A systematic review of sample size calculations in high-profile surgical trials that use patient-reported outcome measures.

Jacklin C[1], Rodrigues JN (corresponding)[2,3], Collins J[1], Cook J[1], Harrison CJ[1]

1. Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
2. Department of Plastic Surgery, Stoke Mandeville Hospital, Buckinghamshire Healthcare NHS Trust, Aylesbury, UK
3. Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, UK

Correspondence:

Jeremy Rodrigues

Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, UK

j.rodrigues@warwick.ac.uk

Funding:

Conrad J. Harrison is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (NIHR300684). Jeremy N. Rodrigues is funded by a NIHR postdoctoral fellowship (PDF-2017-10-075). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. There was no involvement of the funder in study design, data collection, data analysis, manuscript preparation or publication decisions. All authors had complete access to the study data that support the publication.

Category:

Short Report

Previous communications:

This data has not been previously presented in a meeting or to a society.

Introduction

Patient reported outcome measures (PROMs) are questionnaires that assess aspects of patients' perceived health. They may be generic, or specific to a condition, disease, or treatment[1]. The use of PROMs has gained credibility and popularity[2], and this trajectory is set to continue with increasing recognition from governing and advisory bodies[3,4].

Determining a meaningful difference between PROM scores is not intuitive because, without context, PROM scores are challenging to interpret. This is relevant to setting target differences for sample size calculations in randomised controlled trials (RCTs). Flawed target differences can lead to type 1 and type 2 errors, and/or excessive recruitment which incurs unnecessary cost and risk to participants[5].

For trials using PROMs as primary outcome measures, the target difference should be the PROM's minimal important difference (MID). A popular definition of MID is "the smallest difference in score in the domain of interest which patients perceived as beneficial and which would mandate, in the absence of troublesome side-effects and excessive cost, a change in the patient's management"[6]. The MID should not be confused with the minimal detectable change (MDC) - the smallest change in score that cannot be completely attributed to the instrument's inherent test-retest error.

There are several methods to estimate MIDs, outlined by the Difference ELicitation in TriAls (DELTA[2]) guidelines for RCT sample size calculations, which vary in methodological rigour (Box 1). When selecting a target difference for a clinical trial, best practice is to triangulate relevant MID estimates[7–9]. MIDs are specific for given

populations, treatments, and follow-up durations[7], because they balance the benefits and disadvantages of an intervention in context. An MID applied out of context can compromise a trial's results.

Target differences are often determined on a convenient or arbitrary basis[10], which introduces the risk of over- or underpowering. The DELTA[2] guidelines aim to address this by standardising sample size calculations[5,11,12] (Box 2). This is relevant to surgical research where criticism of low quality evidence has been met with initiatives to fund to surgical trial units to support patient-centred research[13,14]. Increasingly, these trials are using PROMs as primary outcome measures[15].

Our aim was to appraise sample size calculations according to DELTA[2] guidance[11], with a focus on target difference determination, in surgical RCTs that used a PROM as its primary outcome, and were published in high-impact journals between 2013 and 2021.

Methods

Protocol and registration

The protocol was registered on PROSPERO (CRD42020180233).

Eligibility criteria

Eligible studies were published between 1st January 2013 and 10th August 2021 as in 2013 the UK launched an initiative to improve surgical research[14].

Inclusion criteria:

- Randomised controlled trial where the intervention and/or comparator was surgery, defined as using instrumentation to change macro-anatomy with the aim of improving health
- All forms of RCT irrespective of type (parallel/cluster) and framework (superiority/non-inferior).
- The primary outcome was a PROM

Exclusion criteria:

- The primary outcome was a composite

Information sources and search

A systematic search of PubMed was conducted on 10th August 2021, using a bespoke search strategy comprising index and freetext terms (Suppl. 1). The search was limited to the five highest impact medical journals and five highest impact surgical journals as determined by Thomson Reuters[12], which have large international readerships of clinicians, academics and policy makers.

Study selection and data extraction

Title/abstract and full-text screening was conducted by author CJ against the eligibility criteria. Disagreements for inclusion were resolved by consensus. Data extraction was performed independently by authors CJ and CJH using piloted forms. Following piloting, an additional column to extract the methods used to determine the target difference, as per DELTA[2], was added. Variables extracted included: study characteristics, DELTA[2] reporting items, methods for determining target difference (Box 1), and funding details (Suppl. Table 1). Disagreements were resolved by a

third author (JNR). Where the method to determine the target difference was not explicitly justified the method was assigned to 'Author determined, not otherwise specified (NOS)'. DELTA^2 is primarily intended for superiority trials, we therefore summarised non-inferiority and equivalence trials separately.

Results

Systematic search

A total of 1528 studies were identified by the search strategy, of which 14 duplicates were removed. The remaining 1514 were double screened by their title and abstract, at which point 1392 were excluded. 122 full-text studies were assessed. A further 65 were excluded (justifications in Suppl. Fig. 1), leaving 57 for appraisal.

Reporting of DELTA^2 items

Of the eligible studies, 51 had a superiority design. Statistical significance level, power, target difference, and adjustments to the sample size calculation were consistently reported (Table 1). No study reported a sensitivity analyses for the sample size, and 9/51 studies did not justify their target difference.

Target difference determination in superiority trials

Superiority trials used a range of methods to determine the target difference (Table 2, Suppl. Table 2, Suppl. Table 3). This included 12 studies that used a standardised effect size and three that used an anchor-based method. The target difference was triangulated in 11/51 (Suppl. Table 4). No study used an adaptive trial design.

Non-inferiority and equivalence trials

Four trials had a non-inferiority design, and two trials were designed to measure equivalence (Suppl. Table 5). The methods for determining the target difference included using MDCs and MIDs.

Funding

The funding source was reported in 53/57 studies. The grant total was readily accessible in 18 studies. Of those, 16 were funded by the UK National Institute of Health and Research (NIHR). In this sample of NIHR-funded trials, £28 million of UK public research funding was spent on trials with sample size calculations that did not adhere to DELTA^2 guidelines.

Discussion

In this sample of RCTs, the target difference for sample size calculations were generally not determined to DELTA^2 standards[11]. This risks over- or under-powering, and even erroneous trial conclusions. Specific weaknesses included: suboptimal methods (standardised effect size and measurement error) to determine the target difference; unclear justification for the target difference; and the application of MIDs calculated in different contexts. There were 14 studies that did use superior methods of anchor-based MIDs, consulted expert panels and triangulation.

Our findings may be explained by demands for timely answers to clinical research questions with suboptimal but readily-available MIDs, and pressure to drive down research costs by selecting larger target differences.

While we acknowledge the difficult balance between timely answers to clinical questions versus investment in measurement science, there are solutions. Adaptive trial designs allow trialists to dynamically refine trial-specific MIDs[16–18]. Funding bodies, ethics committees and journals are the gateway to research, and could promote alternative trial designs and enforce careful MID determination. Barriers to this include rigid budgets and risk aversion of commissioners and funding applicants.

We did not contact trialists therefore this leaves open the possibility that target difference determination was more sophisticated than reported.

Conclusion

In these studies, sample size calculations were generally not reported to DELTA^2 standards and target differences were frequently determined with suboptimal techniques. This can jeopardise trial findings, lead to unnecessary participant risks, and cause excess costs. In our sample, £28 million of UK taxpayer funding was spent on trials that did not meet DELTA^2 standards. These deficiencies can be addressed through investment in PROM research, closer scrutiny by funding bodies, acceptance of alternative trial designs, and greater awareness amongst trialists and clinicians.

Acknowledgements

Data used in the analysis will be made available upon reasonable request. No analytic code or study materials were used in this study. The protocol for this study was registered on PROSPERO (CRD42020180233). No analysis plan was pre-registered as this was not applicable to this study.

References

1. van der Willik EM, Terwee CB, Bos WJW, et al. Patient-reported outcome measures (PROMs): making sense of individual PROM scores and changes in PROM scores over time. *Nephrology*. 2021;26(5):391-399. doi:10.1111/NEP.13843
2. Black N. Patient reported outcome measures could help transform healthcare. *BMJ*. 2013;346(jan28 1):f167-f167. doi:10.1136/bmj.f167
3. Bottomley A, Jones D, Claassens L. Patient-reported outcomes: Assessment and current perspectives of the guidelines of the Food and Drug Administration and the reflection paper of the European Medicines Agency. *Eur J Cancer*. 2009;45(3):347-353. doi:10.1016/j.ejca.2008.09.032
4. Services H. Guidance for industry: Patient-reported outcome measures: Use in medical product development to support labeling claims: Draft guidance. *Health Qual Life Outcomes*. 2006;4:1-20. doi:10.1186/1477-7525-4-79
5. Cook JA, Julious SA, Sones W, et al. DELTA 2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ*. 2018. doi:10.1136/bmj.k3750
6. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10(4):407-415. doi:10.1016/0197-2456(89)90005-6
7. Rodrigues JN. Different terminologies that help the interpretation of outcomes. *J Hand Surg Eur Vol*. 2020;45(1):97-99. doi:10.1177/1753193419870100
8. Rodrigues JN, Mabvuure NT, Nikkhah D, Shariff Z, Davis TRC. Minimal important changes and differences in elective hand surgery. *J Hand Surg Eur*

Vol. 2015. doi:10.1177/1753193414553908

9. 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. *CMAJ*. 2001.
10. Cook JA, Julious SA, Sones W, et al. Choosing the target difference ('effect size') for a randomised controlled trial - DELTA2 guidance protocol. *Trials*. 2017. doi:10.1186/s13063-017-1969-5
11. Cook JA, Hislop JM, Altman DG, et al. Use of methods for specifying the target difference in randomised controlled trial sample size calculations: Two surveys of trialists' practice. *Clin Trials*. 2014;11:300-308. doi:10.1177/1740774514521907
12. Clavirate. https://jcr.clarivate.com/. InCites.
13. McCall B. UK implements national programme for surgical trials. *Lancet*. 2013;382(9898):1083-1084. doi:10.1016/S0140-6736(13)62009-7
14. England RC of S. Surgical Trials Initiative — Royal College of Surgeons. https://www.rcseng.ac.uk/standards-and-research/research/surgical-trials-initiative/. Accessed October 15, 2021.
15. Phillips JD, Wong SL. Patient-Reported Outcomes in Surgical Oncology: An Overview of Instruments and Scores. *Ann Surg Oncol*. 2020;27(1):45. doi:10.1245/S10434-019-07752-7
16. Dimairo M, Pallmann P, Wason J, et al. The Adaptive designs CONSORT Extension (ACE) statement: A checklist with explanation and elaboration guideline for reporting randomised trials that use an adaptive design. *BMJ*. 2020. doi:10.1136/bmj.m115
17. Thorlund K, Haggstrom J, Park JJ, Mills EJ. Key design considerations for

adaptive clinical trials: A primer for clinicians. *BMJ*. 2018. doi:10.1136/bmj.k698

18. Park JJH, Thorlund K, Mills EJ. Critical concepts in adaptive clinical trials. *Clin Epidemiol*. 2018. doi:10.2147/CLEP.S156708

Tables

Table 1 – Number of eligible superiority trials that reported the DELTA[2] recommended reporting items (/51).

DELTA[2] Recommended Reporting Item	**Number of Eligible Superiority Studies that Reported Item (/51)**
Outcome	51 (100%)
Statistical Significance Level	50 (98%)
Power	49 (96%)
Target difference	42 (82%)
Allocation Ratio	51 (100%)
Adjustments to Sample Size (e.g. Attrition rate)	41 (80%)
Sensitivity Analysis	0 (0%)
Basis for Target Difference	40 (78%)
Explanation of Choice of Target Difference	43 (84%)
Reference to Trial Protocol	39 (76%)

Table 2 - The primary outcome measure and method used to determine target difference for the sample size calculations in eligible superiority trials. NOS: Not otherwise specified.

Study	**Primary outcome measure**	**Target difference determination method**
Frobell, 2013	Knee Injury and Osteoarthritis Outcome Score, four of five subscales (KOOS4)	Distribution - measurement error
Grant, 2013	Reflux Questionnaire Score	Standardised effect size
Katz, 2013	Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC)	Triangulation
Rogmark, 2013	Short Form-36 (SF-36)	Standardised effect size
Brittenden, 2014	Aberdeen Varicose Veins Questionnaire (AVVQ)	Standardised effect size
Costa, 2014	Patient Rated Wrist Evaluation (PRWE)	Standardised effect size
Fassov, 2014	Gastrointestinal Syndrome Rating Scale-IBS questionnaire (GSRS-IBS)	Author determined, NOS
Griffin, 2014	Kerr-Atkins Score	Triangulation
Bulian, 2015	Numeric Rating Scale of Pain 1-10 (NRS-11)	Anchor-based
Lane, 2015	Aberdeen Varicose Veins Questionnaire (AVVQ)	Review of evidence base
Lurje, 2015	Validated Cosmesis and Body Image Score	Review of evidence base
Muller-Stich, 2015	Ingestion Syndrome Subscore (ISS) of the Gastrointestinal Symptom Rating Scale questionnaire (GSRS)	Author determined, NOS
Rangan, 2015	Oxford Shoulder Score (OSS)	Triangulation
Ronka, 2015	Visual Analogue Scale (VAS, 0-10)	Author determined, NOS
Skou, 2015	Knee Injury and Osteoarthritis Outcome Score, four of five subscales (KOOS4)	Distribution - measurement error
Burgmans, 2016	Numerical Rating Scale of Pain 4-10 (NRS)	Triangulation
Forsth, 2016	Oswestry Disability Index (ODI)	Author determined, NOS
Ghogawala, 2016	Physical-Component Summary Score of the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36)	Triangulation

Mi Kim, 2016	European Organization for Research and Treatment of Cancer (EORTC) gastric module (STO22)	Author determined, NOS
Kise, 2016	Knee Injury and Osteoarthritis Outcome Score, four of five subscales (KOOS4)	Distribution - measurement error
Watson, 2016	EuroQol 5 Dimensions 3 Level Score (EQ-5D-3L)	Author determined, NOS
Westin, 2016	Inguinal Pain Questionnaire (IPQ)	Author determined, NOS
Costa, 2017	Disability Rating Index (DRI)	Triangulation
Diener, 2017	European Organisation for Research and Treatment of Cancer's Quality of Life Questionnaire (EORTC QLQ-C30)	Review of evidence base
Glazener, 2017	Pelvic Organ Prolapse Symptom Score (POP-SS)	Anchor-based
Juch, 2017	Numeric Rating Scale of Pain 1-10 (NRS-11)	Triangulation
Molegraaf, 2017	Verbal Rating Scale (VRS)	Author determined, NOS
Overdevest, 2017	Roland-Morris Disability Questionnaire for Sciatica (RDQ)	Author determined, NOS
Beard, 2018	Oxford Shoulder Score (OSS)	Standardised effect size
Firanescu, 2018	Visual Analogue Score (VAS)	Triangulation
Griffin, 2018	International Hip Outcome Tool (iHOT-33)	Triangulation
So, 2018	Gastrointestinal (GI) Symptoms Score	Author determined, NOS
Verhagen, 2018	Visual Analogue Scale (VAS)	Review of evidence base
Beard, 2019	Oxford Knee Score (OKS)	Author determined, NOS
Blomström-Lundqvist, 2019	General Health Subscale Score from the Medical Outcomes Study 36-item Short-Form health survey (SF-36)	Review of evidence base
Gazzard, 2019	EuroQol 5 Dimensions 5 Levels (EQ-5D-5L) Utility Scores	Review of evidence base
Palmer, 2019	Hip Outcome Score Activities of Daily Living Subscale (HOS ADL)	Author determined, NOS
Ponds, 2019	Eckardt Symptom Score	Review of evidence base
Spechler, 2019	Gastroesophageal Reflux Disease-Health Related Quality of Life (GERD-HRQL)	Review of evidence base
Sung, 2019	Urogenital Distress Inventory (UDI)	Triangulation

van Egmond, 2019	Glasgow Health Status Inventory (GHSI)	Author determined, NOS
MacKay, 2020	Epworth Sleepiness Scale (ESS)	Author determined, NOS
Ramo, 2020	Disabilities of the Arm, Shoulder and Hand (DASH)	Anchor
Bolkenstein, 2020	Gastro-intestinal Quality of Life Index (GIQLI)	Standardised effect size
Rangan, 2020	Oxford Shoulder Score (OSS)	Standardised effect size
Dias, 2020	Patient-rated wrist evaluation (PRWE)	Standardised effect size
Manyonda, 2020	Quality-of-life domain of the Uterine Fibroid Symptom and Quality of Life (UFS-QOL) questionnaire	Standardised effect size
Jayne, 2021	Faecal Incontinence Quality of Life (FIQoL)	Standardised effect size
Reijman, 2021	International Knee Documentation Committee Score	Standardised effect size
Ghogawala, 2021	Short form 36 physical component summary (SF-36 PCS)	Triangulation
King, 2021	Visual Function Questionnaire-25 (VFQ-25)	Standardised effect size

Box 1 - Methods used to determine the target difference

Anchor-based

Target difference in the outcome measure is related (anchored) to a real-world clinical difference (e.g., requirement vs no requirement for opiate analgesia, or known meaningful difference in Global Rating of Change scores).

Opinion seeking

Expert panel determines what would be a realistic and/or important target difference.

Distribution

Measurement error: target difference is greater than inherent imprecision in the outcome measure.

Rule of thumb: target difference is greater than a substantial fraction of the scale, without further justification.

Health economic

Target difference determined on cost-effectiveness scale.

Pilot trial

Study methodology tested on small scale to inform the target difference.

Review of evidence base

Review outcomes of existing, similar studies to inform the target difference.

Standardised effect size

Target difference is determined as an arbitrary proportion on a standardised scale, for example, Cohen's d metric for a medium effect size is 0.5 of the standard deviation of baseline scores.

Box 2 – A description of the DELTA^2 recommended reporting items

Authors are recommended to report the inputs (statistical significance level, power, and target difference) of the sample size calculation. They should provide any details of any assessment of the sensitivity of the sample size to these inputs.

Authors should report the outcome used for the basis of the sample size calculation.

Authors should report the underlying basis used to specify the target difference (eg a clinically meaningful or realistic difference), and explain the choice of target difference including any formal method used or relevant previous research.

Authors should report the allocation ratio used and any adjustments to the sample size (eg attrition rate).

Authors should reference the trial protocol.