# Examining the empirical evidence for IDEAL 2b studies: the effects of preceding prospective collaborative cohort studies on the quality and impact of subsequent randomized controlled trials of surgical innovations – protocol for a systematic review and case–control analysis

Mudathir Ibrahim [1,2] Arsenio Paez [2,3] Jiajie Yu,[2,4] Baptiste Vasey [5]
Joel Horovitz,[1] Peter McCulloch [2]

## ABSTRACT

Randomized controlled trials (RCTs) in surgery face methodological challenges, which often result in low quality or failed trials. The Idea, Development, Exploration, Assessment and Long-term (IDEAL) framework proposes preliminary prospective collaborative cohort studies with specific properties (IDEAL 2b studies) to increase the quality and feasibility of surgical RCTs. Little empirical evidence exists for this proposition, and specifically designed 2b studies are currently uncommon. Prospective collaborative cohort studies are, however, relatively common, and might provide similar benefits. We will, therefore, assess the association between prior 'IDEAL 2b-like' cohort studies and the quality and impact of surgical RCTs.

We propose a systematic review using two parallel case–control analyses, with surgical RCTs as subjects and study quality and journal impact factor (IF) as the outcomes of interest. We will search for surgical RCTs published between 2015 and 2019 and and prior prospective collaborative cohort studies authored by any of the RCT investigators. RCTs will be categorized into cases or controls by (1) journal (IF ≥or <5) and (2) study quality (PEDro score ≥or < 7). The case/control OR of exposure to a prior '2b like' study will be calculated independently for quality and impact. Cases will be matched 1: 1 with controls by year of publication, and confounding by peer-reviewed funding, author academic affiliation and trial protocol registration will be examined using multiple logistic regression analysis.

This study will examine whether preparatory IDEAL 2b-like studies are associated with higher quality and impact of subsequent RCTs.

## INTRODUCTION

The field of surgery is constantly evolving with the introduction and modification of surgical procedures and therapeutic devices. These require adequate evaluation to ensure their effectiveness and safety. According to the University of Oxford's Center for Evidence-based Medicine, high-quality randomized controlled trials (RCTs) provide the highest standard of evidence for the effectiveness of healthcare interventions.[1] Despite poor initial adoption, RCTs are now widely used to evaluate surgical innovations' effectiveness.[2 3] However, methodological challenges unique to surgical RCTs require attention and continue to need exploration. Among these are: difficulties with participant recruitment and randomization due to a lack of patient or clinician equipoise, challenges with blinding patients and providers, poor quality control of delivery and especially inadequate assessment of operator learning curve effects, and poor standardization of interventions.[4] Chapman et al found that one in five surgical RCTs are discontinued early on, with poor recruitment being the reason for discontinuation half of the time.[5] In a cross-sectional survey of surgical RCTs, Yu et al found that majority are of low quality, calling for significant improvement in the design, conduction and reporting of surgical RCT.[6 7]

McCulloch et al introduced the IDEAL (Idea, Development, Exploration, Assessment and Long-term follow-up) framework and recommendation to provide a pathway for rigorous stepwise development

and evaluation of surgical innovations.[8] The exploration stage (IDEAL 2b) recommends carefully planned, prospective, collaborative multicenter, non-randomized studies or small-scale randomized feasibility studies prior to conducting RCTs.[9 10] Stage 2b studies provide researchers an opportunity to address factors that could hinder the proper design and conduct of a subsequent RCT, and experientially appear to increase trust and co-operation between investigators. There are currently too few published IDEAL 2b studies to allow analysis of their effects on the success of subsequent RCTs,[11] so the hypothesis that they should improve the success of authors in developing better RCTs remains untested. However, studies of the same basic design without the additional features recommended by IDEAL (prospective collaborative cohort studies, here termed 'IDEAL 2b-like studies') are much more common, and could be used as a proxy for IDEAL 2b studies to test their possible influence on subsequent RCTs. This study uses a novel approach to systematic review, taking the form of a case–control analysis of surgical trials to estimate the degree of association between the existence of prior IDEAL 2b-like studies and the quality and impact of subsequent RCTs of the same procedure or device by the same authors.

## AIMS AND OBJECTIVES

I. Systematically review RCTs evaluating any surgical procedure and devices and determine the prevalence of preceding IDEAL 2b-like studies in a random sample of surgical RCTs.

II. Determine whether the presence of preceding IDEAL 2b-like studies is associated with publication of subsequent RCTs in high-impact journals.

Hypothesis preceding IDEAL 2b-like studies are associated with increased publication of subsequent RCTs in high-impact journals.

III. Evaluate the quality of surgical RCTs and determine whether preceding IDEAL 2b-like studies are associated with higher study quality in subsequent RCTs.

Hypothesis preceding IDEAL 2b-like studies are associated with higher methodological quality in subsequent RCTs.

## METHODS

We will perform a systematic review to identify all surgical RCTs (RCTs where the intervention is a surgical operation or procedure, or a surgically implanted therapeutic device) published between 2015 and 2019. We will then divide this population of RCTs into: (1) Those Published in high-impact journals (impact factor, IF ≥5, cases) versus low-impact journals (IF <5, controls); and (2) Those with high study quality (PEDro scale ≥7, cases) versus low study quality (PEDro scale <7, controls).

We will conduct a case–control analysis to compare the prevalence and odds of the outcomes of interest (journal IF or study quality) in RCTs with and without the exposure of interest (a preceding IDEAL 2b-like study). In each group, cases and controls will be matched by year of publication at a 1:1 ratio.

The study PECO is:

**P**: RCTs evaluating any surgical procedures and implantable therapeutic devices in surgical patients.

**E:** Prior IDEAL 2b-like studies, that is, prospective collaborative non-randomized studies or randomized pilot or feasibility studies of the same device or procedure as the RCT but published prior to the RCT, and with at least one author in common.

**C:** RCTs evaluating surgical procedures and implantable therapeutic devices in surgical patients without a preceding IDEAL 2b-like study.

**O:** Association between prior IDEAL 2b-like studies and RCT quality and publication IF.

The study will be reported using PRISMA guidelines.[12]

## Search strategy

We will search for published surgical RCTs on Ovid MEDLINE, using a specific strategy developed with the assistance of specialist librarian and a methodology expert (AP and JY) (online supplemental appendix 1). Our search will be limited to studies published between January 2015 and December 2019, enabling us to identify a large proportion of recent surgical trials while providing sufficient time for MeSH indexing of all publications within the sample period. There will be no search restrictions based on language or country of publication, age, race or sex that was studied. We will use EndNote reference manager to retrieve citations and remove duplicates.

## Screening and selection process

Independent reviewers will conduct title and abstract screening, and inter-rater reliability testing will be performed to assess their consistency. An exploratory sample to estimate the prevalence of prior 2b like studies will be performed to allow sample size calculations. We will then randomly sample the RCTs found at the search stage for screening until the desired sample size of eligible papers is reached. Full texts of the selected studies with be requested and further reviewed for eligibility and data extraction. Disagreements will be resolved by consensus, or by a third reviewer when needed. Reasons for exclusion will be tracked and reported in a PRISMA flow diagram.

## Eligibility criteria
### Inclusion criteria

- ► Parallel group, RCTs published between 2015 and 2019. In case of multiple articles reporting analyses from the same RCT, the first article which falls within the study period, and which reports data for the primary outcome will be selected.
- ► Evaluating any surgical procedures (defined as a process where access is gained via an incision, natural orifice or percutaneous puncture or one that involves a device being used inside the body) or therapeutic device (defined as a medical device with therapeutic

effect generally caused through physical and mechanical effect when used on human body) in humans.

- ► In which at least one patient/group received any kind of surgical procedure or implantable therapeutic device, and a control group received alternative surgical procedures or implantable therapeutic devices, non-surgical management, or no treatment.
- ► Assessing technical, physiological, efficacy, safety or patient-reported outcomes.

## Exclusion criteria
- ► Studies without available full text.
- ► Non-randomized comparative trials, randomized pilot and feasibility trials, RCTs not in surgery, animal studies, Comments, editorials, letters, reviews of surgical RCTs.
- ► RCTs evaluating diagnostic, endoscopic, and radiologic procedures performed without any therapeutic intervention, procedures testing injection or acupuncture, medical or anesthetic therapies in surgical patients.

## Data collection and management
A standardized data collection tool will be created with Microsoft Excel and pilot tested before formal data extraction begins. Independent reviewers will extract data from a random sample of eligible RCTs until the desired sample size for analysis is reached. The following data will be extracted from each included study:

### General
Authors' names (first, second and last authors), number of authors, corresponding authors' email ID, number of centers, country of study, study settings (national, international), surgical specialty, study interest (prevention, screening, treatment, etc), funding source (government, institution, industry, none, unclear), lead or senior author based in academic center (yes/no), produced on behalf of a professional society (yes/no), prior systematic review of literature published by members of the same group/ year of publication, type of journal (general medical journal or surgical).

### RCT specific
study type (two-arm parallel, three-arm parallel), trial type (superiority, non-inferiority, equivalence), sample size, intervention type (surgical procure or device), comparison group (alternate procedure or device, standard of care, physiotherapy, medical, placebo, none), trial registration (yes or no), published study protocol available (yes or no), primary outcome type (efficacy, technical, safety, patient-reported outcome, etc).

## Identifying IDEAL 2b-like study
Preceding IDEAL stage 2b-like studies are defined as:
- ► Any prior, prospective collaborative data collection or registry of relevant patient or procedural outcomes of an operation and/or therapeutic device.
- ► Which was the subject of a subsequent RCT.

- ► Authored by any member of the listed authorship of the subsequent RCT.

We will include prospective non-randomized cohort studies as well as randomized small-scale pilot or feasibility studies, following the IDEAL Recommendations.[8–10] Collaborative studies among multiple trialists within a single center will also be included. Retrospective and single-surgeon studies will be excluded.

We will first search for IDEAL 2b-like studies by screening the reference list of each included RCT. If no possible 2b-like studies are identified, we will search Google Scholar using the following strategy: ("name of first author" OR "name of second author" OR "name of last author" AND "keyword 1 from title" AND "keyword 2 from title"). If no meaningful information is found, we will attempt to contact authors by email to request information about any preceding IDEAL 2b-like studies before the RCT, if any. If all the above-mentioned strategy fails, we will consider the information as absent.

Once we have identified IDEAL 2b-like studies, we will extract the following data to characterize them: number of centers, names of authors, number of authors, type of collaborative work (prospective data registry, prospective large non-randomized study, prospective small-scale randomized pilot/feasibility study), sample size calculation for subsequent RCT included in published report (yes or no), evaluation of feasibility of RCT performed or discussed (yes or no), mention made of progression to RCT (yes or no).

## Study quality
Independent reviewers will assess all included studies with the PEDro scale. The PEDro scale is an 11-item scale that was based on the list developed using a Delphi consensus technique.[13] It is designed for rating methodological quality of RCTs, with each item (except for one) contributing one point to the total PEDro score of 10.[13] Higher score denotes better RCT quality. In cases of disagreement, consensus will be reached through discussion and where not possible, a third reviewer will assess the study in question. For the purposes of this study, a pragmatic decision on the cut-off between high versus low quality has been made based on face validity that is, in alignment with the expectations of experienced researchers accustomed to using PEDro and with previous papers using the scale.

## Journals' IF assessment
To divide the RCTs in the search population into high-impact and low-impact journal publications, the IF of the journal in which they were published was used (based on the published IF as at October 2020). The cut-off point was selected pragmatically using the face validity criteria described for study quality assessment above. A journal IF of 5 was selected on this basis of the responses of a convenience sample of clinical researchers shown the IF table for all high-impact general clinical journals and all surgical journals.

## Data synthesis and statistical plan

The study will be separated into two parts. The first part will evaluate the associations between preceding IDEAL 2b-like studies and RCT impact, while the second part will evaluate the associations between preceding IDEAL 2b-like studies and RCT quality. All data will be inputted and analysed with STATA statistical analysis package (V. 16.1).[14] Descriptive statistics, including frequencies and proportions, will be used to describe categorical variables. Depending on data distribution, mean (SD) or median (IQR/percentile) will be used for continuous variables.

### Preceding IDEAL 2b-like studies and RCT impact

To analyse the odds of publication in high-impact journal, a matched case–control study design will be used. The presence of a preceding IDEAL 2b-like study will be the exposure and RCT publication journal impact will be the outcome. RCTs in high-impact journal (cases) will be matched to those in low-impact (controls) by the year of publication at a ratio of 1:1, to negate any differences in outcome that may be associated with publication time difference. ORs for publication in high-impact journals will be calculated for the (prior IDEAL2b-like study) and (no prior 2b like study) groups. Multiple logistic regression analysis will be used to investigate the influence of study quality (PEDro score), peer-reviewed grant funding, author academic affiliation, trial registration on the high IF publication.

To calculate the power and sample size needed for this study, a random sample of eligible surgical RCTs was assessed to determine the prevalence of IDEAL 2-b like studies among cases (high impact: IF ≥5) and controls (low impact: IF <5). This yielded 10/18 for the probability of a study being a case if exposed, and 13/54 not exposed, with a significant $\chi^2$ test result of $X^2$=8.63 indicating that 2b studies are more likely to be cases (IF ≥5). With these, a sample and power calculation were carried out for paired or matched proportions with alpha 0.05 and a power (beta) of 90%, resulting in 82 studies needed, with 41 studies in each arm.

### Preceding IDEAL 2b-like studies and RCT quality

To calculate the power and sample size needed for this study, a random sample of eligible studies was assessed to determine their study quality with the PEDRO tool and the prevalence of IDEAL 2-b like studies in the sample. With a probability of exposure (prior 2b-like study) among controls (low-moderate quality) of 0.15, and odds of exposure among cases (high quality) of OR 4.0, power of 90% at p=0.05, 102 surgical RCTs are needed, with 51 studies in each arm.

The prevalence of preceding 2b-like studies will be assessed in both the case and control group. A 1:1 case–control comparison with matching for year of publication will be carried out as described for IF above. Odds of high study quality for the groups with and without a prior 2b-like study and the OR calculated. The covariates used to estimate the adjusted OR will be: peer-reviewed grant funding, author academic affiliation and trial registration.

## DISCUSSION

RCTs are widely acknowledged to provide the highest quality evidence for the effectiveness of interventions.[1] Only well-designed and conducted RCTs can reduce bias when studying intervention effect and causality.[15] They are now widely used in the evaluation of surgical interventions, despite the many special challenges presented by surgical trials and historic hesitancy.[2 3] However, RCTs are not immune to biases which may result from poor methodological design, trial preparation or implementation.[4–7]

The IDEAL framework was developed to address the challenges of developing surgical RCTs by providing guidance on the questions which need to be addressed in preliminary studies before such RCTs are ready to proceed.[8] The exploration stage (IDEAL 2b) in particular, proposes that carefully planned prospective collaborative studies prior to the development of an RCT (IDEAL stage 3) will improve the quality of the RCTs by improving consensus between investigators through collection of data using the same dataset, allowing prestudy agreement on subgroup analyses to settle controversial issues around patient inclusion criteria or variations in technique, development of quality of delivery measures, and use of these to evaluate operator learning curves, as well as qualitative research to explore the acceptability of the trial question to patients and investigators.[9 10] Although, the IDEAL 2b-like studies might not have the above-mentioned features, we believe they are acceptable surrogate for a specifically designed IDEAL 2b study in this work.

However, there is little empirical evidence to confirm these hypothesized benefits. In particular, the effect of prior collaborative cohort studies as envisaged in IDEAL stage 2b on RCT quality and journal impact is yet to be investigated. As the closest available approximation to actual IDEAL 2b studies (of which there are currently too few to allow meaningful analysis), we chose to focus on prior prospective collaborative cohort studies of outcomes, and determine whether they are associated with superior quality or high-impact journal publication in subsequent RCTs. A positive result would provide supporting evidence for IDEAL's proposal that IDEAL 2b studies are valuable and could improve RCTs in surgery.

Important potential limitation of this study are: (1) the incompleteness of available information for the preceding IDEAL 2b-like studies that were conducted by authors of RCTs. In the initial phase of the study design, we conducted a feasibility search for IDEAL 2b-like studies using the search engines: Web of Science; Google Scholar; Scopus and ELTEFIND. We chose Google Scholar for this study because it yielded better results in our exploratory studies compared with the others. Despite Google Scholar being very versatile, many preparatory studies

might not be published and reply rates may be low when authors are contacted. However, this may produce a bias in favour of IDEAL 2b studies only if RCTs published after unreported 2b-like studies had lower quality and in lower impact journal than RCTs published with no such prior studies; (2) Searching only MEDLINE for interested study sample. While the failure to search a range of databases may decrease the yield of retrieved articles, and likely limit the quality of this study from the theoretical optimum, it is pragmatic and provides enough articles to meet the required sample size for the study. Also, the characteristics of MEDLINE (in terms of comprehensiveness and the nature of its selectivity) has been previously reported.[16] (3) The unavoidable degree of subjectivity involved in categorizing studies as high quality or low quality, even with standardized tools in frequent use, such as the PEDRO. An alternative is the Cochrane Risk of Bias Tool (C-ROB).[17] Although, ROB is often used interchangeably with quality assessment, it is important to note that they remain distinct.[18] C-ROB tool is designed to assess ROB, while the PEDRO scale is designed to assess study quality, which was the primary focus of our study. This study's case–control design requires a tool that yields an assessment of study quality that could be quantified and categorized distinctly, (high quality=case or low quality=control). While the C-ROB tool yields three result categories (high, low and unclear ROB), the 'unclear risk' of bias category would be very difficult to characterize objectively. Further, because of the subjectivity of the C-ROB tool, a high proportion of studies fall in the unclear ROB category.[19] These features make the C-ROB tool a less suitable option for our study. The PEDRO scale was designed to assess the quality of studies evaluating complex intervention, such as surgical procedures and devices, and can relatively easily be dichotomized or categorized numerically, as we have planned for this study. It also allows for a more granular analysis of individual components of study quality than the C-ROB, including quality items better suited to complex interventions such as surgery. The validity and reliability of the PEDRO tool have also been shown previously.[20] We will attempt to minimize any likely bias introduced through dichotomization of studies into high and low quality by also examining study quality categorically, measured with the PEDRO scale as low (0–3), medium (4–6), good (7–8) or excellent (9+) and separately as low (0–3), medium (4-6) and high (7+) in the analysis of the influence of prior 2b-like studies on IF; 4. It is impossible to exclude the possibility that an apparent association between prior 2b-like studies and either quality or high-impact publication may be due to confounding from a covariate not included in our analysis.

## Dissemination

The protocol will be registered on Open Science Framework and published in a peer-reviewed journal. The study result will be disseminated at national and international conferences and published in a peer-reviewed journal.

**Author affiliations**
[1]Department of Surgery, Maimonides Medical Center, Brooklyn, New York, USA
[2]Nuffield Department of Surgical Science, University of Oxford, Oxford, UK
[3]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[4]Chinese Evidence-based Medicine Center, Sichuan University West China Medical Center, Chengdu, Sichuan, China
[5]Surgery, Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK

**Twitter** Arsenio Paez @DrArsenioPaez

**ORCID iDs**
Mudathir Ibrahim http://orcid.org/0000-0002-6345-1533
Arsenio Paez http://orcid.org/0000-0002-7984-514X
Baptiste Vasey http://orcid.org/0000-0002-0017-8891
Peter McCulloch http://orcid.org/0000-0002-3210-8273

## REFERENCES

1 Center of Evidence-Based Medicine. Available: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009/ [Accessed 30 Sep 2020].
2 McCulloch P, Taylor I, Sasako M, et al. Randomised trials in surgery: problems and possible solutions. *BMJ* 2002;324:1448–51.
3 Cook JA. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. *Trials* 2009;10:9.
4 Robinson NB, Fremes S, Hameed I, et al. Characteristics of randomized clinical trials in surgery from 2008 to 2020: a systematic review. *JAMA Netw Open* 2021;4:e2114494.
5 Chapman SJ, Shelton B, Mahmood H, et al. Discontinuation and non-publication of surgical randomised controlled trials: observational study. *BMJ* 2014;349:g6870.

6   Yu J, Chen W, Chen S, *et al*. Design, conduct, and analysis of surgical randomized controlled trials: a cross-sectional survey. *Ann Surg* 2019;270:1065–9.

7   Yu J, Li X, Li Y, *et al*. Quality of reporting in surgical randomized clinical trials. *Br J Surg* 2017;104:296–303.

8   McCulloch P, Altman DG, Campbell WB, *et al*. No surgical innovation without evaluation: the ideal recommendations. *Lancet* 2009;374:1105–12.

9   McCulloch P, Cook JA, Altman DG, *et al*. Ideal framework for surgical innovation 1: the idea and development stages. *BMJ* 2013;346:f3012.

10  Ergina PL, Barkun JS, McCulloch P, *et al*. Ideal framework for surgical innovation 2: observational studies in the exploration and assessment stages. *BMJ* 2013;346:f3011.

11  Khachane A, Philippou Y, Hirst A, *et al*. Appraising the uptake and use of the ideal framework and recommendations: a review of the literature. *Int J Surg* 2018;57:84–90.

12  Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.

13  Yamato TP, Maher C, Koes B, *et al*. The PEDro scale had acceptably high convergent validity, construct validity, and interrater reliability in evaluating methodological quality of pharmaceutical trials. *J Clin Epidemiol* 2017;86:176–81.

14  StataCorp. *Stata statistical software: release 14*. College Station, TX: StataCorp LP, 2015.

15  Fanaroff AC, Califf RM, Harrington RA, *et al*. Randomized Trials Versus Common Sense and Clinical Observation: JACC Review Topic of the Week. *J Am Coll Cardiol* 2020;76:580–9.

16  Rathbone J, Carter M, Hoffmann T, *et al*. A comparison of the performance of seven key bibliographic databases in identifying all relevant systematic reviews of interventions for hypertension. *Syst Rev* 2016;5:27.

17  Higgins JPT, Altman DG, Gøtzsche PC, *et al*. The Cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928.

18  Furuya-Kanamori L, Xu C, Hasan SS, *et al*. Quality versus Risk-of-Bias assessment in clinical research. *J Clin Epidemiol* 2021;129:172–5.

19  Jørgensen L, Paludan-Müller AS, Laursen DRT, *et al*. Evaluation of the Cochrane tool for assessing risk of bias in randomized clinical trials: overview of published comments and analysis of user practice in Cochrane and non-Cochrane reviews. *Syst Rev* 2016;5:80.

20  Cashin AG, McAuley JH. Physiotherapy evidence database (PEDro) scale. *J Physiother* 2020;66:59.