bode R, Repucci N, Becker J, Sprangers MAG, Fayers PM. The clinical significance of adaptation to changing health: a meta-analysis of response shift. *Qual Life Res* 2006;**15**:1533–50.
15. Sheldrick RC, Kendall PC, Heimberg RG. The clinical significance of treatments: a comparison of three treatments for conduct disordered children. *Clin Psychol* 2001;**8**:418–30.
16. Smith MK, Marshall S. A Bayesian design and analysis for dose–response using informative prior information. *J Biopharm Stat* 2006;**16**:695–709.
17. Stein J, Lippa M, Brähler E, König HH, Riedel-Heller SG. The assessment of changes in cognitive functioning: reliable change indices for neuropsychological instruments in the elderly– a systematic review. *Dement Geriatr Cogn Disord* 2010;**29**:275–86.
18. Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. *Stat Med* 2007;**26**:2479–500.
19. Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. *BMC Med Res Methodol* 2009;**9**:29.
20. Thomas JR, Lochbaum MR, Landers DM, He C. Planning significant and meaningful research in exercise science: estimating sample size. *Res Q Exerc Sport* 1997;**68**:33–43.
21. Woods SW, Stolar M, Sernyak MJ, Charney DS. Consistency of atypical antipsychotic superiority to placebo in recent clinical trials. *Biol Psychiatry* 2001;**49**:64–70.

22. Zanen P, Lammers JW. Sample sizes for comparative inhaled corticosteroid trials with emphasis on showing therapeutic equivalence. *Eur J Clin Pharmacol* 1995;**48**:179–84.

Standardised effect size ($n = 37$)

1. Andrew MK, Rockwood K. A five-point change in Modified Mini-Mental State Examination was clinically meaningful in community-dwelling elderly people. *J Clin Epidemiol* 2008;**61**:827–31.
2. Bain BA, Dollaghan CA. The notion of clinically significant change. *Lang Speech, Hear Serv Sch* 1991;**22**:264–70.
3. Basoglu M, Livanou M, Salcioglu E, Kalender D. A brief behavioural treatment of chronic post-traumatic stress disorder in earthquake survivors: results from an open clinical trial. *Psychol Med* 2003;**33**:647–54.
4. Burton HJ, Kline SA, Cooper BS, Rabinowitz A, Dodek A. Assessing risk for major depression on patients selected for percutaneous transluminal coronary angioplasty: is it a worthwhile venture? *Gen Hosp Psychiatry* 2003;**25**:200–8.
5. Cheung YB, Goh C, Thumboo J, Khoo KS, Wee J. Variability and sample size requirements of quality-of-life measures: a randomized study of three major questionnaires. *J Clin Oncol* 2005;**23**:4936–44.
6. Cramer JA, Cuffel BJ, Divan V, Al-Sabbagh A, Glassman M. Patient satisfaction with an injection device for multiple sclerosis treatment. *Acta Neurol Scand* 2006;**113**:156–62.
7. Cribbie RA, Arpin-Cribbie CA. Evaluating clinical significance through equivalence testing: extending the normative comparisons approach. *Psychother Res* 2009;**19**:677–86.
8. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg Br* 1996;**78**:185–90.
9. Diehr P, Psaty BM, Patrick DL. Effect size and power for clinical trials that measure years of healthy life. *Stat Med* 1997;**16**:1211–23.
10. Dixon T, Lim LL, Oldridge NB. The MacNew heart disease health-related quality of life instrument: reference data for users. *Qual Life Res* 2002;**11**:173–83.
11. Drinkwater EJ, Pritchett EJ, Behm DG. Effect of instability and resistance on unintentional squat-lifting kinetics. *Int J Sports Physiol Perform* 2007;**2**:400–13.
12. Dumas HM, Haley SM, Bedell GM, Hull EM. Social function changes in children and adolescents with acquired brain injury during inpatient rehabilitation. *Pediatr Rehabil* 2001;**4**:177–85.
13. Fayers PM, Langston AL, Robertson C. Implicit self-comparisons against others could bias quality of life assessments. *J Clin Epidemiol* 2007;**60**:1034–9.
14. Gompertz P, Pound P, Ebrahim S. Validity of the extended activities of daily living scale. *Clin Rehabil* 1994;**8**:275–80.
15. Gordon JE, Powell C, Rockwood K. Goal attainment scaling as a measure of clinically important change in nursing-home patients. *Age Ageing* 1999;**28**:275–81.
16. Harris MA, Greco P, Wysocki T, White NH. Family therapy with adolescents with diabetes: a litmus test for clinically meaningful change. *Fam Syst Health* 2001;**19**:159–68.
17. Haymes SA, Johnston AW, Heyes AD. Preliminary investigation of the responsiveness of the Melbourne Low Vision ADL index to low-vision rehabilitation. *Optom Vis Sci* 2001;**78**:373–80.
18. Konst EM, Prah C, Weersink-Braks H, De Boo T, Prah-Andersen B, Kuijpers-Jagtman AM, et al. Cost-effectiveness of infant orthopedic treatment regarding speech in patients with complete unilateral

- cleft lip and palate: a randomized three-center trial in the Netherlands (Dutchcleft). *Cleft Palate Craniofac J* 2004;**41**:71–7.
19. Krakow B, Melendrez D, Sisley B, Warner TD, Krakow J, Leahigh L, *et al*. Nasal dilator strip therapy for chronic sleep-maintenance insomnia and symptoms of sleep-disordered breathing: a randomized controlled trial. *Sleep Breath* 2006;**10**:16–28.
 20. Leon AC, Marzuk PM, Portera L. More reliable outcome measures can reduce sample size requirements. *Arch Gen Psychiatry* 1995;**52**:867–71.
 21. Matza LS, Johnston JA, Faries DE, Malley KG, Brod M. Responsiveness of the Adult Attention-Deficit/Hyperactivity Disorder Quality of Life Scale (AAQoL). *Qual Life Res* 2007;**16**:1511–20.
 22. McKee G. Are there meaningful longitudinal changes in health related quality of life – SF36, in cardiac rehabilitation patients? *Eur J Cardiovasc Nurs* 2009;**8**:40–7.
 23. Merkies IS, Schmitz PI, van der Meché, Samijn JP, van Doorn PA, Inflammatory Neuropathy Cause and Treatment (INCAT) group. Quality of life complements traditional outcome measures in immune-mediated polyneuropathies. *Neurology* 2002;**59**:84–91.
 24. Myles PS, Hunt JO, Fletcher H, Solly R, Woodward D, Kelly S. Relation between quality of recovery in hospital and quality of life at 3 months after cardiac surgery. *Anesthesiology* 2001;**95**:862–7.
 25. Nilsdotter AK, Roos EM, Westerlund JP, Roos HP, Lohmander LS. Comparative responsiveness of measures of pain and function after total hip replacement. *Arthritis Rheum* 2001;**45**:258–62.
 26. O’Carroll RE, Cossar JA, Couston MC, Hayes PC. Sensitivity to change following liver transplantation. A comparison of three instruments that measure quality of life. *J Health Psychol* 2000;**5**:69–74.
 27. Pritchett YL, Marciniak MD, Corey-Lisle PK, Berzon RA, Desai D, Detke MJ. Use of effect size to determine optimal dose of duloxetine in major depressive disorder. *J Psychiatr Res* 2007;**41**:311–18.
 28. Pyne JM, Sullivan G, Kaplan R, Williams DK. Comparing the sensitivity of generic effectiveness measures with symptom improvement in persons with schizophrenia. *Med Care* 2003;**41**:208–17.
 29. Rajagopalan R, Laitinen D, Dietz B. Impact of lipoatrophy on quality of life in HIV patients receiving anti-retroviral therapy. *AIDS Care* 2008;**20**:1197–201.
 30. Rockwood K, Stolee P, Fox RA. Use of goal attainment scaling in measuring clinically important change in the frail elderly. *J Clin Epidemiol* 1993;**46**:1113–18.
 31. Rockwood K, Stolee P. Responsiveness of outcome measures used in an antidementia drug trial. *Alzheimer Dis Assoc Disord* 2000;**14**:182–5.
 32. Rockwood K, Fay S, Song X, MacKnight C, Gorman M, Video-Imaging Synthesis of Treating Alzheimer’s Disease (VISTA) Investigators. Attainment of treatment goals by people with Alzheimer’s disease receiving galantamine: a randomized controlled trial. *CMAJ* 2006;**174**:1099–105.
 33. Tuzun EH, Eker L, Aydar A, Daskapan A, Bayramoglu M. Acceptability, reliability, validity and responsiveness of the Turkish version of WOMAC osteoarthritis index. *Osteoarthritis Cartilage* 2005;**13**:28–33.
 34. van de Port IG, Ketelaar M, Schepers VP, Van den Bos GA, Lindeman E. Monitoring the functional health status of stroke patients: the value of the Stroke-Adapted Sickness Impact Profile-30. *Disabil Rehabil* 2004;**26**:635–40.
 35. van der Putten JJ, Hobart JC, Freeman JA, Thompson AJ. Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel index and the Functional Independence Measure. *J Neurol Neurosurg Psychiatr* 1999;**66**:480–4.

36. van Tubergen A, Landew R, Heuft-Dorenbosch L, Spoorenberg A, Van Der Heijde D, *et al.* Assessment of disability with the World Health Organisation Disability Assessment Schedule II in patients with ankylosing spondylitis. *Ann Rheum Dis* 2003;**62**:140–5.
37. Winkelman C. Effect size: utility and application in neuroscience nursing. *J Neurosci Nurs* 2001;**33**:216–18.

Multiple methods (*n* = 216)

1. Abrams P, Kelleher C, Huels J, Quebe-Fehling E, Omar MA, Steel M. Clinical relevance of health-related quality of life outcomes with darifenacin. *BJU Int* 2008;**102**:208–13.
2. Ad Hoc Committee Response Criteria For Cutaneous SLE. Response criteria for cutaneous SLE in clinical trials. *Clin Exp Rheumatol* 2007;**25**:666–71.
3. American College of Rheumatology Ad Hoc Committee on Systemic Lupus Erythematosus Response Criteria. The American College of Rheumatology response criteria for systemic lupus erythematosus clinical trials: measures of overall disease activity. *Arthritis Rheum* 2004;**50**:3418–26.
4. Andersson EI, Lin CC, Smeets RJEM. Performance tests in people with chronic low back pain: responsiveness and minimal clinically important change. *Spine* 2010;**35**:E1559–63.
5. Arbuckle RA, Humphrey L, Vardeva K, Arondekar B, Danten-Viala M, Scott JA, *et al.* Psychometric evaluation of the Diabetes Symptom Checklist-Revised (DSC-R) – a measure of symptom distress. *Value Health* 2009;**12**:1168–75.
6. Askew RL, Xing Y, Palmer JL, Cella D, Moye LA, Cormier JN. Evaluating minimal important differences for the FACT-Melanoma quality of life questionnaire. *Value Health* 2009;**12**:1144–50.
7. Bagó J, Pérez-Gruoso FJ, Les E, Hernández P, Pellisé F. Minimal important differences of the SRS-22 Patient Questionnaire following surgical treatment of idiopathic scoliosis. *Eur Spine J* 2009;**18**:1898–904.
8. Barber MD, Spino C, Janz NK, Brubaker L, Nygaard I, Nager CW, *et al.* The minimum important differences for the urinary scales of the Pelvic Floor Distress Inventory and Pelvic Floor Impact Questionnaire. *Am J Obstet Gynecol* 2009;**200**:580–7.
9. Barnes ML, Vaidyanathan S, Williamson PA, Lipworth BJ. The minimal clinically important difference in allergic rhinitis. *Clin Exp Allergy* 2010;**40**:242–50.
10. Barrett B, Brown R, Mundt M. Comparison of anchor-based and distributional approaches in estimating important difference in common cold. *Qual Life Res* 2008;**17**:75–85.
11. Beaton DE, van Eerd D, Smith P, van der Velde G, Cullen K, Kennedy CA, *et al.* Minimal change is sensitive, less specific to recovery: a diagnostic testing approach to interpretability. *J Clin Epidemiol* 2011;**64**:487–96.
12. Bellamy N, Bell MJ, Carette S, Fam AG, Haraoui BW, McCain GA, *et al.* Estimation of observer reliability and sample size calculation parameters for outcome measures in fibromyalgia clinical trials. *Inflammopharmacology* 1993;**2**:345–60.
13. Bilbao A, Quintana JM, Escobar A, Garcia S, Andradas E, Baré M, *et al.* Responsiveness and clinically important differences for the VF-14 index, SF-36, and visual acuity in patients undergoing cataract surgery. *Ophthalmology* 2009;**116**:418–24.
14. Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. North American Orthopaedic Rehabilitation Research Network. *Phys Ther* 1999;**79**:371–83.

15. Bols EM, Hendriks EJ, Deutekom M, Berghmans BC, Baeten CG, de Bie RA. Inconclusive psychometric properties of the Vaizey score in fecally incontinent patients: a prospective cohort study. *Neurourol Urodyn* 2010;**29**:370–7.
16. Bouffouix E, Arnould C, Vandervelde L, Thonnard JL. Changes in satisfaction with activities and participation between acute, post-acute and chronic stroke phases: a responsiveness study of the SATIS-Stroke questionnaire. *J Rehabil Med* 2010;**42**:944–8.
17. Boyles RE, Walker MJ, Young BA, Strunce JB, Wainner RS. The addition of cervical thrust manipulations to a manual physical therapy approach in patients treated for mechanical neck pain: a secondary analysis. *J Orthop Sport Phys Ther* 2010;**40**:133–40.
18. Brach JS, Perera S, Studenski S, Katz M, Hall C, Verghese J. Meaningful change in measures of gait variability in older adults. *Gait Posture* 2010;**31**:175–9.
19. Broom R, Du H, Clemons M, Eton D, Dranitsaris G, Simmons C, *et al.* Switching breast cancer patients with progressive bone metastases to third-generation bisphosphonates: measuring impact using the Functional Assessment of Cancer Therapy-Bone Pain. *J Pain Symptom Manage* 2009;**38**:244–57.
20. Brouwer CN, Schilder AG, van Stel HF, Rovers MM, Veenhoven RH, Grobbee DE, *et al.* Reliability and validity of functional health status and health-related quality of life questionnaires in children with recurrent acute otitis media. *Qual Life Res* 2007;**16**:1357–73.
21. Brunner HI, Klein-Gitelman MS, Miller MJ, Barron A, Baldwin N, Trombley M, *et al.* Minimal clinically important differences of the childhood health assessment questionnaire. *J Rheumatol* 2005;**32**:150–61.
22. Brunner HI, Higgins GC, Klein-Gitelman MS, Lapidus SK, Olson JC, Onel K, *et al.* Minimal clinically important differences of disease activity indices in childhood-onset systemic lupus erythematosus. *Arthritis Care Res* 2010;**62**:950–9.
23. Bruynesteyn K, Van Der Heijde D, Boers M, Lassere M, Boonen A, Edmonds J, *et al.* Minimal clinically important difference in radiological progression of joint damage over 1 year in rheumatoid arthritis: preliminary results of a validation study with clinical experts. *J Rheumatol* 2001;**28**:904–10.
24. Bruynesteyn K, Van Der Heijde D, Boers M, Saudan A, Peloso P, Paulus H, *et al.* Determination of the minimal clinically important difference in rheumatoid arthritis joint damage of the Sharp/van der Heijde and Larsen/Scott scoring methods by clinical experts and comparison with the smallest detectable difference. *Arthritis Rheum* 2002;**46**:913–20.
25. Carreon LY, Glassman SD, Campbell MJ, Anderson PA. Neck Disability Index, short form-36 physical component summary, and pain scales for neck and arm pain: the minimum clinically important difference and substantial clinical benefit after cervical spine fusion. *Spine J* 2010;**10**:469–74.
26. Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. *J Pain Symptom Manage* 2002;**24**:547–61.
27. Cella D, Eton DT, Fairclough DL, Bonomi P, Heyes AE, Silberman C, *et al.* What is a clinically meaningful change on the Functional Assessment of Cancer Therapy-Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. *J Clin Epidemiol* 2002;**55**:285–95.
28. Cella D, Yount S, Du H, Dhanda R, Gondek K, Langefeld K, *et al.* Development and validation of the Functional Assessment of Cancer Therapy-Kidney Symptom Index (FKSI). *J Support Oncol* 2006;**4**:191–9.
29. Cella D, Yount S, Brucker PS, Du H, Bukowski R, Vogelzang N, *et al.* Development and validation of a scale to measure disease-related symptoms of kidney cancer. *Value Health* 2007;**10**:285–93.

30. Cella D, Nichol MB, Eton D, Nelson JB, Mulani P. Estimating clinically meaningful changes for the Functional Assessment of Cancer Therapy – Prostate: results from a clinical trial of patients with metastatic hormone-refractory prostate cancer. *Value Health* 2009;**12**:124–9.
31. Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. *Spine* 2005;**30**:1331–4.
32. Cole JC, Lin P, Rupnow MF. Minimal important differences in the Migraine-Specific Quality of Life Questionnaire (MSQ) version. *Cephalalgia* 2009;**29**:1180–7.
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.
34. Coteur G, Feagan B, Keininger DL, Kosinski M. Evaluation of the meaningfulness of health-related quality of life improvements as assessed by the SF-36 and the EQ-5D VAS in patients with active Crohn's disease. *Aliment Pharmacol Ther* 2009;**29**:1032–41.
35. Coyne KS, Matza LS, Thompson CL, Kopp ZS, Khullar V. Determining the importance of change in the overactive bladder questionnaire. *J Urol* 2006;**176**:627–32.
36. Dawson J, Doll H, Coffey J, Jenkinson C, Oxford and Birmingham Foot and Ankle Clinical Research Group. Responsiveness and minimally important change for the Manchester-Oxford foot questionnaire (MOXFQ) compared with AOFAS and SF-36 assessments following surgery for hallux valgus. *Osteoarthritis Cartilage* 2007;**15**:918–31.
37. Dawson J, Doll H, Boller I, Fitzpatrick R, Little C, Rees J, *et al.* Comparative responsiveness and minimal change for the Oxford Elbow Score following surgery. *Qual Life Res* 2008;**17**:1257–67.
38. de Morton NA, Davidson M, Keating JL. The de Morton Mobility Index (DEMMI): an essential health index for an ageing world. *Health Qual Life Outcomes* 2008;**6**:63.
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.
40. de Vet HC, Bouter LM, Bezemer PD, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. *Int J Technol Assess Health Care* 2001;**17**:479–87.
41. de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, *et al.* Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res* 2007;**16**:131–42.
42. de Groot, V, Beckerman H, Uitdehaag BM, de Vet HC, Lankhorst GJ, Polman CH, *et al.* The usefulness of evaluative outcome measures in patients with multiple sclerosis. *Brain* 2006;**129**:10–59.
43. Demers L, Desrosiers J, Nikolova R, Robichaud L, Bravo G. Responsiveness of mobility, daily living, and instrumental activities of daily living outcome measures for geriatric rehabilitation. *Arch Phys Med Rehabil* 2010;**91**:233–40.
44. Demoulin C, Ostelo R, Knottnerus JA, Smeets RJE. Quebec back pain disability scale was responsive and showed reasonable interpretability after a multidisciplinary treatment. *J Clin Epidemiol* 2010;**63**:1249–55.
45. Drossman D, Morris CB, Hu Y, Toner BB, Diamant N, Whitehead WE, *et al.* Characterization of health related quality of life (HRQOL) for patients with functional bowel disorder (FBD) and its response to treatment. *Am J Gastroenterol* 2007;**102**:1442–53.

46. Dubois D, Gilet H, Viala-Danten M, Tack J. Psychometric performance and clinical meaningfulness of the Patient Assessment of Constipation-Quality of Life questionnaire in prucalopride (RESOLOR) trials for chronic constipation. *Neurogastroenterol Motil* 2010;**22**:e54–63.
47. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, *et al.* Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *J Pain* 2008;**9**:105–21.
48. Ekeberg OM, Bautz-Holter E, Keller A, Tveitt EK, Juel NG, *et al.* A questionnaire found disease-specific WORC index is not more responsive than SPADI and OSS in rotator cuff disease. *J Clin Epidemiol* 2010;**63**:575–84.
49. Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, *et al.* A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. *J Clin Epidemiol* 2004;**57**:898–910.
50. Eton DT, Cella D, Bacik J, Motzer RJ. A brief symptom index for advanced renal cell carcinoma. *Health Qual Life Outcomes* 2006;**4**:68.
51. Eton DT, Cella D, Yount SE, Davis KM. Validation of the functional assessment of cancer therapy – lung symptom index-12 (FLSI-12). *Lung Cancer* 2007;**57**:339–47.
52. Fairchild CJ, Chalmers RL, Begley CG. Clinically important difference in dry eye: change in IDEEL-symptom bother. *Optom Vis Sci* 2008;**85**:699–707.
53. Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, *et al.* American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. *Arthritis Rheum* 1995;**38**:727–35.
54. Fragala-Pinkham MA, Haley SM, Goodgold S. Evaluation of a community-based group fitness program for children with disabilities. *Pediatr Phys Ther* 2006;**18**:159–67.
55. Fritz JM, Irrgang JJ. A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale. *Phys Ther* 2001;**81**:776–88.
56. Funk GF, Karnell LH, Smith RB, Christensen AJ. Clinical significance of health status assessment measures in head and neck cancer: what do quality-of-life scores mean? *Arch Otolaryngol Head Neck Surg* 2004;**130**:825–9.
57. Gilbert C, Brown MC, Cappelleri JC, Carlsson M, McKenna SP. Estimating a minimally important difference in pulmonary arterial hypertension following treatment with sildenafil. *Chest* 2009;**135**:137–42.
58. Gold SM, Schulz H, Stein H, Solf K, Schulz KH, Heesen C. Responsiveness of patient-based and external rating scales in multiple sclerosis: head-to-head comparison in three clinical settings. *J Neurol Sci* 2010;**290**:102–6.
59. Groditzky GR, Tafrate RC. Imaginal exposure for anger reduction in adult outpatients: a pilot study. *J Behav Ther Exp Psychiatry* 2000;**31**:259–79.
60. Hagg O, Fritzell P, Nordwall A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. *Eur Spine J* 2003;**12**:12–20.
61. Harman JS, Manning WG, Lurie N, Liu CF. Interpreting results in mental health research. *Ment Health Serv Res* 2001;**3**:91–7.
62. Hawkes WG, Williams GR, Zimmerman S, Lapuerta P, Li T, Orwig D, *et al.* A clinically meaningful difference was generated for a performance measure of recovery from hip fracture. *J Clin Epidemiol* 2004;**57**:1019–24.

63. Hendriks EJ, Bernards AT, Berghmans BC, de Bie RA. The psychometric properties of the PRAFAB-questionnaire: a brief assessment questionnaire to evaluate severity of urinary incontinence in women. *Neurourol Urodyn* 2007;**26**:998–1007.
64. Hendriks EJ, Bernards AT, de Bie RA, de Vet HC. The minimal important change of the PRAFAB questionnaire in women with stress urinary incontinence: results from a prospective cohort study. *Neurourol Urodyn* 2008;**27**:379–87.
65. Holland AE, Hill CJ, Conron M, Munro P, McDonald CF. Small changes in six-minute walk distance are important in diffuse parenchymal lung disease. *Respir Med* 2009;**103**:1430–5.
66. Holland AE, Hill CJ, Rasekaba T, Lee A, Naughton MT, McDonald CF. Updating the minimal important difference for six-minute walk distance in patients with chronic obstructive pulmonary disease. *Arch Phys Med Rehabil* 2010;**91**:221–5.
67. Howard R, Phillips P, Johnson T, O'Brien J, Sheehan B, Lindsay J, *et al*. Determining the minimum clinically important differences for outcomes in the DOMINO trial. *Int J Geriatr Psychiatry* 2011;**26**:812–17.
68. Hsieh YW, Wang CH, Wu SC, Chen PC, Sheu CF, Hsieh CL. Establishing the minimal clinically important difference of the Barthel Index in stroke patients. *Neurorehabil Neural Repair* 2007;**21**:233–8.
69. Huang IC, Liu JH, Wu AW, Wu MY, Leite W, Hwang CC. Evaluating the reliability, validity and minimally important difference of the Taiwanese version of the diabetes quality of life (DQOL) measurement. *Health Qual Life Outcomes* 2008;**6**:87.
70. Hurst H, Bolton J. Assessing the clinical significance of change scores recorded on subjective outcome measures. *J Manipulative Physiol Ther* 2004;**27**:26–35.
71. Husted JA, Gladman DD, Cook RJ, Farewell VT. Responsiveness of health status instruments to changes in articular status and perceived health in patients with psoriatic arthritis. *J Rheumatol* 1998;**25**:2146–55.
72. Hvidsten K, Carlsson M, Stecher VJ, Symonds T, Levinson I. Clinically meaningful improvement on the quality of erection questionnaire in men with erectile dysfunction. *Int J Impot Res* 2010;**22**:45–50.
73. Jenkinson C, Peto V, Jones G, Fitzpatrick R. Interpreting change scores on the Amyotrophic Lateral Sclerosis Assessment Questionnaire (ALSAQ-40). *Clin Rehabil* 2003;**17**:380–5.
74. Jones M, Talley NJ. Minimum clinically important difference for the Nepean Dyspepsia Index, a validated quality of life scale for functional dyspepsia. *Am J Gastroenterol* 2009;**104**:1483–8.
75. Jones PW. Interpreting thresholds for a clinically significant change in health status in asthma and COPD. *Eur Respir J* 2002;**19**:398–404.
76. Jordan K, Dunn KM, Lewis M, Croft P. A minimal clinically important difference was derived for the Roland–Morris Disability Questionnaire for low back pain. *J Clin Epidemiol* 2006;**59**:45–52.
77. Juniper EF, Gruffydd-Jones K, Ward S, Svensson K. Asthma Control Questionnaire in children: validation, measurement properties, interpretation. *Eur Respir J* 2010;**36**:1410–16.
78. Kaplan RM. The minimally clinically important difference in generic utility-based measures. *COPD* 2005;**2**:91–7.
79. Karsten J, Hartman CA, Ormel J, Nolen WA, Penninx BWJH. Subthreshold depression based on functional impairment better defined by symptom severity than by number of DSM-IV symptoms. *J Affect Disorder* 2010;**123**:230–7.

80. Kawata AK, Revicki DA, Thakkar R, Jiang P, Krause S, Davidson MH, *et al.* Flushing ASsessment Tool (FAST): psychometric properties of a new measure assessing flushing symptoms and clinical impact of niacin therapy. *Clin Drug Invest* 2009;**29**:215–29.
81. Kelleher CJ, Pleil AM, Reese PR, Burgess SM, Brodish PH. How much is enough and who says so? *BJOG* 2004;**111**:605–12.
82. Kemmler G, Zubernigg A, Gattringer K, Rumpold G, Giesinger J, Sperner-Unterweger B, *et al.* A new approach to combining clinical relevance and statistical significance for evaluation of quality of life changes in the individual patient. *J Clin Epidemiol* 2010;**63**:171–9.
83. Khanna D, Pope JE, Khanna PP, Maloney M, Samedi N, Norrie D, *et al.* The minimally important difference for the fatigue visual analog scale in patients with rheumatoid arthritis followed in an academic clinical practice. *J Rheumatol* 2008;**35**:2339–43.
84. Kocks JW, Tuinenga MG, Uil SM, van den Berg JW, Ståhl E, van der Molen T. Health status measurement in COPD: the minimal clinically important difference of the clinical COPD questionnaire. *Respirator Res* 2006;**7**:62.
85. Kovacs FM, Abaira V, Royuela A, Corcoll J, Alegre L, Cano A, *et al.* Minimal clinically important change for pain intensity and disability in patients with nonspecific low back pain. *Spine* 2007;**32**:2915–20.
86. Kozma CM, Slaton TL, Monz BU, Hodder R, Reese PR. Development and validation of a patient satisfaction and preference questionnaire for inhalation devices. *Treat Respir Med* 2005;**4**:41–52.
87. Kulkarni AV. Distribution-based and anchor-based approaches provided different interpretability estimates for the Hydrocephalus Outcome Questionnaire. *J Clin Epidemiol* 2006;**59**:176–84.
88. Kupferberg DH, Kaplan RM, Slymen DJ, Ries AL. Minimal clinically important difference for the UCSD Shortness of Breath Questionnaire. *J Cardiopulm Rehabil* 2005;**25**:370–7.
89. Kvam AK, Fayers P, Wisloff F. What changes in health-related quality of life matter to multiple myeloma patients? A prospective study. *Eur J Haematol* 2010;**84**:345–53.
90. Kwon S, Perera S, Pahor M, Katula JA, King AC, Groessl EJ, *et al.* What is a meaningful change in physical performance? Findings from a clinical trial in older adults (the LIFE-P study). *J Nutr Health Aging* 2009;**13**:538–44.
91. Lai JS, Cella D, Kupst MJ, Holm S, Kelly ME, Bode RK, *et al.* Measuring fatigue for children with cancer: development and validation of the pediatric Functional Assessment of Chronic Illness Therapy-Fatigue (pedsFACIT-F). *J Pediatr Hematol Oncol* 2007;**29**:471–9.
92. Lang CE, Edwards DF, Birkenmeier RL, Dromerick AW. Estimating minimal clinically important differences of upper-extremity measures early after stroke. *Arch Phys Med Rehabil* 2008;**89**:1693–700.
93. Las HC, Quintana JM, Padierna JA, Bilbao A, Munoz P, Francis CE. Health-Related Quality of Life for Eating Disorders questionnaire version-2 was responsive 1-year after initial assessment. *J Clin Epidemiol* 2007;**60**:825–33.
94. Lasch K, Joish VN, Zhu Y, Rosa K, Qiu C, Crawford B. Validation of the sleep impact scale in patients with major depressive disorder and insomnia. *Curr Med Res Opin* 2009;**25**:1699–710.
95. Lauridsen HH, Manniche C, Korsholm L, Grunnet-Nilsson N, Hartvigsen J. What is an acceptable outcome of treatment before it begins? Methodological considerations and implications for patients with chronic low back pain. *Eur Spine J* 2009;**18**:1858–66.
96. Laviolette L, Bourbeau J, Bernard S, Lacasse Y, Pepin V, Breton MJ, *et al.* Assessing the impact of pulmonary rehabilitation on functional status in COPD. *Thorax* 2008;**63**:115–21.

97. Leidy NK, Wyrwich KW. Bridging the gap: using triangulation methodology to estimate minimal clinically important differences (MCIDs). *COPD* 2005;**2**:157–65.
98. Lemieux J, Beaton DE, Hogg-Johnson S, Bordeleau LJ, Goodwin PJ. Three methods for minimally important difference: no relationship was found with the net proportion of patients improving. *J Clin Epidemiol* 2007;**60**:448–55.
99. Liang MH. The American College of Rheumatology response criteria for systemic lupus erythematosus clinical trials – measures of overall disease activity. *Arthritis Rheum* 2004;**50**:3418–26.
100. Lin KC, Hsieh YW, Wu CY, Chen CL, Jang Y, Liu JS. Minimal detectable change and clinically important difference of the Wolf Motor Function Test in stroke patients. *Neurorehabil Neural Repair* 2009;**23**:429–34.
101. Lin KC, Fu T, Wu CY, Wang YH, Liu JS, Hsieh CJ, et al. Minimal detectable change and clinically important difference of the Stroke Impact Scale in stroke patients. *Neurorehabil Neural Repair* 2010;**24**:486–92.
102. Locker D, Jokovic A, Clarke M. Assessing the responsiveness of measures of oral health-related quality of life. *Community Dent Oral Epidemiol* 2004;**32**:10–18.
103. Machado P, Landewe R, Lie E, Kvien TK, Braun J, Baker D, et al. Ankylosing Spondylitis Disease Activity Score (ASDAS): defining cut-off values for disease activity states and improvement scores. *Ann Rheum Dis* 2011;**70**:47–53.
104. Malden PE, Thomson WM, Jokovic A, Locker D. Changes in parent-assessed oral health-related quality of life among young children following dental treatment under general anaesthetic. *Community Dent Oral Epidemiol* 2008;**36**:108–17.
105. Maringwa JT, Quinten C, King M, Ringash J, Osoba D, Coens C, et al. Minimal important differences for interpreting health-related quality of life scores from the EORTC QLQ-C30 in lung cancer patients participating in randomized controlled trials. *Support Care Cancer* 2011;**19**:1753–60.
106. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. *Soc Sci Med* 2005;**60**:1571–82.
107. Martin RL, Irrgang JJ, Burdett RG, Conti SF, Van Swearingen JM. Evidence of validity for the Foot and Ankle Ability Measure (FAAM). *Foot Ankle Int* 2005;**26**:968–83.
108. Martínez-Martin P, Carod-Artal FJ, da Silveira RL, Ziolkowski S, Vargas AP, Kummer W, et al. Longitudinal psychometric attributes, responsiveness, and importance of change: an approach using the SCOPA-Psychosocial questionnaire. *Mov Disord* 2008;**23**:1516–23.
109. Mathias SD, Gao SK, Rutstein M, Snyder CF, Wu AW, Cella D. Evaluating clinically meaningful change on the ITP-PAQ: preliminary estimates of minimal important differences. *Curr Med Res Opin* 2009;**25**:375–83.
110. Maughan EF, Lewis JS. Outcome measures in chronic low back pain. *Eur Spine J* 2010;**19**:1484–94.
111. McLeod LD, Fehnel SE, Brandman J, Symonds T. Evaluating minimal clinically important differences for the acne-specific quality of life questionnaire. *Pharmacoeconomics* 2003;**21**:1069–79.
112. McNair PJ, Prapavessis H, Collier J, Bassett S, Bryant A, Larmer P. The lower-limb tasks questionnaire: an assessment of validity, reliability, responsiveness, and minimal important differences. *Arch Phys Med Rehabil* 2007;**88**:993–1001.
113. Meads DM, McKenna SP, Kahler K. The quality of life of parents of children with atopic dermatitis: interpretation of PIQoL-AD scores. *Qual Life Res* 2005;**14**:2235–45.

114. Meads DM, McKenna SP, Doughty N, Das C, Gin-Sing W, Langley J, *et al.* The responsiveness and validity of the CAMPHOR Utility Index. *Eur Respir J* 2008;**32**:1513–19.
115. 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.
116. Michener LA, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: reliability, validity, and responsiveness. *J Shoulder Elbow Surg* 2002;**11**:587–94.
117. Middel B, Stewart R, Bouma J, van Sonderen E, van den Heuvel WJ. How to validate clinically important change in health-related functional status. Is the magnitude of the effect size consistently related to magnitude of change as indicated by a global question rating? *J Eval Clin Pract* 2001;**7**:399–410.
118. Miller KL, Walt JG, Mink DR, Satram-Hoang S, Wilson SE, Perry HD, *et al.* Minimal clinically important difference for the ocular surface disease index. *Arch Ophthalmol* 2010;**128**:94–101.
119. Mills K, Blanch P, Vicenzino B. Identifying clinically meaningful tools for measuring comfort perception of footwear. *Med Sci Sport Exerc* 2010;**42**:1966–71.
120. Morris C, Doll H, Davies N, Wainwright A, Theologis T, Willett K, *et al.* The Oxford Ankle Foot Questionnaire for children: responsiveness and longitudinal validity. *Qual Life Res* 2009;**18**:1367–76.
121. Moser JS, Barker KL, Doll HA, Carr AJ. Comparison of two patient-based outcome measures for shoulder instability after nonoperative treatment. *J Shoulder Elbow Surg* 2008;**17**:886–92.
122. Nasrallah H, Morosini P, Gagnon DD. Reliability, validity and ability to detect change of the Personal and Social Performance scale in patients with stable schizophrenia. *Psychiatry Res* 2008;**161**:213–24.
123. Nichol MB, Epstein JD. Separating gains and losses in health when calculating the minimum important difference for mapped utility measures. *Qual Life Res* 2008;**17**:955–61.
124. Nieves JW, Li T, Zion M, Gussekloo J, Pahor M, Bernabei R, *et al.* The clinically meaningful change in physical performance scores in an elderly cohort. *Aging* 2007;**19**:484–91.
125. Norquist JM, Fitzpatrick R, Jenkinson C. Health-related quality of life in amyotrophic lateral sclerosis: determining a meaningful deterioration. *Qual Life Res* 2004;**13**:1409–14.
126. Oeffinger D, Bagley A, Rogers S, Gorton G, Kryscio R, Abel M, *et al.* Outcome tools used for ambulatory children with cerebral palsy: responsiveness and minimum clinically important differences. *Dev Med Child Neurol* 2008;**50**:918–25.
127. Ornetti P, Brandt K, Hellio-Le Graverand MP, Hochberg M, Hunter DJ, Kloppenburg M, *et al.* OARSI-OMERACT definition of relevant radiological progression in hip/knee osteoarthritis. *Osteoarthritis Cartilage* 2009;**17**:856–63.
128. Paltamaa J, Sarasoja T, Leskinen E, Wikström J, Mälkiä E. Measuring deterioration in international classification of functioning domains of people with multiple sclerosis who are ambulatory. *Phys Ther* 2008;**88**:176–90.
129. Partridge MR, Miravittles M, Stahl E, Karlsson N, Svensson K, Welte T. Development and validation of the Capacity of Daily Living during the Morning questionnaire and the Global Chest Symptoms Questionnaire in COPD. *Eur Respir J* 2010;**36**:96–104.
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

- schizophrenia: interpretation of results of a pooled analysis of three phase III trials of paliperidone extended-release tablets. *Clin Ther* 2010;**32**:275–92.
131. Pejtersen JH, Bjorner JB, Hasle P. Determining minimally important score differences in scales of the Copenhagen Psychosocial Questionnaire. *Scand J Public Health* 2010;**38**:33–41.
 132. Pepin V, Laviolette L, Brouillard C, Sewell L, Singh SJ, Revill SM, *et al*. Significance of changes in endurance shuttle walking performance. *Thorax* 2011;**66**:115–20.
 133. Perera S, Mody SH, Woodman RC, Studenski SA. Meaningful change and responsiveness in common physical performance measures in older adults. *J Am Geriatr Soc* 2006;**54**:743–9.
 134. Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. *Health Qual Life Outcomes* 2007;**5**:70.
 135. Pike E, Landers MR. Responsiveness of the physical mobility scale in long-term care facility residents. *J Geriatr Phys Ther* 2010;**33**:92–8.
 136. Polson K, Reid D, McNair PJ, Larmer P. Responsiveness, minimal importance difference and minimal detectable change scores of the shortened disability arm shoulder hand (QuickDASH) questionnaire. *Man Ther* 2010;**15**:404–7.
 137. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HC. Minimal clinically important change of the Neck Disability Index and the Numerical Rating Scale for patients with neck pain. *Spine* 2007;**32**:3047–51.
 138. Puhan MA, Mador MJ, Held U, Goldstein R, Guyatt GH, Schünemann HJ. Interpretation of treatment changes in 6-minute walk distance in patients with COPD. *Eur Respir J* 2008;**32**:637–43.
 139. Puhan MA, Frey M, Büchi S, Schünemann HJ. The minimal important difference of the hospital anxiety and depression scale in patients with chronic obstructive pulmonary disease. *Health Qual Life Outcomes* 2008;**6**:46.
 140. Quintana JM, Escobar A, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement. *Osteoarthritis Cartilage* 2005;**13**:1076–83.
 141. Quittner AL, Modi AC, Wainwright C, Otto K, Kiriara J, Montgomery AB. Determination of the minimal clinically important difference scores for the Cystic Fibrosis Questionnaire-Revised respiratory symptom scale in two populations of patients with cystic fibrosis and chronic *Pseudomonas aeruginosa* airway infection. *Chest* 2009;**135**:1610–18.
 142. Rabeneck L, Cook KF, Wristers K, Soucek J, Menke T, Wray NP. SODA (severity of dyspepsia assessment): a new effective outcome measure for dyspepsia-related health. *J Clin Epidemiol* 2001;**54**:755–65.
 143. Raj AA, Pavord DI, Birring SS. Clinical cough IV: what is the minimal important difference for the Leicester Cough Questionnaire? *Handbook Exp Pharmacol* 2009;**187**:311–20.
 144. Rejas J, Gil A, San Isidro C, Palacios G, Carrasco P. [Sensitivity to change and minimally important difference of the Spanish version of the life-satisfaction questionnaire LISAT-8 in male patients with erectile dysfunction.] *Med Clin (Barc)* 2005;**124**:165–71.
 145. Rejas J, Ruiz M, Pardo A. [Standard error of measurement: an alternative to minimally important difference to access changes in patient-reported-health-outcomes.] *An Med Interna* 2007;**24**:415–20.
 146. Rejas J, Pardo A, Ruiz MA. Standard error of measurement as a valid alternative to minimally important difference for evaluating the magnitude of changes in patient-reported outcomes measures. *J Clin Epidemiol* 2008;**61**:350–6.

147. Rendas-Baum R, Yang M, Cattelin F, Wallenstein GV, Fisk JD. A novel approach to estimate the minimally important difference for the Fatigue Impact Scale in multiple sclerosis patients. *Qual Life Res* 2010;**19**:1349–58.
148. Rentz AM, Matza LS, Secnik K, Swensen A, Revicki DA. Psychometric validation of the child health questionnaire (CHQ) in a sample of children and adolescents with attention-deficit/hyperactivity disorder. *Qual Life Res* 2005;**14**:719–34.
149. Ries AL. Minimally clinically important difference for the UCSD Shortness of Breath Questionnaire, Borg Scale, and Visual Analog Scale. *COPD* 2005;**2**:105–10.
150. Robinson D Jr, Zhao N, Gathany T, Kim LL, Cella D, Revicki D. Health perceptions and clinical characteristics of relapsing-remitting multiple sclerosis patients: baseline data from an international clinical trial. *Curr Med Res Opin* 2009;**25**:1121–30.
151. Rossi MD, Eberle T, Roche M, Waggoner M, Blake R, Burwell B, *et al*. Delaying knee replacement and implications on early postoperative outcomes: a pilot study. *Orthopedics* 2009;**32**:885–93.
152. Roy KM, Roberts MC, Vernberg EM, Randall CJ. Measuring treatment outcome for children with serious emotional disturbances: discriminant validity and clinical significance of the child and adolescent functioning assessment scale. *J Child Fam Stud* 2008;**17**:232–40.
153. Samsa G, Edelman D, Rothman ML, Williams GR, Lipscomb J, Matchar D. Determining clinically important differences in health status measures: a general approach with illustration to the Health Utilities Index Mark II. *Pharmacoeconomics* 1999;**15**:141–55.
154. Schatz M, Kosinski M, Yarlas AS, Hanlon J, Watson ME, Jhingran P. The minimally important difference of the Asthma Control Test. *J Allergy Clin Immunol* 2009;**124**:719–23.
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.
156. Shi H-Y, Lee K-T, Lee H-H, Uen Y-H, Na H-L, Chao F-T, *et al*. The minimal clinically important difference in the Gastrointestinal Quality-of-Life Index after cholecystectomy. *Surg Endosc* 2009;**23**:2708–12.
157. Shikier R, Harding G, Leahy M, Lennox RD. Minimal important difference (MID) of the Dermatology Life Quality Index (DLQI): results from patients with chronic idiopathic urticaria. *Health Qual Life Outcomes* 2005;**3**:36.
158. Shikier R, Willian MK, Okun MM, Thompson CS, Revicki DA. The validity and responsiveness of three quality of life measures in the assessment of psoriasis patients: results of a phase II study. *Health Qual Life Outcomes* 2006;**4**:71.
159. Shulman LM, Gruber-Baldini AL, Anderson KE, Fishman PS, Reich SG, Weiner WJ. The clinically important difference on the unified Parkinson's disease rating scale. *Arch Neurol* 2010;**67**:64–70.
160. Sim J, Jordan K, Lewis M, Hill J, Hay EM, Dziedzic K. Sensitivity to change and internal consistency of the Northwick Park Neck Pain Questionnaire and derivation of a minimal clinically important difference. *Clin J Pain* 2006;**22**:820–6.
161. Sim J, Jordan K, Lewis M, Hill J, Hay EM, Dziedzic K. Sensitivity to change and internal consistency of the Northwick Park Neck Pain Questionnaire and derivation of a minimal clinically important difference. *Clin J Pain* 2006;**22**:820–6.
162. Smith M, Wells J, Borrie M. Treatment effect size of memantine therapy in Alzheimer disease and vascular dementia. *Alzheimer Dis* 2006;**20**:133–7.

163. Spiegel BM, Younossi ZM, Hays RD, Revicki D, Robbins S, Kanwal F. Impact of hepatitis C on health related quality of life: a systematic review and quantitative assessment. *Hepatology* 2005;**41**:790–800.
164. Spies-Dorgelo MN, Terwee CB, Stalman WA, van der Windt DA. Reproducibility and responsiveness of the Symptom Severity Scale and the hand and finger function subscale of the Dutch arthritis impact measurement scales (Dutch-AIMS2-HFF) in primary care patients with wrist or hand problems. *Health Qual Life Outcomes* 2006;**4**:87.
165. Staples MP, Forbes A, Green S, Buchbinder R. Shoulder-specific disability measures showed acceptable construct validity and responsiveness. *J Clin Epidemiol* 2010;**63**:163–70.
166. Stargardt T, Gonder-Frederick L, Krobot KJ, Alexander CM. Fear of hypoglycaemia: defining a minimum clinically important difference in patients with type 2 diabetes. *Health Qual Life Outcomes* 2009;**7**:91.
167. Steel JL, Eton DT, Cella D, Olek MC, Carr BI. Clinically meaningful changes in health-related quality of life in patients diagnosed with hepatobiliary carcinoma. *Ann Oncol* 2006;**17**:304–12.
168. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. *J Clin Epidemiol* 1995;**48**:1369–78.
169. Stull DE, Vernon MK, Canonica GW, Crespi S, Sandor D. Using the Congestion Quantifier Seven-Item Test to assess change in patient symptoms and their impact. *Allergy Asthma Proc* 2008;**29**:295–303.
170. Submacular Surgery Trials Research Group. Evaluation of minimum clinically meaningful changes in scores on the National Eye Institute Visual Function Questionnaire (NEI-VFQ) SST Report Number 19. *Ophthalmic Epidemiol* 2007;**14**:205–15.
171. Suner IJ, Kokame GT, Yu E, Ward J, Dolan C, Bressler NM. Responsiveness of NEI VFQ-25 to changes in visual acuity in neovascular AMD: validation studies from two phase 3 clinical trials. *Invest Ophthalmol Vis Sci* 2009;**50**:3629–35.
172. Sutherland ER, Make BJ. Maximum exercise as an outcome in COPD: minimal clinically important difference. *COPD* 2005;**2**:137–41.
173. Swigris JJ, Wamboldt FS, Behr J, Du Bois RM, King TE, Raghu G, et al. The 6 minute walk in idiopathic pulmonary fibrosis: longitudinal changes and minimum important difference. *Thorax* 2010;**65**:173–7.
174. Swigris JJ, Brown KK, Behr J, Du Bois RM, King TE, Raghu G, et al. The SF-36 and SGRQ: validity and first look at minimum important differences in IPF. *Respir Med* 2010;**104**:296–304.
175. Symonds T, Spino C, Sisson M, Soni P, Martin M, Gunter L, et al. Methods to determine the minimum important difference for a sexual event diary used by postmenopausal women with hypoactive sexual desire disorder. *J Sex Med* 2007;**4**:1328–35.
176. Tamber AL, Wilhelmsen KT, Strand LI. Measurement properties of the Dizziness Handicap Inventory by cross-sectional and longitudinal designs. *Health Qual Life Outcomes* 2009;**7**:101.
177. Terwee CB, Dekker FW, Mourits MP, Gerding MN, Baldeschi L, Kalmann R, et al. Interpretation and validity of changes in scores on the Graves' ophthalmopathy quality of life questionnaire (GO-QOL) after different treatments. *Clin Endocrinol (Oxf)* 2001;**54**:391–8.
178. Terwee CB, Roorda LD, Knol DL, De Boer MR, de Vet HC. Linking measurement error to minimal important change of patient-reported outcomes. *J Clin Epidemiol* 2009;**62**:1062–7.
179. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, et al. Mind the MIC: large variation among populations and methods. *J Clin Epidemiol* 2010;**63**:524–34.

180. Tsai CL, Hodder RV, Page JH, Cydulka RK, Rowe BH, Camargo CA Jr. The short-form chronic respiratory disease questionnaire was a valid, reliable, and responsive quality-of-life instrument in acute exacerbations of chronic obstructive pulmonary disease. *J Clin Epidemiol* 2008;**61**:489–97.
181. Tsakos G, Bernabe E, D'Aiuto F, Pikhart H, Tonetti M, Sheiham A, *et al.* Assessing the minimally important difference in the Oral Impact on Daily Performances index in patients treated for periodontitis. *J Clin Periodontol* 2010;**37**:903–9.
182. Tsang RCC. Measurement properties of the Hong Kong Chinese version of the Roland–Morris Disability Questionnaire. *Hong Kong Physiother J* 2004;**22**:40–9.
183. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, *et al.* The minimal detectable change cannot reliably replace the minimal important difference. *J Clin Epidemiol* 2010;**63**:28–36.
184. Twiss J, Doward LC, McKenna SP, Eckert B. Interpreting scores on multiple sclerosis-specific patient reported outcome measures (the PRIMUS and U-FIS). *Health Qual Life Outcomes* 2010;**8**:117.
185. van der Roer N, Ostelo RW, Bekkering GE, van Tulder MW, de Vet HC. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. *Spine* 2006;**31**:578–82.
186. van Grootel RJ, van der Glas HW. Statistically and clinically important change of pain scores in patients with myogenous temporomandibular disorders. *Eur J Pain*:2009;**13**:506–10.
187. van Stel HF, Mailler, AR, Colland VT, Everaerd W. Interpretation of change and longitudinal validity of the quality of life for respiratory illness questionnaire (QoLRIQ) in inpatient pulmonary rehabilitation. *Qual Life Res* 2003;**12**:133–45.
188. Vernon MK, Rentz AM, Wyrwich KW, White MV, Grienemberger A. Psychometric validation of two patient-reported outcome measures to assess symptom severity and changes in symptoms in hereditary angioedema. *Qual Life Res* 2009;**18**:929–39.
189. Vernon MK, Revicki DA, Awad AG, Dirani R, Panish J, Canuso CM, *et al.* Psychometric evaluation of the Medication Satisfaction Questionnaire (MSQ) to assess satisfaction with antipsychotic medication among schizophrenia patients. *Schizophr Res* 2010;**118**:271–8.
190. Viala-Danten M, Dubois D, Gilet H, Martin S, Peeters K, Cella D. Psychometric evaluation of the functional assessment of HIV Infection (FAHI) questionnaire and its usefulness in clinical trials. *Qual Life Res* 2010;**19**:1215–27.
191. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health Qual Life Outcomes* 2003;**1**:4.
192. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Qual Life Res* 2005;**14**:1523–32.
193. Wang YC, Hart DL, Stratford PW, Mioduski JE. Clinical interpretation of a lower-extremity functional scale-derived computerized adaptive test. *Phys Ther* 2009;**89**:957–68.
194. Wang YC, Hart DL, Stratford PW, Mioduski JE. Clinical interpretation of computerized adaptive test-generated outcome measures in patients with knee impairments. *Arch Phys Med Rehabil* 2009;**90**:1340–8.
195. Wang YC, Hart DL, Stratford PW, Mioduski JE. Clinical interpretation of computerized adaptive test outcome measures in patients with foot/ankle impairments. *J Orthop Sports Phys Ther* 2009;**39**:753–64.
196. Wang YC, Hart DL, Werneke M, Stratford PW, Mioduski JE. Clinical interpretation of outcome measures generated from a lumbar computerized adaptive test. *Phys Ther* 2010;**90**:1323–35.

197. Wang YC, Hart DL, Cook KF, Mioduski JE. Translating shoulder computerized adaptive testing generated outcome measures into clinical practice. *J Hand Ther* 2010;**23**:372–83.
198. Wells G, Li T, Maxwell L, MacLean R, Tugwell P. Determining the minimal clinically important differences in activity, fatigue, and sleep quality in patients with rheumatoid arthritis. *J Rheumatol* 2007;**34**:280–9.
199. Wiebe S, Matijevic S, Eliasziw M, Derry PA. Clinically important change in quality of life in epilepsy. *J Neurol Neurosurg Psychiatry* 2002;**73**:116–20.
200. Wiersinga WM, Prummel MF, Terwee CB. Effects of Graves' ophthalmopathy on quality of life. *J Endocrinol Invest* 2004;**27**:259–64.
201. Williams VS, Morlock RJ, Feltner D. Psychometric evaluation of a visual analog scale for the assessment of anxiety. *Health Qual Life Outcomes* 2010;**8**:57.
202. Wolfe F, Michaud K. Assessment of pain in rheumatoid arthritis: minimal clinically significant difference, predictors, and the effect of anti-tumor necrosis factor therapy. *J Rheumatol* 2007;**34**:1674–83.
203. Wright P, Marshall L, Smith AB, Velikova G, Selby P. Measurement and interpretation of social distress using the social difficulties inventory (SDI). *Eur J Cancer* 2008;**44**:1529–35.
204. Wuang YP, Su CY. Reliability and responsiveness of the Bruininks–Oseretsky Test of Motor Proficiency-Second Edition in children with intellectual disability. *Res Dev Disabil* 2009;**30**:847–55.
205. Wyrwich K, Harnam N, Revicki DA, Locklear JC, Svedsäter H, Endicott J. Assessing health-related quality of life in generalized anxiety disorder using the Quality Of Life Enjoyment and Satisfaction Questionnaire. *Int Clin Psychopharmacol* 2009;**24**:289–95.
206. Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Med Care* 1999;**37**:469–78.
207. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol* 1999;**52**:861–73.
208. Wyrwich KW, Tierney WM, Wolinsky FD. Using the standard error of measurement to identify important changes on the Asthma Quality of Life Questionnaire. *Qual Life Res* 2002;**11**:1–7.
209. Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Interpreting quality-of-life data: methods for community consensus in asthma. *Ann Allergy Asthma Immunol* 2006;**96**:826–33.
210. Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Measuring patient and clinician perspectives to evaluate change in health-related quality of life among patients with chronic obstructive pulmonary disease. *J Gen Intern Med* 2007;**22**:161–70.
211. Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Triangulating patient and clinician perspectives on clinically important differences in health-related quality of life among patients with heart disease. *Health Serv Res* 2007;**42**:2257–74.
212. Yang M, Morin CM, Schaefer K, Wallenstein GV. Interpreting score differences in the Insomnia Severity Index: using health-related outcomes to define the minimally important difference. *Curr Med Res Opin* 2009;**25**:2487–94.
213. Yost KJ. Using multiple anchor- and distribution-based estimates to evaluate clinically meaningful change on the Functional Assessment of Cancer Therapy. *Value Health* 2005;**8**:117–27.
214. Yost KJ, Cella D, Chawla A, Holmgren E, Eton DT, Ayanian JZ, *et al.* Minimally important differences were estimated for the Functional Assessment of Cancer Therapy-Colorectal (FACT-C)

- instrument using a combination of distribution- and anchor-based approaches. *J Clin Epidemiol* 2005;**58**:1241–51.
215. Young BA, Walker MJ, Strunce JB, Boyles RE, Whitman JM, Childs JD. Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. *Spine J* 2009;**9**:802–8.
216. Yount S, List M, Du H, Yost K, Bode R, Brockstein B, *et al.* A randomized validation study comparing embedded versus extracted FACT Head and Neck Symptom Index scores. *Qual Life Res* 2007;**16**:1615–26.

Appendix 4 Survey form sent to UK- and Ireland-based trialists



**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

Oral and Gastrointestinal

Primary Care

Renal and Urogenital

Reproductive Health

Respiratory

Skin

Stroke

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	<input type="checkbox"/>
Distribution methods	<input type="checkbox"/>
Health economic methods	<input type="checkbox"/>
Opinion seeking methods	<input type="checkbox"/>
Pilot study	<input type="checkbox"/>
Review of evidence base	<input type="checkbox"/>
Standardised effect size approach	<input type="checkbox"/>
Other, <i>please state</i>	<input type="checkbox"/>
We have not used any of these methods	<input type="checkbox"/>

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

Anchor methods	<input type="checkbox"/>
Distribution methods	<input type="checkbox"/>
Health economic methods	<input type="checkbox"/>
Opinion seeking methods	<input type="checkbox"/>
Pilot study	<input type="checkbox"/>
Review of evidence base	<input type="checkbox"/>
Standardised effect size approach	<input type="checkbox"/>
Other, <i>please state</i>	<input type="checkbox"/>
I would not recommend any method	<input type="checkbox"/>

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)

A realistic difference given the interventions under evaluation

☐

A difference which would led to an achievable sample size

☐

A difference that would be viewed as important by a relevant stakeholder group (e.g. clinicians)

☐

Other, *please state*.....

☐

.....

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?

Yes

☐

No

☐

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/

A decorative graphic consisting of numerous thin, parallel green lines that curve from the left side of the page towards the right, creating a sense of movement and flow.

EME
HS&DR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

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

Published by the NIHR Journals Library