Randomised trials in surgery: problems and possible solutions

Peter McCulloch, Irving Taylor, Mitsuru Sasako, Bryony Lovett, Damian Griffin

The quality and quantity of randomised trials of surgical techniques is acknowledged to be limited. According to Peter McCulloch and colleagues, however, some aspects of surgery present special difficulties for randomised trials. In this article they analyse what these difficulties are and propose some solutions for improving the standards of clinical research in surgery

Academic Unit of Surgery, University of Liverpool, Clinical Sciences Centre, University Hospital Aintree, Liverpool L9 7AL
Peter McCulloch *senior lecturer in surgery*

Department of Surgery, Royal Free and University College Medical School, Charles Bell House, London W1W 7EJ
Irving Taylor *professor of surgery*

Gastric Surgery Division, National Cancer Centre Hospital, Tsukiji, 5-1-1 Chuo-Ku, Tokyo, Japan
Mitsuru Sasako *professor of surgery*

Basildon Hospital, Nethermayne, Basildon SS16 5NL
Bryony Lovett *lecturer in surgery*

Nuffield Department of Orthopaedic Surgery, Orthopaedic Centre, Oxford OX3 7LD
Damian Griffin *clinical reader*

Correspondence to: P McCulloch, Academic Unit of Surgery, University of Liverpool, Clinical Sciences Centre, University Hospital Aintree, Long Lane, Liverpool L9 7AL petermcculloch@cs.com

BMJ 2002;324:1448–51

References cited in the text with the prefix "w" are available on bmj.com

The improvement in the quality of clinical research in the past decade is to be welcomed, but it carries its own dangers. Some have extrapolated the advantages of the randomised controlled trial (RCT) into the dogma that it is the only valid method for comparing treatments,[1] ignoring the difficulties that have hampered the use of RCTs in some disciplines. The RCT has theoretical advantages over other study designs, but experimental studies comparing treatment effect estimates in randomised and non-randomised studies have not consistently confirmed this,[2 3 w1-w3] and the superiority of RCTs should not therefore be accepted as axiomatic.

Small, poorly conducted RCTs are more likely to result when RCTs are difficult to conduct, and these may then be misleading because their design affords them unwarranted credibility. Surgery seems to be such an area. Until recently, most studies of operations were retrospective case series, with RCTs accounting for less than 10% of the total.[w4-w6] RCTs declined from 14% of research articles in the *British Journal of Surgery* in 1985 to 5% in 1992.[4 5] Treatments in general surgery are half as likely to be based on RCT evidence as treatments in internal medicine.[6 7] Methodological quality was poor in 56% of RCTs comparing cancer surgery techniques.[8] Only 58% of these studies described satisfactory randomisation, and few significant outcome differences were found, probably because of type II statistical errors.

Why is surgery so deficient? Some of the obstacles militate against all scientific studies, but in view of previous specific criticism,[w7] we focus on randomised trials and try to evaluate the problems and suggest potential solutions.

Obstacles to randomised trials in surgery

Historical, structural, and cultural

History

History did not favour the validation of surgery by RCTs. After the invention of anaesthesia and antiseptic techniques, surgical treatments were rapidly developed for many previously untreatable conditions. Many current operations were therefore introduced well before randomised trials became established in medicine—unlike most modern drugs. Once a treatment is accepted as standard, testing it against placebo becomes difficult. Rarely, treatment benefits are so obvious that a trial would clearly be unethical,[9] but often lack of equipoise (see below) simply prevents studies. This problem applies equally to old drugs—for example, digoxin—which are also difficult to study in RCTs using placebo. For fields such as cardiac surgery, transplantation, orthopaedics, and neurosurgery, however, which have developed rapidly since 1950, surgeons cannot fall back on history to explain the lack of rigour in surgical research.

Summary points

- Research in surgery is disadvantaged by the limited quality and quantity of randomised trials of surgical techniques
- Some aspects of surgery present special difficulties for randomised trials
- The existence and nature of these difficulties needs to be recognised, with strategies developed to overcome them
- A proposed strategy involves the integration of modified randomised trials with prospective audit and quality control studies

Commercial competition and personal prestige

Doctors can be tempted to ignore evidence that threatens their personal interests. Objectivity about procedures central to a surgeon's reputation is difficult, and RCTs may seem threatening. Private sector competition may affect surgeons particularly strongly, and it arguably influenced the introduction of laparoscopic cholecystectomy. A consensus conference in 1994[10] quoted many reports of increased bile duct injuries and only two RCTs.[11 12] The benefits that these showed were not overwhelming against this evidence of possible harm, but further RCTs were declared infeasible because the technique was already so widespread. Surgeons' eagerness to learn the operation seemed related more to commercial concerns than to concern for patients.

Surgeons' equipoise

Other doctors regard surgeons as making up in self confidence for what they lack in patience, a stereotype containing a kernel of truth. Career surgeons are selected for traits that include comfort with making important clinical decisions quickly with incomplete information. This quality, required for decisive action during operations, may make it difficult for them to be consciously uncertain which of two treatments is better. This state of equipoise, however, is a prerequisite for performing RCTs.

Box 1: Problems of performing randomised trials in surgery

- Structural, cultural, and psychological resistance exists to the use of randomisation
- The inherent variability of surgery requires precise definition of interventions and close monitoring of quality
- Surgical learning curves cause difficulty in timing and performing randomised trials of new techniques
- Comparisons of surgical and non-surgical treatments with greatly different risks causes difficulties with patients' equipoise
- Rare conditions and urgent and life threatening situations cause difficulties with recruitment, consent, and randomisation

Lack of funding, infrastructure, and experience of data collection

These are real and major problems for surgical trials.[w8] The difficulty is partly self inflicted as funding bodies are influenced by the poor quality of much previous surgical research.[w9]

Lack of education in clinical epidemiology

Subjectively, surgeons' knowledge of clinical epidemiology remains poor despite relevant publications in surgical journals[w10-w17]: we have no objective evidence that they receive less specific education than other doctors.[13 w15] Surgeons recruit patients for cancer chemotherapy trials[14 w18] but less readily for trials of surgical technique. Whether lack of education can explain this is unclear.

Rare conditions and life threatening and urgent situations

Emergency surgery often occurs outside normal working hours and involves urgent lifesaving treatment, making consent and randomisation difficult. Uncommon conditions are difficult to investigate when accrual of patients takes over two years.[13]

Special technical problems

The learning curve

Some authors suggest that RCTs of new operations should begin with the first patient.[15 w19] Operations, however, are complex procedures, and quality in performance requires frequent repetition over time. Learning curves of similar lengths are reported for disparate operations.[16 17 w20] During the learning curve, errors and adverse outcomes are more likely. Randomising between a familiar and an unfamiliar operation therefore introduces bias against the latter, as observed for gastrectomy.[18] This problem for surgical RCTs has few parallels in drug trials.

Definition

Variations on an operation are common and may influence success rates. When comparing operations, clear definitions are therefore needed of the limits on acceptable technical variation. A standard description may be necessary, proscribing all modifications. If definitions are not precise, the treatments delivered may overlap, whereas in drug trials, treatments are usually simple to define exactly.

Quality control monitoring

The technical quality of operations undoubtedly affects outcome. Poor quality surgery represents failure to deliver the intended treatment, causing a difference between efficacy and effectiveness. Trials then measure deliverability, not efficacy.[w21] Quality control failures may narrow important differences in the surgery received—for example, for gastric cancer[19 20]—and may influence outcomes.[w22 w23] Defining and enforcing minimum quality standards may be difficult for surgical trials.

Development versus research

RCTs consume substantial resources and are therefore not justified for some questions about small modifications to treatments. Surgical technique typically progresses via such modifications, which individually are unlikely to produce detectable benefits, but which collectively may do so. During the historical progression through hand washing via the use of antiseptics to the aseptic surgical environment, the change in morbidity from surgical infection was huge, but the increment with each step was small enough to allow persistent scepticism.[21] Small randomised trials of components of this progression showed no benefit.[22 w24] If a positive RCT were required before adopting each small improvement, most would be rejected, and progress would be slowed. RCTs are appropriate where a clear, clinically important choice exists between contrasting alternatives. For smaller changes, an industrial paradigm may be needed.

Patients' equipoise

Three types of RCT are commonly described as "surgical." Type 1 trials—standard RCTs comparing medical treatments in surgical patients—account for 75% of "surgical trials."[23] Type 2 trials—comparing surgical techniques—pose the problems described above. Type 3 trials—comparing surgical and non-surgical treatments—pose particular difficulties with the equipoise of patients[w25]: patients often reject RCTs because they do not wish their treatment to be decided by chance.[w26] Type 3 trials increase this discomfort because the adverse effects of the options often differ enormously and the surgical option is irreversible. Eighty two per cent of problems preventing type 3 trials are related to patients' equipoise.[13] Examples of choices include aspirin versus carotid endarterectomy to prevent embolic stroke[24] and goserelin versus castration for prostate cancer.[25 w27] Such trials may recruit slowly, or select an unusual subgroup of patients, making them impractical or their results difficult to generalise.[w28]

Blinding

Blinding is particularly difficult in surgical trials, although creative solutions—such as the use of standardised wound dressings—can succeed.[w29] Only a third of surgical trials examined by Solomon et al had adequate blinding of patients and/or surgeons.[23]

Proposed solutions

History—A comprehensive review of the evidence base is needed to indicate areas warranting new trials of old techniques.

Commercial competition and prestige may be less obstructive in a framework of comprehensive continuous performance evaluation (see below).

Surgeons' equipoise, if confirmed, may need to be accommodated by including parallel, non-randomised, preference arms alongside RCTs.

Lack of funding, infrastructure, and experience of data collection require a change to a culture of cooperation rather than competition. This would facilitate the creation of large groups to perform specific trials, thereby attracting funding and developing the infrastructure. This change would require support from bodies responsible for funding clinical research.

Lack of education in clinical epidemiology needs to be investigated and if necessary corrected through the bodies responsible for postgraduate surgical education and training.

Rare conditions and life threatening and urgent situations will always be challenging areas for RCTs, but have been successfully studied in other disciplines.[26 w30] Paediatric oncologists have illustrated the enormous value of cooperation through their success in trials on childhood leukaemia.[27 w31]

The learning curve needs to be recognised and evaluated using appropriate statistical techniques.[28] Trial methodology will need modification—for example, to show completion of the curve before beginning randomisation,[w32] as in two recent trials.[29 30] In theory, patients could also be randomised not to operations but to surgeons, who would perform their operation of preference, although this option remains untested in practice.

Definition of intervention and quality control monitoring—Precisely defined photographic or video evidence and/or pathological specimens could document the nature and quality of the treatment delivered, as in a recent trial of total mesorectal excision in rectal cancer.[31] Norms for pre-trial success rates and complications could provide a basis for defining acceptable quality, making reliable surgical audit data essential for participation in RCTs.

Development v research—Surgeons should adopt industrial quality assessment techniques to evaluate changes in technique where RCTs are inappropriate.[32] The Japanese term "kaizen" defines an evaluative system akin to the classical audit loop.[w33] Sequential approaches such as CUSUM[33] and the "control curve"[32] are also applicable to surgical innovation.

Patients' equipoise in type 3 trials may be helped by decision analysis techniques[w34] and carefully designed composite end points[w35] to reflect the contrasting possible outcomes of trial arms.

Blinding will always be difficult for surgical treatments,[34] but blinded observers should be used routinely for evaluating outcomes.[w36]

Proposed framework for clinical research in surgery

This analysis of the problems shows why current practices are not working. We need a framework that reflects the difficulties of evaluation in surgery.

Audit data collection

The baseline for the scientific study of surgery is routine collection of comprehensive data about practice and outcomes. The culture and organisation necessary for this should permit easy participation in trials, whereas where these are absent, trialists have to develop the trial infrastructure and run it simultaneously. Surgeons need the resources to record a meaningful audit dataset, entailing considerable investment in data acquisition and management resources.

MICHAEL DONNE/SPL

Continuous performance evaluation

Systems for continuous quality control, using instruments such as CUSUM, CRAM or VLAD plots[33 35 36] or control curves[32] should be used for the analysis of technical innovations. Indications of outcome changes from this surveillance should lead to an audit or kaizen assessment, using decision analysis techniques to determine whether an RCT is warranted.[w37] Where it is not, continuing prospective data collection and regular re-evaluation using bayesian analysis[w38] provide the best available data on outcome changes and allow reconsideration of the need for an RCT.

Conduct of RCTs

When RCTs are necessary, they should routinely be preceded by preliminary phase 2S (phase 2 surgical) studies. These would develop satisfactory definition criteria for the procedure, test measures of surgical quality, define suitable end points, estimate the required sample size, and analyse the learning curve of participants. Such studies would reduce the problems of timing surgical RCTs, and randomisation could be introduced early using "tracker" designs if desired.[w39] During randomised data entry, continuous quality control should be linked to preplanned interim analyses by the trial review committee and appropriate stopping rules. Objective validation of quality should evaluate images, pathological specimens, and outcome data against criteria drawn up in the phase 2S study. Parallel preference arms may be used to improve overall power and evaluate generalisability. For type 3 trials, end point design and decision analysis tools to help patients understand their choices may be important.

Other sources of evidence

Historically, the surgical literature is poor in RCTs. Meta-analysis of non-randomised evidence should therefore be used wherever appropriate. Where RCTs are difficult for sound reasons, prospective non-randomised designs that minimise known biases should be considered sympathetically by journals and funding bodies.

Box 2: Suggestions for progress in surgical research

- Detailed prospective "audit" data collection is essential for surgical research
- Continuous quality control techniques should be used to help determine whether randomised trials are appropriate
- Larger randomised trials are needed, requiring better cooperation
- Learning curves and variations in technique and in quality of surgery must be measured and controlled
- Trials should incorporate a non-randomised initial phase to permit these evaluations, determine suitable end points, and allow sample size calculations
- The need for study types other than randomised trials should be recognised

Conclusion

The substantial obstacles to RCTs of surgical techniques should be recognised. Alternative methods of studying operations should be based on comprehensive prospective audit data. Where RCTs are appropriate they require attention to the issues of the learning curve, intervention definition, and quality control; a preliminary non-randomised phase is also recommended.

This work was partly inspired by interactions with members of the Cochrane Non-randomised Studies Methodology Group and by the activities of its surgical subgroup. We thank Laurent Audige and Barney Reeves in particular for their helpful criticisms. The final article is the responsibility of the authors and not of the surgical subgroup.

Funding: None.

Competing interests: PMcC and DG are members of the Cochrane Non-randomised Studies Methodology Group and its surgical subgroup. PMcC is a member of the Centre for Evidence Based Medicine and is paid to facilitate at its Oxford teaching courses once a year.

1 Doll R. Summation of conference. Doing more good than harm: the evaluation of health care interventions. *Ann N Y Acad Sci* 1994;703:313.
2 Benson K, Harz AJ. A comparison of observational studies and randomised controlled trials. *N Engl J Med* 2000;342:1878-86.
3 Concato J, Shah N, Horwitz RI. Randomised controlled trials, observational studies and the hierarchy of research designs. *N Engl J Med* 2000;342:1887-92.
4 Pollock AV. The rise and fall of the random controlled trial in surgery. *Theoretical Surgery* 1989;4:163-70.
5 Pollock AV. Surgical evaluation at the crossroads. *Br J Surg* 1993;80:964-6.
6 Ellis J, Mulligan I, Rowe J, Sackett DL. Inpatient general medicine is evidence based. *Lancet* 1995;364:407-10.
7 Howes N, Chagla L, Thorpe M, McCulloch P. Surgical practice is evidence based. *Br J Surg* 1997;84:1220-3.
8 Lovett B, Sawyer W, Houghton J, Taylor I. Systematic review of the methodological quality of randomized controlled trials of the surgical excision of cancer [abstract]. *Eur J Surg Oncol* 2000;26:840.
9 Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215-8.
10 Neugebauer E, Troidl H, Kum CK, Eypasch E, Miserez M. The EAES consensus development conferences on laparoscopic cholecystectomy, appendectomy and hernia repair. *Surg Endosc* 1995;9:550-63.
11 Barkun JS, Barkun AN, Sampalis JS, Fried G, Taylor B, Wexler MJ, et al. Randomised controlled trial of laparoscopic versus mini-cholecystectomy. The McGill gallstone treatment group. *Lancet* 1992;340:1116-9.
12 McMahon AJ, Russell IT, Baxter JN, Ross S, Anderson JR, Morran CG, et al. Laparoscopic versus mini-laparotomy cholecystectomy: a randomised controlled trial. *Lancet* 1994;343:135-8.
13 Solomon MJ, McLeod RS. Should we be performing more randomized controlled trials evaluating surgical operations? *Surgery* 1995;118:459-67.
14 Comparison of fluorouracil with additional levamisole, higher-dose folinic acid, or both, as adjuvant chemotherapy for colorectal cancer: a randomised trial. QUASAR Collaborative Group. *Lancet* 2000;355:1588-96.
15 Chalmers TC. Randomization of the first patient. *Med Clin North Am* 1975;59:1035-8.
16 Parikh D, Chagla L, Johnson M, Lowe D, McCulloch P. D2 gastrectomy: lessons from a prospective audit of the learning curve. *Br J Surg* 1996;83:1595-9.
17 Testori M, Bartolomei M, Grana C, Mezzetti M, Chinol M Mazzarol G, et al. Sentinel node localization in primary melanoma: learning curve and results. *Melanoma Res* 1999;9:587-93.
18 Bonenkamp JJ, Songun I, Hermans J, Sasako M, Welvaart K, Plukker JTM, et al. Randomised comparison of morbidity and mortality after D1 and D2 dissection for gastric cancer in Dutch patients. *Lancet* 1995;345:745-8.
19 Bonenkamp JJ, Hermans J, Sasako M, van de Velde CJH. Extended lymph node dissection for gastric cancer. *N Engl J Med* 1999;340:908-14.
20 Cuschieri A, Weeden S, Fielding J, Bancewicz J, Craven J, Joypaul V, et al. Patient survival after D1 and D2 resecctions for gastric cancer: long term results of the MRC randomised surgical trial. *Br J Cancer* 1999;79:1522-30.
21 Wangensteen OH, Wangensteen SD. *The rise of surgery*. Minneapolis, MN: University of Minnesota Press, 1978:425-31.
22 Tunevall TG. Postoperative wound infections and surgical face masks: a controlled study. *World J Surg* 1991;15:383-7.
23 Solomon MJ, Laxamana A, Devore L, McLeod RS. Randomized controlled trials in surgery. *Surgery* 1994;115:707-12.
24 Endarterectomy for asymptomatic carotid artery stenosis. Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. *JAMA* 1995;273:1421-8.
25 Vogelzang NJ, Chodak GW, Soloway MS, Block NL, Schellhammer PF, Smith JA Jr, et al. Goserelin versus orchiectomy in the treatment of advanced prostate cancer: final results of a randomized trial. Zoladex Prostate Study Group. *Urology* 1995;46:220-6.
26 Gausche M, Lewis RJ, Stratton SJ, Haynes BE, Gunter CS, Goodrich SM, et al. Effect of out-of-hospital pediatric endotracheal intubation on survival and neurological outcome: a controlled clinical trial. *JAMA* 2000;283:783-90.
27 Nesbit ME, Sather H, Robison LL, Donaldson M, Littman P, Ortega JA, et al. Sanctuary therapy: a randomized trial of 724 children with previously untreated acute lymphoblastic leukemia: a report from Children's Cancer Study Group. *Cancer Res* 1982;42:674-80.
28 Ramsay CR, Grant AM, Wallace SA, Garthwaite PH, Monk AF, Russell IT. Statistical assessment of the learning curves of health technologies. *Health Technology Assess* 2001;5:1-79.
29 Deguili M, Sasako M, Ponti A, Soldati T, Danese F, Calvo F. Morbidity and mortality after D2 gastrectomy for gastric cancer: results of the Italian Gastric Cancer Study Group prospective multicenter surgical study. *J Clin Oncol* 1998;16:1-6.
30 Clarke D, Khonji NI, Mansel RE. Sentinel node biopsy in breast cancer: almanac trial. *World J Surg* 2001;25:819-22.
31 Kapiteijn E, Kranenbarg EK, Steup WH, Taat CW, Rutten HJ, Wiggers T, et al. Total mesorectal excision (TME) with or without preoperative radiotherapy in the treatment of primary rectal cancer. Prospective randomised trial with standard operative and histopathological techniques. Dutch ColoRectal Cancer Group. *Eur J Surg* 1999;165:410-20.
32 Mohammed MA, Cheng KK, Rouse A, Marshall T. Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. *Lancet* 2001;357:463-7.
33 Van Rij AM, McDonald JR, Pettigrew RA, Putterill MJ, Reddy CK, Wright JJ. CUSUM as an aid to early assessment of the surgical trainee. *Br J Surg* 1995;82:1500-3.
34 Van Der Linden W. Pitfalls in randomized surgical trials. *Surgery* 1980;7:258-62.
35 Poloniecki J, Valencia O, Littlejohns P. Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *BMJ* 1998;316:1697-700.
36 Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* 1997;350:1128-30.

Endpiece

The physician's objectives

Declare the past, diagnose the present, foretell the future; practice these acts.
As to diseases, make a habit of two things—to help, or at least do no harm.
The art has there factors, the disease, the patient, the physician.
The physician is the servant of the art. The patient must co-operate with the physician in combating the disease.

Hippocrates, *Epidemics I, xi.*
Translated by WHS Jones.
Cambridge: Harvard University Press, 1995

Submitted by Jeremy Hugh Baron,
honorary professorial lecturer,
Mount Sinai School of Medicine, New York