Tools to Assess the External Validity of Randomized Controlled Trials in Systematic Reviews: a Systematic Review of Measurement Properties

Andres Jung (a.jung@uni-luebeck.de)
Institute of Health Sciences, Department of Physiotherapy, Pain and Exercise Research Luebeck (P.E.R.L), University of Luebeck, Ratzeburger Allee 160, 23562 Luebeck

Julia Balzer
Faculty of Applied Public Health, European University of Applied Sciences, Werftstr. 5, 18057 Rostock

Tobias Braun
Department of Applied Health Sciences, Hochschule für Gesundheit (University of Applied Sciences), Division of Physiotherapy, Gesundheitscampus 6-8, 44801 Bochum

Kerstin Luedtke
Institute of Health Sciences, Department of Physiotherapy, Pain and Exercise Research Luebeck (P.E.R.L), University of Luebeck, Ratzeburger Allee 160, 23562 Luebeck

Research Article

Keywords: External validity, generalizability, applicability, measurement properties, tools, randomized controlled trial

Posted Date: September 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-830181/v1

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on April 6th, 2022. See the published version at https://doi.org/10.1186/s12874-022-01561-5.

Abstract

Background: Internal and external validity are the most relevant components when critically appraising randomized controlled trials (RCTs) for systematic reviews. However, there is no gold standard to assess external validity. This might be related to the heterogeneity of terminology as well as to unclear evidence of the measurement properties of available tools. The aim of this review was to identify tools to assess the external validity of RCTs in systematic reviews and to evaluate the quality of evidence regarding their measurement properties.

Methods: A two-phase systematic literature search was performed in four databases: MEDLINE via PubMed, Scopus, PsycINFO via OVID, and CINAHL via EBSCO. First, tools to assess the external validity of RCTs were identified. Second, studies aiming to investigate the measurement properties of these tools were selected. The measurement properties of each included tool were appraised using an adapted version of the COnsensus based Standards for the selection of health Measurement INstruments (COSMIN) guidelines.

Results: 34 publications reporting on the development or validation of 26 included tools were included. For 62% of the included tools, there was no evidence of any measurement property. For the remaining tools, reliability was assessed most frequently. Reliability was judged as "*sufficient*" for three tools (very low quality of evidence). Content validity was rated as "*sufficient*" for one tool (moderate quality of evidence).

Conclusions: Based on these results, no available tool can be fully recommended to assess the external validity of RCTs in systematic reviews. Several steps are required to overcome the identified difficulties to either adapt and validate available tools or to develop a new one. There is a need for more research for this purpose.

Trial registration: Prospective registration at Open Science Framework (OSF): https://doi.org/10.17605/OSF.IO/PTG4D

Background

Systematic reviews are powerful research formats to summarize and synthesize the evidence from primary research in health sciences [1, 2]. In clinical practice, their results are often applied for the development of clinical guidelines and treatment recommendations [2]. Consequently, the methodological quality of systematic reviews is of great importance. In turn, the informative value of systematic reviews depends on the overall quality of the included controlled trials [3, 4]. Accordingly, the evaluation of internal and external validity is considered the key step in systematic review methodology [3].

Internal validity relates to the systematic error or bias in clinical trials [5] and expresses how methodologically robust the study was conducted. External validity is the inference about the extent to which "a causal relationship holds over variations in persons, settings, treatments and outcomes" [6]. There are plenty of definitions for external validity and a variety of different terms. However, external validity, generalizability, applicability, and transferability, among others, are used interchangeably in the literature [7]. Schünemann et al. [8] suggested that: 1) generalizability "may refer to whether or not the evidence can be generalized from the population from which the actual research evidence is obtained to the population for which a healthcare answer is required"; 2) applicability may be interpreted as "whether or not the research evidence answers the healthcare question asked by a clinician or public health practitioner" and 3) transferability is often interpreted as to "whether research evidence can be transferred from one setting to another". Four essential dimensions are proposed to evaluate the external validity of controlled clinical trials in systematic reviews: patients, treatment variables, settings, and outcome modalities [3]. Its evaluation depends on the specificity of the reviewers´ research question, the review´s inclusion and exclusion criteria compared to the trial´s population, the setting of the study, as well as the quality of reporting these four dimensions.

In health research, however, external validity is often neglected when critically appraising clinical studies [9, 10]. One possible explanation might be the lack of a gold standard for assessing external validity of clinical trials. Systematic and scoping reviews have examined published frameworks and tools for assessing the external validity of clinical trials in health research

[7, 9, 11–15]. A substantial heterogeneity of terminology and criteria as well as a lack of guidance on how to assess the external validity of intervention studies was found [7, 9, 12–15].

Although some tools for the evaluation of external validity exist, no comprehensive evaluation of the measurement properties of these tools has been performed. RCTs are considered the most suitable research design for investigating cause and effect mechanisms of interventions [16]. However, the study design of a RCT is susceptible for a lack of external validity due to the randomization, the use of exclusion criteria and poor willingness of eligible participants to participate [17, 18]. There is evidence that the reliability of external validity evaluations with the same measurement tool differs between randomized and non-randomized trials [19]. In addition, due to differences in requested information from reporting guidelines (e.g. CONSORT, STROBE), respective items used for assessing the external validity can vary between research designs. Acknowledging the importance of RCTs in the medical field, this review focused only on tools to assess the external validity of RCTs. The aim was to identify tools to assess the external validity of RCTs in systematic reviews and to evaluate the quality of evidence regarding their measurement properties. Objectives: 1) to identify published measurement tools to assess the external validity of RCTs in systematic reviews ; 2) to evaluate the quality of evidence of their measurement properties; 3) to formulate recommendations for or against the use of tools to assess the external validity of RCTs in future systematic reviews based on evidence of their measurement properties.

Methods

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) Statement [20] and an adapted version of the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) methodology for systematic reviews of measurement instruments in health sciences [21–23]. The COSMIN methodology was chosen since this method is comprehensive and validation processes do not differ substantially between patient-reported outcome measures (PROMs) and measurement instruments of other latent constructs. According to the COSMIN authors, it is acceptable to use this methodology for non-PROMs [22]. Furthermore, because of its flexibility, it has already been used in systematic reviews assessing measurement tools which are not health measurement instruments [24–26]. However, adaptations or modifications may be necessary [22]. Pilot tests and adaptation-processes of the COSMIN methodology are described below (see section “Quality assessment and evidence synthesis”). The review protocol was prospectively registered on March 6, 2020 in the Open Science Framework (OSF) with the registration DOI: https://doi.org/10.17605/OSF.IO/PTG4D [27].

Deviations from the preregistered protocol

One of the aims listed in the review protocol was to evaluate the characteristics and restrictions of measurement tools in terms of terminology and criteria for assessing external validity. This issue has been addressed in two recent reviews with a similar scope [7, 14]. Although our eligibility criteria differed, it was concluded that no novel data was available for the present review to extract, since authors of included tools did not describe the definition or construct of interest or cited the same reports. Therefore, this objective was omitted.

Literature search and screening

A search of the literature was conducted in four databases: MEDLINE via PubMed, Scopus, PsycINFO via OVID, and CINAHL via EBSCO. The eligibility criteria and search strategy were predefined in collaboration with a research librarian and is detailed in Table S1 [see Additional file 1]. The search strategy was designed according to the COSMIN methodology and consists of the following four key elements: 1) construct (external validity of RCTs from review authors´perspective), 2) population(s) (RCTs), 3) type of instrument(s) (measurement tools, checklists, surveys etc.), and 4) measurement properties (e.g. validity and reliability) [28]. The four key elements were divided in two main searches (as performed by others [29]): the phase 1 search contained the first three key elements to identify measurement tools to assess the external validity of RCTs. The phase 2 search aimed to identify studies evaluating the measurement properties of each tool, which was identified and included during phase 1 search, using the fourth key element. For this second search, a sensitive PubMed search filter developed by

Terwee et al. [30] was applied. Translations of this filter for the remaining databases were taken from the COSMIN website and from other published COSMIN reviews [31, 32] with permission from the authors. The phase 1 search was conducted in February and March 2020 and the phase 2 search was conducted in December 2020. Both searches were performed without restriction regarding the time of publication (databases were searched from inception). In addition, reference lists of the retrieved articles were scanned manually for any additional studies. An update of both literature searches was conducted in March 2021.

Title and abstract screening for both searches and the full-text screening during phase 2 were performed independently by two researchers (AJ, KL & TB). Full-text screening and tool/data extraction in phase 1 was performed by one reviewer (AJ) and checked by a second reviewer (TB). Data extraction for both searches was performed with an extraction sheet based on the recommendations of the COSMIN user manual [28]. The Rayyan QCRI web app [33] was used to facilitate the screening process (both searches) according to a priori defined eligibility criteria.

Eligibility criteria

Phase 1 search (identification of tools)

Records were considered for inclusion based on their title and abstract according to the following criteria: 1) records describing the development and or implementation (application), e.g. manual or handbook, of any tool to assess the external validity of RCTs; 2) systematic reviews that applied tools to assess the external validity of RCTs and which explicitly mentioned the tool in the title or abstract; 3) systematic reviews or any other publications in which it was assumed that a tool was used for external validity assessment, but the tool was not explicitly mentioned in the title or abstract; 4) records that gave other references to, or dealt with, tools for the assessment of external validity of RCTs, e.g. method papers, commentaries.

The full-text screening was performed to extract or to find references to potential tools. If a tool was cited, but not presented or available in the full-text version, the internet was searched for websites on which this tool was presented, to extract and review for inclusion. Potential tools were extracted and screened for eligibility as follows: measurement tools aiming to assess the external validity of RCTs and designed for implementation in systematic reviews of intervention studies. Since the terms external validity, applicability, generalizability, relevance and transferability are used interchangeably in the literature [8, 34], tools aiming to assess one of these constructs were eligible. Exclusion criteria: 1) The multidimensional tool includes at least one item related to external validity, but it is not possible to assess and interpret external validity separately. 2) The tool was developed exclusively for study designs other than RCTs. 3) The tool contained items assessing information not requested in the CONSORT-Statement [35] (e.g. cost-effectiveness of the intervention, salary of health care provider) and these items could not be separated from external validity. 4) The tool was published in a language other than English or German. 5) The tool was explicitly designed for a specific medical profession or field and cannot be used in other medical fields.

Phase 2 search (identification of studies on the measurement properties of included tools)

For the phase 2 search, reports evaluating the measurement properties of at least one of the included measurement tools were selected. Reports only using the measurement tool as an outcome measure without the evaluation of at least one measurement property were excluded. If a report did not provide evidence of measurement properties of the tool or dimension of interest individually, it was also excluded. Hence, studies providing data on validity or reliability of sum-scores of multidimensional tools with a dimension relevant to external validity, but not showing any data on the measurement properties of the individual dimension of interest, were excluded.

Quality assessment and evidence synthesis

All included reports were systematically evaluated by two independent reviewers (AJ & JB): 1) for their methodological quality by using the COSMIN Risk of Bias (RoB) checklist [21]; 2) against the updated criteria for good measurement properties [22,

23]; and 3) for their level of evidence, according to a modified GRADE approach [22, 23]. In case of disagreement, a third reviewer (TB) was consulted to reach consensus.

The COSMIN RoB checklist is a tool [21, 23, 36, 37] designed for the systematic evaluation of the methodological quality of studies assessing the measurement properties of health measurement instruments [21]. Although this checklist was specifically developed for systematic reviews of PROMs, it can also be used for reviews of non-PROMs [22] or measurement tools of other latent constructs [24]. As mentioned in the COSMIN user manual, adaptations for some items in the COSMIN RoB checklist might be necessary, in relation to the construct being measured [28]. Therefore, pilot tests were performed for the assessment of measurement properties of tools assessing the quality of RCTs before data extraction, aiming to ensure feasibility during the planned evaluation of the included tools. The pilot tests were performed with a random sample of publications on measurement instruments of potentially relevant tools. After each pilot test, results and problems regarding the comprehensibility, relevance and feasibility of the instructions, items, and response options in relation to the construct of interest were discussed. Where necessary, adaptations and/or supplements were added to the instructions of the evaluation with the COSMIN RoB checklist. Saturation was reached after two rounds of pilot testing. Substantial adaptations or supplements were required for Box 1 ('development process') and Box 10 ('responsiveness') of the COSMIN RoB checklist. Minor adaptations were necessary for the remaining boxes. The specification list, including the adaptations, can be seen in Table S2 [see Additional file 2]. The methodological quality of included studies was rated via the four-point rating scale of the COSMIN RoB checklist as "inadequate", "doubtful", "adequate", or "very good" [21].

After the RoB-assessment, the result of each single report on a measurement property was rated against the updated criteria for good measurement properties for content validity [23] and for the remaining measurement properties[22] as "sufficient" (+), "insufficient" (-), or "indeterminate" (?). These ratings were summarized and an overall rating for each measurement property was given as "sufficient" (+), "insufficient" (-), "inconsistent" (±), or "indeterminate" (?). However, the overall rating criteria for good content validity was adapted to the research topic of the present review. This method usually requires an additional subjective judgement from reviewers [38]. Since one of the biggest limitations on the topic of interest is the lack of consensus on terminology and criteria as well as how to assess the external validity [7, 9], a reviewers' subjective judgement was considered inappropriate. After this issue was also discussed with one leading member of the COSMIN steering committee, the reviewers' rating was omitted. A "sufficient" (+) overall rating was given if there was evidence of face or content validity of the final version of the measurement tool assessed by a user or expert panel. Otherwise, the rating "indeterminate" (?) or "insufficient" (-) was used for the content validity.

The quality of evidence was graded as "high", "moderate", "low", or "very low" according to the modified GRADE approach for content validity [23] and for the remaining measurement properties [22]. The modified GRADE approach distinguishes between four factors influencing the quality of evidence: risk of bias, inconsistency, indirectness, and imprecision. The starting point for all measurement properties is high quality of evidence and is subsequently downgraded by one to three levels per factor when there is risk of bias, (unexplained) inconsistency, imprecision, or indirect results [22, 23]. If there is no report on the content validity of a tool, the starting point for this measurement property is "moderate" and is subsequently downgraded depending on the quality of the development process [23]. Selective reporting bias or publication bias is not taken into account in the modified GRADE approach, because of a lack of registries for studies on measurement properties [22].

The evidence synthesis was performed qualitatively according to the COSMIN methodology [22]. If several reports revealed homogenous quantitative data (e.g. same statistics, population) on internal consistency, reliability, measurement error or hypotheses testing of a measurement tool, pooling the results was considered using generic inverse variance (random effects) methodology and weighted means as well as 95% confidence intervals were calculated for each measurement property [28]. No subgroup analysis was planned. However, statistical pooling was not possible in the present review.

Results

Literature search and selection process

Figure 1 shows the selection process. The “PRISMA flow diagram of systematic search strategy used to identify clinimetric papers” was adapted from Clark et al. [29] In the phase 1 search, from 4670 non-duplicate records, 4330 irrelevant records were excluded. 345 full-text articles were screened, and 71 potential tools were extracted. After reaching consensus, 45 tools were excluded (reasons for exclusion are presented in Table S3 [see Additional file 3]) and 26 were included.

In the phase 2 search, 1188 non-duplicate records were screened for title and abstract. 1147 records were excluded as they did not assess any measurement property of the included tools. Of 41 full-text records, 6 were included. The most common reason for exclusion was that records evaluating the measurement properties of multidimensional tools did not evaluate external validity as a separate dimension. For example, studies assessing the interrater reliability of the GRADE method [39, 40] were identified during reference screening, but had to be excluded, since they did not provide separate data on the reliability of the indirectness domain (representing external validity).

34 publications on the development or evaluation of the measurement properties of 26 included tools were included for quality appraisal according to the COSMIN guidelines.

Methods to assess the external validity of RCTs

During full-text screening in phase 1, several concepts to assess the external validity of RCTs were found (Table 1). Two main concepts were identified: experimental/statistical methods and non-experimental methods. The experimental/statistical methods were summarized and collated into five subcategories giving a descriptive overview of the different approaches used to assess the external validity. However, according to our eligibility criteria these methods were excluded, since they were not developed for use in systematic reviews of interventions. In addition, a comparison of these methods as well as appraisal of risk of bias with the COSMIN RoB checklist would not have been feasible. Therefore, the experimental/statistical methods described below were not included for further evaluation.

Table 1
Experimental/statistical methods to evaluate the EV of RCTs

1. Comparing differences of characteristics and/or NNT analysis from not-enrolled eligible patients with enrolled patients [41–44]
2. Conduction of observational studies to assess the “real world” applicability of RCTs [17, 45, 46]
3. Meta-analysis of patient characteristics data from RCTs [47, 48]
4. Comparison of data from RCTs with data from health record database and/or other epidemiological data: **a)** retrospectively [47–51] **b)** simulation-based (a priori and retrospective) [52, 53]
5. Review of exclusion criteria in RCTs which would limit the EV [54]
Abbreviations: EV = external validity, NNT = numbers needed to treat; RCT = randomized controlled trial For non-experimental methods, please refer to Table 2

Characteristics of included measurement tools

The included tools and their characteristics are listed in Table 2. Overall, the tools were heterogenous with respect to the number of items or dimensions, response options and development processes. The number of items varied between one and 26 items and the response options varied between 2-point-scales to 5-point-scales. Most tools used a 3-point-scale (n = 19, 73%). For 13 (50%) of the tools, the development was not described in detail [55, 56, 65–67, 57–64]. Seven review authors appear to have developed their own tool but did not provide any information on the development process [55, 56, 58–61, 64].

Table 2
Characteristics of included tools

Dimension and/or tool	Authors	Construct(s), as described by the authors	Target population	Domains, nr. of items	Response options	Development and validation
"Applicability"-dimension of LEGEND	Clark et al. [74]	Applicability of results to treating patients	P1: RCTs and CCTs P2: reviewers and clinicians	3 items	3-point-scale	Deductive and inductive item-generation. Tool was pilot tested among an interprofessional group of clinicians.
“Applicability”-dimension of Carr´s evidence-grading scheme	Carr et al. [55]	Generalizability of study population	P1: clinical trials P:2 authors of SRs	1 item	3-point-classification-scale	No specific information on tool development.
Bornhöft´s checklist	Bornhöft et al. [98]	External validity (EV) and Model validity (MV) of clinical trials	P1: clinical trials P2: authors of SRs	4 domains with 26 items for EV and MV each	4-point-scale	Development with a comprehensive, deductive item-generation from the literature. Pilot-tests were performed, but not for the whole scales.
Cleggs´s external validity assessment	Clegg et al. [56]	Generalizability of clinical trials to England and Wales	P1: clinical trials P2: authors of SRs and HTAs	5 items	3-point-scale	No specific information on tool development
Clinical applicability	Haraldsson et al. [59]	Report quality and applicability of intervention, study population and outcomes	P1: RCTs P2: reviewers	6 items	3-point-scale and 4-point-scale	No specific information on tool development
Clinical Relevance Instrument	Cho & Bero [72]	Ethics and Generalizability of outcomes, subjects, treatment and side effects	P1: clinical trials P2: reviewers	7 items	3-point-scale	Tool was pilot tested on 10 drug studies. Content validity was confirmed by 7 reviewers with research experience. - interrater reliability of overall score: ICC = 0.41 (n = 10) for pilot version

Dimension and/or tool	Authors	Construct(s), as described by the authors	Target population	Domains, nr. of items	Response options	Development and validation
“Clinical Relevance” according to the CCBRG	Van Tulder et al. [78]	Applicability of patients, interventions and outcomes	P1: RCTs P2: authors of SRs	5 items	3-point-scale (Staal et al., 2008)	Deductive item-generation for Clinical Relevance. Results were discussed in a workshop. After two rounds, a final draft was circulated for comments among editors of the CCBRG.
Clinical Relevance Score	Karjalainen et al. [71]	Report quality and applicability of results	P1: RCTs P2: reviewers	3 items	3-point-scale	No specific information on tool development.
EVAT (External Validity Assessment Tool)	Khorsan & Crawford [75]	External validity of participants, intervention, and setting	P1: RCTs and non-randomized studies P2: reviewers	3 items	3-point-scale	Deductive item-generation. Tool developed based on the GAP-checklist [57] and the Downs and Black-checklist [19]. Feasibility was tested and a rulebook was developed but not published.
"External validity"-dimension of the Downs & Black-Checklist	Downs & Black [19]	Representativeness of study participants, treatments and settings to source population or setting	P1: RCTs and non-randomised studies P2: reviewers	3 items	3-point-scale	Deductive item-generation, pilot test and content validation of pilot version. Final version tested for: - internal consistency: KR-20 = 0.54 (n = 20), - test-retest reliability: k = -0.05-0.48 and 10–15% disagreement (measurement error) (n = 20), - interrater reliability: k = -0.08-0.00 and 5–20% disagreement (measurement error) (n = 20) [19]; ICC = 0.76 (n = 20) [79]

Dimension and/or tool	Authors	Construct(s), as described by the authors	Target population	Domains, nr. of items	Response options	Development and validation
"External validity"-dimension of Foy´s quality checklist	Foy et al. [58]	External validity of patients, settings, intervention and outcomes	P1: intervention studies P2: reviewers	6 items	not clearly described	Deductive item-generation. No further information on tool development.
"External validity"-dimension of Liberati´s quality assessment criterias	Liberati et al. [62]	Report quality and generalizability	P1: RCTS P2: reviewers	9 items	dichotomous and 3-point-scale	Tool is a modified version of a previously developed checklist [99] with additional inductive item-generation. No further information on tool development.
"External validity"-dimension of Sorg´s checklist	Sorg et al. [64]	External validity of population, interventions, and endpoints	P1: RCTs P2: reviewers	4 domains with 11 items	not clearly described	Developed based on Bornhöft et al. [98] No further information on tool development.
“external validity”-criteria of the USPSTF	USPSTF Procedure manual [66]	Generalizability of study population, setting and providers for US primary care	P1: clinical studies P2: USPSTF reviewers	3 items	Sum-score-rating: 3-point-scale	Tool developed for USPSTF reviews. No specific information on tool development. - interrater reliability: ICC = 0.84 (n = 20) [79]
FAME (Feasibility, Appropriateness, Meaningfulness and Effectiveness) scale	Averis et al. [63]	Grading of recommendation for applicability and ethics of intervention	P1: intervention studies P2: reviewers	4 items	5-point-scale	The FAME framework was created by a national group of nursing research experts. Deductive and inductive item-generation. No further information on tool development.
GAP (Generalizability, Applicability and Predictability) checklist	Fernandez-Hermida et al. [57]	External validity of population, setting, intervention and endpoints	P1: RCTs P2: Reviewers	3 items	3-point-scale	No specific information on tool development.

Dimension and/or tool	Authors	Construct(s), as described by the authors	Target population	Domains, nr. of items	Response options	Development and validation
Gartlehner´s tool	Gartlehner et al. [68]	To distinguish between effectiveness and efficacy trials	P1: RCTs P2: reviewers	7 items	Dichotomous	Deductive and inductive item-generation. - Criterion validity testing with studies selected by 12 experts as gold standard.: Specificity = 0.83, sensitivity = 0.72 (n = 24) - Measurement error: 78.3% agreement (n = 24) - Interrater reliability: k = 0.42 (n = 24) [68]; k = 0.11–0.81 (n = 151) [80]
Green & Glasgow´s external validity quality rating criteria	Green & Glasgow [70]	Report quality for generalizability	P1: trials (not explicitly described) P2: reviewers	4 Domains with 16 items	Dichotomous	Deductive item-generation. Mainly based on the Re-Aim framework.[100] - interrater reliability: ICC = 0.86 (n = 14) [84] - discriminative validity: TREND studies report on 77% and non-TREND studies report on 54% of scale items (n = 14) [84] - ratings across included studies (n = 31) [86], no hypothesis was defined

Dimension and/or tool	Authors	Construct(s), as described by the authors	Target population	Domains, nr. of items	Response options	Development and validation
"Indirecntess"-dimension of the GRADE handbook	Schünemann et al. [87]	Differences of population, interventions, and outcome measures to research question	P1: intervention studies P2: authors of SRs, clinical guidelines and HTAs	4 items	Overall: 3-point-scale (downgrading options)	Deductive and inductive item-generation, pilot-testing with 17 reviewers [77]. - interrater reliability: k = 0.00–0.82 (n = 12) of pilot-version [77]
Modified "Indirectness" of the Checklist for GRADE	Meader et al. [76]	Differences of population, interventions, and outcome measures to research question.	P:1 meta-analysis of RCTs P:2 authors of SRs, clinical guidelines and HTAs	5 items	Item-level: 2- and 3-point-scale Overall: 3-point-scale (grading options)	Developed based on GRADE method, two phase pilot-tests, - interrater reliability: kappa was poor to almost perfect on item-level [76] and k = 0.69 for overall rating of indirectness (n = 29) [81]
external validity checklist of the NHMRC handbook	NHMRC handbook [67]	external validity of an economic study	P1: clinical studies P2: clinical guideline developers, reviewers	6 items	3-point-scale	No specific information on tool development.
revised GATE in NICE manual (2012)	NICE manual [65]	Generalizability of population, interventions and outcomes	P1: intervention studies P2: reviewers	2 domains with 4 items	3-point-scale and 5-point-scale	Based on Jackson et al. [101] No specific information on tool development.

Dimension and/or tool	Authors	Construct(s), as described by the authors	Target population	Domains, nr. of items	Response options	Development and validation
RITES (Rating of Included Trials on the Efficacy-Effectiveness Spectrum)	Wieland et al. [69]	To characterize RCTs on an efficacy-effectiveness continuum.	P1: RCTs P2: reviewers	4 items	5-point-likert-scale	Deductive and inductive item-generation, modified Delphi procedure with 69–72 experts, pilot testing in 4 Cochrane reviews, content validation with Delphi procedure and core expert group (n = 14) [69], - interrater reliability: ICC = 0.235–0.942 (n = 18) [69] ICC = 0.54-1.0 (n = 22) [82, 85] - convergent validity with PRECIS 2 tool: r = 0.55 correlation (n = 59) [82]
Section A (Selection Bias) of EPHPP (Effective Public health Practice Project) tool	Thomas et al. [73]	Representativeness of population and participation rate.	P1: clinical trials P2: reviewers	2 items	Item-level: 4-point-scale and 5-point-scale Overall: 3-point-scale	Deductive item-generation, pilot-tests, content validation by 6 experts, - convergent validity with Guide to Community Services (GCPS) instrument: 52.5–87.5% agreement (n = 70) [73] - test-retest reliability: k = 0.61–0.74 (n = 70) [73] k = 0.60 (n = 20) [83]
Section D of the CASP checklist for RCTs	CASP Programme [102]	Applicability to local population and outcomes	P1: RCTs P2: participants of workshops, reviewers	2 items	3-point-scale	Deductive item-generation, development and pilot-tests with group of experts.

Dimension and/or tool	Authors	Construct(s), as described by the authors	Target population	Domains, nr. of items	Response options	Development and validation
Whole Systems research considerations´ checklist	Hawk et al. [60]	Applicability of results to usual practice	P1: RCTs P2: Reviewers (developed for review)	7 domains with 13 items	Item-level: dichotomous Overall: 3-point-scale	Deductive item-generation. No specific information on tool development.

Abbreviations: CASP = Critical Appraisal Skills Programme; CCBRG = Cochrane Collaboration Back Review Group; CCT = controlled clinical trial; GATE = Graphical Appraisal Tool for Epidemiological Studies; GRADE = Grading of Recommendations Assessment, Development and Evaluation; HTA = Health Technology Assessment; ICC = intraclass correlation; LEGEND = Let Evidence Guide Every New Decision; NICE = National Institute for Health and Care Excellence; PRECIS = PRagmatic Explanatory Continuum Indicator Summary; RCT = randomized controlled trial; TREND = Transparent Reporting of Evaluations with Nonrandomized Designs; USPSTF = U.S. Preventive Services Task Force

The constructs aimed to be measured by the tools or dimensions of interest are diverse. Two of the tools focus on the characterization of RCTs on an efficacy-effectiveness continuum [68, 69], two tools focus predominantly on the report quality of factors essential to external validity [62, 70] (rather than the external validity itself), 17 tools aim to assess the representativeness, generalizability or applicability of population, setting, intervention, and/or outcome measure to usual practice, and five tools seem to measure a mixture of these different constructs related to external validity [59, 65, 71–73]. However, the construct of interest of most tools was not described adequately (see below).

Measurement properties

The results of the methodological quality assessment according to the COSMIN RoB checklist are detailed in Table 3. The results of the ratings against the updated criteria for good measurement properties and the overall level of evidence, according to the modified GRADE approach, can be seen in Table 4. The detailed grading is described in Table S4 [see Additional file 4].

Table 4
Criteria for good measurement properties & level of evidence according to the modified GRADE method

Tool or dimension	Content validity	Internal consistency	Reliability	Measurement error	Criterion validity	Construct validity
"Applicability"-dimension of LEGEND [74]						
Quality criteria	(?)					
Level of evidence	Low					
“Applicability”-dimension of Carr´s evidence-grading scheme [55]						
Quality criteria	(?)					
Level of evidence	Very Low					
Bornhöft´s checklist [98]						
Quality criteria	(?)					
Level of evidence	Very Low					
Cleggs´s external validity assessment [56]						
Quality criteria	(?)					
Level of evidence	Very Low					
Clinical Applicability [59]						
Quality criteria	(?)					
Level of evidence	Very Low					
Clinical Relevance Instrument [72]						
Quality criteria	(?)		(-)			
Level of evidence	Moderate		Very Low			
Clinical Relevance according to the CCBRG [78]						
Quality criteria	(?)					
Level of evidence	Moderate					
Clinical relevance scores [61]						
Quality criteria	(?)					
Level of evidence	Very Low					
External Validity Assessment Tool (EVAT) [75]						
Quality criteria	(?)					
Level of evidence	Low					
"External validity"-dimension of the Downs & Black Checklist [19]						

Tool or dimension	Content validity	Internal consistency	Reliability	Measurement error	Criterion validity	Construct validity
Quality criteria	(?)	(?)	(±)	(?)		(-)
Level of evidence	Moderate	Very Low	Low	Very Low		Very Low
“External validity”-dimension of Foy´s quality checklist [58]						
Quality criteria	(?)					
Level of evidence	Very Low					
"External validity"-dimension of Liberati´s quality assessment criteria [62]						
Quality criteria	(?)					
Level of evidence	Very Low					
“External validity”-dimension of Sorg´s checklist [64]						
Quality criteria	(?)					
Level of evidence	Very Low					
“External validity”-criteria of the USPSTF manual [66]						
Quality criteria	(?)		(+)			
Level of evidence	Very Low		Very Low			
Feasibility, Appropriateness, Meaningfulness and Effectiveness (FAME) scale [63]						
Quality criteria	(?)					
Level of evidence	Very Low					
Generalizability, Applicability and Predictability (GAP) checklist [57]						
Quality criteria	(?)					
Level of evidence	Very Low					
Gartlehner´s tool [68]						
Quality criteria	(?)		(-)	(?)	(+)	
Level of evidence	Very Low		Moderate	Very Low	Very Low	
Green & Glasgow´s external validity quality rating criteria [70]						
Quality criteria	(?)		(+)			(?)
Level of evidence	Very Low		Very Low			Very Low
"Indirecntess"-dimension from the GRADE Handbook [87]						
Quality criteria	(?)		(-)			

Tool or dimension	Content validity	Internal consistency	Reliability	Measurement error	Criterion validity	Construct validity
Level of evidence	Moderate		Very Low			
modified "Indirectness" of the Checklist for GRADE [76]						
Quality criteria	(?)		(±)			
Level of evidence	Low		Very Low			
External validity checklist of the National Health & Medical Research Council (NHMRC) Handbook [67]						
Quality criteria	(?)					
Level of evidence	Very Low					
revised Graphical Appraisal Tool for Epidemiological Studies (GATE) [65]						
Quality criteria	(?)					
Level of evidence	Very Low					
Rating of Included Trials on the Efficacy-Effectiveness Spectrum (RITES) [69]						
Quality criteria	(+)		(+)			(+)
Level of evidence	Moderate		Very Low			Low
“Selection Bias”-dimension (Section A) of EPHPP [73]						
Quality criteria	(?)		(-)			(+)
Level of evidence	Moderate		Low			Very Low
Section C of the CASP checklist for RCTs [102]						
Quality criteria	(?)					
Level of evidence	Very Low					
Whole Systems research considerations´checklist [60]						
Quality criteria	(?)					
Level of evidence	Very Low					

Abbreviations: CCBRG = Cochrane Collaboration Back Review Group; EPHPP = Effective Public Health Practice Project; GRADE = Grading of Recommendations Assessment, Development and Evaluation; LEGEND = Let Evidence Guide Every New Decision; NICE = National Institute for Health and Care Excellence; USPSTF = U.S. Preventive Services Task Force;

Criteria for good measurement properties: (+) = sufficient; (?) = indeterminate; (-) = insufficient, (±) or inconsistent;

Level of evidence according to the modified GRADE approach: high, moderate, low, or very low evidence.

Note: the measurement properties “structural validity” and “cross-cultural validity” are not presented in this table, since they were not assessed in any of the included studies.

Fields left blank indicate that those measurement properties were not assessed by the study authors.

Content validity

The methodological quality of the development process was “inadequate” for 20 (77%) of the included tools. This was mainly due to insufficient description of the construct to be measured, the target population, or missing pilot tests. Three development studies had a “doubtful” methodological quality [19, 74, 75] and three had an “adequate” methodological quality [69, 76, 77].

There was evidence for content validation of five tools [19, 69, 72, 73, 78]. However, the methodological quality of the content validity studies was “adequate” and “very good” only for the Rating of Included Trials on the Efficacy-Effectiveness Spectrum (RITES) tool [69] and “doubtful” for the Clinical Relevance Instrument [72], the “external validity”-dimension of the Downs & Black-checklist [19], the “Selection Bias”-dimension of the Effective Public Health Practice Project (EPHPP) tool [73], and the “Clinical Relevance” tool [78]. The overall level of evidence for content validity was “very low” for 17 tools, “low” for three tools and “moderate” for six tools. All but one tool had an “indeterminate” content validity. The RITES tool [69] had “moderate” quality of evidence for “sufficient” content validity.

Internal consistency

Internal consistency was assessed for one tool (“external validity”-dimension of the Downs & Black-checklist) [19]. The methodological quality was “doubtful” due to a lack of evidence on unidimensionality or structural validity. Thus, this tool had a “very low” quality of evidence for “indeterminate” internal consistency. Reasons for downgrading were a very serious risk of bias and imprecision.

Reliability

Out of 13 studies assessing the reliability of 9 tools, eleven evaluated the interrater reliability [68, 69, 84, 72, 76, 77, 79–83], one the test-retest reliability [73], and one evaluated both [19]. Three studies had an “inadequate” [69, 77, 85], two had a “doubtful” [73, 83], four had an “adequate” [72, 76, 81, 86], and four had a “very good” methodological quality [19, 68, 79, 80]. The overall quality of evidence was “very low” for six tools [66, 69, 70, 72, 76, 87]. The quality of evidence was “low” for the “external validity”-dimension of the Downs & Black-checklist as well as for the “Selection Bias”-dimension of the EPHPP tool [19, 73] and “moderate” for Gartlehner´s tool [68]. Reasons for downgrading were risk of bias, indirectness and imprecision due to small sample sizes.

Out of nine evaluated tools, the Downs & Black-checklist [19] and the modified indirectness-checklist [76] had “inconsistent” results on reliability [19, 76]. The Clinical Relevance Instrument [72], Gartlehner´s tool [68], the “Selection Bias”-dimension of the EPHPP [73] and the indirectness-dimension of the GRADE handbook [87] had an “insufficient” rating for reliability [68, 72, 73, 87]. Green & Glasgow´s tool [70], the external validity dimension of the U.S. Preventive Services Task Force (USPSTF) manual [66] and the RITES tool [69] had a “very low” quality of evidence for “sufficient” reliability.

Measurement error

Measurement error was reported for two tools. Gartlehner´s tool [68] had an “adequate” and the Downs & Black-checklist [19] had an “inadequate” methodological quality. However, both had a “very low” quality of evidence for “indeterminate” measurement error. Reasons for downgrading were risk of bias, indirectness, and imprecision due to small sample sizes.

Criterion validity

Criterion validity was reported only for Gartlehner´s tool [68]. Although there was no gold standard available to assess the criterion validity of this tool, the authors used expert opinion as the reference standard. This measurement property had an “adequate” methodological quality, but was downgraded to a “very low” quality of evidence for “sufficient” criterion validity due to risk of bias, imprecision, and indirectness.

Construct validity

Five studies [19, 73, 82, 84, 86] reported on the construct validity of four tools. Three studies had a “doubtful” [73, 84, 86], one had an “adequate” [19] and one had a “very good” [82] methodological quality. The overall level of evidence was “very low” for

three tools and “low” for one tool. The “Selection-Bias”-dimension of the EPHPP tool [73] had “very low” quality of evidence for “sufficient” construct validity and the RITES tool [69] had “low” quality of evidence for “sufficient” construct validity. Green & Glasgow´s tool [70] had “very low” quality of evidence for “indeterminate” and the Downs & Black-checklist [19] had an “insufficient” construct validity.

Structural validity and cross-cultural validity were not assessed in any of the included studies.

Discussion

Summary and interpretation of results

To our knowledge this is the first systematic review identifying and evaluating the measurement properties of tools to assess the external validity of RCTs. A total of 26 tools were included. Overall, for more than half (62%) of the included tools the measurement properties were not reported. Only five tools had at least one “sufficient” measurement property. Moreover, the development process was not described in 13 (50%) of the included tools. The RITES tool [69] was the measurement tool with the strongest evidence for validity and reliability. Its content validity, based on international expert-consensus, was “sufficient” with “moderate” quality of evidence, while reliability and construct validity were rated as “sufficient” with “very low” and “low” quality of evidence, respectively.

According to the COSMIN guidelines, there are three criteria for the recommendation of a measurement tool: (A) “Evidence for sufficient content validity (any level) and at least low-quality evidence for sufficient internal consistency” for a tool to be recommended; (B) tool “categorized not in A or C” and further research on the quality of this tool is required to be recommended; and (C) tool with “high quality evidence for an insufficient psychometric property” and this tool should not be recommended [22]. Following these criteria, all included tools were categorized as ‘B’. This is interpreted as “further research is required before any recommendation for or against any of the included tools can be given” according to COSMIN [22]. Sufficient internal consistency may not be relevant for the assessment of external validity, as the measurement models might not be fully reflective. However, none of the authors/developers did specify the measurement model of their measurement tool.

Specification of the measurement model is considered a requirement of the appropriateness for the latent construct of interest during scale or tool development [88]. It could be argued that researchers automatically expect their tool to be a reflective measurement model. E.g., Downs and Black [19] assessed internal consistency without prior testing for unidimensionality or structural validity of the tool. Structural validity or unidimensionality is a prerequisite for internal consistency [22] and both measurement properties are only relevant for reflective measurement models [89, 90]. Misspecification as well as lack of specification of the measurement model can lead to potential limitations when developing and validating a scale or tool [88, 91]. Hence, the specification of measurement models should be considered in future research.

Content validity is the most important measurement property of health measurement instruments [23] and a lack of face validity is considered a strong argument for not using or to stop further evaluation of a measurement instrument [92]. Only the RITES tool [69] had evidence of “sufficient” content validity. Nevertheless, this tool does not directly measure the external validity of RCTs. The RITES tool [69] was developed to classify RCTs on an efficacy-effectiveness continuum. An RCT categorized as highly pragmatic or as having a “strong emphasis on effectiveness” [69] implies that the study design provides rather applicable results, but it does not automatically imply high external validity or generalizability of a trial´s characteristics to other specific contexts and settings [93]. Even a highly pragmatic/effectiveness study might have little applicability or generalizability to a specific research question of review authors. An individual assessment of external validity may still be needed by review authors in accordance with the research question and other contextual factors.

Another tool which might have some degree of content or face validity is the indirectness-dimension of the GRADE method [87]. This method is a widely used and accepted method in research synthesis in health science [94]. It has been evolved over the years based on work from the GRADE Working Group and on feedback from users worldwide [94]. Thus, it might be

assumed that this method has a high degree of face validity, although it has not been systematically tested for content validity.

If all tools are categorized as 'B' in a review, the COSMIN guidelines suggests that the measurement instrument "with the best evidence for content validity could be the one to be provisionally recommended for use, until further evidence is provided." [28]. In accordance with this suggestions, the use of the RITES tool [69] as an provisionally solution might therefore be justified until more research on this topic is available. However, users should be aware of its limitations, as described above.

Implication for future research

This study affirms and supplements what is already known from previous reviews [7, 9, 11–15]. The heterogeneity of characteristics of tools included in those reviews was also observed in the present review. Although Dyrvig et al. [13] did not assess the measurement properties of available tools, they reported a lack of empirical support of items included in measurement tools.

One major challenge on this topic is the serious heterogeneity regarding the terminology, criteria and guidance to assess the external validity of RCTs. Development of new tools and/or further revision (and validation) of available tools may not be appropriate before consensus-based standards are developed. Generally, it may be argued whether these methods to assess the external validity in systematic reviews of interventions are suitable [7, 9]. The experimental/statistical methods presented in Table 1 may offer a more objective approach to evaluate the external validity of RCTs. However, they are not feasible to implement in the conduction of systematic reviews. Furthermore, they focus mainly on the characteristics and generalizability of the study populations, which is insufficient to assess the external validity of clinical trials [95], since they do not consider other relevant dimensions of external validity such as intervention settings or treatment variables etc. [3, 95].

The methodological possibilities in tool/scale development and validation regarding this topic have not been exploited, yet. More than 20 years ago, there was no consensus regarding the definition of quality of RCTs. In 1998, Verhagen et al. [4] performed a Delphi study to achieve consensus regarding the definition of quality of RCTs and to create a quality criteria list. Until now, these criteria list has been a guidance in tool development and their criteria are still being implemented in methodological quality or risk of bias assessment tools (e.g. the Cochrane Collaboration risk of bias tool 1 & 2.0, the Physiotherapy Evidence Database (PEDro) scale etc.). Consequently, it seems necessary to seek consensus in order to overcome the issues regarding the external validity of RCTs in a similar way. After reaching consensus, further development and validation is needed following standard guidelines for scale/tool development (e.g. de Vet et al. [92]; Streiner et al. [96]; DeVellis [97]). Since the assessment of external validity seems highly context-dependent [7, 9], this should be taken into account in future research. A conventional checklist approach seems inappropriate [7, 9, 95] and a more comprehensive but flexible approach might be necessary. The experimental/statistical methods (Table 1) may offer a reference standard for convergent validity testing of the dimension "patient population" in future research.

This review has highlighted the necessity for more research in this area. Published studies and evaluation tools are important sources of information and should inform the development of a new tool or approach.

Limitations

There are some limitations of the present review. First, a search for grey literature was not performed. Second, we focused on RCTs only and did not include assessment tools for non-randomized or other observational study design. Third, authors were not contacted if there was missing data. However, there was only one study [77] with missing data for our analysis. Fourth, due to heterogeneity in terminology, we might have missed some tools with our electronic literature search strategy. We tried to address this potential limitation by performing a comprehensive reference screening and snowballing.

Conclusions

Based on the results of this review, no available measurement tool can be fully recommended for the use in systematic reviews to assess the external validity of RCTs. Several steps are required to overcome the identified difficulties before a new

tool is developed or available tools are further revised and validated.

Abbreviations

CASP = Critical Appraisal Skills Programme; CCBRG = Cochrane Collaboration Back Review Group; CCT = controlled clinical trial; COSMIN = COnsensus based Standards for the selection of health Measurement Instruments; EPHPP = Effective Public Health Practice Project; EVAT = External Validity Assessment Tool; FAME = Feasibility, Appropriateness, Meaningfulness and Effectiveness; GATE = Graphical Appraisal Tool for Epidemiological Studies; GAP = Generalizability, Applicability and Predictability; GRADE = Grading of Recommendations Assessment, Development and Evaluation; HTA = Health Technology Assessment; ICC = intraclass correlation; LEGEND = Let Evidence Guide Every New Decision; NICE = National Institute for Health and Care Excellence; PEDro = Physiotherapy Evidence Database; PRECIS = PRagmatic Explanatory Continuum Indicator Summary; RCT = randomized controlled trial; RITES = Rating of Included Trials on the Efficacy-Effectiveness Spectrum; TREND = Transparent Reporting of Evaluations with Nonrandomized Designs; USPSTF = U.S. Preventive Services Task Force

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data generated or analyzed during this study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

Open Access funding enabled and organized by Projekt DEAL.

The authors do not receive any financial support for the research, authorship, and/or publication of this research project.

Authors' contributions

All authors contributed to the design of the study. AJ designed the search strategy and conducted the systematic search. AJ and TB screened titles and abstracts as well as full-text reports in phase 1. AJ and KL screened titles and abstracts as well as full-text reports in phase 2. Data extraction was performed by AJ and checked by TB. Quality appraisal and data analysis was performed by AJ and JB. AJ drafted the manuscript. JB, TB and KL critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Sven Bossmann and Sarah Tiemann for their assistance with the elaboration of the search strategy.

References

1. Bastian H, Glasziou P, Chalmers I (2010) Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 7:e1000326. https://doi.org/10.1371/journal.pmed.1000326.
2. Mulrow CD (1994) Rationale for systematic reviews. BMJ 309:597–599. https://doi.org/10.1136/bmj.309.6954.597.
3. Jüni P, Altman DG, Egger M (2001) Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 323:42–46. https://doi.org/10.1136/bmj.323.7303.42.
4. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Bouter LM, Knipschild PG (1998) The Delphi List: A Criteria List for Quality Assessment of Randomized Clinical Trials for Conducting Systematic Reviews Developed by Delphi Consensus. J Clin Epidemiol 51:1235–1241. https://doi.org/10.1016/S0895-4356(98)00131-0.
5. Boutron I, Page MJ, Higgins JPT, Altman DG, Lundh A, Hróbjartsson A, Group CBM (2021) Considering bias and conflicts of interest among the included studies. Cochrane Handb. Syst. Rev. Interv. version 6.2 (updated Febr. 2021). . Accesed 15 Apr 2021.
6. Cook TD, Campbell DT, Shadish W (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin Boston, MA.
7. Weise A, Büchter R, Pieper D, Mathes T (2020) Assessing context suitability (generalizability, external validity, applicability or transferability) of findings in evidence syntheses in healthcare-An integrative review of methodological guidance. Res Synth Methods 11:760–779. https://doi.org/10.1002/jrsm.1453.
8. Schunemann HJ, Tugwell P, Reeves BC, Akl EA, Santesso N, Spencer FA, Shea B, Wells G, Helfand M (2013) Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. Res Synth Methods 4:49–62. https://doi.org/10.1002/jrsm.1078.
9. Burchett HED, Blanchard L, Kneale D, Thomas J (2018) Assessing the applicability of public health intervention evaluations from one setting to another: a methodological study of the usability and usefulness of assessment tools and frameworks. Heal Res policy Syst 16:88. https://doi.org/10.1186/s12961-018-0364-3.
10. Dekkers OM, von Elm E, Algra A, Romijn JA, Vandenbroucke JP (2010) How to assess the external validity of therapeutic trials: a conceptual approach. Int J Epidemiol 39:89–94. https://doi.org/10.1093/ije/dyp174.
11. Burchett H, Umoquit M, Dobrow M (2011) How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. J Health Serv Res Policy 16:238–244. https://doi.org/10.1258/jhsrp.2011.010124.
12. Cambon L, Minary L, Ridde V, Alla F (2012) Transferability of interventions in health education: a review. BMC Public Health 12:497. https://doi.org/10.1186/1471-2458-12-497.
13. Dyrvig A-K, Kidholm K, Gerke O, Vondeling H (2014) Checklists for external validity: a systematic review. J Eval Clin Pract 20:857–864. https://doi.org/10.1111/jep.12166.
14. Munthe-Kaas H, Nøkleby H, Nguyen L (2019) Systematic mapping of checklists for assessing transferability. Syst Rev 8:22. https://doi.org/10.1186/s13643-018-0893-4.
15. Nasser M, van Weel C, van Binsbergen JJ, van de Laar FA (2012) Generalizability of systematic reviews of the effectiveness of health care interventions to primary health care: concepts, methods and future research. Fam Pract 29 Suppl 1:i94–i103. https://doi.org/10.1093/fampra/cmr129.
16. Hariton E, Locascio JJ (2018) Randomised controlled trials - the gold standard for effectiveness research: Study design: randomised controlled trials. BJOG 125:1716. https://doi.org/10.1111/1471-0528.15199.
17. Pressler TR, Kaizar EE (2013) The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. Stat Med 32:3552–3568. https://doi.org/10.1002/sim.5802.
18. Rothwell PM (2005) External validity of randomised controlled trials: “to whom do the results of this trial apply?” Lancet 365:82–93. https://doi.org/10.1016/S0140-6736(04)17670-8.
19. Downs SH, Black N (1998) The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 52:377–384. https://doi.org/10.1136/jech.52.6.377.

20. Page MJ, Moher D, Bossuyt PM, et al (2021) PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 372:n160. https://doi.org/10.1136/bmj.n160.
21. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, Terwee CB (2018) COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual life Res an Int J Qual life Asp Treat care Rehabil 27:1171–1179. https://doi.org/10.1007/s11136-017-1765-4.
22. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, Terwee CB (2018) COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 27:1147–1157. https://doi.org/10.1007/s11136-018-1798-3.
23. Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, Bouter LM, de Vet HCW, Mokkink LB (2018) COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual life Res an Int J Qual life Asp Treat care Rehabil 27:1159–1170. https://doi.org/10.1007/s11136-018-1829-0.
24. Glover PD, Gray H, Shanmugam S, McFadyen AK (2021) Evaluating collaborative practice within community-based integrated health and social care teams: a systematic review of outcome measurement instruments. J Interprof Care 1–15. https://doi.org/10.1080/13561820.2021.1902292.
25. Maassen SM, Weggelaar Jansen AMJW, Brekelmans G, Vermeulen H, van Oostveen CJ (2020) Psychometric evaluation of instruments measuring the work environment of healthcare professionals in hospitals: a systematic literature review. Int J Qual Heal care J Int Soc Qual Heal Care 32:545–557. https://doi.org/10.1093/intqhc/mzaa072.
26. Yaqoob Mohammed Al Jabri F, Kvist T, Azimirad M, Turunen H (2021) A systematic review of healthcare professionals' core competency instruments. Nurs Health Sci 23:87–102. https://doi.org/10.1111/nhs.12804.
27. Jung A, Balzer J, Braun T, Luedtke K (2020) Psychometric properties of tools to measure the external validity of randomized controlled trials: a systematic review protocol. https://doi.org/10.17605/OSF.IO/PTG4D
28. Mokkink LB, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, de Vet HCW, Terwee CB (2018) COSMIN manual for systematic reviews of PROMs, user manual. 1–78. https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018-1.pdf. Accessed 3 Feb 2020.
29. Clark R, Locke M, Hill B, Wells C, Bialocerkowski A (2017) Clinimetric properties of lower limb neurological impairment tests for children and young people with a neurological condition: A systematic review. PLoS One 12:e0180031. https://doi.org/10.1371/journal.pone.0180031.
30. Terwee CB, Jansma EP, Riphagen II, De Vet HCW (2009) Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 18:1115–1123. https://doi.org/10.1007/s11136-009-9528-5.
31. Sierevelt IN, Zwiers R, Schats W, Haverkamp D, Terwee CB, Nolte PA, Kerkhoffs GMMJ (2018) Measurement properties of the most commonly used Foot- and Ankle-Specific Questionnaires: the FFI, FAOS and FAAM. A systematic review. Knee Surg Sports Traumatol Arthrosc 26:2059–2073. https://doi.org/10.1007/s00167-017-4748-7.
32. van der Hout A, Neijenhuijs KI, Jansen F, et al (2019) Measuring health-related quality of life in colorectal cancer patients: systematic review of measurement properties of the EORTC QLQ-CR29. Support care cancer Off J Multinatl Assoc Support Care Cancer 27:2395–2412. https://doi.org/10.1007/s00520-019-04764-7.
33. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A (2016) Rayyan-a web and mobile app for systematic reviews. Syst Rev 5:210. https://doi.org/10.1186/s13643-016-0384-4.
34. Atkins D, Chang SM, Gartlehner G, Buckley DI, Whitlock EP, Berliner E, Matchar D (2011) Assessing applicability when comparing medical interventions: AHRQ and the Effective Health Care Program. J Clin Epidemiol 64:1198–1207. https://doi.org/10.1016/j.jclinepi.2010.11.021.
35. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG (2012) CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg 10:28–55. https://doi.org/10.1016/j.ijsu.2011.10.001.

36. Mokkink LB, Terwee CB (2010) The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. 539–549. https://doi.org/10.1007/s11136-010-9606-8.
37. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW (2010) The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 63:737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006.
38. Terwee CB, Prinsen CA, Chiarotto A, De Vet H, Bouter LM, Alonso J, Westerman MJ, Patrick DL, Mokkink LB (2018) COSMIN methodology for assessing the content validity of PROMs–user manual. Amsterdam VU Univ. Med. Cent. https://cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf. Accessed 3 Feb 2020.
39. Hartling L, Fernandes RM, Seida J, Vandermeer B, Dryden DM (2012) From the trenches: a cross-sectional study applying the GRADE tool in systematic reviews of healthcare interventions. PLoS One 7:e34697. https://doi.org/10.1371/journal.pone.0034697.
40. Mustafa RA, Santesso N, Brozek J, et al (2013) The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. J Clin Epidemiol 66:735–736. https://doi.org/10.1016/j.jclinepi.2013.02.004.
41. Abraham NS, Wieczorek P, Huang J, Mayrand S, Fallone CA, Barkun AN (2004) Assessing clinical generalizability in sedation studies of upper GI endoscopy. Gastrointest Endosc 60:28–33. https://doi.org/10.1016/S0016-5107(04)01307-0.
42. Arabi YM, Cook DJ, Zhou Q, et al (2015) Characteristics and Outcomes of Eligible Nonenrolled Patients in a Mechanical Ventilation Trial of Acute Respiratory Distress Syndrome. Am J Respir Crit Care Med 192:1306–1313. https://doi.org/10.1164/rccm.201501-0172OC.
43. Williams AC de C, Nicholas MK, Richardson PH, Pither CE, C. FAC de (1999) Generalizing from a controlled trial: The effects of patient preference versus randomization on the outcome of inpatient versus outpatient chronic pain management. Pain 83:57–65. https://doi.org/10.1016/S0304-3959(99)00074-3.
44. De Jong Z, Munneke M, Jansen LM, Ronday K, Van Schaardenburg DJ, Brand R, Van Den Ende CHM, Vliet Vlieland TPM, Zuijderduin WM, Hazes JMW (2004) Differences between participants and nonparticipants in an exercise trial for adults with rheumatoid arthritis. Arthritis Care Res 51:593–600. https://doi.org/10.1002/art.20531.
45. Hordijk-Trion M, Lenzen M, Wijns W, et al (2006) Patients enrolled in coronary intervention trials are not representative of patients in clinical practice: Results from the Euro Heart Survey on Coronary Revascularization. Eur Heart J 27:671–678. https://doi.org/10.1093/eurheartj/ehi731.
46. Wilson A, Parker H, Wynn A, Spiers N (2003) Performance of hospital-at-home after a randomised controlled trial. J Heal Serv Res Policy 8:160–164. https://doi.org/10.1258/135581903322029511.
47. Smyth B, Haber A, Trongtrakul K, Hawley C, Perkovic V, Woodward M, Jardine M (2019) Representativeness of Randomized Clinical Trial Cohorts in End-stage Kidney Disease: A Meta-analysis. JAMA Intern Med 179:1316–1324. https://doi.org/10.1001/jamainternmed.2019.1501.
48. Leinonen A, Koponen M, Hartikainen S (2015) Systematic Review: Representativeness of Participants in RCTs of Acetylcholinesterase Inhibitors. PLoS One 10:e0124500–e0124500. https://doi.org/10.1371/journal.pone.0124500.
49. Chari A, Romanus D, Palumbo A, Blazer M, Farrelly E, Raju A, Huang H, Richardson P (2020) Randomized Clinical Trial Representativeness and Outcomes in Real-World Patients: Comparison of 6 Hallmark Randomized Clinical Trials of Relapsed/Refractory Multiple Myeloma. Clin Lymphoma Myeloma Leuk 20:8. https://doi.org/10.1016/j.clml.2019.09.625.
50. Susukida R, Crum RM, Ebnesajjad C, Stuart EA, Mojtabai R (2017) Generalizability of findings from randomized controlled trials: application to the National Institute of Drug Abuse Clinical Trials Network. Addiction 112:1210–1219. https://doi.org/10.1111/add.13789.
51. Zarin DA, Young JL, West JC (2005) Challenges to evidence-based medicine: a comparison of patients and treatments in randomized controlled trials with patients and treatments in a practice research network. Soc Psychiatry Psychiatr

Epidemiol 40:27–35. https://doi.org/10.1007/s00127-005-0838-9.

52. Gheorghe A, Roberts T, Hemming K, Calvert M (2015) Evaluating the Generalisability of Trial Results: Introducing a Centre- and Trial-Level Generalisability Index. Pharmacoeconomics 33:1195–1214. https://doi.org/10.1007/s40273-015-0298-3.
53. He Z, Wang S, Borhanian E, Weng C (2015) Assessing the Collective Population Representativeness of Related Type 2 Diabetes Trials by Combining Public Data from ClinicalTrials.gov and NHANES. Stud Health Technol Inform 216:569–573. https://doi.org/10.3233/978-1-61499-564-7-569.
54. Schmidt AF, Groenwold RHH, van Delden JJM, van der Does Y, Klungel OH, Roes KCB, Hoes AW, van der Graaf R (2014) Justification of exclusion criteria was underreported in a review of cardiovascular trials. J Clin Epidemiol 67:635–644. https://doi.org/10.1016/j.jclinepi.2013.12.005.
55. Carr DB, Goudas LC, Balk EM, Bloch R, Ioannidis JP, Lau J (2004) Evidence report on the treatment of pain in cancer patients. J Natl Cancer Inst Monogr 23–31. https://doi.org/10.1093/jncimonographs/lgh012.
56. Clegg A, Bryant J, Nicholson T, et al (2001) Clinical and cost-effectiveness of donepezil, rivastigmine and galantamine for Alzheimer's disease: a rapid and systematic review. Health Technol Assess (Rockv) 5:1–136.
57. Fernandez-Hermida JR, Calafat A, Becoña E, Tsertsvadze A, Foxcroft DR, Fernandez-Hermida JR, Calafat A, Becoña E, Tsertsvadze A, Foxcroft DR (2012) Assessment of generalizability, applicability and predictability (GAP) for evaluating external validity in studies of universal family-based prevention of alcohol misuse in young people: systematic methodological review of randomized controlled trials. Addiction 107:1570–1579. https://doi.org/10.1111/j.1360-0443.2012.03867.x.
58. Foy R, Hempel S, Rubenstein L, Suttorp M, Seelig M, Shanman R, Shekelle PG (2010) Meta-analysis: effect of interactive communication between collaborating primary care physicians and specialists. Ann Intern Med 152:247–258. https://doi.org/10.7326/0003-4819-152-4-201002160-00010.
59. Haraldsson BG, Gross AR, Myers CD, Ezzo JM, Morien A, Goldsmith C, Peloso PM, Bronfort G (2006) Massage for mechanical neck disorders. Cochrane database Syst Rev CD004871. https://doi.org/10.1002/14651858.CD004871.pub3.
60. Hawk C, Khorsan R, AJ L, RJ F, MW E (2007) Chiropractic care for nonmusculoskeletal conditions: a systematic review with implications for whole systems research. J Altern Complement Med 13:491–512. https://doi.org/10.1089/acm.2007.7088.
61. Karjalainen K, Malmivaara A, van Tulder M, et al (2000) Multidisciplinary rehabilitation for fibromyalgia and musculoskeletal pain in working age adults. Cochrane Database Syst Rev. https://doi.org/10.1002/14651858.CD001984
62. Liberati A, Himel HN, Chalmers TC (1986) A quality assessment of randomized control trials of primary treatment of breast cancer. J Clin Oncol 4:942–951. https://doi.org/10.1200/JCO.1986.4.6.942.
63. Averis A, Pearson A (2003) Filling the gaps: identifying nursing research priorities through the analysis of completed systematic reviews. Jbi Reports 1:49–126. https://doi.org/10.1046/j.1479-6988.2003.00003.x.
64. Sorg C, Schmidt J, Büchler MW, Edler L, Märten A (2009) Examination of external validity in randomized controlled trials for adjuvant treatment of pancreatic adenocarcinoma. Pancreas 38:542–550. https://doi.org/10.1097/MPA.0b013e31819d7370.
65. National Institute for Health and Care Excellence. (2012) Methods for the development of NICE public health guidance, Third edit. National Institute for Health and Care Excellence. https://www.nice.org.uk/process/pmg4/chapter/introduction. Accessed 15 Apr 2020
66. U.S. Preventive Services Task Force (2017) Criteria for Assessing External Validity (Generalizability) of Individual Studies. US Prev Serv Task Force Appendix VII. https://uspreventiveservicestaskforce.org/uspstf/about-uspstf/methods-and-processes/procedure-manual/procedure-manual-appendix-vii-criteria-assessing-external-validity-generalizability-individual-studies. Accessed 15 Apr 2020.
67. National Health and Medical Research Council NHMRC handbooks. https://www.nhmrc.gov.au/about-us/publications/how-prepare-and-present-evidence-based-information-consumers-health-services#block-views-block-file-attachments-content-block-1. Accessed 15 Apr 2020.

68. Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS (2006) A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 59:1040–1048. https://doi.org/10.1016/j.jclinepi.2006.01.011.
69. Wieland LS, Berman BM, Altman DG, et al (2017) Rating of Included Trials on the Efficacy-Effectiveness Spectrum: development of a new tool for systematic reviews. J Clin Epidemiol 84:95–104. https://doi.org/10.1016/j.jclinepi.2017.01.010.
70. Green LW, Glasgow RE (2006) Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Eval Health Prof 29:126–153. https://doi.org/10.1177/0163278705284445.
71. KA K, Malmivaara A, MW van T, Roine R, Jauhiainen M, Hurri H, BW K (2000) Multidisciplinary rehabilitation for fibromyalgia and musculoskeletal pain in working age adults. Cochrane Database Syst Rev. https://doi.org/10.1002/14651858.CD001984
72. Cho MK, Bero LA (1994) Instruments for assessing the quality of drug studies published in the medical literature. JAMA J Am Med Assoc 272:101–104. https://doi.org/10.1001/jama.272.2.101.
73. Thomas BH, Ciliska D, Dobbins M, Micucci S (2004) A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. Worldviews evidence-based Nurs 1:176–184. https://doi.org/10.1111/j.1524-475X.2004.04006.x.
74. Clark E, Burkett K, Stanko-Lopp D (2009) Let Evidence Guide Every New Decision (LEGEND): an evidence evaluation system for point-of-care clinicians and guideline development teams. J Eval Clin Pract 15:1054–1060. https://doi.org/10.1111/j.1365-2753.2009.01314.x.
75. Khorsan R, Crawford C (2014) How to assess the external validity and model validity of therapeutic trials: a conceptual approach to systematic review methodology. Evid Based Complement Alternat Med 2014:694804. https://doi.org/10.1155/2014/694804.
76. Meader N, King K, Llewellyn A, Norman G, Brown J, Rodgers M, Moe-Byrne T, Higgins JPT, Sowden A, Stewart G (2014) A checklist designed to aid consistency and reproducibility of GRADE assessments: Development and pilot validation. Syst Rev. https://doi.org/10.1186/2046-4053-3-82. https://doi.org/10.1186/2046-4053-3-82.
77. Atkins D, Briss PA, Eccles M, et al (2005) Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. BMC Health Serv Res 5:25. https://doi.org/10.1186/1472-6963-5-25.
78. van Tulder M, Furlan A, Bombardier C, Bouter L (2003) Updated method guidelines for systematic reviews in the cochrane collaboration back review group. Spine (Phila Pa 1976) 28:1290–1299. https://doi.org/10.1097/01.BRS.0000065484.95996.AF.
79. O'Connor SR, Tully MA, Ryan B, Bradley JM, Baxter GD, McDonough SM (2015) Failure of a numerical quality assessment scale to identify potential risk of bias in a systematic review: a comparison study. BMC Res Notes 8:224. https://doi.org/10.1186/s13104-015-1181-1
80. Zettler LL, Speechley MR, Foley NC, Salter KL, Teasell RW (2010) A scale for distinguishing efficacy from effectiveness was adapted and applied to stroke rehabilitation studies. J Clin Epidemiol 63:11–18. https://doi.org/10.1016/j.jclinepi.2009.06.007.
81. Llewellyn A, Whittington C, Stewart G, Higgins JP, Meader N (2015) The Use of Bayesian Networks to Assess the Quality of Evidence from Research Synthesis: 2. Inter-Rater Reliability and Comparison with Standard GRADE Assessment. PLoS One 10:e0123511. https://doi.org/10.1371/journal.pone.0123511.
82. Aves T (2017) The Role of Pragmatism in Explaining Heterogeneity in Meta-Analyses of Randomized Trials: A Methodological Review. McMaster University. http://hdl.handle.net/11375/22212. Accessed 12 Jan 2021.
83. Armijo-Olivo S, Stiles CR, Hagen NA, Biondo PD, Cummings GG (2012) Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. J Eval Clin Pract 18:12–18. https://doi.org/10.1111/j.1365-2753.2010.01516.x.

84. Mirza NA, Akhtar-Danesh N, Staples E, Martin L, Noesgaard C (2014) Comparative Analysis of External Validity Reporting in Non-randomized Intervention Studies. Can J Nurs Res = Rev Can Rech en Sci Infirm 46:47–64. https://doi.org/10.1177/084456211404600405.
85. Aves T, Allan KS, Lawson D, Nieuwlaat R, Beyene J, Mbuagbaw L (2017) The role of pragmatism in explaining heterogeneity in meta-analyses of randomised trials: a protocol for a cross-sectional methodological review. BMJ Open 7:e017887. https://doi.org/10.1136/bmjopen-2017-017887.
86. Laws RA, St George AB, Rychetnik L, Bauman AE (2012) Diabetes prevention research: a systematic review of external validity in lifestyle interventions. Am J Prev Med 43:205–214. https://doi.org/10.1016/j.amepre.2012.04.017.
87. Schünemann H, Brożek J, Guyatt G, Oxman A (2013) Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach (updated October 2013). GRADE Work. Gr. https://gdt.gradepro.org/app/handbook/handbook.html. Accessed 15 Apr 2020.
88. Diamantopoulos A, Riefler P, Roth KP (2008) Advancing formative measurement models. J Bus Res 61:1203–1218. https://doi.org/10.1016/j.jbusres.2008.01.009.
89. Fayers PM, Hand DJ (1997) Factor analysis, causal indicators and quality of life. Qual Life Res. https://doi.org/10.1023/A:1026490117121. https://doi.org/10.1023/A:1026490117121.
90. Streiner DL (2003) Being Inconsistent About Consistency: When Coefficient Alpha Does and Doesn't Matter. J Pers Assess 80:217–222. https://doi.org/10.1207/S15327752JPA8003_01.
91. MacKenzie SB, Podsakoff PM, Jarvis CB (2005) The Problem of Measurement Model Misspecification in Behavioral and Organizational Research and Some Recommended Solutions. J Appl Psychol 90:710–730. https://doi.org/10.1037/0021-9010.90.4.710.
92. De Vet HCW, Terwee CB, Mokkink LB, Knol DL (2011) Measurement in medicine: a practical guide. https://doi.org/10.1017/CBO9780511996214.
93. Dekkers OM, Bossuyt PM, Vandenbroucke JP (2017) How trial results are intended to be used: is PRECIS-2 a step forward? J Clin Epidemiol 84:25–26. https://doi.org/10.1016/j.jclinepi.2016.01.033.
94. Brozek JL, Canelo-Aybar C, Akl EA, et al (2021) GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence-An overview in the context of health decision-making. J Clin Epidemiol 129:138–150. https://doi.org/10.1016/j.jclinepi.2020.09.018.
95. Burchett HED, Kneale D, Blanchard L, Thomas J (2020) When assessing generalisability, focusing on differences in population or setting alone is insufficient. Trials 21:286. https://doi.org/10.1186/s13063-020-4178-6.
96. Streiner DL, Norman GR, Cairney J (2015) Health measurement scales: a practical guide to their development and use, Fifth edit. Oxford University Press, USA, Oxford.
97. DeVellis RF (2017) Scale development: Theory and applications, Fourth edi. Sage publications, Los Angeles.
98. Bornhöft G, Maxion-Bergemann S, Wolf U, Kienle GS, Michalsen A, Vollmar HC, Gilbertson S, Matthiessen PF (2006) Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity. BMC Med Res Methodol 6:56. https://doi.org/10.1186/1471-2288-6-56.
99. Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A (1981) A method for assessing the quality of a randomized control trial. Control Clin Trials 2:31–49. https://doi.org/https://doi.org/10.1016/0197-2456(81)90056-8.
100. Glasgow RE, Vogt TM, Boles SM (1999) Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 89:1322–1327. https://doi.org/10.2105/AJPH.89.9.1322.
101. Jackson R, Ameratunga S, Broad J, Connor J, Lethaby A, Robb G, Wells S, Glasziou P, Heneghan C (2006) The GATE frame: critical appraisal with pictures. Evid Based Med 11:35 LP – 38. https://doi.org/10.1136/ebm.11.2.35.
102. Critical Appraisal Skills Programme (2020) CASP Randomised Controlled Trial Standard Checklist. https://casp-uk.net/casp-tools-checklists/. Accessed 10 Dec 2020.

Tables

Due to technical limitations, table 3 is only available as a download in the Supplemental Files section.

Figures

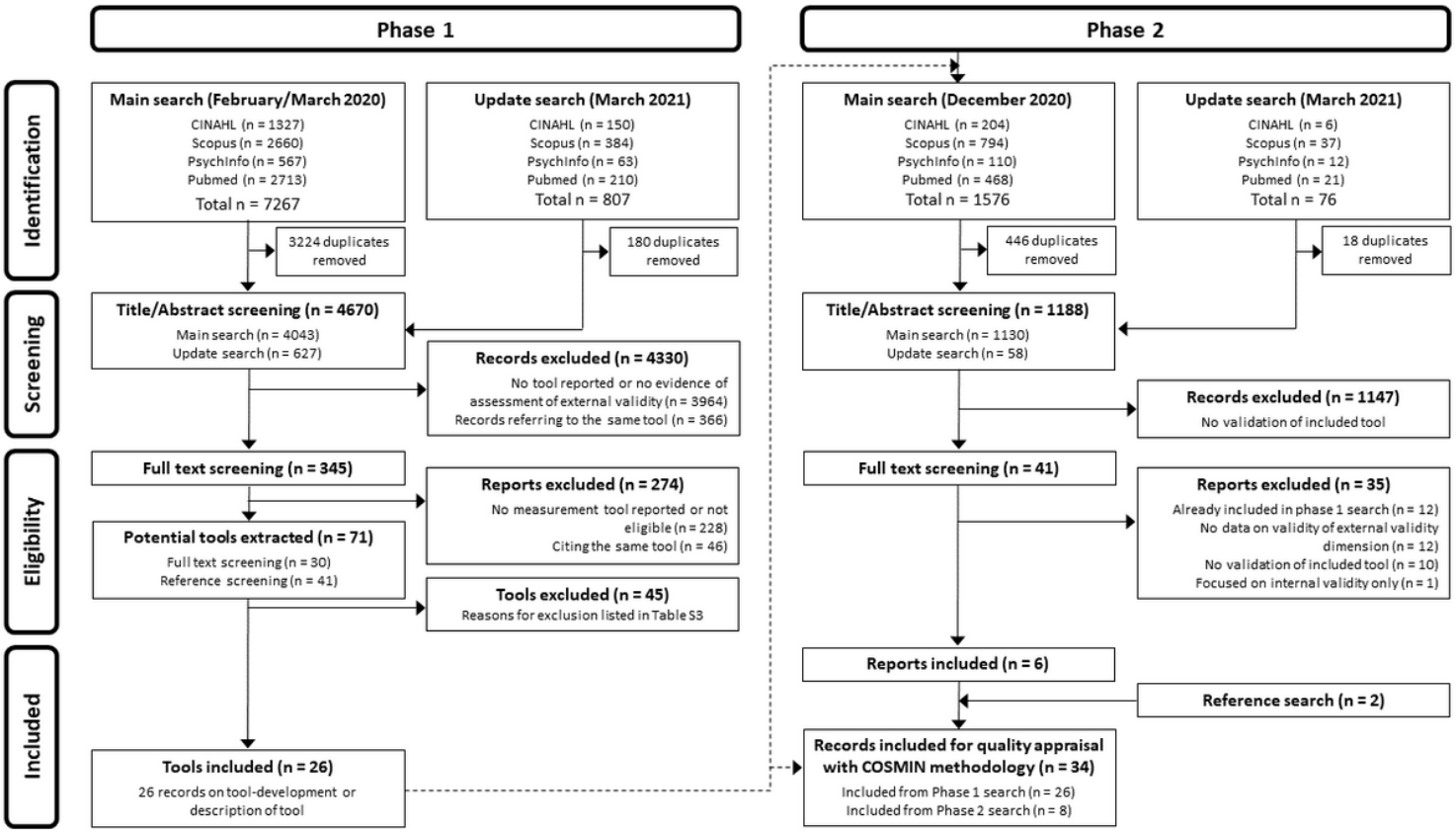

Figure 1

Flow diagram “of systematic search strategy used to identify clinimetric papers”[29]

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1TableS1searchstratgy.docx
- Additionalfile2TableS2COSMINadaptations.docx
- Additionalfile3TableS3excludedtoolsandreports.docx
- Additionalfile4TableS4GRADElevelofevidenceindetail.docx
- Table03.png