Journal of Clinical Epidemiology 107 (2019) 36−41

Journal of Clinical Epidemiology

REVIEW

Generalizability of findings from randomized controlled trials is limited in the leading general medical journals

Antti Malmivaara*

Chief Physician, Centre for Health and Social Economics, National Institute for Health and Welfare, Mannerheimintie 166, 00270 Helsinki, Finland

Accepted 14 November 2018; Published online 17 November 2018

Abstract

Objectives: To document reporting of study characteristics of randomized controlled trials (RCTs) in the four leading general medical journals and to appraise the generalizability of the evidence.

Study Design and Setting: All RCTs in BMJ, JAMA, Lancet, and NEJM from January 1, 2017 to September 30, 2017 were searched by hand, and data were extracted according to the benchmarking method.

Results: Hundred sixty-one RCTs were found; 67% assessed pharmacological therapy. The percentages of adequate documentation were patients' path before randomization 3% to 33% of trials; characteristics of the health care settings 0% to 75%; at least two comorbid conditions 25% to 50%; at least one measure was reported of functioning 42% to 54%, of behavioral factors 25% to 58%, of environmental factors 3% to 25%, and of inequity-related factors 28% to 68%; cointerventions 6% to 25%; and reasons for dropping out of follow-up 39% to 100%.

Conclusion: Almost all RCTs showed deficiencies in description of patient selection and study setting and in reporting of patient characteristics related to functioning, comorbidities, and to behavioral, environmental, and inequity factors. The findings indicate that generalizability of this evidence may be limited. The benchmarking method can be used for planning and appraisal of clinical trials and systematic reviews.

Keywords: Randomized controlled trial; Generalizability; Risk of bias; Benchmarking method; Medical journals; Systematic review

1. Introduction

There is literature on how to describe the essential patient, intervention, control intervention, outcome (PICO) elements in randomized controlled trials (RCTs) and on how to assess risk of bias and generalizability of evidence in RCTs [1−4]. However, comprehensive reporting of other patient-related characteristics and those related to the health care system are not usually recommended for the assessments of RCTs but are considered essential in the assessment of observational effectiveness studies, the benchmarking controlled trials (BCTs) [5]. The benchmarking method (BM) is based on BCTs and comprises five main categories: patient selection, baseline characteristics of patients, interventions and cointerventions, outcomes, and statistical issues. There are several subcategories under each of the main five categories [6−11] (Table 1).

The aim of the present article was to assess comprehensiveness of reporting of patient population, interventions, control interventions, and outcomes using the novel method (BM) for the recent RCTs reported in the four leading general medical journals and to appraise the generalizability of the evidence.

2. Methods

The BM is based on supplementing current methods intended for the assessment of validity of RCTs and generalizability of their evidence with those intended for observational effectiveness studies, BCTs [5]. In

Funding: No outside funding.

Conflicts of interest statement: 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.

Author contribution: The author has developed the idea for the article and written it solely.

* Corresponding author. Tel.: +358 40 554 5435; fax: +358 29 524 6111.

E-mail address: antti.malmivaara@thl.fi.

https://doi.org/10.1016/j.jclinepi.2018.11.014

What is new?

Key findings

- A detailed description of patient selection and study setting, as well as patient characteristics related to functioning, comorbidities, and behavioral, environmental, and equity was deficient in almost all of the randomized controlled trials (RCTs) published in the leading general medical journals.

What this adds to what was known?

- The findings indicate that generalizability of evidence from the leading general medical journals RCTs may be limited.

What is the implication and what should change now?

- Better reporting of patient characteristics beyond the biomedical is needed for generalizable evidence. The benchmarking method can be used for planning of clinical trials and systematic reviews and for assessing validity and generalizability of evidence on effectiveness of interventions.

observational effectiveness studies, the lack of randomization risks baseline comparability between treatment groups and necessitates a very detailed description of patient characteristics. These characteristics are also needed when assessing generalizability of evidence [5,12,13]. The BM is based on five categories (selection, baseline characteristics, process factors, outcome measures, and statistical issues) and several subcategories [5,13] (Table 1). A modification of the BM was used for this study. Statistical assessments were not undertaken because of the diversity in the substance and methods of the 161 RCTs.

Literature was hand-searched to find all RCTs published in BMJ, JAMA, Lancet, and NEJM from January 1, 2017 to September 30, 2017 by inspecting all the weekly issues. Completeness of the search was ensured by hand searching again all the issues of the four journals. In addition, PubMed database was sought using the following key words: RCT and name of each journal to ensure completeness of hand search. No further articles were found based on the PubMed search.

The descriptive information was extracted by the author concerning selection of patients; completeness and validity of data of baseline characteristics, of treatment processes, and of outcome; and statistical analysis. The accuracy of the extracted data was checked. The information was gathered uniformly for the four journals from the main articles without examining possible supplementary material.

3. Results

Altogether 161 RCTs fulfilling the inclusion criteria were found, four in BMJ, 50 in JAMA, 71 in Lancet, and 36 in NEJM. Of the single country trials, 79% (72 of 91) were conducted either in Europe or North America (USA or Canada) (Table 2). Share of multinational trials was 43% and was highest in NEJM (69%). Most of the trials (68%) recruited from over 10 centers. Considering all comparisons, pharmacological interventions' share was 67% (108/161) of the trials, other conservative treatments' 27% (43/161), surgical interventions' 9% (15/161), and rehabilitation interventions' 4% (7/161).

The assessment of reporting of patient, intervention, and outcome characteristics by the BM is shown in

Table 1. The categories and subcategories in the benchmarking method

1. Selection of patients/population to the study
1.1 Description of patients' clinical eligibility criteria
1.2. Description of patients' clinical path before being eligible for the study
1.3. Preintervention therapy
1.4. Comprehensiveness of patient population of the catchment area
1.5. Place and time of recruitment. Number of patients per recruiting unit per year
1.6. Number of patients declining participation
2. Validity and completeness of baseline data
2.1. Number of patients
2.2. Clinically important data relevant to the particular disorder/disease (e.g., age, gender, severity by outcome variables)
2.3. General health/risk status
2.4. Comorbid conditions
2.5. Behavioral factors (e.g., on health-related lifestyle)
2.6. Environmental factors (e.g., work conditions)
2.7. Inequality (e.g., socioeconomic status)
2.8. Other potential predictors (e.g., genetic factors), confounders, and effect modifiers
3. Validity and completeness of process data
3.1. Content of the index treatment
3.2. Content of the control intervention
3.3. Staff competence
3.4. Health care system features (e.g., resources, clinical paths)
3.5. Adherence to index treatments
3.6. Adherence to comparison treatments
3.7. Use of other health care services
4. Validity and completeness of outcome data
4.1. Primary and secondary outcomes
4.2. Percentage of and reasons for dropping out of follow-up
5. Statistical analysis
5.1. Description of power calculations
5.2. Description and appropriateness of all primary and secondary statistical analyses

Table 2. Baseline characteristics (*N*) of randomized controlled trials published in the BMJ, JAMA, Lancet, and NEJM from January 1, 2017 to September 30, 2017

Journal (number of RCTs)	BMJ (*N* = 4)	JAMA (*N* = 50)	Lancet (*N* = 71)	NEJM (*N* = 36)
Study characteristics ↓				
Single country trials (*N*)				
Africa	0	0	2	0
Asia	0	4	6	1
Australia	1	1	4	0
Central and South America	0	0	0	0
Europe	3	12	17	4
North America (US, Canada)	0	23	7	6
Multinational trials	0	10	35	25
Number of centers (*N*)				
1–10	3	20	24	1
11–50	1	17	20	11
51–100	0	4	10	8
101–	0	6	14	9
Not applicable, unclear	0	3	3	7
Type of intervention (*N*)				
Pharmacological	2	26	44	27
Conservative nonpharmacological	2	16	14	2
Surgical	0	2	6	3
Rehabilitation	0	3	2	0
Conservative vs. pharmacological	0	2	3	1 0
Surgical vs. pharmacological	0	0	0	3
Conservative vs. rehabilitation	0	0	2	0
Conservative vs. surgical	0	1	0	0

Table 3. Because of only four trials published in BMJ, mostly the findings in the three other journals are considered here.

Patients' path before assessment of eligibility was described for 33% of JAMA trials and for 7% and 3% of the Lancet and NEJM trials, respectively. Reporting of reasons for exclusions was reported in 82% of JAMA articles, in 65% of Lancet articles, and in 25% of NEJM articles. Share of eligible patients declining participation was 72%, 66%, and 22%, in the three journals, respectively. Characteristics of health care settings were described in 75%, 60%, 8%, and 0% of BMJ, JAMA, Lancet, and NEJM articles, respectively.

Demographic- and disorder-specific baseline characteristics were described in all BMJ and JAMA articles, and in 97% and 94% of Lancet and NEJM articles, respectively. Functioning of patients was described using at least one characteristic in around half of the articles. All three categories of functioning were described in 6% of the JAMA and Lancet articles, but in none of the articles published in BMJ and NEJM. At least two comorbid conditions or a comorbidity index was reported in 25%, 50%, 26%, and 47% of BMJ, JAMA, Lancet, and NEJM articles, respectively. At least one behavioral factor was described from 25% to 58% of the journals, but three or more factors (smoking, alcohol/substance consumption, exercise, obesity) were reported only in 0%, 2%, 3%, and 7% of BMJ, JAMA, Lancet, and NEJM articles, respectively. At least one environmental factor (work conditions, living conditions, marital status) was reported in between 3% and 22%, but all three factors were reported only in one of the four BMJ articles and in none of the other journals. At least one factor related to inequity (socioeconomic status, education, deprivation, ethnicity) was reported in 50% in BMJ, in 68% in JAMA, in 52% in Lancet, and in 28% in NEJM. Lancet was the only journal reporting all three categories of potential inequity but only in 3% of the articles.

Index and control interventions were completed according to the protocol from 75% to 88% of the articles in the four journals. Crossover between the treatment arms was from 2% to 6% in the three journals (not considering BMJ). Cointerventions (use of other health services besides the experimental interventions) were reported in 25%, 10%, 10%, and 6% of BMJ, JAMA, Lancet, and NEJM articles, respectively. Data on patients' follow-up percentages were available in all BMJ and JAMA articles, and in 90% and 72% of the Lancet and NEJM articles, respectively. Reasons for dropping out were in most cases reported, except in NEJM (39% of articles).

Table 3. Appropriate reporting (%) on generalizability of evidence in single subject randomized controlled trials published in the BMJ, JAMA, Lancet, and NEJM from the January 1, 2017 to the September 30, 2017

Journal (number of RCTs)	BMJ (*N* = 4)	JAMA (*N* = 50)	Lancet (*N* = 71)	NEJM (*N* = 36)
Study characteristics ↓				
1. Selection of patients; health care system features				
1.1. Description of patients' path before assessment of eligibility (%)	25	33.3	7.0	3
1.2. Reporting of reasons for exclusions before randomization (%)	100	82	65	25
1.3. Percentage of eligible patients declining participation documented (%)	100	72	66	22
1.4. Description of characteristics of all the health care settings where the data were collected (%)	75	60	8	0
2. Baseline characteristics of patients				
2.1. Demographic- and disorder-specific clinical data (%)	100	100	97	94.4
2.2. Functioning (disease specific, or generic, and health-related quality of life (% of at least 1; at least 2; all three items described)	50 0 0	54 32 6	49 15 6	41.6 2.8 0
2.3. Comorbidity, at least two comorbid conditions reported or a comorbidity index (%)	25	50	26	47.2
2.4. Behavioral factors (smoking, alcohol/substance consumption, or exercise reported); and obesity. In children: parents data. (% of 1; 2;3; 4 items described)	25 0 0 0	50 22 2 0	58 15 3 1	47.1 16 7 0
2.5. Environmental factors (work or living conditions; marital status). In children: parents data. (% of 1; 2; 3 items described)	25 25 25	14 4 0	11 6 0	2.8 0 0
2.6. Potential inequity (socioeconomic status, education, deprivation, ethnicity). In children: data from parents. (% of 1; 2–3; 4 items described)	50 0 0	68 14 0	52 6 3	27.8 0 0
3. Interventions				
3.1. Completed index intervention(s) according to protocol among all recruited or adherence rate (%); data available (%)	75 75	87.5 94	87 84.5	83.7 55.6
3.2. Completed control intervention according to protocol among all recruited or adherence rate (%); data available (%)	75 75	86.5 94	84 84.5	82.6 50
3.3. Crossover to index intervention more than 5% (%)[a]; data available (%)	75 75	4.3 100	3 87.3	2.8 50
3.4. Crossover to control intervention more than 5% (%)[a]; data available (%)	75 75	2.1 100	6 88.3	2.8 44.4
3.5. Cointerventions (use of other health services) reported within each intervention arm (%)	25	10.0	10	5.6
4. Follow-up				
4.1. Follow-up percentage (of those randomized) for the primary outcome at primary follow-up time (mean); data available (%)	75 100	89.6 100	88.1 90.1	87.7 72.2
4.2. Reasons for dropping out/withdrawal reported in each group or no dropouts (%); data available (%)	75 100	100.0 100	97.2 100	38.9 94.4

[a] Excluding crossover trials.

4. Discussion

This article utilizes methodology of observational effectiveness studies, the benchmarking controlled trials [5,13], as well as earlier studies and recommendations on how to describe the PICO in RCTs, and how to assess validity and generalizability of findings [1,2,4]. The BM proposes three main issues (domain, cause, and effect) in the design, conduct, and appraisal of RCTs (Fig. 1).

According to the BM, generalizability of evidence was hampered by several reasons. Regarding the study domain, there were shortcomings in the description of patient selection (patients' path to trial, share of those declining

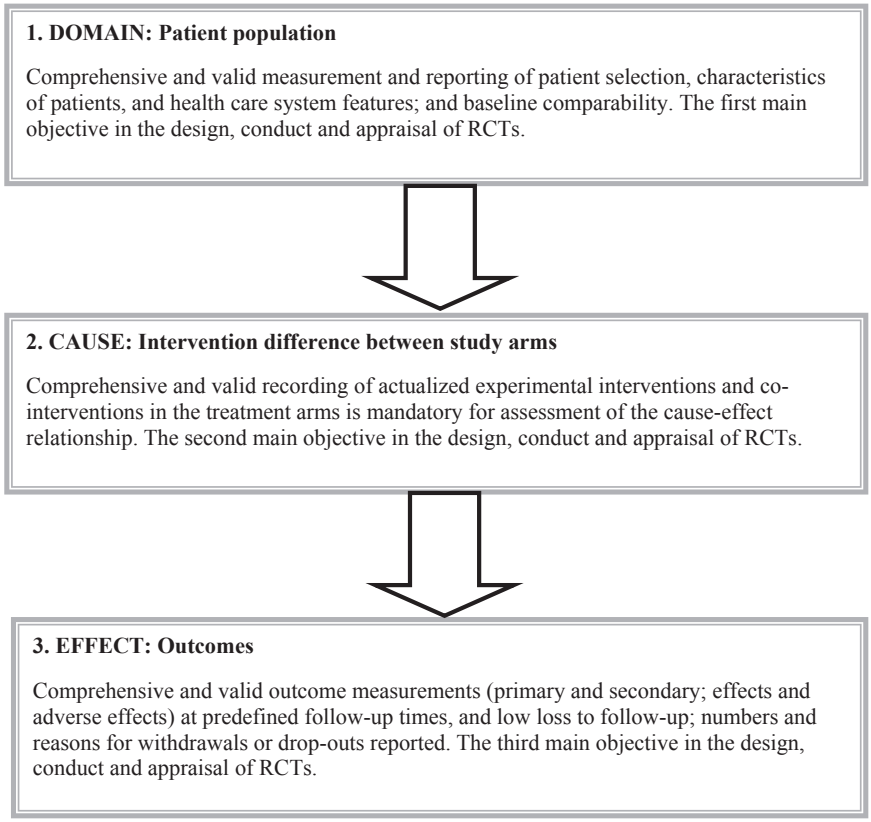

Fig. 1. The three main issues (domain, cause, effect) in the design, conduct, and appraisal of RCTs.

participation and reporting of reasons for declining), health care system features, and reporting of patients' functioning, comorbidities, and behavioral, environmental and equity related factors. There were notable differences in reporting between the four journals. The BM indicates that the study characteristics needed for the assessment of generalizability of evidence should be described more comprehensibly than hitherto.

A limitation of the study is that the data were extracted solely from the main text and tables; supplements and appendices were not considered. The purpose was to assess the generalizability of evidence of each article based on material readily available for the readers. There is evidence indicating that clinicians read mainly the abstracts of the articles, and only in about one-third of the cases look at the main texts of the articles [14]. A limitation is also that it was not possible in a context-specific manner to examine whether the patient characteristics in each of the 161 RCTs actually covered the spectrum of patients needed for making conclusions on generalizability. Another limitation may be that the treatment-modifying effect of patient characteristics such as functioning, comorbidities, and behavioral-, environmental-, and equity-related factors may be considered small for some biologically specific study questions. However, there is evidence showing that these patient characteristics may have a direct therapeutic effect on outcome or may modify the treatment effect [6–8,10,11,15–17]. Furthermore, without documenting these characteristics—even in cases where the effects are assumed to be minor—their importance as modifiers of treatment effects will remain unknown. Only one person, the author, performed the data extraction, which is an obvious limitation, although the author checked the accuracy of the extracted data. The literature search was repeated to ensure accuracy. The lack of statistical assessment was because of the huge diversity in the substance and methods of the 161 trials; a comprehensive analysis and collaboration with statisticians specialized on RCTs would have been needed to appraise these issues.

5. Conclusions

When assessed with the BM, the RCTs published in the leading general medical journals show deficiencies hampering assessment of generalizability of evidence. The findings indicate that generalizability of the evidence from these RCTs may be limited. A detailed description of patient selection and study setting, as well as the patients characteristics (demographic and disorder specific, functioning, comorbidities, and behavioral-, environmental-, and equity-related factors), of all interventions, and of outcome assessments are needed to ensure generalizability

of evidence from future RCTs. The BM is suggested for planning clinical trials and for assessing validity and generalizability of evidence on effectiveness of interventions. The systematic reviews could also use BM for providing a detailed description of study characteristics; for assessing appropriateness of a meta-analysis (based on clinical homogeneity); and finally, for assessment of generalizability of evidence.

References

[1] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 2009;151:W65−94.

[2] 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.

[3] Furlan AD, Malmivaara A, Chou R, Maher CG, Deyo RA, Schoene M, et al. 2015 Updated method guideline for systematic reviews in the Cochrane back and neck group. Spine (Phila Pa 1976) 2015;40:1660−73.

[4] Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.

[5] Malmivaara A. Benchmarking controlled trial-a novel concept covering all observational effectiveness studies. Ann Med 2015;47:332−40.

[6] Stringhini S, Carmeli C, Jokela M, Avendano M, Muennig P, Guida F, et al. Socioeconomic status and the 25 x 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1.7 million men and women. Lancet 2017;389:1229−37.

[7] Marmot M, Allen J, Bell R, Goldblatt P. Building of the global movement for health equity: from Santiago to Rio and beyond. Lancet 2012;379:181−8.

[8] Ngandu T, Lehtisalo J, Solomon A, Levalahti E, Ahtiluoto S, Antikainen R, et al. A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. Lancet 2015;385:2255−63.

[9] Olazaran J, Reisberg B, Clare L, Cruz I, Pena-Casanova J, Del Ser T, et al. Nonpharmacological therapies in Alzheimer's disease: a systematic review of efficacy. Dement Geriatr Cogn Disord 2010;30:161−78.

[10] GBD 2015 Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016;388:1659−724.

[11] Makaroun LK, Brown RT, Diaz-Ramirez LG, Ahalt C, Boscardin WJ, Lang-Brown S, et al. Wealth-associated disparities in death and disability in the United States and England. JAMA Intern Med 2017;177:1745−53.

[12] Malmivaara A. System impact research - increasing public health and health care system performance. Ann Med 2016;48:211−5.

[13] Malmivaara A. Assessing validity of observational intervention studies - the benchmarking controlled trials. Ann Med 2016;48:440−3.

[14] Saint S, Christakis DA, Saha S, Elmore JG, Welsh DE, Baker P, et al. Journal reading habits of internists. J Gen Intern Med 2000;15:881−4.

[15] Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med 2001;344:1343−50.

[16] Mackenbach JP, Kulhanova I, Artnik B, Bopp M, Borrell C, Clemens T, et al. Changes in mortality inequalities over two decades: register based study of European countries. BMJ 2016;353:i1732.

[17] Corraini P, Olsen M, Pedersen L, Dekkers OM, Vandenbroucke JP. Effect modification, interaction and mediation: an overview of theoretical insights for clinical investigators. Clin Epidemiol 2017;9:331−8.