ANN MED

Annals of Medicine, 2015; 47: 332–340
ISSN 0785-3890 print/ISSN 1365-2060 online
DOI: 10.3109/07853890.2015.1027255

informa
healthcare

ORIGINAL ARTICLE

Benchmarking Controlled Trial—a novel concept covering all observational effectiveness studies

Antti Malmivaara

Centre for Health and Social Economics, National Institute for Health and Welfare, Helsinki, Finland

The Benchmarking Controlled Trial (BCT) is a novel concept which covers all observational studies aiming to assess effectiveness. BCTs provide evidence of the comparative effectiveness between health service providers, and of effectiveness due to particular features of the health and social care systems. BCTs complement randomized controlled trials (RCTs) as the sources of evidence on effectiveness. This paper presents a definition of the BCT; compares the position of BCTs in assessing effectiveness with that of RCTs; presents a checklist for assessing methodological validity of a BCT; and pilot-tests the checklist with BCTs published recently in the leading medical journals.

Key words: benchmarking controlled trial, cost-effectiveness, effectiveness, inequality, real-effectiveness medicine

Key messages

- The Benchmarking Controlled Trial (BCT) is a novel concept which covers all observational studies aiming to assess effectiveness.
- BCTs assess difference in effectiveness between single or a set of intervention(s), between clinical pathways, or between interventions targeting health care system factors with an aim to increase effectiveness.
- Published BCTs have currently several methodological limitations, some of which could be avoided, and others should be acknowledged.
- BCTs support both clinical and policy decisions, and should be given a high priority in research and in improvement activities.

Introduction

The experimental studies, randomized controlled trials (RCTs), provide the least biased information of the efficacy of medical interventions and create the basis for systematic reviews on effectiveness of interventions (1). However, RCTs mostly assess effectiveness of interventions in ideal settings, and they focus on specific interventions rather than considering how effective is the whole clinical pathway (from the first treatment through all interventions during e.g. a 1-year follow-up time)—the latter is crucial for overall effectiveness. Thus there is a need for valid observational data on actual performance in routine settings, particularly as all educational, research, and leadership activities in medicine are intended to advance the health of the general population and care of ordinary patients (2,3).

The first aim of this paper is to assess the need for the new concept of Benchmarking Controlled Trials (BCTs), provide a definition of the BCT, and to present the two main categories (clinical, and health and social care system-related), and the respective subcategories of BCTs. The second aim is to present a checklist for assessing the methodological validity of a BCT and to point out methodological differences between RCTs and BCTs. The third aim is to pilot-test the checklist with BCTs published recently in the leading medical journals.

Methods

The previous international recommendations on how to report observational studies and systematic reviews of them (4,5) provide guidance on studies that investigate associations between exposures and health outcomes and address three types of observational studies: cohort, case-control, and cross-sectional studies. The author's idea was that there is a need for a framework which starts from the study question of effectiveness in observational settings. When the aim is to assess effectiveness of interventions, there are two options: experimental design (randomized controlled trials) or observational design. This paper concentrates on observational designs, and presents a comprehensive framework for them within the novel concept of Benchmarking Controlled Trials (BCTs).

When assessing effectiveness in an observational (real-world) setting, the index and comparator groups must have a priori as similar groups of patients as possible in order to allow adjusting for the potential baseline incomparability. Therefore, the comparisons have to be made between peers treating similar patients and thus *there is always an element of benchmarking involved*. This is the reason for the concept Benchmarking Controlled Trial. In addition, using e.g. a term such as observational controlled trial

Correspondence: Antti Malmivaara, MD, PhD, Chief Physician, Centre for Health and Social Economics, National Institute for Health and Welfare, Mannerheimintie 166, 00270 Helsinki, Finland. E-mail: antti.malmivaara@thl.fi

(Received 18 December 2014; accepted 4 March 2015)

would probably have connotations that do not coincide with the present new idea (6).

Differentiating the two main BCT categories—clinical and health care system determinants for effectiveness—was based on the author's idea that the requirement for baseline comparability for the clinical comparisons is, indeed, equally much needed when studying interventions aimed to make changes in the health care system (and through these changes increase effectiveness of interventions).

The pertinent clinical subcategories were consequently: 1) effectiveness of a particular single or set of interventions during a limited time frame (like surgery, or 3 months' rehabilitation period) and 2) effectiveness of the whole clinical pathway from start (e.g. acute myocardial infarction) through all various health (and social) care interventions (diagnostic, treatment, rehabilitation; primary, secondary, tertiary care) which happen during e.g. a 1-year follow-up time. The health care system intervention subcategories were defined further according to recent literature (Figure 1) (7). For health care system interventions no universally established categories exist, but, regardless of what they are, any change in the health care system aiming to increase effectiveness falls into the category of a BCT.

The checklist for methodological validity issues of BCTs, as well as the appraisal of methodological issues inherent to BCTs, was based on the author's previous work with randomized controlled trials and observational studies (8–13), and with methodological issues in RCTs and observational effectiveness studies, including work within the Cochrane Collaboration Back Review Group (1,9,12,14–16). Previous checklists for observational studies and systematic reviews of them were also utilized (STROBE (4), MOOSE (5)), as well as scientific literature on particular characteristics of observational studies relevant in the assessment of effectiveness of interventions (17).

For piloting the checklist, the 10 most recent BCTs published in the leading medical journals (*New England Journal of Medicine*, *Lancet*, *Journal of American Medical Association*, *British Medical Journal*, and *Annals of Internal Medicine*) were identified through a PubMed search and by the author searching the articles directly from the journals. The search terms were: benchmarking, registries, effectiveness, and name of the journal. All the included articles had to have an observational design, and aim to assess effectiveness of an intervention directed to patients or directed to the health care system. Five articles assessing clinical features and five assessing health care system-related features as determinants of effectiveness from January 2010 to October 2014 were included. Data extraction was rechecked, and errors were corrected by the author to reach the final appraisal.

Results

Definition and categories of the Benchmarking Controlled Trial

There is a clear need for the new concept Benchmarking Controlled Trial (BCT) as there is no previous systematic guidance on methodological issues in planning and reporting an observational effectiveness study (4,5). Furthermore, the idea of the author that, in addition to clinical interventions, any intervention directed to the health care system must be studied in a BCT is a new one. The term benchmarking is accurate because all comparisons have to be between peers and thus include an element of benchmarking. Furthermore, the results of BCTs should be exploited in the effort to increase effectiveness using the comparative data between peers—which is benchmarking (2).

A BCT is defined as an observational study aiming to provide non-biased estimates of comparative differences in outcomes and costs in real-world circumstances due to a single or a set of intervention(s) or throughout the clinical pathway between two or more health service providers for a well-defined group of patients; or an observational study aiming to provide evidence of the comparative effectiveness of the health care system or parts of it among a well-defined group of patients. Data on disadvantaged

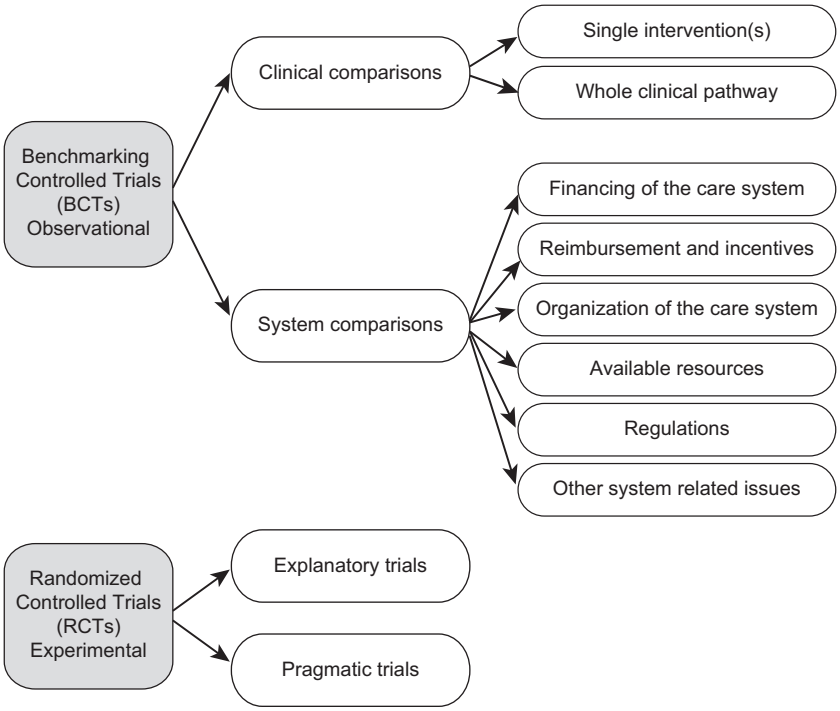

Figure 1. Categories and subcategories of Benchmarking Controlled Trials (BCTs). Randomized Controlled Trials (RCTs) constitute the category of experimental effectiveness studies (shown in the figure only to illustrate that all effectiveness studies are either BCTs or RCTs).

patient groups should be included always when feasible, because their prognosis often differs from that of non-disadvantaged groups. Therefore, inability to control for the differences between disadvantaged and non-disadvantaged populations may lead to biased estimates. Furthermore, data on prevailing inequality will be go unnoticed.

The study question in BCTs should ideally be defined according to the PICO principle (patient, intervention, comparison intervention, and outcome) taking into consideration interventions during the whole clinical pathway. The health care service providers can be individuals, health care units, hospitals, health care districts, or countries.

Features of BCTs in the two main categories (clinical effectiveness and factors related to the health care system) and in their subcategories are presented in Table I. Figure 1 illustrates the categories and subcategories of BCTs covering all observational study designs on effectiveness. In order to illustrate the entity of effectiveness studies, also RCTs are shown in the picture, as well as their subcategories explanatory (ideal circumstances) and pragmatic (ordinary health care circumstances). It must be emphasized that although pragmatic RCTs provide evidence on effectiveness in routine settings, they seldom cover the whole clinical pathway, and generalizability to other settings is limited.

Characteristics of the checklist and methodological issues in BCTs

The main categories and their subcategories of methodological issues in BCTs are presented in Table II. The pilot-testing of the checklist shows also main contents of the 10 studies.

It is noteworthy that there is an overall methodological difference between experimental trials and benchmarking trials. In experimental trials (RCTs) the data collection in each treatment arm is determined in a uniform way, and researchers' obligation is that the conduct of an RCT adheres to the protocol. In observational settings—comparing different service providers—the accrual of the data may not be determined beforehand as strictly as in an RCT, or the quality assurance during data gathering may not be as rigorous. Therefore *validity assessment in BCTs must usually be undertaken separately for all the health care service providers*. Even if there has been uniform instructions on how to collect the data, the success of doing so may differ between the providers.

Another notable methodological issue is that when assessing the comparative effectiveness of a particular intervention or the whole clinical pathway in BCTs, appropriate baseline adjustment is a major challenge. Obtaining proper information of the interventions during the clinical pathway is also most important for two reasons: Firstly, to get further evidence supporting the plausibility of differences in effectiveness estimates, and secondly to have information to be used for improving the treatment processes.

When assessing the effectiveness of interventions targeting the health care system there are four major challenges. Firstly, sufficient data are needed to obtain information indicating whether the health care system factors (e.g. related to an economic incentive) may have led to selection of patients and thus to differences in baseline characteristics. The second challenge is to obtain data of the patients' clinical pathways to know in what degree the intervention targeting the system may have changed

Table I. Categories, subcategories, and characteristics of Benchmarking Controlled Trials (BCTs).

BCT categories and subcategories	Study objective	Design issues	Causal and effect factors	Implications for
1. Clinical comparison (as determinants for effectiveness and efficiency) Subtypes: 1.1. single or set of intervention(s) 1.2. whole clinical pathway	To assess differences in outcome between health care providers (individual, hospital, district, country) who treat similar patients but their way of treating patients (from single intervention to clinical pathway) potentially differs	1. Between-group differences at baseline must be adjusted for 2. Diagnostics and treatment procedures during the clinical pathway should be properly documented (i) to appraise how plausible the differences in outcome are, and (ii) to make decisions on how to improve treatment of patients	Causal factor: differences in single or sets of interventions or in clinical pathways between the comparator arms Effect factor: differences in all relevant outcomes between the comparator arms	Clinicians, policy-makers
2. System comparisons of the health and social care system (as determinants for effectiveness and efficiency) Subtypes: 2.1. related to the financing of the care system (e.g. tax-based or insurance-based system) 2.2. related to the reimbursement and incentives (e.g. fee for service, bonus for quality) 2.3. related to how and by whom the services are organized/provided (e.g. centralized versus decentralized) 2.4. related to the regulations (e.g. on uptake of new technology) 2.5. related to the available resources for health care (e.g. amount of personnel, GDPs of the countries). 2.6. related to other system or structure-related issues (e.g. freedom of choice)	To assess differences in outcome between health care providers due to reasons related to the health and social care system	1. Between-group differences at baseline must be adjusted for 2. Diagnostic and treatment procedures during the clinical pathway should be properly documented and analyzed as mediators of effectiveness	Causal factor: differences in features related to the health care system or part of it Effect factor: differences in all relevant outcomes between the comparator arms: clinical effects and effects on the health care system itself	Policy-makers

the way patients are treated. The third challenge is to adjust for differences in baseline characteristics between the comparators, and analyze differences in treatment processes as mediators of the effects posed by the health care system factors. The fourth challenge is to try to document all the effects the intervention causes to the health care system including unintended unfavorable effects. However, this major challenge of observing a complex system goes beyond the present treatise.

A big difference between benchmarking controlled trials (BCTs) and randomized controlled trials (RCTs) is selection of patients. In the former, patients entering the study in each treatment arm may differ due to selection, while in the latter random allocation to treatment arms (regardless of selection) leads often to comparable treatment groups. To decrease potential for selection bias in BCTs, a two-step procedure is suggested: 1) eligibility criteria should be chosen so that they lead to a homogeneous patient population (e.g. only patients having their first-time acute ischemic stroke will be included) (13,18), and 2) the residual baseline differences have to be statistically adjusted. Instrumental variables may be feasible in some cases to compensate partially for the lack of randomization (19), and the propensity score method may enhance baseline comparability in BCTs (8). Exploitation of a natural experiment may provide an excellent opportunity to increase baseline comparability in BCTs; e.g. in a previous study the health effects of becoming unemployed were studied in a situation when due to nationwide recession suddenly half of construction workers become unemployed, and the allocation to unemployment occurred mainly by chance (20).

Concerns of sufficient clinical information and validity of the data are usually greater in BCTs than in RCTs—particularly if the data for a BCT have been gathered retrospectively, and thus no a-priori protocol has been used. A high number of dropouts is a validity concern for both RCTs and BCTs, as well as the importance of using valid outcome measures. Selective outcome reporting by researchers within a RCT may lead to biased conclusions, but in BCTs selective reporting may occur also during the data collection—often undertaken by the health care providers themselves. There are a number of statistical analysis issues that are characteristic to BCTs (Table II).

Pilot-testing of the checklist

All the 10 articles were from the *New England Journal of Medicine* and *Lancet*, as eligible studies were not found from the other journals (Table II) (21–30).

In the five studies assessing clinical effectiveness, the diagnoses included treatments for selected cancers, non-cardiac surgery, bariatric surgery, rupture of an aortic aneurysm, and acute myocardial infarction. The main outcomes were mortality in four studies, and complication rates in one study. In the five studies assessing effectiveness in relation to health care system-related factors, the indications were more varied than in the clinical effectiveness studies and included a set of surgical indications (two BCTs), a set of indications treated conservatively, intensive care patients, and ambulatory care patients. The determinants for the outcome were the size of the centers providing the service, quality improvement program, presence of a night-time intensivist in the hospital, pay-for-performance, and workload and qualifications of nurses. The main outcomes were mortality in four studies, and health care spending and quality of care in one study.

Concerning methodological issues in the 10 studies several limitations were observed. No study provided a description of patients' clinical path prior to eligibility for the study. No study exploited an opportunity provided by a natural experiment. Valid diagnostic information at baseline was presented by four studies with a clinical research question, and in two of the studies with a health care system-related objective. There were deficiencies in other clinical baseline factors; and factors indicating lifestyle or environment were lacking in all the studies. Information of diagnostics and treatment procedures was lacking altogether in one clinical study and in three studies with focus on the health care system. No study assessed outcomes among disadvantaged patient groups. No study utilized instrumental variables, and only two studies provided power calculations for determining size of the study sample.

Discussion

This paper presents a novel concept, the Benchmarking Controlled Trial (BCT). There are several new ideas involved, particularly 1) that an element of benchmarking is always involved when making observational comparisons in real-world circumstances, and 2) that assessment of effectiveness due to any health care system intervention faces the same methodological challenges as clinical comparisons. Because of the risk for more than one connotation for one concept, a new term of e.g. observational controlled trial did not seem to be appropriate (6).

In those BCTs which pursue evidence on clinical effectiveness, information of baseline patient characteristics, of diagnostic procedures and treatments, and of the outcomes is needed *for the comparisons between providers*. If baseline imbalances between patients treated by different providers can be satisfactorily adjusted for, also comparisons based on treatment outcomes may be justified (31). If feasible, all clinically important patient-relevant outcomes should be documented. However, it is most important to obtain data also of the treatment processes—how well these concord with current scientific evidence (32). Benchmarking controlled trials should aim to assess quality (appropriate interventions), effectiveness and costs of services, as well as issues related to potential inequality in obtaining services shown effective (3).

In BCTs which pursue evidence on effectiveness due to health care system-related factors, there must be a homogeneous target population, and if there are several diagnoses, they should preferentially be differentiated and evidence presented separately for each diagnosis. If there is insufficient data of the diagnoses and related baseline characteristics, the evidence on effectiveness may remain very uncertain.

Previous checklists for advancement of better reporting of observational studies give guidance for studies aiming to assess causal relationship between exposure and outcome. The checklist developed for and described in this paper is intended for supporting planning, conducting, reporting, and peer reviewing manuscripts of observational studies assessing effectiveness of interventions, the BCTs.

The pilot-testing of the checklist using recent articles published in leading medical journals showed a wide variety of methodological strengths and limitations in the original studies. No study provided a description of patients' clinical path before entering the study. Description of baseline characteristics was deficient or even lacking, causing uncertainness in between-group comparability. Information of diagnostics and treatment procedures was scarce. Instrumental variables were not utilized, and power calculations were rare.

Conclusions

The new concept of the BCT provides guidance for studies assessing comparative effectiveness between single or sets

Table II. Methodological characteristics of Benchmarking Controlled Trials (BCTs) in 10 studies published between January 2010 and October 2014 in leading medical journals. Assessment is based solely on each particular paper; if information is not reported, the issue is assessed as unclear. Each characteristic is recorded as yes, partial, unclear, or no; yes indicates that the criterion has been met.

Study characteristics	Coleman et al., Lancet, 8 Jan 2011[a]	Pearse et al., Lancet, 22 Sep 2012	Birkmeyer et al., NEJM, 10 Oct 2013	Karthikesalinam et al., Lancet, 15 Mar 2014
1. Research question and study design	To produce up-to-date survival estimates for selected cancers, to establish whether international differences (Australia, Canada, Denmark, Norway, Sweden, UK) in survival have changed, and to investigate the causes of survival deficits	Describe mortality rates and patterns of critical care resource use for patients undergoing non-cardiac surgery across several European nations	To assess the effect of surgical skill as a determinant for complication rates after bariatric surgery	To compare the in-hospital mortality of patients with rupture of an abdominal aortic aneurysm in England and USA
1.1. clinical or system comparison	Clinical	Clinical	Clinical	Clinical
1.2. subcategory of comparison	Whole clinical pathway	Whole clinical pathway	Single intervention	Whole clinical pathway
1.3. conceptually pertinent and clear	Yes	Yes	Yes	Yes
1.4. natural experiment (allocation to study groups apparently by chance)	No	No	No	No
1.5. operationalized according to the PICO principle (patient; treatment; comparison treatment; outcomes)	No	No	Yes	Yes
2. Selection of patients/population to the study and measures to increase comparability (all studies have individual patient data)				
2.1. population-based cohort, administrative database, or clinical register	Population-based register	Clinical sample	Clinical register	Administrative databases
2.2. prospective or retrospective design	Retrospective	Prospective	Prospective	Unclear
2.3. level of health care provider (e.g. individual, health care center, hospital, district, country)	Country level	Country level	Individual provider	Country level
2.4. description of patients' clinical path before eligible for the study	No	No	NA	No
2.5. description of patients' clinical eligibility criteria	Yes	No	Yes	Yes
2.6. comprehensive patient population of the catchment area	Yes	Yes	No	Yes
2.7. restriction of patients to a particular group in order to increase homogeneity (e.g. first episode ever of ischemic stroke)	No	No	No	No
2.8. use of instrumental variables to compensate for lack of randomization	No	No	No	No
3. Validity and completeness of baseline data + Comparability ensured between groups at baseline (e.g. Validity: Yes; Comparability: No → Yes/No)				
3.1. diagnostics	Yes/Unclear	No/Unclear	Yes/Unclear	Yes/Yes
3.2. other clinically important data relevant to the particular disorder/ disease (e.g. severity)	No/Unclear	No/Unclear	Yes/Unclear	Yes/Unclear
3.3. general health/risk status	No/Unclear	No/Unclear	No/Unclear	No/Unclear
3.4. co-morbid conditions	No/Unclear	No/Unclear	Yes/Unclear	Yes/Yes
3.5. behavioral factors (e.g. on health-related lifestyle)	No/Unclear	No/Unclear	No/Unclear	No/Unclear
3.6. environmental factors (e.g. work conditions)	No/Unclear	No/Unclear	No/Unclear	No/Unclear
3.7. potential inequality (e.g. socio-economic status)	No/Unclear	No/Unclear	No/Unclear	Yes/Yes
3.8. other potential predictors (e.g. genetic factors), confounders, and effect modifiers	No/Unclear	No/Unclear	No/Unclear	Unclear/Unclear
4. Validity and completeness of process data (also unrelated to the disorder in question) throughout the clinical pathway				
4.1. diagnostics	Yes	No	Yes	Yes
4.2. treatment procedures	No	No	Yes	Yes

Chung et al., Lancet, 12 April 2014	Finks et al., NEJM, 2 June 2011	Song et al., NEJM, 9 Aug 2011	Wallace et al., NEJM, 31 May 2012[b]	Sutton et al., NEJM, 8 Nov 2012[a]	Aiken et al., Lancet, 24 May 2014
To compare crude and casemix-standardized 30-day mortality for acute myocardial infarction between UK and Sweden	To evaluate the extent to which decreases in mortality after esophagectomy, pancreatectomy, lung resection, cystectomy, and abdominal aortic aneurysm repair could be associated with a concentration of surgical care in high-volume hospitals	To assess the effect of the Alternative Quality Contract system on health care spending and on measures of the quality of ambulatory care in 2009	To assess the relationship between night-time intensivist physician staffing and mortality among intensive care patients	To analyze the association of a hospital pay-for-performance program with patient mortality among patients with pneumonia, heart failure, or acute myocardial infarction	To assess whether differences in patient-to-nurse workloads and nurses' educational qualifications in nine countries with similar patient discharge data are associated with variation in hospital mortality after common surgical procedures
Clinical Whole clinical pathway	System comparison Related to how and by whom the services are organized / provided	System comparison Related to the reimbursement and incentives	System comparison Related to how and by whom the services are organized / provided, and to the resources available for health care	System comparison Related to the reimbursement and incentives	System comparison Related to the resources available for health care
Yes No	Yes No	Yes No	Yes No	Yes No	Yes No
Yes	Yes	No	No	No	No
Clinical register	Administrative databases	Clinical register	Clinical register	Clinical register	Administrative databases
Retrospective Country level	Retrospective Hospital level	Unclear Provider organization	Retrospective Hospital level	Unclear Hospital level	Retrospective Hospital level
No	No	No	No	No	No
Partial	Yes	No	No	No	No
Unclear	Unclear	No	Unclear	Yes	Unclear
Partial	No	No	No	No	No
No	No	No	No	No	No
Yes/Yes Yes/Yes	No/Unclear No/Unclear	No/Unclear No/Unclear	No/Unclear No/Unclear	Yes/Yes No/Unclear	No/Unclear No/Unclear
Yes/Yes Yes/Yes No/Unclear	No/Unclear Yes/Yes No/Unclear	Yes/Yes No/Unclear No/Unclear	Yes/Yes Yes/Yes No/Unclear	No/Unclear Yes/Yes No/Unclear	No/Unclear Yes/Yes No/Unclear
No/Unclear	No/Unclear	No/Unclear	No/Unclear	No/Unclear	No/Unclear
No/Unclear	Yes/Yes	No/Unclear	No/Unclear	No/Unclear	No/Unclear
Yes/Yes	No/Unclear	No/Unclear	No/Unclear	No/Unclear	No/Unclear
Yes Yes	No No	Yes Yes	Yes Yes	No No	No No

(Continued)

Table II. (Continued)

Study characteristics	Coleman et al., Lancet, 8 Jan 2011[a]	Pearse et al., Lancet, 22 Sep 2012	Birkmeyer et al., NEJM, 10 Oct 2013	Karthikesalinam et al., Lancet, 15 Mar 2014
4.3. rehabilitation	No	No	NA	NA
4.4. hospitalizations and health care visits	No	No	Yes	Yes
4.5. individual behavior (e.g. lifestyle-related to health)	No	No	No	NA
4.6. adherence to treatments	Yes	No	Yes	NA
4.7. characteristics of the clinical pathway	No	No	NA	NA
5. Validity and completeness of outcome data (related to the disorder in question)				
5.1. validity of the outcomes	Yes	Yes	Yes	Yes
5.2. outcomes assessed also among disadvantaged patients	No	No	No	No
5.3. comparability (similarity) of follow-up time points	Yes	Yes	Yes	Yes
5.4. percentage of dropouts during follow-up documented and acceptable (< e.g. 10%)	Yes	Yes	Yes	Yes
5.5. data at each comparator arm free of suggestion of selective outcome reporting	Yes	Yes	Yes	Yes
6. Statistical and data issues				
6.1. description of power calculations and rationale on how the study size was arrived at or post-analysis power calculation	No	Yes	No	No
6.2. documentation of how data were classified and coded (e.g. blinding, multiple raters, inter-rater reliability)	Unclear	Unclear	Yes	Unclear
6.3. measures to increase reliability of data classification and coding (e.g. blinding, multiple raters, inter-rater reliability)	Yes	Unclear	Yes	Unclear
6.4. description of all primary statistical methods, including those used to control for confounding	Yes	Yes	Yes	Yes
6.5. description and use of methods to examine subgroups and interactions	No	Yes	Yes	Yes
6.6. description and use of propensity score or other methods to improve comparability at baseline	No	No	No	No
6.7. adjustment for the characteristic outcomes of each health care provider (e.g. differences in general life expectancy in each country)	Yes	No	NA	No
6.8. incomplete outcome data adequately addressed. If no missing data or appropriate imputation: Yes	Yes	Yes	Yes	Yes
6.9. use of multilevel modeling or survival modeling	Yes	Yes	Yes	Yes
In studies on health care system as determinant of outcome:				
6.10. diagnostic and treatment processes analyzed as the mediators of the effects	No	No	No	No
Study having comparisons also with cohorts in time, i.e. changes in outcomes between follow-up years (each have patients of their own). For this design there are three additional methodological issues:				
7.1. Documentation of changes in patient characteristics over time	No			
7.2. Documentation of changes in treatment practices over time	No			
7.3. Documentation of changes in patient outcomes over time	Yes			

NA = not applicable.
[a]Study having comparisons also with cohorts in time: study characteristics 7.1.–7.3.
[b]Study assessing the effect of one factor related to the organization of the system (e.g. presence of night-time intensivist).

Chung et al., Lancet, 12 April 2014	Finks et al., NEJM, 2 June 2011	Song et al., NEJM, 9 Aug 2011	Wallace et al., NEJM, 31 May 2012[b]	Sutton et al., NEJM, 8 Nov 2012[a]	Aiken et al., Lancet, 24 May 2014
Yes	No	NA	NA	No	No
NA	No	NA	NA	No	No
No	No	NA	NA	No	No
No	No	NA	NA	No	No
No	No	NA	NA	No	No
Yes	Yes	Yes	Yes	Yes	Yes
No	No	No	No	No	No
Yes	Yes	Yes	No	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes
No	No	Yes	No	No	No
Yes	Unclear	Unclear	Yes	Unclear	Unclear
Partial	Unclear	Unclear	Yes	Unclear	Unclear
Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes
Yes	No	Yes	No	No	No
No	No	No	No	No	No
Yes	Yes	Yes	Yes	Yes	Yes
Unclear	Yes	Yes	No	No	Yes
No	Yes	No	No	No	No
	No				
	No				
	Yes				

of interventions, between clinical pathways, or between health care systems or factors related to the system. Benchmarking controlled trials cover the whole area of observational effectiveness research.

A checklist for assessing the methodological validity of BCTs has here been subjected to preliminary pilot-testing, but should be properly validated. However, the checklist can readily be used in planning, conducting, reporting, and appraising BCTs.

Current BCTs seem to have several methodological limitations, some of which could be avoided in planning and conducting phases of the studies, and others should be acknowledged in discussion.

Benchmarking controlled trials—supporting both clinical and policy decisions—should be given a high priority in research, and their results should be used in improvement activities provided they have sufficient methodological rigor and generalizability. The proposed methodology is suggested also for non-scientific quality improvement and benchmarking undertakings.

Funding: No outside funding.

Declaration of interest: The author declares no support from any organization for the submitted work; no financial relationships with any organization that might have an interest in the submitted work; and no other relationships or activities that could appear to have influenced the submitted work.

References

1. Furlan AD, Pennick V, Bombardier C, van Tulder M; Editorial Board, Cochrane Back Review Group. 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. Spine (Phila Pa 1976). 2009;34:1929–41.
2. Malmivaara A. Real-effectiveness medicine-pursuing the best effectiveness in the ordinary care of patients. Ann Med. 2013;452:103–6.
3. Malmivaara A. On decreasing inequality in health care in a cost-effective way. BMC Health Serv Res. 2014;14:79.
4. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg. 2014;12:1495–9.
5. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000;283:2008–12.
6. Larsen KR, Voronovich ZA, Cook PF, Pedro LW. Addicted to constructs: science in reverse? Addiction. 2013;108:1532–3.
7. Klazinga N, Li L. Comparing health services outcomes. In: Papanicolas I, Smith P, editors. Health system performance comparison. An agenda for policy, information and research. 1st ed. Maidenhead, England: Open University Press, McGraw-Hill Education; 2013. p. 157–82.
8. Häkkinen U, Malmivaara A. The PERFECT project: measuring performance of health care episodes. Ann Med. 2011;43(Suppl 1):S1–3.
9. Sihvonen R, Paavola M, Malmivaara A, Itala A, Joukainen A, Nurmi H, et al. Arthroscopic partial meniscectomy versus sham surgery for a degenerative meniscal tear. N Engl J Med. 2014;370:1260–1.
10. Viljanen M, Malmivaara A, Uitti J, Rinne M, Palmroos P, Laippala P. Effectiveness of dynamic muscle training, relaxation training, or ordinary activity for chronic neck pain: randomised controlled trial. BMJ. 2003;327:475.
11. Torkki M, Malmivaara A, Seitsalo S, Hoikka V, Laippala P, Paavolainen P. Surgery vs orthosis vs watchful waiting for hallux valgus: a randomized controlled trial. JAMA. 2001;285:2474–80.
12. Malmivaara A, Hakkinen U, Aro T, Heinrichs ML, Koskenniemi L, Kuosma E, et al. The treatment of acute low back pain—bed rest, exercises, or ordinary activity? N Engl J Med. 1995;332:351–5.
13. Malmivaara A, Meretoja A, Peltola M, Numerato D, Heijink R, Engelfriet P, et al. Comparing ischaemic stroke in six European countries. The Euro-HOPE register study. Eur J Neurol. 2015;22:284–91.
14. Croft P, Malmivaara A, van Tulder M. The pros and cons of evidence-based medicine. Spine. 2011;36:E1121–5.
15. Sihvonen R, Paavola M, Malmivaara A, Jarvinen TL. Finnish Degenerative Meniscal Lesion Study (FIDELITY): a protocol for a randomised, placebo surgery controlled trial on the efficacy of arthroscopic partial meniscectomy for patients with degenerative meniscus injury with a novel 'RCT within-a-cohort' study design. BMJ Open. 2013;33.
16. Malmivaara A, Koes BW, Bouter LM, van Tulder MW. Applicability and clinical relevance of results in randomized controlled trials: the Cochrane review on exercise therapy for low back pain as an example. Spine (Phila Pa 1976). 2006;31:1405–9.
17. Vandenbroucke J. When are observational studies as credible as randomised trials? Lancet. 2004;363:1728–31.
18. Peltola M, Juntunen M, Häkkinen U, Rosenqvist G, Seppälä TT, Sund R. A methodological approach for register-based evaluation of cost and outcomes in health care. Ann Med. 2011;43:S4–13.
19. Vandenbroucke JP. Observational research, randomised trials, and two views of medical science. PLoS Med. 2008;53:e67.
20. Leino-Arjas P, Liira J, Mutanen P, Malmivaara A, Matikainen E. Predictors and consequences of unemployment among construction workers: prospective cohort study. BMJ. 1999;319:600–5.
21. Coleman MP, Forman D, Bryant H, Butler J, Rachet B, Maringe C, et al. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. Lancet. 2011;377:127–38.
22. Pearse R, Moreno RP, Bauer P, Pelosi P, Metnitz P, Spies C, et al. Mortality after surgery in Europe: a 7 day cohort study. Lancet. 2012;380:1059–1065.
23. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med. 2013;369:1434–42.
24. Karthikesalingam A, Holt PJ, Vidal-Diez A, Ozdemir BA, Poloniecki JD, Hinchliffe RJ, et al. Mortality from ruptured abdominal aortic aneurysms: clinical lessons from a comparison of outcomes in England and the USA. Lancet. 2014;383:963–9.
25. Chung SC, Gedeborg R, Nicholas O, James S, Jeppsson A, Wolfe C, et al. Acute myocardial infarction: a comparison of short-term survival in national outcome registries in Sweden and the UK. Lancet. 2014;383:1305–12.
26. Finks JF, Osborne NH, Birkmeyer JD. Trends in hospital volume and operative mortality for high-risk surgery. N Engl J Med. 2011;364:2128–37.
27. Song Z, Safran DG, Landon BE, He Y, Ellis RP, Mechanic RE, et al. Health care spending and quality in year 1 of the alternative quality contract. N Engl J Med. 2011;365:909–18.
28. Wallace DJ, Angus DC, Barnato AE, Kramer AA, Kahn JM. Nighttime intensivist staffing and mortality among critically ill patients. N Engl J Med. 2012;366:2093–101.
29. Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. N Engl J Med. 2012;367:1821–8.
30. Aiken LH, Sloane DM, Bruyneel L, Van den Heede K, Griffiths P, Busse R, et al. Nurse staffing and education and hospital mortality in nine European countries: a retrospective observational study. Lancet. 2014;383:1824–30.
31. Hakkinen U, Iversen T, Peltola M, Seppala TT, Malmivaara A, Belicza E, et al. Health care performance comparison using a disease-based approach: the EuroHOPE project. Health Policy. 2013;112-2:100–9.
32. Hermans MP, Elisaf M, Michel G, Muls E, Nobels F, Vandenberghe H, et al. Benchmarking is associated with improved quality of care in type 2 diabetes: the OPTIMISE randomized, controlled trial. Diabetes Care. 2013;36:3388–95.