
Publisher's PDF, also known as Version of record

License (if available):
CC BY-NC-ND

Link to published version (if available):
10.1016/j.jclinepi.2019.05.007

Link to publication record in Explore Bristol Research
PDF-document

This is the final published version of the article (version of record). It first appeared online via Elsevier at https://www.sciencedirect.com/science/article/pii/S0895435619300630 . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/user-guides/explore-bristol-research/ebr-terms/
Funders improved the management of learning and clustering effects through design and analysis of randomized trials involving surgery

Elizabeth J. Conroy a,*, Anna Rosala-Hallas a, Jane M. Blazeby b, Girvan Burnside c, Jonathan A. Cook c, Carrol Gamble a,d

a Department of Biostatistics, University of Liverpool, A member of Liverpool Health Partners, Liverpool, UK
b Centre for Surgical Research, Population Health Sciences, University of Bristol, Bristol, UK
c Centre for Statistics in Medicine, University of Oxford, Oxford, UK
d North West Hub for Trials Methodology Research, University of Liverpool, Liverpool, UK

Accepted 15 May 2019; Published online 20 May 2019

Abstract

Objective: The objective of this study was to provide insight into current practice in planning for, and acknowledging, the presence of learning and clustering effects, by treating center and surgeon, when developing randomized surgical trials.

Study Design and Setting: Complexities associated with delivering surgical interventions, such as clustering effects, by center or surgeon, and surgical learning should be considered at trial design. Main trial publications, within the wider literature, under-report these considerations. Funded applications, within a 4-year period, from a leading UK funding body were searched. Data were extracted on considerations for learning and clustering effects and the driver, funder, or applicant, behind these.

Results: Fifty trials were eligible. Managing learning through establishing predefined center and surgeon credentials was common. One planned exploratory analysis of learning within center, and two within surgeon. Clustering, by site and surgeon, was often managed through stratifying randomization, with 81% and 60%, respectively, also planning to subsequently adjust analysis. One-third of responses to referees contained funder led changes accounting for learning and/or clustering.

Conclusion: This review indicates that researchers do consider impact of learning and clustering, by center and surgeon, during trial development. Furthermore, the funder is identified as a potential driver of considerations.

Keywords: Randomized controlled trials; Surgery; Clustering; Learning curve; Statistics; Surgical intervention

1. Introduction

Randomized controlled trials (RCTs) are recognized as providing the highest level of evidence, second only to systematic reviews of such trials [1]. The need for surgical randomized trials is well recognized [2,3], and this has led to a push for growth in recent years [3,4]. Leading research organizations are supporting this growth through establishing a number of initiatives and research objectives, ultimately aiming to improve the global surgical evidence base [5–10]. One such initiative, set up by the UK’s leading publicly funded health research body, the National Institute
What is new?

Key findings
- Results indicate that while considerations of surgical learning and clustering are underreported in main trial publications, funders and researchers alike appear to be aware of the need to manage these aspects at trial design stage.

What this adds to what is known?
- A novel assessment of the decision making behind intended design and analysis with respect to the management of surgical learning and clustering is presented.
- This review investigates successful funding applications comprising a wide variety of trials, both by surgical discipline and by geographic location, by a leading UK funder.
- This review is timely as it comprises successful funding applications following a call by this funder recognizing a need for an increase in evidence-based surgical research.

What is the implication and what should change now?
- Insight into the promising role of the funder as a driver to improving the, long criticized, surgical evidence base is provided.

2. Materials and methods

2.1. Included studies

We sought to examine trials that had received funding from the NIHR from two funding streams, the Health Technology Assessment (HTA) program [17] and Efficacy and Mechanism Evaluation (EME) [18] program, in the United Kingdom, from 2012 to 2016. Research projects funded by these programs are either in response to a commissioning brief or an open investigator-led call. These funding streams were chosen as they are known to endorse high-quality research and were actively funding surgical research during this time [7]. An initial unpublished search indicated that this period would provide a reasonable cohort size to establish current practice. All randomized trials where the patient pathway involves a surgical intervention of any kind were eligible for inclusion.

2.2. Documents for review

The NIHR HTA and NIHR EME funding process involves a two-stage, peer-reviewed application process.Protocols and the commissioning brief (where applicable) were obtained from the open-access NIHR Journals Library [19]. The NIHR Evaluation, Trials and Studies Coordinating Centre (NETSCC) provided documentation not publicly available: project descriptions and applicant responses to reviewer comments.

2.3. Data extraction

A previously developed extraction form [16] was adapted for use on this cohort by E.J.C. and C.G. and approved by G.B., J.M.B., and J.A.C. (see Supplement A1). The extraction form was piloted on five applications initially and, as no further amendments were required, subsequently
used on all applications by a single assessor (E.J.C.). Data extracted were quality-checked through double data extraction by a second reviewer (A.R.H.) on 10% of all applications. A discrepancy rate was specified a priori such that if greater than 5% across all fields then a further 10% would be checked until the rate was below 5%. Discrepancies were jointly reviewed and agreement was reached; if agreement could not be reached, then a third reviewer (C.G.) was consulted.

Details on trial design, randomization stratification, sample size adjustment, predetermined center and surgeon credentials, outcomes, and planned statistical analyses that adjusted for center and surgeon were collected.

2.4. Statistical analysis

Quantitative items were summarized using descriptive statistics; no formal statistical comparisons were undertaken. Data were analyzed using SAS 9.3 (SAS Institute Inc., Cary, NC, USA). Open textual data items were categorized using NVivo qualitative data analysis software (QSR International Pty Ltd., Doncaster, Victoria, Australia, version 10, 2012). A confidentiality agreement with the NIHR Evaluation, Trials and Studies Coordinating Centre was signed prior to receiving the documentation. The raw data cannot therefore be made publicly available, and text extracts have been anonymized by removal of treatment identifiers. Deleted text is denoted by [...] and the addition of words or replaced words is denoted by [words] to aid understanding.

3. Results

3.1. Cohort details

The NETSCC compiled a report listing all surgery randomized controlled trials funded by the NIHR HTA and NIHR EME funding streams within the eligible period. Sixty potentially eligible studies were identified, of which 49 (82%) met the eligibility criteria following further central screening (Fig. A1).

3.2. Double data extraction

Five articles were randomly selected from the eligible studies for double data extraction. Of 155 variables checked, two discrepancies were identified (1% error rate).

3.3. Cohort summary

The majority of the applications were funded by the NIHR HTA (n = 44/49, 89%) and had start dates from 2014 onward (n = 37/49, 76%) (see Table 1).

Documents for review consisted of commissioning briefs (n = 15/49, 31%), project descriptions (n = 40/49, 82%), applicant responses to board and peer review comments (n = 40/49, 82%) and protocols (n = 42/49, 86%). Either the protocol or project description was available for all applications (see Table 1).

One application consisted of two distinct RCTs, herein treated as two separate trials.

3.4. Trial demographics

Trials were primarily two-armed (n = 45/50, 90%) and of a parallel design (n = 49/50, 98%). Eight did not use a pilot or feasibility study (n = 8/50, 16%) [20]. In 11 studies (n = 11/50, 22%), surgery was not the intervention of interest and delivered as part of the patient pathway. Where surgery was the intervention of interest (n = 39/50, 78%), 21 compared against surgery, for example, minimal access vs. open surgery (n = 21/39, 54%). The remaining 18 compared surgery against a nonsurgical comparator (medical comparator, e.g., injection vs. surgery: n = 7/39, other e.g., active monitoring and surgery vs. active monitoring only: n = 11/39) (see Tables A1 and 2).

3.5. Recruitment and randomization

Patients were the randomization unit in all trials and were primarily allocated to equal groups (n = 48/50, 96%). The majority stratified randomization (n = 46/50, 92%). In trials comparing two surgeries, there were no expertise-based designs [21]. Table A2 provides more detail.

Almost all studies were multicentre (n = 49/50, 98%), with over half stratifying by center (n = 28/49, 57%). Of the 21 that did not stratify by center, one provided justification for not stratifying by centre and this was related to concern over allocation concealment:

“To reduce the risk of the randomization sequence being predictable we will not stratify by center, which in addition to using randomly selected permuted blocks, will make the allocation sequence unpredictable for individual trial centers.”

Twenty-two trials had multiple surgeons within each center, of which eight stratified the randomization accordingly (n = 8/22, 36%). Two surgeon-stratified trials followed funder recommendation.

“We have made a number of changes since the first application…randomization will be stratified according to [stratification 1], [stratification 2], and according to consultant surgeon.”

In trials reported as multicenter and multisurgeon (n = 21), two stratified for both center and surgeon, eleven center only, six surgeon only, and two stratified for neither. Three trials were international, of which one stratified randomization on randomized within a UK, or non-UK, center.

Table 2 provides more detail.
3.6. Surgeon and center credentials

Center and surgeon credentials, or inclusion criteria of those delivering the intervention, were provided in 41 (n = 41/50, 82%) and 36 (n = 36/50, 72%) trials respectively (Table 3). Most common center credentials were case volume (n = 20) and required fields of expertise within center (n = 13). Examples of surgeon credentials were
grade or experience \((n = 16)\) and study-specific training \((n = 13)\).

### 3.7. Trial outcomes related to learning and clustering

Forty-one applications explored outcomes that may reflect variability in center or surgeon skill \((82\%, \text{Table } 4)\). Common outcomes were safety events \((n = 36)\); recovery from surgery \((n = 13)\), and operative time \((n = 6)\).

Surgeon level outcomes were experience of surgeons in trial, established through qualitative methods \((n = 3)\); surgeon accuracy as a main trial outcome \((n = 1)\); and expertise \((n = 1)\), more specifically:

“\text{The first [feasibility] phase will establish \{words\} and a measure of surgical expertise.}”

### 3.8. Statistical considerations

#### 3.8.1. Sample size calculation

There were no examples of sample size adjustment for clustering at a center level. Three applications adjusted the sample size for surgeon using an intraclass correlation coefficient and a fourth chose not to adjust although provided justification:

“As this study is not evaluating surgery per-se, surgical experience is not a criterion for participation (all participants will be under the care of a consultant surgeon). In the context of [this] study, clustering by surgeon is not relevant to the sample size and can be ignored (on the basis that the intraclass correlation is negligible).”

#### 3.8.2. Exploratory analysis

Eight applications planned exploratory analysis considering differences by center. Three analyzed using descriptive statistics and three via a subgroup analysis: the first conducting a trial center-by-treatment effect analysis, the second comparing outcomes between more and less experienced centers, and the third exploring trends within centers over time. A sensitivity analysis adjusting for center effects was planned in one application. Learning within center was described in another.

“The effect of experience in [comparator intervention] at each recruitment center will be studied to characterise the effect of the learning curve on clinical effectiveness, and also the effect on [standard intervention] outcomes.”

### Table 3. Center and surgeon credentials

<table>
<thead>
<tr>
<th>Center level</th>
<th>Surgeon level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center credential provided</td>
<td>41</td>
</tr>
<tr>
<td>Case volume</td>
<td>20 (48%)</td>
</tr>
<tr>
<td>Fields of expertise within center</td>
<td>13 (32%)</td>
</tr>
<tr>
<td>Experience required without definition</td>
<td>9 (22%)</td>
</tr>
<tr>
<td>Experience required with definition</td>
<td>8 (20%)</td>
</tr>
<tr>
<td>Good recruiting reputation</td>
<td>8 (20%)</td>
</tr>
<tr>
<td>Experience required with definition</td>
<td>8 (20%)</td>
</tr>
<tr>
<td>Access to equipment required</td>
<td>7 (17%)</td>
</tr>
<tr>
<td>Center to undertake trial-specific training</td>
<td>2 (5%)</td>
</tr>
<tr>
<td>Demonstrated ability to participate</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Interest expressed in specific treatment</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Prior number of cases required</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Center delivers one treatment only</td>
<td>1 (2%)</td>
</tr>
</tbody>
</table>

### Table 4. Outcomes

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Relevant outcome reported</th>
<th>Safety measures</th>
<th>Recovery from surgery time</th>
<th>Operative time</th>
<th>Patient satisfaction with surgery</th>
<th>Infection</th>
<th>Experience of surgeons in trial$^a$</th>
<th>Surgeon accuracy</th>
<th>Surgeon expertise$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relevant outcome reported</td>
<td>41</td>
<td>36 (88%)</td>
<td>13 (32%)</td>
<td>6 (15%)</td>
<td>4 (10%)</td>
<td>3 (7%)</td>
<td>1 (2%)</td>
<td>1 (2%)</td>
</tr>
</tbody>
</table>

$^a$ Established using qualitative methods.  
$^b$ Feasibility outcome.
sampled for observations in theater according to their treating surgeons’ grade. As with center, one application planned a sensitivity analysis that adjusted for surgeon.

3.8.3. Formal adjustment

Formal adjustment for multiple center or surgeon effect was planned in 21 and 15 applications, respectively. Table 5 provides more detail. When formally adjusting for center, nine planned to use a random effect and thirteen did not specify. Similarly, six planned to adjust for surgeon using a random effect and nine did not specify. Of the applications planning a formal adjustment, 17 ($n = 17/21, 81\%$) of applications adjusting for center and nine ($n = 9/15, 60\%$) adjusting for surgeon did so in addition to stratifying randomization by these variables.

The two applications that planned to stratify by both center and surgeon (Table 3) also planned formally adjusting analysis by these variables.

3.9. Funder-led considerations

3.9.1. Commissioning briefs

Of the 15 commissioning briefs, one permitted single-center studies and one required a multicenter setting. No other brief gave guidance with respect to number of centers. Two briefs identified surgical learning considerations as an issue to address: the first indicating outcomes may be independent of surgeon grade and the second:

“The Board suggested that the team should consider the addition of a second center to demonstrate generalisability and help with recruitment.”

In one application, funders requested applicants increase homogeneity in treatments and the applicants argued against this.

“To ensure homogeneity in treatments we have consulted with our participating surgeons [and] the National […] Registry and agreed to specify the use of a CE marked [device]…there are three main devices. Surgical trials that specify a single type of [device] are notoriously difficult to conduct and we do not believe such a design could recruit surgeons, nor would the outputs be generalisable.”

Further considerations with regard to surgeon credentials ($n = 3$) and the impact of surgeon equipoise on recruitment ($n = 1$) were also funder driven.

“The sample size has been increased from a total of $[n]$ patients to a total of $[1.4n]$ to take into account clustering of surgeon as per the feedback from the first stage.”

4. Conclusions

This review has investigated the decision making behind intended design and analysis of 50 randomized surgical trials funded by the NIHR EME and NIHR HTA programs from 2012 to 2016. These results show frequent consideration of centers and surgeon impact during design, and these may be funder led, due to concerns around homogeneity or generalizability of results. This review provides a cross-sectional insight into current practice of researchers, and expectations of reviewers and funders, during trial design within two streams of a major UK funder [17,18].

The need for transparency around learning curves and clustering is highlighted within reporting of nonpharmacological interventions guidelines [22,23], and a review of the published literature identified a deficiency in adherence to these [16]. By contrast, this review identifies that

---

**Table 5. Planned statistical adjustments through analysis in multicenter and multisurgeon trials**

<table>
<thead>
<tr>
<th>Adjustment made</th>
<th>Center</th>
<th></th>
<th>Surgeon</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$N$</td>
<td>$n/N%$</td>
<td>$n$</td>
<td>$N$</td>
</tr>
<tr>
<td>Adjustment made</td>
<td>21</td>
<td>49</td>
<td>43%</td>
<td>15</td>
</tr>
<tr>
<td>Approach to adjustment (type of effect)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td>0</td>
<td>21</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>Random</td>
<td>9</td>
<td>21</td>
<td>43%</td>
<td>6</td>
</tr>
<tr>
<td>Time varying</td>
<td>0</td>
<td>21</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>Not specified</td>
<td>12</td>
<td>21</td>
<td>57%</td>
<td>9</td>
</tr>
<tr>
<td>Randomization stratified by and adjustment made</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>17</td>
<td>21</td>
<td>81%</td>
<td>9</td>
</tr>
</tbody>
</table>
considerations to manage learning and clustering are made, by both researchers and funders, during development of trials funded by a prestigious body. For example, 30% of multicenter and 12% of multisurgeon studies reported a statistical adjustment of these within published manuscripts. This was 43% and 68%, respectively, in this cohort. When randomization was stratified by center or surgeon, this was accounted for in the analysis in 30% of multicenter and 40% of multisurgeon trials in the published manuscripts, as opposed to 81% and 60% in this cohort. In drawing this comparison, it is important to differentiate between the intended audiences. The detail required for a funding application, assessed by clinicians and methodologists/statisticians, may exceed that required to communicate results to a clinical audience. This demonstrates benefit in exploring unpublished trial documentation to understand approaches to trial design and analysis and highlights the need for improvements to transparent reporting.

The cohort included successful applications to the NIHR 2012 call for Applied Health in Surgery [7]. This call recognized the need to increase research-based evidence in surgery. Applications that evaluated technology-driven implanted or implantable medical devices, surgical procedures, or surgical services were invited. As a clinical trial is typically a major financial investment [24], applicants need to assure funders that their proposal is important, is well designed, and demonstrates scientific value to add to the current evidence base. Each application undergoes a peer review process, where “experts” critically review the trial to ensure standards are met in terms of design, quality, feasibility, acceptability, and importance of the topic [17,18]. A strength of this review is the insight into the designs proposed to funders, and impact of feedback on subsequently funded studies.

While the degree of learning and clustering will vary trial-to-trial, many interventions require surgical skill in their delivery regardless of whether the surgery is the intervention of interest. The impact of any potential imbalance in delivery on comparing interventions should be considered at trial outset routinely. Early and careful consideration will ensure that procedures are standardized as completely as possible such that, in severe cases, the trial team can alleviate any doubts about homogeneity raised by the medical community should the trial results be questioned [12]. These results indicate funder awareness of this early consideration, with one of the two examples of balancing randomization surgeon following recommendation being in a trial where surgery was not the intervention of interest.

When interpreting these results, it is important to consider the limitations of this review. First, only successful applications could be included because of confidentiality constraints. It is therefore not possible to determine whether the management of learning and clustering contributes to the success of the application. However, given that the application process consists of iterations whereby peer reviewers are able to request that researchers address paucities in their application, it is unlikely that a promising application, lacking in the appropriate considerations, would be deemed unsuitable for funding outright. More likely, researchers would be given the opportunity to make these considerations during this iteration process. Second, as part of this iterative review, it is possible that additional discussions at the funder board meetings did not make it in to the comments fed back to applicants. This could mean that funders raised these issues more frequently than this review suggests. Third, due to the nature of the grant application process, the funder impact observed may be in part due to an increased awareness of the reviewers involved. Fourth, this work has focused on a single funding body that primarily supports UK-based research. However, trials supported span a wide range of surgical specialities and healthcare conditions and results from this review will be generalizable to other funding bodies with a similar peer review process.

Fundamental to trial design and analysis is understanding the objectives. While considerations relating to clustering and learning effects are not widely reported in main trial publications, these results indicate both funders and researchers consider these aspects to address a specific research question. Such issues may have varying relevance depending on the overall design of the trial. A very pragmatic study may deliberately include surgeons and centers of all types and have less emphasis on expertise and learning, whereas the delivery of the intervention in more explanatory studies is critical and requires consideration during design and analysis. Another approach to overcoming these issues is to provide quality assurance of the intervention. Early work to develop methods to achieve this have been developed, and it is expected that this will expand in the future [25]. Furthermore, these results provide insight into the promising role of the funder as a driver to improving the, long criticized, surgical evidence base. The funder, who has influence over whether and how studies are carried out and has been suggested as a driver for improving the quality of research during the period of growth for surgical trials [3], can play a valuable role in ensuring that future trials do not have the same shortfalls as those in the past.

CRediT authorship contribution statement

Elizabeth J. Conroy: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. Anna Rosala-Hallas: Validation, Writing - review & editing. Jane M. Blazeby: Conceptualization, Methodology, Supervision, Writing - review & editing. Girvan Burnside: Conceptualization, Methodology, Supervision, Writing - review & editing. Jonathan A. Cook: Conceptualization, Methodology, Supervision, Writing - review & editing. Carrol Gamble:
Conceptualization, Data curation, Methodology, Supervision, Writing - original draft, Writing - review & editing.

Acknowledgments

The authors would like to thank the National Institute for Health Research (NIHR) Evaluation, Trials and Studies Coordinating Centre (NETSCC) for collating and permitting access to the data used for this review.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2019.05.007.

References