J Rehabil Med 2020; 52: jrm00031

REVIEW ARTICLE

GENERALIZABILITY OF FINDINGS FROM SYSTEMATIC REVIEWS AND META-ANALYSES IN THE LEADING GENERAL MEDICAL JOURNALS

Antti MALMIVAARA, MD, PhD
From the Centre for Health and Social Economics, National Institute for Health and Welfare, Helsinki, Finland

Objective: **To assess how items relevant for the assessment of the generalizability of findings from randomized controlled trials were recorded in systematic reviews published in leading general medical journals.**

Methods: **All systematic reviews and meta-analyses published in the Annals of Internal Medicine, *BMJ*, *JAMA (The Journal of the American Medical Association)* and *Lancet* from 1 January 2016 to 28 February 2019 were searched via PubMed. Reporting of the characteristics of randomized controlled trials in the systematic reviews was documented by the benchmarking method.**

Results: **A total of 115 systematic reviews were found. Of these, 71% included pharmacological interventions, 35% included other conservative treatments, 13% included surgical interventions, and 0% included rehabilitation interventions. None of the systematic reviews assessed patient selection, 35% reported disorder-specific clinical features, 25 % reported comorbid conditions, and 21% reported patients' behavioural factors in randomized controlled trials. Functioning, environmental factors and inequity-related factors were recorded in 3%, 0% and 9%, respectively, of the systematic reviews; and adherence to interventions, crossovers, and co-interventions in 7%, 0% and 2%, respectively; follow-up percentages in 8%; and adequacy of statistical analyses in 3%.**

Conclusion: **In all systematic reviews the recording of characteristics of patients, adherence to interventions, follow-up, and statistical analyses in the RCTs was insufficient. The data did not allow assessment of the clinical homogeneity of the randomized controlled trials, or provide justification for meta-analysis, or allow generalizability of the findings.**

Key words: systematic review and meta-analysis; generalizability; risk of bias; benchmarking method; medical journal; systematic review.

Accepted Feb 14, 2020; Epub ahead of print Feb 27, 2020

J Rehabil Med 2020; 52: jrm00031

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

LAY ABSTRACT

In a systematic review it is important that all characteristics of randomized controlled trials are reported, so that clinicians can determine to which patients the results of a systematic review can be applied. This study assessed how comprehensively these characteristics of randomized controlled trials were recorded in the systematic reviews published in leading general medical journals. A search of the literature found a total of 115 systematic reviews. Of these, 71% were on pharmacological interventions, 35% were on other conservative treatments, 13% were on surgical interventions, and 0% were on rehabilitation interventions. None of the systematic reviews assessed how patients were selected to the study; 35% reported relevant clinical features; 25% comorbid conditions; and 21% patients' behavioural factors. Functioning, environmental factors, inequity-related factors; how interventions were carried out; how well the patients were followed-up; and the adequacy of statistical analyses were reported in only 0–9% of the systematic reviews. In conclusion, the reporting of study characteristics in the systematic reviews does not make it possible to assess how similar the different studies had been, or to which patients these study findings could be generalized. In future, randomized controlled trials should be described better in systematic reviews. Further studies are needed on this subject.

Regarding the scientific literature, there are guidelines for how to comprehensively describe the essential PICO (patient, intervention, control intervention/comparator, outcome) elements of randomized controlled trials (RCTs) in systematic reviews (SRs) and meta-analyses, and how to assess the risk of bias and evaluate the generalizability of evidence of RCTs in SRs and meta-analyses (1–4). In addition to a description of characteristics of the main diagnosis, it is recommended that SRs also include comprehensive reporting of other patient characteristics, including equity-related factors, and those related to the healthcare system (1, 5, 6). Even more comprehensive documentation is essential in the assessment of studies on observational effectiveness, the benchmarking controlled trials (BCTs) (7). The benchmarking method (BM) used in BCTs comprises 5 main categories and several subcategories related to PICO and statistical issues (8–13). The conceptualization and operationalization of the BM has been based on the CONSORT (Consolidated Standards of Reporting Trials) statement for RCTs, and on scientific studies on methodology, as well as the relevance of items in experimental and observational effectiveness studies. The BM referred to in this study has been used previously

doi: 10.2340/16501977-2659

for assessing the generalizability of evidence from RCTs published in the leading medical journals (14).

The aims of the current paper were to evaluate how comprehensively the relevant items for the assessment of the generalizability of findings from RCTs were recorded in SRs in the 4 leading general medical journals publishing SRs; and, based on the SRs, to evaluate the clinical homogeneity of the RCT studies, the justification for meta-analysis of the RCT data, and the generalizability of evidence from the SRs.

METHODS

The BM was used to assess the degree of completeness of how the SRs documented reporting of their study object, the RCTs. The BM is based on 5 categories (selection, baseline characteristics, intervention factors, outcome assessments, and statistical issues) and several subcategories (7, 14, 15), which were modified for the assessment of how comprehensively the authors of the SRs had recorded the essential features of the RCTs (Table I). Considering an item as recorded in an SR was based on whether there was any information on that particular item. For example, regarding competence, it was assessed whether the competence of the staff had been rated and documented, whether any characterization was included in the SR, and it was rated as not-documented if there was no mention of competence.

Table I. Categories and subcategories of the benchmarking method (BM) for assessment of the capability of systematic reviews to record essential items in randomized controlled trials (RCTs)

1. Selection of patients/population of the study
1.1 Description of intended patient population: inclusion and exclusion criteria
1.2. Description of patients' clinical path before being eligible for the study
1.3. Reporting of reasons for exclusions before randomization
1.4. Percentage of eligible patients declining participation
1.5. Description of consecutiveness of patient recruitment
1.6. Description of characteristics of the healthcare settings
1.8. Description of competence of the staff
2. Validity and completeness of baseline data
2.1. Demographic and clinical data (duration, quality and severity of indication)
2.2. Functioning (disease-specific or generic, health-related quality of life)
2.3. Comorbidity or a comorbidity index
2.4. Behavioural factors (smoking, alcohol/substance consumption, exercise, obesity)
2.5. Environmental factors (work, living conditions, marital status).
2.6. Potential inequity (socioeconomic status, education, deprivation, ethnicity).
2.7. Assessment of baseline comparability
3. Validity and completeness of intervention data
3.1. Description of intended index intervention: dose, frequency, duration
3.2. Description of intended control intervention: dose, frequency, duration
3.3. Completion rate of index intervention(s) according to protocol among all recruited, or adherence rate
3.4. Completion rate of control intervention(s) according to protocol among all recruited, or adherence rate
3.5. Proportion of patients' crossing over to index intervention
3.6. Proportion of patients' crossing over to control intervention
3.7. Co-interventions (use of other health services)
4. Validity and completeness of outcome data
4.1. Description of primary outcome variable(s) and secondary outcome variables
4.2. Follow-up percentage for the primary outcome at primary follow-up time
4.3. Reasons for dropping out/withdrawal reported in each group
4.4. Assessment of validity of outcome variables
5. Statistical analysis
5.1. Description of sample size calculation
5.2. Description and assessment of adequateness of statistical analysis

PubMed was searched to find all SRs and meta-analyses of cluster randomized and individually randomized controlled trials aiming at the assessment of the effectiveness of treatment or rehabilitation interventions, which were published in the *Annals of Internal Medicine, BMJ, JAMA (The Journal of the American Medical Association)* and *Lancet* from 1 January 2016 to 28 February 2019.

The inclusion criterion was that "systematic review" was stated in the title or in the abstract of the paper, and when the full-text article was obtained it conformed with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement definition of a SR. The New England Journal of Medicine was not included, because it does not publish SRs. Reviews based on individual patient data (IPD) were included.

The exclusion criteria were: if there were fewer than 2 RCTs; if the review assessed the effectiveness of diagnostic or screening procedures; if the review assessed the effects of healthcare-system-related interventions (e.g. the effects of resources or the organization of care); umbrella reviews, i.e. SRs of SRs. The following key words were used: SR, meta-analysis, name of each journal, and time-frame from 1 January 2016 to 28 February 2019. PubMed's advanced search tool was used. The search was repeated to ensure that all eligible papers were included. Both literature searches were performed in March 2019. No additional papers were found in the replicated PubMed search. Any paper with potential eligibility was examined according to the title and abstract, and for the potentially eligible ones the full-text papers were retrieved. Final decisions on eligibility were based on the full-text paper.

Descriptive information was extracted on the selection of patients, completeness and validity of the data for the baseline characteristics, interventions, outcomes, and statistical analysis. For patient characteristics, the numbers of subcategories recorded in the SRs were documented separately, e.g. for the behavioural factors, the percentages of SRs documenting 1, 2, 3, or 4 of the subcategories of exercise, smoking, alcohol use, and obesity were documented. Data extraction was checked twice (3 assessments in total) to ensure the accuracy of the data. The information was gathered both from the main texts and from all supplementary material provided alongside the article. Categorization of the type of intervention was made based on both the index and control interventions.

RESULTS

A total of 115 SRs and meta-analyses fulfilling the inclusion criterion were identified; 36 in *Annals of Internal Medicine*, 32 in *BMJ*, 26 in *JAMA*, and 21 in *Lancet*. The included and excluded RCTs are shown in Table SI[1]. Considering all the comparisons, pharmacological interventions were included in 71% of the SRs, other conservative treatments in 35%, and surgical interventions in 13%. There were no SRs on rehabilitation interventions. A meta-analysis was included in all of the SRs published

Table II. Baseline characteristics of systematic reviews (SRs) and meta-analyses published in the *Annals of Internal Medicine, BMJ, JAMA (The Journal of the American Medical Association)* and *Lancet* from 1 January 2016 to 28 February 2019

	Journal (number of SRs)				
SR characteristics	*Annals of Internal Medicine* (*n* = 36)	*BMJ* (*n* = 32)	*JAMA* (*n* = 26)	*Lancet* (*n* = 21)	Total (*n* = 115)
Type of intervention, *n* (%; some overlapping)					
Pharmacological	25 (69)	21 (66)	22 (85)	14 (67)	82 (71)
Conservative	15 (42)	10 (31)	10 (38)	5 (24)	40 (35)
Surgical	7 (19)	4 (13)	0 (0)	4 (19)	15 (13)
Rehabilitation	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Meta-analysis included, %	75	100	92	100	90
Inequity addressed as a study question in the systematic review, %	0	0	0	0	0

in the *BMJ* and the *Lancet*, and in 75% and 92% of the SRs published in *Annals of Internal Medicine* and *JAMA*, respectively. None of the SRs addressed inequity as a study question (Table II).

A description of patient inclusion and exclusion criteria in the RCTs under review was available in 11% (range 5–16% in the 4 journals) of the SRs (Table III). A description of index and control interventions was provided in 55% (range 29–69%) and 50% (range 24–62%), respectively. Primary outcomes were described in 23% of the SRs (range 19–31%). All 4 components of the PICO in individual RCTs were described in 3 of the SRs from the *BMJ*, 2 from *JAMA*, and none from the *Annals of Internal Medicine or Lancet* (Table SI[1]). A total of 5 out of 115 SRs reported all PICO characteristics, i.e. 4% of all SRs.

None of the 115 SRs reported a description of the patients' paths prior to assessment of their eligibility, or the reasons for exclusion before randomization, or percentages of eligible patients declining participation (Table IV). A description of the consecutiveness of the patient recruitment, the characteristics of healthcare system features, and the competence of staff were reported in 1%, 6% and 1% of the SRs, respectively.

Demographic and disorder specific clinical data was reported in 35% (range 0–50%) of the SRs (Table IV). Functioning of the patients (at least one item describing disease-specific or generic disability or health-related quality of life) was reported in 3% of the SRs. Co-morbid conditions were reported in 25% of the SRs (range 19–48%).

Any behavioural factor was reported in 21% (range 13–29%) of the SRs; any environmental factor in 0% of the SRs; and any factor related to potential inequity in 9% of the SRs (Table IV).

The baseline comparability of patients in the index and control groups in the RCTs was assessed in 7% of the SRs.

Adherence of patients in the RCTs to the index and control intervention(s) according to the protocol was reported in 7% of the SRs (range 5–8%). Cross-over to index interventions and control interventions was reported in 0% of the SRs. The use of other healthcare services besides the experimental interventions was reported in 2% of the SRs (Table IV).

The validity of outcome assessment was reported in 5% of the SRs (range 0–12%). None of the SRs reported reasons for dropping out of the RCT. None of the SRs assessed whether power calculations were performed in the RCTs. The adequateness of statistical analysis in the RCTs was assessed in 3% of the SRs (range 0–6%) (Table IV).

The results of assessment of individual SRs are shown in Table SI[1].

[1]https://doi.org/10.2340/16501977-2659

Table III. Reporting (%) of intended PICO (patients, index interventions, control interventions and outcomes) of the randomized controlled trials (RCTs) in systematic reviews (SRs) and meta-analyses published in the *Annals of Internal Medicine, BMJ, JAMA (The Journal of the American Medical Association)* and *Lancet* from 1 January 2016 to 28 February 2019

	Journal (number of SRs)				
PICO characteristics	*Annals of Internal Medicine* (*n* = 36) %	*BMJ* (*n* = 32) %	*JAMA* (*n* = 26) %	*Lancet* (*n* = 21) %	Total (*n* = 115) %
1. Description of intended patient population: inclusion and exclusion criteria reported	11	16	12	5	11
2. Description of intended index intervention: dose, frequency, duration reported	50	66	69	29	55
3. Description of intended control intervention: dose, frequency, duration reported	50	56	62	24	50
4. Description of primary and secondary outcome variables	22	31	19	19	23

Table IV. Reporting (%) of characteristics of patients, interventions and outcomes in the randomized controlled trials (RCTs) included in systematic reviews (SRs) and meta-analyses published in the *Annals of Internal Medicine, BMJ, JAMA* (*The Journal of the American Medical Association*) and *Lancet* from 1 January 2016 to 30 April 2019

	Journal (number of SRs)				
Study characteristics	*Annals of Internal Medicine* (*n* = 36) %	*BMJ* (*n* = 32) %	*JAMA* (*n* = 26) %	*Lancet* (*n* = 21) %	Total (*n* = 115) %
1. Selection of patients; healthcare system features					
1.1. Description of patients' path prior to assessment of eligibility	0	0	0	0	0
1.2. Reporting of reasons for exclusion before randomization	0	0	0	0	0
1.3. Percentage of eligible patients declining participation documented	0	0	0	0	0
1.4 Description of consecutiveness of patient recruitment	0	3	0	0	1
1.5. Description of characteristics of all the healthcare settings where the data was collected	6	0	15	5	6
1.5. Description of staff competence	0	0	4	0	1
2. Baseline characteristics of patients					
2.1. Demographic and disorder specific clinical data	50	0	46	48	35
2.2. Functioning (disease specific (D) or generic (G) and health-related quality of life (Q) (% of at least 1; at least 2; all 3 items described)	6	0	8	0	3
	0	0	0	0	0
	0	0	0	0	0
2.3. Comorbidity, at least 2 comorbid conditions reported or a comorbidity index	22	19	19	48	25
2.4. Behavioural factors (smoking (S), alcohol/substance consumption (A) or exercise reported (E)); and obesity (O (BMI)). In children: parent data. (% of at least 1; 2; 3 or 4 items described)	19	13	27	29	21
	11	0	8	14	8
	0	0	0	0	0
	0	0	0		0
2.5. Environmental factors (work (W) or living conditions (L); marital status (M)). In children: parent data. (% of 1; 2; 3 items described)	0	0	0	0	0
	0	0	0	0	0
	0	0	0	0	0
2.6. Potential inequity (socioeconomic status (S), education ("E"), deprivation (D), ethnicity (Et). In children: data from parents. (% of 1; 2–3; 4 items described)	8	3	19	5	9
	6	0	0	0	2
	0	0	0	0	0
2.7. Assessment of baseline comparability of the treatment arms	8	3	15	0	7
3. Interventions					
3.1. Completed index intervention(s) according to protocol among all recruited or adherence rate	8	6	8	5	7
3.2. Completed control intervention according to protocol among all recruited or adherence rate	8	6	8	5	7
3.3. Crossover to index intervention	0	0	0	0	0
3.4. Crossover to control intervention	0	0	0	0	0
3.5. Co-interventions (use of other health services) reported within each intervention arm	0	3	4	0	2
4. Follow-up					
4.1. Assessment of validity of outcome variables	6	3	12	0	5
4.2. Follow-up percentage (of those randomized) for the primary outcome at the primary follow-up time	17	0	8	5	8
4.3. Reasons for dropping out/withdrawal reported in each group or no drop-outs	0	0	0	0	0
5. Statistical analyses					
5.1. Assessment of power calculations	0	0	0	0	0
5.2. Assessment of the adequateness of statistical analysis, %	3	6	0	0	3

DISCUSSION

The aim of this review was to determine how comprehensively items relevant to the assessment of the generalizability of findings from RCTs were assessed in SRs published in leading general medical journals. It was found that the original RCTs often do not report these features (14), but *a systematic review can assess the presence or non-presence of reporting of all the items, irrespective of whether they have been reported in the RCTs.*

The BM referred to in this study was originally designed for the assessment of observational effectiveness studies where there is no randomization of the comparison groups, and therefore baseline comparability between groups is usually not satisfactory and necessitates statistical adjustment. In order to be able to adjust for baseline differences between groups, a very detailed description of patient characteristics is needed (7, 15, 16). The BM is in harmony with the PRISMA recommendations; but it makes explicit the items to be considered in SRs when evaluating the RCTs comprehensively (Table I). The BM has also been used for the assessment of the validity of RCTs and the generalizability of their evidence (7).

For the current paper, some modifications were made to the BM in order to best answer the current study questions. Most of the SRs assessed the effectiveness of pharmacological therapies, which is in agreement with the majority of RCTs being focused on these interventions (14) (Table I). Surprisingly none of the SRs assessed the effectiveness of rehabilitation, although the proportion of the elderly population in need of rehabilitative interventions is increasing rapidly worldwide, particularly in lower-middle income and upper-middle income countries (17). To be considered a form of rehabilitation, the current paper considered only those interventions that unequivocally fulfilled the criterion

of also using efforts besides or beyond biomedical interventions to increase the functioning of patients. Thus, for example, single physical medicine interventions were not considered to be a form of rehabilitation. None of the SRs focussed on equity (Table III), even though the PRISMA statement extension to equity has been published in 2012, with the aim of increasing the number of SRs with a focus on equity issues (6).

Reporting of the intended PICO (patients, index interventions, control interventions and outcomes) in the RCTs under study was deficient overall, and this was the case in all 4 journals. Patient inclusion and exclusion criteria were reported in approximately 10% of the SRs overall. A description of the index and control interventions was reported in approximately 50% of the SRs, but varied from 24% to 69% between the journals. Only 4% of the SRs reported all PICO characteristics, although complete reporting of PICO information is necessary for the reader to understand the aim of each individual RCT. Furthermore, without comprehensive PICO information, assessment of the clinical heterogeneity of RCTs, and of the generalizability of the findings of a SR, may not be possible.

None of the SRs reported the description of patient selection in the RCTs (i.e. the patients' path to entering a trial, and the proportion of those declining to participate and reporting the reasons for exclusion). A description of the consecutiveness of patient recruitment was rarely reported. A lack of information on the selection of patients means that the reader is not able to make judgements about whether the study population is representative of all those patients who need help for the disorder in question, or whether the patients are selected in a way that does not allow any generalizations, or that all generalizations are uncertain (14). In essence, the lack of data concerning the selection limits the possibilities to assess the clinical homogeneity of the RCTs and the generalizability of the findings of the SRs.

Only 6% of the SRs assessed the reporting of characteristics of the healthcare systems in the RCTs, and only 1% assessed the competence of staff. Healthcare systems are potential determinants for patient outcomes (16), and there is evidence to show that the safety of healthcare interventions is dependent on staff competence (18). Lack of information on healthcare systems and on the competence of staff impedes the generalizability of the findings from the SRs.

Only 35% of the SRs reported what the medical indication was in the RCTs and only 3% reported respective disability. One of the journals (BMJ) did not report this information at all. Thus, the most crucial data from the RCTs is lacking in most of the SRs. Data on comorbid conditions, often encountered in clinical practice, was lacking in 75% of the SRs. A lack of data regarding these essentials may preclude any assessment of the generalizability of the findings (14).

Reporting of any behavioural factor in the RCTs was assessed in 21% of the SRs, while 0% of the SRs reported environmental factors (work and living conditions, marital status); and only 9% of the SRs reported any factor related to inequity (education, ethnicity). None of the SRs reported a single characteristic belonging to all 3 subcategories of behavioural, environmental or inequity-related factors, although all of these categories may affect the effectiveness of treatment (10, 19). None of the SRs reported data on socioeconomic status, which has been shown to be a strong prognostic indicator for a wide spectrum of disorders (20, 21).

Overall, there is evidence to show that patient characteristics, such as functioning, co-morbidities, as well as behavioural, environmental and equity-related factors may have a direct effect on outcome in RCTs, or may modify the treatment effect (8–10, 12, 13, 19, 22, 23). Even in cases where the effects of these characteristics are assumed to be minor, their importance as modifiers of the treatment effects will remain unknown unless these features are reported and analysed in the RCTs and SRs.

Infrequent reporting of the baseline comparability of the patients in the index and control treatment arms (which was reported in only in 7% of the SRs) impairs the assessment of the validity of the effectiveness estimates of the RCTs in the SRs. Baseline comparability between treatment arms is a shared criteria for BCTs and RCTs. It is also the feature of RCTs that makes it superior to observational effectiveness studies in terms of internal validity (7, 15)

The adherence of patients in the RCTs to the index and control intervention(s) according to the protocol was reported in 7% of SRs, while 0% of the SRs reported the cross-over to index interventions and control interventions; and only 2% of the SRs reported on use of other healthcare services besides the experimental interventions. In experimental studies the interventions (the experimental interventions and interventions due to use of other healthcare services) are the causal factors for the outcomes (24). If adherence to the intervention has been very high in some RCTs, and very low in other RCTs, the latter obviously results in lower estimates of effectiveness than the former. If there is no information on the degree of adherence it is not possible to assess the clinical homogeneity of the RCTs, and neither is it possible to assess the generalizability of the findings from the SRs.

The rare assessment of reporting of the follow-up percentages in the RCTs (in only 8% of the SRs), and total lack of reporting of reasons for patients dropping

out of RCTs weakens the assessment of the generalizability of findings from the SRs.

Baseline comparability (the aim of randomization), description of adherence to interventions (a causal factor), and the validity of the outcome (effect) was assessed in 7%, 7% and 5% of the SRs, respectively. As these vital validity factors of an RCT are lacking, the appraisal of internal validity of individual RCTs is seriously impaired (24, 25).

The adequateness of the statistical analysis in the RCTs was assessed in only 3% of the SRs. There is evidence that there are often errors in statistical analyses of RCTs, which may lead to biased effectiveness estimates (26). In most meta-analyses the effectiveness estimates derive from the results sections of the RCTs (an exception is with individual patient meta-analysis) and if the estimates are not valid, the estimates of the meta-analyses will be biased. Lack of reporting of whether power calculations were made in the RCTs may lead to unwarranted interpretations of no effectiveness in cases where the statistical power has been insufficient to reach statistically significant results. Only clinically homogeneous RCTs can be analysed in a meta-analysis, and only then can the SR increase the statistical power and lead to narrower confidence intervals.

Although there were differences between the 4 journals in how the SRs assessed the comprehensiveness of reporting in the RCTs, the reporting was poor in all the journals. The reporting did not allow the assessment of the clinical homogeneity of the RCTs, the justification for meta-analysis, or inferences concerning the generalizability of the findings. Even in meta-analyses based on individual patient data from each RCT, there must be appropriate information from all the trials to ensure that it is plausible to combine patient data from different trials.

Limitations

A limitation of this paper is that a single person, the author, performed the literature search and data extraction. However, the literature search was repeated to ensure the completeness of the SRs published, and the author checked the accuracy of the extracted data twice. In order to make the judgement as unequivocal as possible as to whether an item was recorded in an SR, the decision was based on whether there was any information on that particular item; for example, the competence of the staff was rated as documented, if any characterization was included in the SR, and rated as not-documented if there was no mention of competence. Moreover, in the case of patient characteristics, the numbers of subcategories recorded in the SRs were documented separately (Table IV). Table SI shows for each RCT whether each of the items was reported, making it readily possible to check the accuracy of any assessment. Although there would be errors in some of the assessments, it is highly unlikely that the majority of the assessments would not be valid after the assessments had been repeated twice. However, there is an urgent need to replicate the findings of this paper, and to assess different sets of criteria for SRs in their ability to assess clinical heterogeneity/homogeneity of the included RCTs enabling assessment of justification of a meta-analysis, and generalizability of findings in different clinical contexts. Persons having the best competence on each particular study question are the most capable of reaching valid inferences on this matter.

The number of items that are important for any particular SR may be fewer than the number of items that are potentially important and thus suggested in the BM. Many journals have word limitations, which may pose limitations for the inclusion of BM characteristics. Since it is obvious that the items important for the generalizability of results should be documented, researchers should provide the information in tables or report some items in web appendices.

Conclusions

The current paper suggests that a comprehensive description of characteristics of RCTs by using the BM would be the primary option for SRs, and in case some items are left out of reporting, the reasons for omitting these would be provided in the SR. A comprehensive description will allow generalizability to any setting, and obviously the authors of SRs cannot be aware of all of these. Documentation of the BM items requires work from researchers carrying out SRs. However, this should not be considered additional work, but as recording of the characteristics of the very study object, the RCT. Compared with the total work involved in a SR, documenting these essentials is not an unsurmountable task.

The main finding of this paper is that the reporting of the characteristics of RCTs in the SRs published in the leading medical journals is extremely poor. The relevance of each item is, to some degree, dependent on the study context. However, every SR should document at least a description of the indication and its severity, the adherence to the intervention, and the proportions of patients crossing over to other treatment arms in each RCT. Considering the vast influence that SRs have in clinical and health policy decision-making, the findings of the current paper may have implications both for research, for critical assessment of SRs, and for clinical practice and policy-making. There is a need for further research on how exactly to define the concepts of BM and how to operationalize these concepts, and whether

it is feasible also to consider minimum standards, even though the variation in context dependence worldwide may make this difficult.

The aim of a SR is to provide a full description of its study object, which, in most cases, is a RCT. When assessed in a comprehensive manner, the SRs and meta-analyses published in the leading general medical journals show a lack of reporting of the essential characteristics of their study objects, i.e. the RCTs. The findings indicate that it is currently not possible to generalize evidence from any of these SRs and meta-analyses. In the future, a detailed description of the RCTs is needed to decide whether a meta-analysis is justified in SRs, and in order to ensure the generalizability of evidence for clinical practice. The comprehensive data extraction, based here on the BM, allows the assessment of the clinical homogeneity of RCTs, and the assessment of the generalizability of the evidence in SRs. There is an urgent need for further research into these questions.

ACKNOWLEDGEMENTS

The author thanks Professor Pekka Jousilahti for valuable comments in the final stages of writing the manuscript; and Gareth Attwood, PG Dip, Licentiate Dip. TESOL, and Riitta Malmivaara, BA, MA for checking the English language.

The author has no conflicts of interest to declare.

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: 1405–1409.
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: 1660–1673.
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. Tugwell P, Maxwell L, Welch V, Kristjansson E, Petticrew M, Wells G, et al. Is Health Equity Considered in Systematic Reviews of the Cochrane Musculoskeletal Group? Arthritis Rheum-Arthritis Care Res 2008; 59: 1603–1610.
6. Welch V, Petticrew M, Tugwell P, Moher D, O'Neill J, Waters E, et al. PRISMA-Equity 2012 extension: reporting guidelines for systematic reviews with a focus on health equity. PLos Med 2012; 9: e1001333.
7. Malmivaara A. Benchmarking controlled trial – a novel concept covering all observational effectiveness studies. Ann Med 2015; 47: 332–340.
8. 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: 1229–1237.
9. 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: 181–188.
10. 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: 2255–2263.
11. 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: 161–178.
12. 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: 1659–1724.
13. 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: 1745–1753.
14. Malmivaara A. Generalizability of findings from randomized controlled trials is limited in the leading general medical journals. J Clin Epidemiol 2019; 107: 36–41.
15. Malmivaara A. Assessing validity of observational intervention studies – the Benchmarking Controlled Trials. Ann Med 2016; 48: 440–443.
16. Malmivaara A. System impact research – increasing public health and health care system performance. Ann Med 2016; 48: 211–215.
17. Stucki G. Olle Hook Lectureship 2015: The World Health Organization's paradigm shift and implementation of the International Classification of Functioning, Disability and Health in rehabilitation. J Rehabil Med 2016; 48: 486–493.
18. 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: 1434–1442.
19. 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: 1343–1350.
20. Malmivaara A. On decreasing inequality in health care in a cost-effective way. BMC Health Serv Res 2014; 14: 79.
21. Marmot M, Friel S, Bell R, Houweling TAJ, Taylor S. Closing the gap in a generation: health equity through action on the social determinants of health. Lancet 2008; 372: 1661–1669.
22. 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.
23. 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: 331–338.
24. Malmivaara A. Pure intervention effect or effect in routine health care – blinded or non-blinded randomized controlled trial. BMC Med Res Methodol 2018; 18: 91.
25. Malmivaara A. Validity and generalizability of findings of randomized controlled trials on arthroscopic partial meniscectomy of the knee. Scand J Med Sci Sports 2018; 28: 1970–1981.
26. Dwan K, Altman DG, Clarke M, Gamble C, Higgins JP, Sterne JA, et al. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. PLoS Med 2014; 11: e1001666.