Developing appropriate methodology for the study of surgical techniques

Peter McCulloch

Nuffield Department of Surgery, University of Oxford, UK

E-mail: peter.mcculloch@nds.ox.ac.uk

DECLARATIONS

Competing interests
None declared

Funding
None

Ethical approval
Not applicable

Guarantor
PM

Contributorship
PM is the sole contributor

Acknowledgements
The author would like to acknowledge Jonathan Meakins and the other attendees of the Balliol Conferences

Introduction

The critical re-examination of practice stimulated by evidence-based medicine revealed that some disciplines were supported by more scientifically valid research than others. Surgery was specifically criticized for failing to adopt randomized controlled trials (RCTs), when the small proportion of practice supported by such evidence was reported. Randomized trials of surgical technique face a range of problems,[1] several shared with other disciplines using complex or therapist-dependent interventions, (e.g. Public Health, Physiotherapy), but the problem is more fundamental than surgeons not doing enough RCTs. The development process of surgical techniques simply does not fit well into our current paradigm of clinical investigation.

This has often prevented adequate evaluation ever occurring, leaving much surgical knowledge in a state of arrested development. We need to recognize that the technical development process differs fundamentally from that for drug treatments, having its own recognizable stages. Only by explicitly acknowledging this natural history can we recognize the weak points in current methodological approaches, and develop more appropriate study designs (Table 1). In this article the phases of surgical technique development are described, together with the problems which commonly affect their study. Some approaches which might result in more effective evaluation of each stage are proposed.

Stages of development

The phases of development for surgical techniques differ significantly from those for new drugs. Operations, for example, are not tested on healthy volunteers, and the dose of surgery cannot be adjusted. Development of surgical techniques begins with description and proof of feasibility (Phase 0) followed by refinement and definition of the procedure (Phase 1). After this, techniques undergo dissemination and evaluation by other surgeons (Phase 2). Ideally, techniques should then be compared with gold standard therapy (Phase 3). Finally, monitoring for long-term adverse effects and deviation from expected outcomes occurs (Phase 4) and is arguably even more important for surgical techniques than for medications.

Phase 0: Proof of principle

The first publication about a new surgical technique is usually a report of a single case or a small case series. This serves the purposes of describing the technique, and of establishing that a satisfactory outcome is feasible. Analogous studies in drug development would be preliminary to a Phase 1 trial.

Phase 1: Refinement and definition

In the next phase, the technique is modified in the light of early experience. Surgical techniques often develop initially via a series of small steps, each unlikely to produce effects large enough to be confirmed statistically without a large trial, yet such a trial is clearly inappropriate for a small modification. We currently lack formal study designs which allow rapid but provisional conclusions about outcomes based on small sample sizes. Consequently this phase is often unreported, as surgeons reasonably hesitate to publish the results of incompletely developed techniques. By the end of this phase the technique is optimized and can be evaluated in larger scale studies.

Table 1
Current weaknesses in the evaluation of surgical techniques

Aspects of methodology for evaluating surgical techniques which are currently lacking or in need of improvement	*Phase of study*
Innovative designs for small sample studies	(I)
Description of disease population	(II)
Description of selection process	(II)
Explanation of modifications of technique (with timescale)	(II)
Standard methods for reporting co-morbidity	(II)
Standard set of reported outcomes	(II)
Standard definitions of reported outcomes	(II)
Learning curve evaluation	(II/III)
Definition of procedures compared	(III)
Quality control	(II/III)
Informed consent based on qualitative assessment of patient values	(III)
Third party randomization to improve equipoise	(III)
Risk adjusted continuous performance monitoring	(IV)

This iterative process might usefully be studied using formal industrial continuous quality improvement methodology.[2] The PDCA ('Plan; Do; Check; Act') cycle, comprising changes in method, careful observation for a pre-specified period, comparison with previous results and a decision to adopt, reject or continue testing seems particularly suitable for new operations. The development of the total hip replacement by Charnley used similar methodology, although the long follow-up required caused problems in identifying unsuccessful modifications early.[3] Statistically, we need methods which allow objective estimates about the likelihood of particular outcomes in samples too small for conventional significance testing. The PDCA study design approximates to the 'interrupted time series', for which either frequentist or Bayesian statistical approaches exist. The latter allows repeated re-estimation of the likely treatment effect, based on 'aliquots' of additional cases,[4] making it easier to rapidly determine whether the likelihood of improvement justifies persisting (Table 2). A parallel control group might be provided by observing patients receiving conventional therapy. Explicit reporting of this 'baizen' process might help development by preventing repetition of unhelpful modifications.

Phase 2: Dissemination

Promising techniques are usually adopted rapidly by other interested surgeons, who report their experience, usually, by publishing personal case series. The defects of case series are well recognized, and their persisting prominence has been a major cause of criticism of the surgical literature.[1] Case series are usually retrospective, and since they are non-comparative, context specific, not hypothesis driven and often based on incomplete data, the reliable conclusions we can derive from them are extremely limited. Comparison of such studies should not be used as a guide to the superiority of one treatment over another. Purists[5] argue that randomized trials should begin from the first report of a new procedure, but surgical innovators' anxiety to optimize procedures and overcome their learning curves before setting up formal comparisons makes such exhortations unrealistic. Learning curves are important in studies of surgical technique, but unfortunately case series do not usually tell us anything about them. The durability of this maligned study 'design' illustrates the powerful need surgeons feel for real outcome data – however biased. Perhaps we should accept this reality, and work to make series as useful as possible.

At present, the variability in reporting details of case selection and outcome makes any summary of findings extremely difficult. In particular, the reporting of postoperative morbidity is usually rendered almost meaningless by the absence of standard definitions. Thus, in a recent study of wound infections,[6] infection rates differed nearly three-fold depending on the definition adopted.

The CONSORT initiative for RCTs and similar initiatives for other study designs[7] have greatly assisted interpretation. No similar initiative has been developed for surgical series, although some suggestions have been made for particular specialties.[8,9] An equivalent effort is needed to develop standards for reporting surgical case series, particularly for reporting patient selection, co-morbid pathology and postoperative morbidity.

Bridging the gap: the Phase IIS study

Once a technique has achieved stability and popularity, it becomes important to determine whether it is better than current treatments, preferably

Table 2
Proposals for modified study designs for surgical techniques

Proposals for modified study designs for evaluating surgical techniques	
Phase 1	Interrupted time series/industrial quality improvement hybrid. Planned repetitive Bayesian analysis and modification.
Phase 2	Development of reporting standards for surgical case series.
Phase 2	Phase IIS prospective non-randomized study progressing to RCT.
Phase 3	Modified RCT with operation definition, quality control, learning curve documentation.
Phase 4	Continuous performance monitoring using statistical process control.

by performing a randomized trial. Adequately powered trials can rarely be developed by single units, but the barriers to multicentre randomized trials of surgery are significant.[1] These include problems of quality control, definition of the intervention, the learning curve, and both practical and psychological/sociological barriers to participation and collaboration. These barriers are not insuperable, as demonstrated by examples of excellent trials of surgical technique.[10]

A loose definition of the surgical procedure can lead to controversy about whether the trial has really tested the intended treatment.[11] The quality of the surgeons' performance obviously affects outcome, but is only rarely monitored in surgical trials.[12] Randomization should prevent bias, but if the effects of surgical quality are large relative to the difference between procedures, 'noise' from quality variations may make differences difficult to detect. The relationship of performance to experience is not wholly predictable,[13] but in general randomizing patients during the rapid learning phase introduces bias against the new procedure as a time-dependent variable, declining with increasing experience.[14] A surgical team's comfort with a standard methodology is important for quality, and once established, makes it difficult to return to previous methods even for a randomized trial. Some have argued that surgeons should only be asked to do procedures with which they are comfortable, eliminating the learning curve,[14] but this design potentially introduces major new confounders, unless the rest of the surgical, nursing and anaesthetic team and the organizational framework are kept constant for pairs of surgeons.

Given these obstacles, it is not surprising that the process of investigation often stops short of an RCT. To move from unit-based non-randomized studies to a multicentre RCT requires the development of a group who are willing to contribute to a common database using common terminology. We therefore need to develop a bridging study methodology based on prospective collaboration, to allow collaborators to achieve consensus on definitions, and to monitor their learning curves, using tools such as CUSUM.[15] A model for these studies could be oncological Phase II studies, which have limited objectives explicitly directed towards a possible future RCT. In surgical 'Phase IIS' studies, the objectives of collaborative non-randomized prospective study would be: (a) to document learning curves; (b) to determine likely treatment effect, permitting power calculations for RCTs; (c) to build consensus on the question for an RCT; and (d) to develop quality measures to confirm delivery of the intended operation. Such markers have been under-developed in surgery, but could include standardized photos, flow measurements, verification samples for histology (e.g. of terminal ileum at colonoscopy) or pH measurement in the oesophagus after anti-reflux surgery. A recent Italian study of gastrectomy provided a successful example of a phase IIS-like investigation which progressed to an RCT after a non-randomized phase during which the quality of surgery was carefully monitored.[16]

Phase 3: Comparison with current standard treatment

RCTs of surgical techniques will be more likely to succeed if they incorporate pre-study documentation of learning curves, a clear definition of the intervention and quality control measures. Where surgery is compared with non-invasive treatment, the asymmetry between the two arms poses problems of equipoise for participants and clinicians[1] making difficulties for both recruitment and interpretation. Surgery for early prostate cancer may

cure the patient but risks causing impotence or incontinence, whereas watchful waiting avoids these risks but involves uncertainty about cancer progression. Such complex choices are psychologically tricky, and suffer severely from 'framing effects' depending on how the choice is presented.

Methods are needed to make this kind of trial less difficult to perform and interpret, by incorporating the values of possible outcomes to the patients and clinicians involved. This requires qualitative research techniques, with which most surgeons are unfamiliar.

A recent trial of prostate cancer treatment exemplified the imaginative use of such techniques to solve a serious recruitment problem.[17] Qualititative analysis of the values placed by patients on possible outcomes allowed the development of a standard methodology for informed consent delivered by nurses not involved in treatment.

Phase 4: Surveillance and quality control

For established techniques, the need is for monitoring of results to ensure that outcomes remain satisfactory among multiple practitioners, to identify rare or long-term adverse effects, and to evaluate factors which modify success. The pharmaceutical industry invests heavily in 'post-marketing surveillance' to achieve these goals, for which randomized trials are inappropriate. No systematic approach exists in surgery, although some registries for minimally invasive techniques fulfil this function.[18]

Standard statistical methods for continuous monitoring of surgical outcomes would greatly assist the surveillance phase. Where no treatment comparison is involved, statistical power is the most important property of a data source. We cannot influence risk factors such as diabetes, and must therefore rely on analysis of outcome in prospectively collected data without random allocation of exposure. A large data-set is very useful, as it allows a precise estimate of effect size, but data-sets based on procedure retain an important risk of selection bias, and disease based registries are therefore more likely to give a reliable picture of the overall impact of a technique on outcome. This approach can allow demonstration and measurement of the strength of important treatment effect modifiers through multivariate analysis.[19] Continuous data collection allows detection of temporal trends, and tentative inferences about their causes. The collection of a rich data-set allows us to analyse multiple influences together, and to utilize mathematical tools to extract further information from the data-set. Continuous performance monitoring tools such as CUSUM[20] and SPRT can allow earlier warning of problems as 'special cause' variation in sequential outcomes. Modelling techniques can help determine which tools best detect changes in outcome without causing unnecessary alarms. Such techniques, integrated with quality improvement techniques adapted from industrial process control could help to significantly reduce the risks of surgery.

Conclusion

There is an identifiable life history to the development of surgical techniques, and the questions which need to be addressed differ in the different phases of development. There appears to be a particular difficulty in the phase between definition and refinement of a new technique and comparison with best current practice in a randomized trial. Methodology for investigating surgery should be re-designed taking into account the nature of the development process to match methods against the specific problems of each phase.

References

1 McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. *BMJ* 2002;**324**:1448–51
2 Heard E. Rapid-fire improvement with short-cycle kaizen. *Hosp Mater Manage Q* 1999;**20**:15–23
3 Waugh W. *Charnley: the man and the hip.* Berlin: Springer-Verlag; 1990
4 Lilford RJ, Thornton JG, Braunholtz D. Clinical trials and rare diseases: a way out of a conundrum. *BMJ* 1995;**311**:1621–5
5 Chalmers TC, Sacks H. Randomized clinical trials in surgery. *N Engl J Med* 1979;**301**:1182
6 Wilson AP, Gibbons C, Reeves BC, *et al*. Surgical wound infection as a performance indicator: agreement of common definitions of wound infection in 4773 patients. *BMJ* 2004;**329**:720
7 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clinical Chemistry* 2003;**49**:1–6
8 Rubino M, Pragnell MVC. Guidelines for reporting case series of tumours of the colon and rectum. *Tech. Coloproctol* 1999;**3**:93–7
9 Jabs DA. Improving the reporting of clinical case series. *Am J Ophthalmol* 2005;**139**:900–5

10 Haynes RB, Taylor DW, Sackett DL, Thorpe K, Ferguson GG, Barnett HJ. Prevention of functional impairment by endarterectomy for symptomatic high-grade carotid stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators. *JAMA* 1994;**271**:1256–9
11 Macdonald JS, Smalley SR, Benedetti J, *et al*. Chemoradiotherapy after surgery compared with surgery alone for adenocarcinoma of the stomach or gastroesophageal junction. *N Engl J Med* 2001;**345**:725–30
12 den Dulk L, Collette C, van de Velde C, *et al*. Quality of surgery in T3-4 rectal cancer: Involvement of circumferential resection margin not influenced by preoperative treatment. Results from EORTC trial 22921. *Eur J Cancer* 2007;**43**:1821–8
13 Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, Russell IT. Statistical assessment of the learning curves of health technologies. *Health Technol Assess* 2001;**5**:1–79
14 van der Linden W. Pitfalls in randomized surgical trials. *Surgery* 1980;**87**:258–62
15 Novick RJ, Fox SA, Stitt LW, *et al*. Assessing the learning curve in off-pump coronary artery surgery via CUSUM failure analysis. *Ann Thorac Surg* 2002;**73**:S358–S362
16 Degiuli M, Sasako M, Calgaro M, *et al*. Italian Gastric Cancer Study Group. Morbidity and mortality after D1 and D2 gastrectomy for cancer: interim analysis of the Italian Gastric Cancer Study Group (IGCSG) randomised surgical trial. *Eur J Surg Oncol* 2004;**30**:303–8
17 Donovan J, Mills N, Smith M, *et al*. Improving design and conduct of randomised trials by embedding them in qualitative research: ProtecT (prostate testing for cancer and treatment) study. *BMJ* 2002;**325**:766–70
18 Vallabhaneni SR, Harris PL. Lessons learnt from the EUROSTAR registry on endovascular repair of abdominal aortic aneurysm repair. *Eur J Radiol* 2001;**39**:34–41
19 McCulloch P, Ward J, Tekkis PP. ASCOT Group of Surgeons; British Oesophago-Gastric Cancer Group. Mortality and morbidity in gastro-oesophageal cancer surgery: initial results of ASCOT multicentre prospective cohort study. *BMJ* 2003;**327**:1192–7
20 Beiles CB, Morton AP. Cumulative sum control charts for assessing performance in arterial surgery. *ANZ J Surg* 2004;**74**:146–51