Can J Anesth/J Can Anesth (2018) 65:611–618
https://doi.org/10.1007/s12630-018-1109-z

REPORTS OF ORIGINAL INVESTIGATIONS

Sample size calculations for randomized clinical trials published in anesthesiology journals: a comparison of 2010 *versus* 2016

Calculs de la taille des échantillons pour les essais cliniques randomisés publiés dans les journaux d'anesthésiologie : comparaison entre 2010 et 2016

Jeffrey T. Y. Chow, MSc · Timothy P. Turkstra, MD, MEng · Edmund Yim, BMSc · Philip M. Jones, MD, MSc

Received: 13 October 2017 / Revised: 14 February 2018 / Accepted: 26 February 2018 / Published online: 22 March 2018

Abstract

Purpose *Although every randomized clinical trial (RCT) needs participants, determining the ideal number of participants that balances limited resources and the ability to detect a real effect is difficult. Focussing on two-arm, parallel group, superiority RCTs published in six general anesthesiology journals, the objective of this study was to compare the quality of sample size calculations for RCTs published in 2010 vs 2016.*

Methods *Each RCT's full text was searched for the presence of a sample size calculation, and the assumptions made by the investigators were compared with the actual values observed in the results. Analyses were only performed for sample size calculations that were amenable to replication, defined as using a clearly identified outcome that was continuous or binary in a standard sample size calculation procedure.*

Results *The percentage of RCTs reporting all sample size calculation assumptions increased from 51% in 2010 to 84% in 2016. The difference between the values observed in the study and the expected values used for the sample size calculation for most RCTs was usually > 10% of the expected value, with negligible improvement from 2010 to 2016.*

Conclusion *While the reporting of sample size calculations improved from 2010 to 2016, the expected values in these sample size calculations often assumed effect sizes larger than those actually observed in the study. Since overly optimistic assumptions may systematically lead to underpowered RCTs, improvements in how to calculate and report sample sizes in anesthesiology research are needed.*

Résumé

Objectif *Même si chaque essai clinique randomisé (ECR) nécessite des participants, la détermination de leur nombre idéal prenant en compte d'une part des ressources limitées et d'autre part la capacité à détecter un effet réel s'avère difficile. Se concentrant sur les ECR à deux groupes, à groupes parallèles, et essais de supériorité publiés dans six journaux d'anesthésiologie, l'objectif de cette étude était de comparer la qualité des calculs de taille d'échantillon pour les ECR publiés en 2010 et en 2016.*

Méthodes *Le texte complet de chaque ECR a été analysé en fonction du calcul de la taille de l'échantillon et les hypothèses faites par les investigateurs ont été comparées aux valeurs réelles observées dans les résultats. Des analyses n'ont été pratiquées que pour les calculs de taille d'échantillon qu'il était possible de répliquer, en utilisant*

This article is accompanied by an editorial. Please see Can J Anesth 2018; 65: this issue.

J. T. Y. Chow, MSc · P. M. Jones, MD, MSc
Department of Epidemiology & Biostatistics, The University of Western Ontario, London, ON, Canada

J. T. Y. Chow, MSc · T. P. Turkstra, MD, MEng ·
E. Yim, BMSc · P. M. Jones, MD, MSc
Schulich School of Medicine & Dentistry, The University of Western Ontario, London, ON, Canada

T. P. Turkstra, MD, MEng · P. M. Jones, MD, MSc
Department of Anesthesia & Perioperative Medicine,
The University of Western Ontario, London, ON, Canada

P. M. Jones, MD, MSc (✉)
Rm C3-110 - University Hospital, London Health Sciences
Centre, 339 Windermere Rd, London, ON N6A 5A5, Canada
e-mail: pjones8@uwo.ca

une mesure clairement identifiée, continue ou binaire, avec une procédure usuelle de calcul de taille d'échantillon.

Résultats *Le pourcentage des ECR indiquant toutes les hypothèses du calcul de taille de l'échantillon est passé de 51 % en 2010 à 84 % en 2016. La différence entre les valeurs observées dans les études et les valeurs attendues utilisées pour les calculs de taille d'échantillon de la majorité des ECR était habituellement plus de 10 % plus élevée que la valeur attendue, sans véritable amélioration entre 2010 et 2016.*

Conclusion *Alors que la présentation des calculs de la taille des échantillons s'est améliorée entre 2010 et 2016, les valeurs attendues dans ces calculs ont souvent supposé des ampleurs d'effet supérieures à celles véritablement observées dans les études. Considérant que des hypothèses excessivement optimistes entraînent un manque de puissance des ECR, des améliorations sur la façon de calculer et présenter la taille des échantillons pour la recherche en anesthésiologie sont nécessaires.*

In evidence-based medicine, randomized clinical trials (RCTs) are considered to provide the highest quality of evidence with the least amount of bias,[1] representing the gold standard for assessing therapeutic interventions.[2] In every study, participants are needed to provide data. Nevertheless, the ideal number of participants cannot be determined without appropriate calculations. A sample size calculation involves determining the minimum number of participants needed to detect a treatment effect that is clinically relevant.[3] An inadequately small sample size may result in the inability to detect a precise effect, if present, while a needlessly large sample size may result in extra participants being exposed to the potential risks of the trial.[4] While there are ethical debates over whether underpowered trials should be conducted, most researchers agree that an *a priori* sample size calculation should still be conducted and reported to ensure methodologic quality.[5]

The Consolidated Standards of Reporting Trials (CONSORT) statement is a set of guidelines aimed at improving the quality of RCT reporting.[6] Item seven in the most recent checklist focusses on sample size, with recommendations to explain how the sample size was determined and, ideally, to possess a high probability of detecting a statistically significant, clinically relevant difference if one exists.[6] The components needed for a sample size calculation are: the expected outcomes for each group (which implies the clinically relevant effect size), the type I (α) error level, the type II (β) error or power (1- β) level, the standard deviation for continuous outcomes, and any allowances made for attrition or non-compliance, if appropriate.[6,7]

Though most RCTs report a sample size calculation, previous studies in general medical journals[8] and anesthesiology specialty journals[9] have found that the quality of reporting is inadequate. While the investigators found differences between the expected and actual values for each sample calculation parameter,[9] only the medians and interquartile ranges of the differences were presented and not the number of RCTs with > 10% difference between the expected and actual values. Since the most recent version of the CONSORT guidelines was published in 2010,[6] the quality of RCT reporting is expected to improve as journals implement corresponding policies and authors become aware of requirements. Nevertheless, sample size calculations for anesthesiology RCTs were still inadequately reported in 2013,[9] suggesting that a longer adjustment period may be needed for actual progress to occur.

The current study was focussed on anesthesiology RCTs, specifically two-arm, parallel group, superiority RCTs published in the top six general anesthesiology journals as determined by impact factor. The first objective of this study was to compare RCTs published in 2010 *vs* 2016 in terms of the proportion of RCTs that performed an *a priori* sample size calculation and the proportion of RCTs that reported the necessary elements to allow for replication of the calculation. This would update the literature using the most recent RCTs and examine potential trends over time in sample size calculation reporting. The second objective was to compare the expected sample size calculation parameters with the actual values observed in the results.

Methods

An electronic database had been created for a previous project[10] to collect all RCTs published in 2010 and 2016 from the top six general (non-pain-centric) anesthesiology journals as determined by impact factor: *Anaesthesia*, *Anesthesia & Analgesia*, *Anesthesiology*, *British Journal of Anaesthesia*, *Canadian Journal of Anesthesia*, and *European Journal of Anaesthesiology*.[11] Detailed screening and data extraction have been described previously,[10] with the study protocol publicly available as well.[12]

An RCT was defined as a prospective study that assessed randomly allocated healthcare interventions in human participants where authors had clearly reported that participants had been randomly allocated to study arms (i.e., the use of the words "random", "randomized",

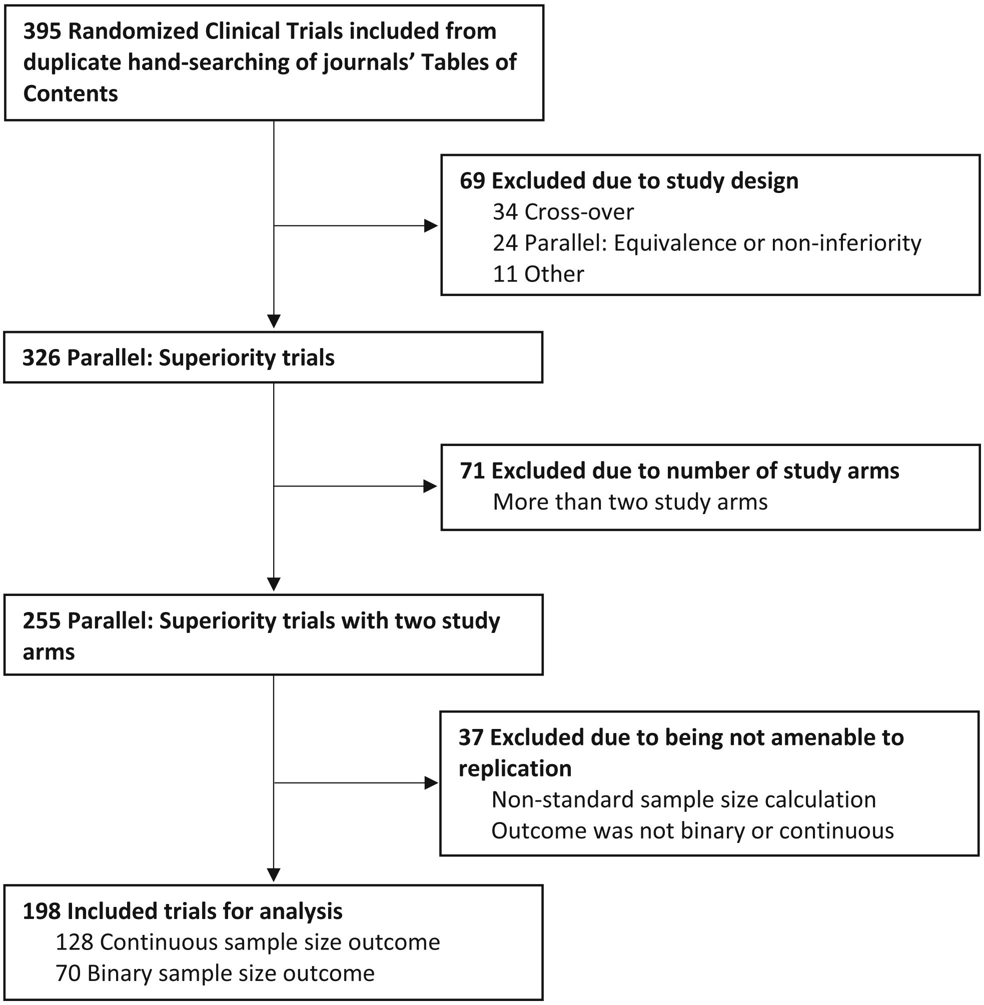
395 Randomized Clinical Trials included from duplicate hand-searching of journals' Tables of Contents
69 Excluded due to study design
34 Cross-over
24 Parallel: Equivalence or non-inferiority
11 Other
326 Parallel: Superiority trials
71 Excluded due to number of study arms
More than two study arms
255 Parallel: Superiority trials with two study arms
37 Excluded due to being not amenable to replication
Non-standard sample size calculation
Outcome was not binary or continuous
198 Included trials for analysis
128 Continuous sample size outcome
70 Binary sample size outcome

Figure Flowchart for inclusion of trials

or "randomised").[10] Randomized clinical trials were only included in the current study if there were two study arms and a parallel group, superiority design. Randomized clinical trials were deemed adequately registered if the trial was registered in a publicly available trials registry prior to the first participant being enrolled with a clearly defined primary outcome.[10] Sample size was defined as the number randomized in the trial. The statistical software used for each RCT was also extracted, with the primary statistical package for quantitative data analysis recorded if more than one was reported.

The primary outcome was to determine the proportion of RCTs that reported all necessary assumptions (expected outcome for treatment group, expected outcome for control group, type I error level, type II error level, and standard deviation for continuous outcomes) in their sample size calculation. The secondary outcomes assessed whether the actual values reported in the results corresponded to these expected sample size calculation parameters.

Data extraction

Data extraction for the current study was performed by one of four reviewers; training was provided to each reviewer to ensure consistency of data extraction. A second reviewer provided input if variables were unclear and a decision was made by consensus. During data extraction, the full text of each included RCT was searched for the presence of a sample size calculation. Analyses were only performed for RCTs that had a sample size calculation amenable to replication, defined as using a clearly identified outcome that was continuous or binary in a standard sample size calculation procedure (i.e., using formulae commonly used in statistical textbooks). For RCTs with sample size calculations amenable to replication, the assumptions made by the investigators were recorded, such as the expected outcome for the treatment group, the expected outcome for the control group, the type I (α) error level, the type II (β) error level, and whether any allowances were made for attrition or non-compliance. The standard deviation was also recorded for RCTs that used a continuous outcome in their sample size calculation. The actual values for the treatment group outcome, control group outcome, and standard deviation were recorded to determine if the difference between actual and expected values was within 10% of the expected value. The justifications made by the authors for these assumptions were recorded using a categorical outcome, with options for previously published trial, meta-analysis, pilot study, observational data, unspecified, or other.

Data analysis

For studies that provided all assumptions needed to replicate the sample size (i.e., expected outcome for treatment group, expected outcome for control group, type I (α) error level, type II (β) error level, and standard deviation for continuous outcomes), the sample size was replicated using both the user-written add-on command, *ssi*,[13] and the *power* command in version 13 Stata statistical software (StataCorp LLC, College Park, TX, USA). When provided in the sample size calculation, the allowance made for attrition or non-compliance was included in the recalculation to ensure accuracy. The two replicated sample sizes were compared with the reported target sample size, with the sample sizes deemed to be equivalent for the purposes of this study if either replicated sample size had a relative difference of ≤ 10% on either side of the target sample size. Consistent with previous studies,[8,9] when the difference (positive or negative) was > 10%, a discrepancy was noted and values in the manuscript checked a second time to ensure accurate data extraction. For one-sided sample size calculations, the type I (α) error level was doubled and the same procedures followed. During data extraction, some trials with binary outcomes were found to not specify whether relative or absolute differences were expected. In those cases, both alternatives were tried with the replicated sample size closest to the reported target sample size used. Descriptive statistics were used to present the raw numbers and percentages for all measured outcomes.

Results

Of the 395 RCTs identified through duplicate hand-searching of the Table of Contents for the top six general anesthesiology journals (*Anaesthesia*, *Anesthesia & Analgesia*, *Anesthesiology*, *British Journal of Anaesthesia*, *Canadian Journal of Anesthesia*, and *European Journal of Anaesthesiology*), 255 RCTs met the inclusion criteria of being parallel, superiority RCTs with two study arms (Figure). There were 143 RCTs from 2010 and 112 RCTs from 2016, with decreases in the number of RCTs from each journal other than *Anesthesiology* and *CJA* (Table 1). Most RCTs were conducted at a single centre. Though SPSS was the most commonly used statistical software, the number of RCTs using SPSS decreased from 2010 to 2016; this coincided with increases in the use of Stata, SAS, and R. The percentage of adequately registered RCTs increased almost seven-fold from 2010 to 2016 even though more than half of RCTs in 2016 were still inadequately registered (Table 1).

Table 1 Characteristics of included randomized clinical trials

	2010 (*n* = 143)	2016 (*n* = 112)
Journal		
Anaesthesia	28 (20%)	22 (20%)
Anesthesia & Analgesia	37 (26%)	21 (19%)
Anesthesiology	11 (8%)	13 (12%)
British Journal of Anaesthesia	24 (17%)	22 (20%)
Canadian Journal of Anesthesia	12 (8%)	17 (15%)
European Journal of Anaesthesiology	31 (22%)	17 (15%)
Multicentricity[1]		
Single-centre	137 (96%)	97 (87%)
Multicentre	6 (4%)	15 (13%)
Statistical software		
SPSS	68 (48%)	46 (41%)
Stata	10 (7%)	14 (13%)
SAS	8 (6%)	18 (16%)
R	1 (1%)	8 (7%)
JMP	2 (1%)	3 (3%)
Other	27 (19%)	14 (13%)
GraphPad (Prism/InStat)	9 (6%)	7 (6%)
SigmaPlot/SigmaStat	2 (1%)	3 (3%)
Statistica	2 (1%)	1 (1%)
StatView	4 (3%)	0 (0%)
Other	10 (7%)	3 (3%)
Not mentioned	27 (19%)	9 (8%)
Trial registration[2]		
Adequately registered	8 (6%)	46 (41%)
Inadequately registered	135 (94%)	66 (59%)
Median sample size [IQR]	63 [41-101]	80 [52-136]

Data presented are the number of trials (%) unless otherwise stated. Percentages may not add up to 100% because of rounding. IQR = interquartile range

[1] Multicentre is defined as participants recruited from more than one institution or clinic

[2] Adequate registration is defined as the trial being registered before the first participant was enrolled with a primary outcome clearly defined in the registry

There were 110 RCTs in 2010 and 88 RCTs in 2016 that had a sample size calculation amenable to replication, using a standard sample size calculation with a clearly identified binary or continuous outcome (Table 2). Of these eligible trials, the percentage of RCTs reporting all assumptions (expected outcome for treatment group, expected outcome for control group, type I error level, type II error level, and standard deviation for continuous outcomes) increased from 51% in 2010 to 84% in 2016. Nevertheless, the difference between the actual and expected values for most RCTs was usually > 10% of the expected value, with negligible improvement from 2010 to 2016. Most RCTs based their sample size calculations on published trials, with the percentage continuing to increase in 2016. When replicating the initial sample size calculation for RCTs with all assumptions reported, 70% of RCTs' originally estimated sample sizes could be reproduced within a 10% margin of error (Table 2).

For RCTs where the primary outcome used in the sample size calculation was binary (and when expected and actual values were reported for both the treatment and control groups), the majority of RCTs expected a larger effect size than the actual effect size observed (Table 3). From 2010 to 2016, there was an increase in the number and percentage of RCTs overestimating their effect sizes for both binary and continuous primary outcomes. Nevertheless, only about half of RCTs underestimated the standard deviation for sample size calculations (Table 3).

Table 2 Reporting and accuracy of sample size assumptions in randomized clinical trials

	2010	2016
Sample size calculation amenable to replication	110/143 (77%)	88/112 (79%)
No sample size calculation	16/33 (48%)	4/24 (17%)
Not amenable to replication	17/33 (52%)	20/24 (83%)
Assumptions		
Expected outcome for treatment group	70/110 (64%)	76/88 (86%)
Actual values within 10%[1]	14/62 (23%)	17/68 (25%)
Expected outcome for control group	70/110 (64%)	76/88 (86%)
Actual values within 10%[1]	26/62 (42%)	24/68 (35%)
Type I (α) error level	108/110 (98%)	88/88 (100%)
Type II (β) error or power level	108/110 (98%)	88/88 (100%)
Standard deviation[2]	36/74 (49%)	50/54 (93%)
Actual values within 10%[1,2]	3/21 (14%)	4/21 (19%)
Allowances made for attrition or non-compliance	30/110 (27%)	47/88 (53%)
All assumptions reported	**56/110 (51%)**	**74/88 (84%)**
Replication of initial sample size calculation[3]	39/56 (70%)	52/74 (70%)
Justification for assumptions[4]		
Published trial	36/110 (33%)	45/88 (51%)
Meta-analysis	3/110 (3%)	1/88 (1%)
Pilot study	30/110 (27%)	21/88 (24%)
Observational data	14/110 (13%)	14/88 (16%)
Unspecified	18/110 (16%)	8/88 (9%)
Other	18/110 (16%)	7/88 (8%)

Bold values used to distinguish them as the culmination of the items

Data presented are the number of trials with the specific criterion reported/number of eligible trials (%) unless otherwise specified. Percentages may not sum to 100% because of rounding

[1] Number of trials where the actual value was within 10% of the assumed value/number of trials with both the specific assumption and the actual value reported (%)

[2] For trials using a continuous outcome in their sample size calculation

[3] Number of trials where the replicated initial sample size was within 10% of the reported target sample size/number of trials with all sample size assumptions reported (%)

[4] Subgroups were not mutually exclusive so percentages sum to over 100%

Discussion

In the six general anesthesiology journals included, the percentage of two-arm parallel group, superiority RCTs with a sample size calculation increased from 2010 to 2016. The assumptions needed for conducting a sample size calculation were well reported, with increases seen in all sample size calculation parameters when comparing RCTs in 2016 *vs* RCTs in 2010. Nevertheless, these sample size calculation parameters often expected values that were significantly different from the actual values reported in the results.

In six high-impact factor general medical journals, 95% of two-arm, parallel group, superiority RCTs reported a sample size calculation but only 53% reported all parameters required for sample size calculation.[8] In ten high-impact-factor anesthesiology journals, similar results were found for RCTs published in 2013, with 92% reporting a sample size calculation. These 2013 findings are in line with the trend observed in the current study, increasing from 89% reported in 2010 to 96% in 2016. Consistent with previous studies in top general medical journals and anesthesiology journals,[8,9] about 30% of the initial size calculation replications were > 10% different than the value estimated by the authors, suggesting that either researchers were calculating their sample sizes incorrectly or using non-standard sample size calculation formulas without specifying as such.

The percentages of each sample size calculation parameter being reported in 2016 was higher than in a

Table 3 Comparison of expected and reported sample size calculation assumptions

	2010	2016
Binary primary outcome—effect size	*n* = 27	*n* = 30
Actual effect size < expected effect size	16 (59%)	23 (77%)
Actual effect size ≥ expected effect size	11 (41%)	7 (23%)
Absolute difference between actual and expected[1]	0.12 (0.05 to 0.20)	0.12 (0.06 to 0.18)
Continuous primary outcome—effect size	*n* = 35	*n* = 38
Actual effect size < expected effect size	16 (46%)	25 (66%)
Actual effect size ≥ expected effect size	19 (54%)	13 (34%)
Absolute difference between actual and expected[1]	2.4 (1.0 to 16.4)	4.7 (0.4 to 16.8)
Continuous primary outcome—standard deviation	*n* = 21	*n* = 21
Actual standard deviation ≤ expected standard deviation	12 (57%)	11 (52%)
Actual standard deviation > expected standard deviation	9 (43%)	10 (48%)
Absolute difference between actual and expected[2]	6.7 (2.3 to 16.3)	12.9 (0.2 to 20.4)

Data presented are the number of trials (%) unless otherwise stated

n = number of eligible trials, both expected and actual values provided for the specific type of primary outcome described

[1] Data presented as median absolute difference (interquartile range) between the actual effect size and the expected effect size (only trials that had the effect size overestimated were included)

[2] Data presented as median absolute difference (interquartile range) between the actual pooled standard deviation and the expected standard deviation (only trials that had the standard deviation underestimated were included)

similar study in 2013,[9] suggesting an increase in sample size calculation reporting quality over time. Though type I and II error levels were reported for all RCTs, the expected values for the treatment and control group were still < 90% in 2016, suggesting that the reason > 10% of RCTs were not explaining how the sample size was calculated was that the expected outcomes for the treatment and control group were not specified.

While this is encouraging, these sample size calculations appear to have often been based on assumptions that were more optimistic than the results actually obtained. Most RCTs that reported both expected and reported values had discrepancies > 10%. Large differences between estimates and actual sample size calculation parameters have also been found in general medical journals.[8] Around 80% of RCTs in general medical journals underestimated the sample size, with the actual standard deviation being greater than the expected standard deviation.[14] This may be due to different techniques being available to estimate the standard deviation using previously published literature or pilot studies, with some methods such as using the standard deviation from a small pilot study having a 50% chance of underestimating the standard deviation and resulting in an underpowered study.[15] Investigators should be cautious when estimating sample size calculation parameters as the benefits of performing a sample size calculation are diminished if the estimates are inaccurate.

Nevertheless, the difficulty with estimating assumptions must also be recognized; if the sample size calculation parameters are known with certainty, there will be no need for a trial to be conducted. Since estimates are usually conducted in one of two ways, by using the treatment effect that is considered clinically meaningful or by using the treatment effect that is expected,[16] the inaccuracy of estimates may be due to the intervention being ineffective or the potential effect being poorly estimated. For RCTs where the clinically important difference is used as the anticipated effect size in the sample size calculation, discrepancies would not imply any incorrect estimation by investigators. Due to the potential for meta-analyses to aggregate the results of RCTs, under-powered trials may still contribute valuable evidence if there is high methodologic rigour, clear reporting, and a lack of publication bias.[5]

Most sample size calculation parameters are poorly estimated and result in inadequately powered RCTs. Two potential solutions include improving estimates or using alternative methods to determine the sample size. Termed “sample size samba” or “delta inflation”, investigators commonly start with the number of available (or fundable) participants and adjust their estimates of the sample size calculation assumptions to justify their sample size.[5,17] Increasing the expected effect size will increase the power,[18] allowing for a smaller sample size to be calculated. For included RCTs with a binary outcome, the majority found a smaller effect size than what was expected, showing one parameter that may be manipulated to produce a feasible sample size. If estimates are being manipulated in this way, it should be no surprise when the actual values are different from what is expected. Among the flaws of the traditional sample size calculation is the difficulty in accurately estimating all parameters and the design-use mismatch resulting from using a *P* value cut-off to design a study that should not be interpreted based on a single *P* value.[19] While performing sample size calculations can provide valuable information, less importance should be placed on these calculations because of the inherent subjectivity of the estimates.[5] Some alternatives include using hybrid approaches, value of information methods, pragmatic methods based on cost or feasibility, sensitivity analyses, or using the same sample size as previous analogous studies.[19-21]

The results of this study must be interpreted considering its limitations. Since this study only included two-arm, parallel group, superiority RCTs published in the top anesthesiology journals as determined by impact factor, the generalizability of the findings may be limited. While this study focussed on the most commonly used trial design, different sample size calculations are used for different trial designs. Since only information from the published manuscript was used, definitive conclusions about sample size calculations being performed *a priori* cannot be made. Previous research has identified discrepancies in sample size calculations between published manuscripts and their original protocols[22] so there is a chance of authors creating a sample size calculation based on the number of participants recruited. Another limitation was that analyses could only be performed for RCTs that reported the relevant values. For example, RCTs that did not report all necessary sample size calculation assumptions could not be replicated to check for accuracy. While two reviewers were used to screen for eligible RCTs, data extraction for each RCT was only performed by a single reviewer, potentially increasing the amount of error if reviewers were inconsistent. To improve quality and reduce potential inconsistency, all reviewers were trained using the same guidelines, with data quality checks throughout the process.

In conclusion, almost all RCTs published in the top six general anesthesiology journals reported a sample size calculation. In 2016, most two-arm, parallel group, superiority RCTs included enough information in the methods section to allow for replication of the sample size calculation. While this represented a large increase from 2010, the outcome values and variability used in the sample size calculation were often different from what was observed, with most actual values having a difference from the expected value $\geq$ 10%. Improvements in how to calculate and report sample sizes in anesthesiology research are needed. This may be accomplished by further research into how sample size calculation parameter estimates can be improved and whether alternative methods should be used to determine a trial's ideal sample size.

Acknowledgements We appreciate the support and advice offered by Neil Klar (PhD, Department of Epidemiology and Biostatistics, The University of Western Ontario) and Janet Martin (PharmD, Department of Anesthesia & Perioperative Medicine and Department of Epidemiology and Biostatistics, The University of Western Ontario).

Conflicts of interest Philip M. Jones is an Associate Editor at the *Canadian Journal of Anesthesia*. No other authors have any conflicts of interest.

Editorial responsibility This submission was handled by Dr. Gregory L. Bryson, Deputy Editor-in-Chief, *Canadian Journal of Anesthesia.*

Author contributions *Jeffrey T. Y. Chow* and *Philip M. Jones* designed the study, analyzed the data, and wrote the manuscript. *Jeffrey T. Y. Chow*, *Timothy P. Turkstra*, *Edmund Yim*, and *Philip M. Jones* conducted the study, collected data, and revised the manuscript. *Philip M. Jones* supervised the study.

Funding Jeffrey T. Y. Chow is supported by an Ontario Graduate Scholarship: Queen Elizabeth II Graduate Scholarship in Science and Technology (OGS: QEIIGSST) and the funder had no role in any part of the study. Philip M. Jones is supported by internal departmental funding. No other authors received funding for this study.

References

1. *Guyatt G*, *Rennie D*, *Meade MO*, *Cook DJ*. Users' Guides to the Medical Literature - A Manual for Evidence-based Clinical Practice. 3rd ed. USA: McGraw-Hill Education; 2015 .
2. *Bothwell LE*, *Greene JA*, *Podolsky SH*, *Jones DS*. Assessing the gold standard — lessons from the history of RCTs. N Engl J Med 2016; 374: 2175-81.
3. *Noordzij M*, *Tripepi G*, *Dekker FW*, *Zoccali C*, *Tank MW*, *Jager KJ*. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant 2010; 25: 1388-93.
4. *Gupta KK*, *Attri JP*, *Singh A*, *Kaur H*, *Kaur G*. Basic concepts for sample size calculation: critical step for any clinical trials! Saudi J Anaesth 2016; 10: 328-31.
5. *Schulz KF*, *Grimes DA*. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005; 365: 1348-53.
6. *Moher D*, *Hopewell S*, *Schulz KF*, *et al*. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c869.
7. *Campbell MJ*, *Julious SA*, *Altman DG*. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in paired comparisons. BMJ 1995; 311: 1145-8.
8. *Charles P*, *Giraudeau B*, *Dechartres A*, *Baron G*, *Ravaud P*. Reporting of sample size calculation in randomised controlled trials: review. BMJ 2009; 338: b1732.
9. *Abdulatif M*, *Mukhtar A*, *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.
10. *Jones PM*, *Chow JT*, *Arango MF*, *et al*. Comparison of registered and reported outcomes in randomized clinical trials published in anesthesiology journals. Anesth Analg 2017; 125: 1292-300.
11. *Thomson Reuters*. Journal Citation Reports — Anesthesiology — 2016. ISI Web Knowl. Available from URL: http://admin-apps.webofknowledge.com/JCR/JCR (accessed February 2018).
12. *Chow J*, *Jones P*. Identifying areas to improve the quality of reporting in randomized clinical trials published in anesthesiology journals: a study protocol for a series of literature surveys assessing quality of trial registration, adherence to abstract reporting guidelines adequacy of sample size calculations, and impact of funding source. Figshare 2016. DOI: https://doi.org/10.6084/m9.figshare.4490582.v1.
13. *Jones PM*. SSI: Stata module to estimate sample size for randomized controlled trials. IDEAS — 2010. Available from URL: https://ideas.repec.org/c/boc/bocode/s457150.html (accessed February 2018).

14. *Vickers AJ*. Underpowering in randomized trials reporting a sample size calculation. J Clin Epidemiol 2003; 56: 717-20.
15. *Chen H*, *Zhang N*, *Lu X*, *Chen S*. Caution regarding the choice of standard deviations to guide sample size calculations in clinical trials. Clin Trials 2013; 10: 522-9.
16. *Wittes J*. Sample size calculations for randomized controlled trials. Epidemiol Rev 2002; 24: 39-53.
17. *Aberegg SK*, *Richards DR*, *O'Brien JM*. Delta inflation: a bias in the design of randomized controlled trials in critical care medicine. Crit Care 2010; 14: R77.
18. *Wilson Van Voorhis CR, Morgan BL.* Understanding power and rules of thumb for determining sample sizes. Tutor Quant Methods Psychol 2007; 3: 43-50.
19. *Bacchetti P*. Current sample size conventions: flaws, harms, and alternatives. BMC Med 2010; 8: 17.
20. *Ciarleglio MM*, *Arendt CD*. Sample size determination for a binary response in a superiority clinical trial using a hybrid classical and Bayesian procedure. Trials 2017; 18: 83.
21. *Bacchetti P*, *McCulloch CE*, *Segal MR*. Simple, defensible sample sizes based on cost efficiency. Biometrics 2008; 64: 577-85.
22. *Chan AW*, *Hrobjartsson A*, *Jorgensen KJ*, *Gotzsche PC*, *Altman DG*. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. BMJ 2008; 337: a2299.