

Sample-size estimation is not reported in 24% of randomised controlled trials of inflammatory bowel disease: A systematic review

United European Gastroenterology

Journal

0(0) 1–6

© Author(s) 2020



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/2050640620967899

journals.sagepub.com/home/ueg

Zipporah Iheozor-Ejiofor¹, Svetlana Lakunina² , Morris Gordon², Daniel Akintelure², Vasiliki Sinopoulou² and Anthony Akobeng³

Abstract

Background: Sample-size estimation is an important factor in designing a clinical trial. A recent study found that 65% of Cochrane systematic reviews had imprecise results.

Objective: This study set out to review the whole body of inflammatory bowel disease (IBD) randomised controlled trials systematically in order to identify the reporting of sample-size estimation.

Methods: We conducted a comprehensive hand search of the Cochrane Library and Cochrane IBD Specialized Trials Register. We extracted information on relevant features and the results of the included studies. We produced descriptive statistics for our results.

Results: A total of 242 randomised controlled trials were included from 44 Cochrane systematic reviews. About 25% of the studies failed to report on sample-size estimation. Of those that did report on sample-size estimation, 33% failed to recruit their target sample size.

Conclusions: Around half of the randomised controlled trials in IBD either do not report sample-size estimation or reach their recruitment target with the level of detail in reporting being limited.

Keywords

Gastroenterology, IBD, inflammatory bowel disease, Crohn's disease, ulcerative colitis

Received: 18 May 2020; accepted: 23 September 2020

Introduction

The number of study participants, or sample size, is an important factor to consider when designing a clinical trial. The larger the sample size, the more precise the results are and the higher the likelihood of detecting statistically significant results.¹ Studies with very small sample sizes may not be sufficiently powered to detect an important difference.² On the other hand, sample sizes that are too large can detect statistically significant differences even when they might not be clinically important.³ This could result in the recommendation of treatments that are not effective. It is therefore important to carry out a sample-size calculation.

Typically, a sample-size estimation (SSE) would require the following components: the probability of

a type I error (concluding that there is an effect when in reality there is not), the probability of a type II error (concluding that there is no effect when in reality there

¹Centre for Biostatistics, University of Manchester, Manchester, UK

²Biomedical Evidence Synthesis and Translation to Practice (BEST) Research Unit, School of Medicine, University of Central Lancashire, National Institute of Health Research, Preston, UK

³Department of Gastroenterology, Sidra Medical Centre, Ar-Rayyan, Qatar

Corresponding author:

Gordon Morris, Biomedical Evidence Synthesis and Translation to Practice (BEST) Research Unit, School of Medicine, University of Central Lancashire, National Institute of Health Research, 135A Adelphi St, Preston, PR1 7BH, UK.
Email: MGordon@uclan.ac.uk

is), minimal clinically important difference (MCID; the smallest difference in means that you regard as being important to be able to detect) and standard error.³ These tests are so sensitive that small differences in any of the components could lead to a wide variation in the estimates.⁴

The reporting of SSE in randomised controlled trials (RCTs) has become a standard requirement since the Consolidated Standards of Reporting Trials (CONSORT) statement was published in 1996.⁵ An improvement in power calculation reporting since the publication of the CONSORT statement has been seen.⁶

Achieving an optimal sample size can improve the precision of trial results. For systematic reviews, a meta-analysis of data from multiple studies has offered the promise of addressing the weaknesses in an evidence base made up of underpowered studies. Proponents of evidence-based medicine maintain that by pooling data from multiple studies, regardless of the sample size of the individual studies, power and the likelihood of achieving precision is enhanced in systematic reviews.⁷ However, the issue of imprecision persists in systematic reviews, as a recent study found that 65% of Cochrane systematic reviews had imprecise results.⁸ Given that current methods (the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach) of assessing the precision of systematic evidence from optimal information sizes tend to rely on adequate reporting of SSE,^{9,10} poor practice in SSE can impact the certainty of outcomes.

Additionally, studies with suboptimal small sample sizes may seem unethical for a number of reasons. Primarily, the risks that participants undergo are not compensated for by the potential of the trial to detect meaningful or clinically important estimates.¹¹ Additionally, the financial costs and practical implications of the time commitment needed by researchers or patients must be based on the assumption that a study is able to address its hypothesis, and in the case of an underpowered study, this will never be the case.

Research investigators fail to recruit the number of participants stipulated in their sample size calculation for various reasons. For inflammatory bowel disease (IBD) trials, this may be due to certain elements of study design such as randomisation and blinding, frequency of visits, invasiveness of intervention or need for colonoscopy/sigmoidoscopy.¹² Most studies on key IBD outcomes usually involve some or all these factors, but as they are essentially predictable, designing studies to mitigate such issues should always be possible. This study set out to review the whole body of published IBD RCTs systematically in order to identify the

reporting of power calculations and the nature of these calculations.

Methods

This review was performed in alignment with Cochrane guidelines¹³ in June 2019 and reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement.¹⁴ A protocol for the review is available for the analysis.¹⁵

Search methods for identification of studies

We conducted a comprehensive search of the Cochrane IBD Specialized Trials Register, CENTRAL and hand searched within the Cochrane library of IBD reviews for further primary RCTs. We included RCTs published since 1996 (after the publication of the CONSORT statement). We excluded cluster RCTs, pilot or feasibility studies, studies with mixed population of people with and without IBD and studies on secondary analyses of follow-up data collection after discontinuation of treatment. We included abstracts if information was available to judge inclusion. If this information was not available, we contacted the authors, and if there was no response, we excluded the study from our analyses. We included studies whose participants were of any ages with IBD, and we included studies of any therapeutic intervention when compared to any other intervention, placebo or no treatment.

Using the above search strategy, two review authors (S.L. and D.A.) identified RCT titles that appeared to be potentially relevant. These were independently screened, and in circumstances of disagreement, a third review author (Z.I.E.) was involved to reach a consensus.

Data extraction and management

We developed a data-extraction form and used it to extract information on the relevant features and results of the included studies. Two review authors (S.L. and D.A.) independently extracted and recorded data on a predefined checklist. When disagreements occurred, a third review author (Z.I.E.) was involved and a consensus was reached. The fourth author (V.S.) then reviewed the completed data-extraction form and checked it with the studies used.

The main outcome was to assess the proportion of studies reporting power calculation, the reproducibility of such calculations. The secondary outcomes were to compare the differences studies used and the sample sizes involved.

Extracted data included: the characteristics of the participants (disease type and state); the presence of

SSE and calculation details (MCID, power, significance level, target sample size); the total number of participants originally assigned to each group; the intervention and control details; and the outcomes: the achievement of target sample size; number of patients recruited and completing the study; the number of treatment success/failures; the MCID proposed and the difference achieved; whether the studies are underpowered and by how many people; adverse events; and definitions of the outcomes.

We resolved inconsistencies in data extraction, and transferred the information above into the characteristics of the included studies table.

Data synthesis

We produced descriptive statistics regarding the overall rates of sample size calculation, and pooled studies with the same population, intervention, comparator and outcomes.

Ethical statement

As all data included already existed within published scholarly output, no ethical approval was sought.

Results

The search performed in June 2019 revealed 765 RCTs (697 after the removal of duplicates). Initial screening excluded 418 studies, leaving 279 articles for further assessment. The reasons for exclusion included articles published before 1996 (117 studies) and the wrong patient group or wrong diagnosis (301 studies). A total of 279 articles were assessed further, and 47 of them were excluded for the following reasons: published as abstracts with insufficient information (30 studies), pilot/feasibility studies (11 studies), non-RCTs (two studies) or not written in the English language (four studies). This left 242 studies (reported in 232 publications) to be included (see Figure 1).

Of the 242 included studies, 116 studies were on ulcerative colitis (UC; 48%; 84 induction and 32 maintenance), 99 were on Crohn's disease (CD; 41%; 54 induction and 45 maintenance) and 27 were on other IBD conditions (11%). There were more studies on UC than CD. The reference list of the included studies can be found in Appendix 1. Full extracted data are available from the authors on request. We carried out a subgroup analysis by disease type, disease state and drug class (Table 1), and performed chi-square analysis between the drugs classes, as well as between induction and maintenance studies (0.05 significance level). There was no difference in reporting of SSE between immunomodulators and microbiome subgroups ($p=0.067797$; 101 SSE/30 no SSE immunomodulators,

49/26 microbiome), maintenance and induction studies ($p=0.360891$; 70/27 maintenance, 119/35 induction) or biologics and immunomodulators ($p=0.50793$; 52/12 biologics, 101/30 immunomodulators). The difference between biologics and microbiome ($p=0.035853$; 52/12 biologics, 49/26 microbiome) was statistically significant. The difference between CD and UC studies was statistically significant ($p=0.003627$; 90/130 CD, 99/32 UC).

About 25% (59/242) of the studies failed to report on SSE. In CD studies, reporting was more common in inactive (80%) compared to active (72%) disease studies. Of the 183 studies which reported SSE, 61 (33%) failed to recruit their target sample size. Studies on UC (67%) were more likely to meet their target sample size than CD studies (61%) though not by a substantial difference (Table 2 and Figure 2). For the studies which failed to meet their recruitment target, the mean sample size deficit was about 31% and ranged from 21% to 40%.

The sample-size calculation reported in the studies was assessed for reproducibility. Most of the studies failed to report sufficient information for their sample size to be replicated. There were 99 two-arm superiority trials of which only 35 (35%) studies reported sufficient information to enable the replication of SSE. However, we managed to replicate sample sizes of 71 (71%) studies using parameters proposed in the protocol. The reported sample size was equal to the recalculated estimate in eight (11%) studies, higher in 43 (61%) studies and lower in 19 (27%) studies. The difference between the reported and recalculated estimate was up to 10% in 20 (28%) studies, 20% in 19 (27%) studies and >20% in 24 (34%) studies.

There was variation across studies in the parameters used in their SSE (Table 3). However, the majority of the studies used 80% power, probability of type I error was 0.05 and the most commonly reported MCID ranged from 20% to 30%.

Discussion

The aim of this study was to examine the reporting of SSEs in IBD trials. To achieve this, we found 242 RCTs (reported in 233 publications) assessing the effectiveness of interventions used in managing IBD. The results showed that SSE was reported in 75% of the studies. This finding is also consistent with previous reports.⁶ However, a third of those that did report SSE failed to meet the recruitment target specified in their study, meaning that half of all included trials did not report SSE or meet their required target. When we examined reporting trends by disease type and purpose of the intervention, the purpose of the intervention (induction or maintenance) appeared to impact on

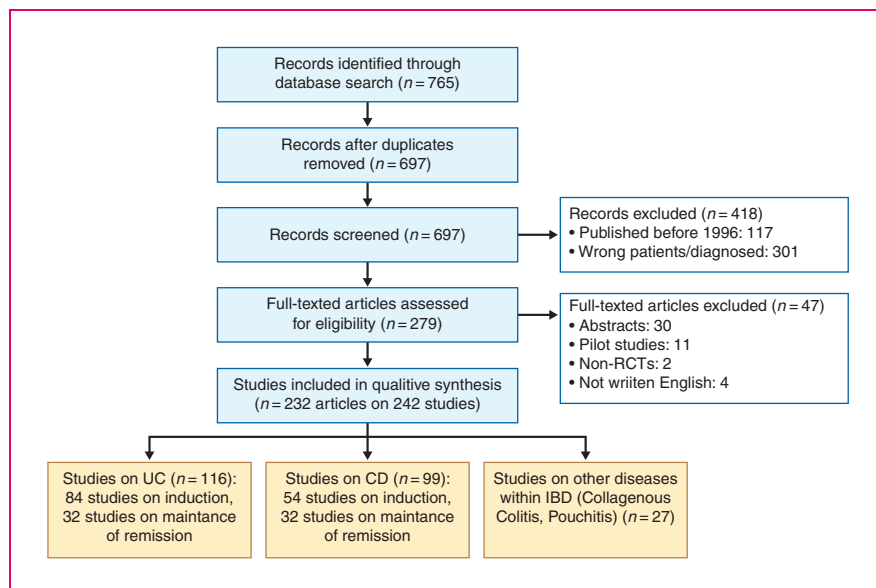


Figure 1. Flow diagram of the study selection process. UC: ulcerative colitis; CD: Crohn's disease; RCTs: randomised controlled trials.

Table 1. Subgroup analysis of the included studies.

Drug categories	CD induction		CD maintenance		UC induction		UC maintenance		Other		Total	
	Total papers (%)	SSE reported (%)	Total papers (%)	SSE reported (%)	Total papers (%)	SSE reported (%)	Total papers (%)	SSE reported (%)	Total papers (%)	SSE reported (%)	Total papers (%)	SSE reported (%)
Biologics	17 (7%)	16 (94%)	8 (3%)	5 (63%)	34 (15%)	28 (82%)	4 (2%)	2 (50%)	1 (1%)	1 (100%)	64 (27%)	52 (81%)
Immunomodulators	21 (9%)	16 (76%)	35 (15%)	30 (86%)	38 (16%)	28 (74%)	27 (12%)	21 (78%)	10 (4%)	6 (60%)	131 (56%)	101 (77%)
Microbiome	27 (12%)	16 (59%)	12 (5%)	7 (58%)	17 (7%)	15 (88%)	11 (5%)	5 (45%)	8 (3%)	6 (75%)	75 (32%)	49 (65%)

SSE: sample-size estimation; CD: Crohn's disease; UC: ulcerative colitis.

Table 2. Reporting of SSE based on disease type and purpose of intervention.

Disease type/ purpose	Total	Estimation		% Reporting	Recruitment success	Recruitment failure	% Recruitment success	% Sample-size deficit
		reported	not reported					
UC/induction	84	69	15	82.1%	48	17	69.6%	29.2%
UC/maintenance	32	21	11	65.6%	13	7	61.9%	37.4%
CD/induction	54	39	15	72.2%	24	15	61.5%	39.6%
CD/maintenance	45	36	9	80.0%	22	13	61.1%	21.4%
Other	27	18	9	66.7%	12	9	66.7%	27.0%
Total	242	183	59	75.6%	119	61	49.2%	31.0%

successful recruitment in UC studies. However, this was not the case in CD studies. This adds to the knowledge on barriers to study recruitment in IBD.¹² In the studies which failed to meet their recruitment target, reported sample-size deficits ranged from 29% to 40%, significantly underpowering the subsequent output. Our chi-square analysis showed that maintenance studies are better at reporting sample sizes than

induction studies are. The reason for this is unclear, and further research on this topic is required. Studies on biologics are better at reporting SSE than studies on immunomodulators are. This could be because studies on biologics are generally newer, and hence they are more likely to report on SSE.

To assess whether the SSEs were reliable, we attempted to recalculate the study sample sizes and

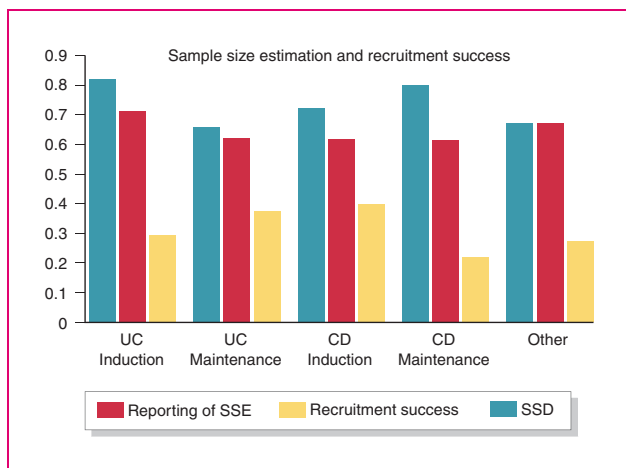


Figure 2. Sample size estimation and recruitment success. SSE: sample-size estimation; SSD: sample-size deficit.

found that the studies rarely (35%) provided full details to enable replication. Although we were able to recalculate study sample sizes for a substantial proportion (71%) of the eligible studies, this was only enabled by our use of agreed default values for the sample-size parameters and hypothesis testing. This finding is also consistent with similar reviews on anaesthesia and osteoarthritis trials which found that only a small proportion of studies reported sufficient details to enable replication of their SSEs.^{16,17} When we recalculated the sample sizes for this review, around 90% of the studies assessed were found to have overestimated or underestimated the required sample size. Overestimation of sample size was expectedly more common, as trial investigators tend to inflate sample sizes to account for drop-out and withdrawal due to adverse events. This finding should be interpreted with caution, as some of the recalculated estimates may not accurately reflect the estimations carried out by the trial investigators due to partial reporting of SSE details in the studies.

These findings support the shift by evidence producers such as Cochrane from emphasising statistical significance to clinical importance. It also shows that having multiple studies with small sample sizes does not eliminate the need for single well-powered RCTs. In most studies, it was unclear whether the parameters for their SSEs were informed by the broader literature or clinical experience. Future research should assess parameters of SSE which determine whether meaningful results will be obtained for specific outcomes. This will determine whether there is any consensus on what is considered a ‘meaningful’ result for specific outcomes in IBD trials and form a useful resource for future researchers. Also, considering if poor reporting of SSE is correlated with other areas of reporting,

Table 3. Details of SSE and parameters reported in studies.

SSE reported	183 (75.6%)
Not reported	59 (24.4%)
Target sample size achieved	119 (65%)
Not achieved	61 (33.3%)
<i>Sample-size deficit (N = 61)</i>	
Up to 10%	15 (24.6%)
>10–20%	10 (16.4%)
>20%	34 (55.7%)
<i>Sample-size recalculation (N = 183)</i>	
Parameters fully reported	35 (19.1%)
Partially reported	65 (35.5%)
Non-inferiority trials	17 (9.3%)
Three arm trials	57 (31.1%)
Studies with continuous outcome:	9 (4.9%)
<i>Power of study (N = 67)</i>	
0.54	1 (1.5%)
0.2	44 (65.7%)
0.19	1 (1.5%)
0.17	1 (1.5%)
0.15	2 (3%)
0.14	1 (1.5%)
0.11	1 (1.5%)
0.10	16 (23.9%)
<i>Type I error (alpha) (N = 63)</i>	
0.05	58 (92%)
0.025	3 (4.8%)
0.017	1 (1.6%)
0.001	1 (1.6%)
<i>Minimal clinically important difference (N = 101)</i>	
Up to 10%	5 (5%)
>10–20%	31 (30.7%)
>20–30%	35 (34.7%)
>30%	30 (29.7%)
<i>Reported versus calculated sample size estimation (N = 71)</i>	
Identical	8 (11.3%)
Less than calculated	19 (26.8%)
More than calculated	43 (60.6%)
<i>Difference between reported and calculated estimation (N = 62)</i>	
Up to 10% difference	20 (32.3%)
>10–20% difference	19 (30.6%)
>20% difference:	24 (38.7%)

comparing with the Cochrane risk of bias tool, for example, would be useful. This would allow the subgroup analysis of other factors such as different disease types or settings.

We were aware of potential biases in the process of conducting this review and put in measures to minimise them. However, there are decisions that were made during the process which may have introduced limitations. As a result, due to the large number of studies found, we attempted to minimise errors by involving two authors at the data-extraction phase, while additional checks were carried out by a third author. We encountered difficulties dealing with a lack of clarity

and incompleteness in the reporting in the studies in ways that were not anticipated at the protocol phase. For instance, we had concerns about two studies which appeared to have estimated sample sizes retrospectively, a study that indicated that SSE was not done statistically and two studies that were described as being 'exploratory' in nature which may have been wrongly included. The decision to include or exclude these studies from the analysis could be regarded as a study limitation. However, given the small numbers, we do not expect these studies to have had a substantial impact on the results. We did not contact authors for clarification due to the number of studies we found, only authors of abstracts, and we excluded four studies that were not in English.

Conclusions

In summary, around half of the RCTs on IBD either do not report SSE or do not reach their recruitment target. When studies do report on SSE, the level of detail in reporting is limited. The results of this study provide an insight into the current practices of reporting SSE, highlighting the need for discussions on how to utilise them better in primary trials and systematic reviews.

Whilst reaching the recruitment target is expected to produce meaningful results in the studies, a third of the studies are not recruiting successfully. Even when studies can successfully reach their target sample size, it is uncertain whether it is sufficient to detect a meaningful result.

Declaration of conflicting interests

The authors have no conflicts of interest to declare.

Ethics approval

The study is a piece of secondary research with no new data collection and as such ethical approval was not sought.


Funding

The authors received no financial support for the research, authorship and/or publication of this article.

Informed consent

Not required.

ORCID iD

Lakunina Svetlana  <https://orcid.org/0000-0002-3180-6336>

Supplemental Material

Supplemental material for this article is available online.

References

1. Biau D, Kernéis S and Porcher R. Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clin Orthop Relat Res* 2008; 466: 2282–2288.
2. Nayak B. Understanding the relevance of sample size calculation. *Indian J Ophthalmol* 2010; 58: 469–470.
3. Cornish R. *Statistics: an introduction to sample size calculations*. Loughborough: Mathematics Learning Support Centre, 2006.
4. Noordzij M, Tripepi G, Dekker F, et al. Sample size calculations: basic principles and common pitfalls. *Nephrol Dial Transplant* 2010; 25: 1388–1393.
5. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT Statement. *JAMA* 1996; 276: 637–639.
6. Sully BG, Julious SA and Nicholl J. A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies. *Trials* 2013; 14: 166.
7. Higgins J, Altman D, Gotzsche P, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011; 343: d5928.
8. Castellini G, Bruschetti M, Gianola S, et al. Assessing imprecision in Cochrane systematic reviews: a comparison of GRADE and Trial Sequential Analysis. *Syst Rev* 2018; 7: 110.
9. Schünemann HJ, Tugwell P, Reeves BC, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Res Synth Methods* 2013; 4: 49–62.
10. Guyatt G, Oxman A, Akl E, et al. GRADE guidelines: 1. Introduction – GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011; 64: 383–394.
11. Bacchetti P, Wolf L, Segal M, et al. Ethics and sample size. *Am J Epidemiol* 2005; 161: 105–110.
12. Ravikoff J, Cole E and Korzenik J. Barriers to enrollment in inflammatory bowel disease randomized controlled trials: an investigation of patient perspectives. *Inflamm Bowel Dis* 2012; 18: 2092–2098.
13. Higgins JPT, Thomas J, Chandler J, et al. (eds) *Cochrane handbook for systematic reviews of interventions*. 2nd edn. Chichester: John Wiley, 2019.
14. Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta – Analyses: the PRISMA statement. *PLoS Med* 2009; 6: e1000097.
15. Gordon M and Lakunina S. Protocol for systematic review 'The Reporting of Sample Size Estimation in Randomised Trials of Inflammatory Bowel Disease: A systematic review', <https://clock.uclan.ac.uk/33088/> (accessed 8th July 2020).
16. Abdulatif M, Mukhtar A and Obayah G. Pitfalls in reporting sample size calculation in randomized controlled trials published in leading anaesthesia journals: a systematic review. *Br J Anaesth* 2015; 115: 699–707.
17. Copsey B, Thompson J, Vadher K, et al. Problems persist in reporting of methods and results for the WOMAC measure in hip and knee osteoarthritis trials. *Qual Life Res* 2019; 28: 335–343.