Qual Life Res (2009) 18:1115–1123
DOI 10.1007/s11136-009-9528-5

Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments

Caroline B. Terwee · Elise P. Jansma · Ingrid I. Riphagen · Henrica C. W. de Vet

Accepted: 6 August 2009 / Published online: 27 August 2009

Abstract
Objectives For the measurement of patient-reported outcomes, such as (health-related) quality of life, often many measurement instruments exist that intend to measure the same construct. To facilitate instrument selection, our aim was to develop a highly sensitive search filter for finding studies on measurement properties of measurement instruments in PubMed and a more precise search filter that needs less abstracts to be screened, but at a higher risk of missing relevant studies.
Methods A random sample of 10,000 PubMed records (01-01-1990 to 31-12-2006) was used as a gold standard. Studies on measurement properties were identified using an exclusion filter and hand searching. Search terms were selected from the relevant records in the gold standard as well as from 100 systematic reviews of measurement properties and combined based on sensitivity and precision. The performance of the filters was tested in the gold standard as well as in two validation sets, by calculating sensitivity, precision, specificity, and number needed to read.
Results We identified 116 studies on measurement properties in the gold standard. The sensitive search filter was able to retrieve 113 of these 116 studies (sensitivity 97.4%, precision 4.4%). The precise search filter had a sensitivity of 93.1% and a precision of 9.4%. Both filters performed very well in the validation sets.
Conclusion The use of these search filters will contribute to evidence-based selection of measurement instruments in all medical fields.

Keywords Information storage and retrieval · Outcome assessment · Psychometrics · Review literature as topic

Abbreviations

NLM	National Library of Medicine
PMID	PubMed unique identifiers numbers
NNR	Number needed to read
WOMAC	Western Ontario and McMaster Universities Osteoarthritis Index
COSMIN	Consensus-based standards for the selection of health measurement instruments

Introduction

For the measurement of patient-reported outcomes, such as (health-related) quality of life, often many measurement instruments exist that intend to measure the same construct. Choosing an appropriate instrument for a certain purpose should be based on the measurement properties of the available instruments. It is, therefore, important to have easy access to data on the measurement properties of the available instruments. Studies on measurement properties

C. B. Terwee (✉) · H. C. W. de Vet
Department of Epidemiology and Biostatistics and the EMGO Institute for Health and Care Research, VU University Medical Center, van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands
e-mail: cb.terwee@vumc.nl

E. P. Jansma · I. I. Riphagen
Medical Library, VU University, Amsterdam, The Netherlands

I. I. Riphagen
Unit for Applied Clinical Research, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

of measurement instruments are, however, often difficult to find in PubMed. This is due to a number of reasons:

1. Indexing by the National Library of Medicine (NLM) is sometimes incomplete and often unpredictable. There are three specific index terms in PubMed for studies on measurement properties: the publication type term "Validation study" and the MeSH terms "Reproducibility of results" and "Psychometrics". However, in many cases studies on measurement properties are not tagged with these terms. For example, in a systematic review on shoulder disability questionnaires, we identified 26 studies on measurement properties in PubMed [1]. None of them were tagged with the publication type term "Validation study", 13 were tagged with the MeSH term "Reproducibility of results", and two were tagged with the MeSH term "Psychometrics". Three studies were tagged with other related MeSH headings such as "Outcome Assessment (Health Care)" (1 study) and "Sensitivity and Specificity" (two studies). These MeSH headings are, however, not specific for studies on measurement properties. Seven of the 26 studies were not tagged with any MeSH heading relevant for measurement properties.
2. There is large variation in terminology for measurement properties. For example, for reliability, many synonyms can be found in the literature, e.g. reproducibility, repeatability, precision, variability, consistency, dependability, stability, agreement, and measurement error. This makes it difficult to find all studies on the reliability of a measurement instrument.
3. Studies on measurement properties are sometimes poorly reported in the abstract. Some authors do not use any commonly used term for measurement properties in the title or abstract of their study.

There is thus a need for a methodological search filter to find studies on measurement properties in PubMed. A methodological search filter is a combination of search terms designed to retrieve studies with a particular type of study design, in this case studies on measurement properties of measurement instruments. Such a filter does not yet exist. There is one filter available for finding outcome measures [2], but this filter was developed to find measurement instruments, and not studies on the measurement properties of these instruments.

The aim of this study was, therefore, to develop a highly sensitive methodological search filter for finding studies on measurement properties in PubMed. A highly sensitive search filter is especially useful for systematic reviews of measurement properties. A second aim was to develop a more precise search filter for a less extensive search, e.g. for researchers who have to choose a measurement instrument, but do not have the time and resources to perform a systematic review. With a more precise filter, less abstracts need to be screened to find a study on measurement properties, but at a slightly higher risk of missing relevant studies.

Methods

The search filters were developed according to four phases, as described by Jenkins [3] and the UK InterTASC Information Specialists' Sub-Group (ISSG), a group of experienced health care information specialists [4]. The first phase concerns identification of a gold standard set of records to evaluate the search filters. The second phase concerns the selection and combination of search terms to develop the search filters. In the third phase, the search filters are evaluated against the gold standard set of records. And in the fourth phase, the search filters are validated by examining the performance of the filters in a new set of records.

Phase 1: Identification of a gold standard

We selected a random sample of PubMed records as a representation of the literature in which the search filters are going to be used. We performed a power analysis to estimate the required number of records based on the estimated prevalence of studies on measurement properties in PubMed and the desired sensitivity. We selected a random set of 500 PubMed records to estimate the prevalence of studies on measurement properties, which was 1%. With a desired sensitivity of 98% with the lower limit of the confidence interval at 95%, we estimated that the gold standard should contain 100 relevant studies. This meant that we had to select a random sample of 10,000 PubMed records as our gold standard. The sample was drawn based on random PubMed unique identifiers numbers (PMIDs; in PubMed every record has a unique number). We selected only records from 1990 onwards, because most relevant studies on measurement properties have been published after this date. We selected records up to December 2006 (the search was performed on March 12, 2007) to include also the records that were not yet indexed by the NLM to simulate a 'real life' search as much as possible. We did not restrict our search to any medical field or journal.

The records in the gold standard were hand searched to find studies on measurement properties. In order to reduce the workload, we first developed an exclusion filter to identify irrelevant research such as editorials, reviews, comments, case reports, and animal research. This filter was developed by combining publication types and MeSH headings, based on experience of the information specialists

[EPJ and IR]. This filter was used on the gold standard to identify irrelevant records. All records identified by this exclusion filter were recorded as being not studies on measurement properties. The remaining records were hand searched to identify studies on measurement properties. The hand search was performed by two reviewers [CBT and EPJ]. Both reviewers, independently, screened all titles, abstracts, and MeSH headings. Disagreements were resolved by a consensus meeting with a third reviewer [HCWdV]. We did not screen the full-text articles to mimic the future situation, because the performance of the filter will be determined by the information included in PubMed, not in the full-text articles. We identified abstracts as studies on measurement properties if they had the aim to develop or evaluate a measurement instrument and that reported at least some information on the measurement properties.

At the end of this phase, all 10,000 records in the gold standard were categorized as being studies on measurement properties or not.

Phase 2a: Search term selection

Five sources were used for search term selection: (1) we searched for relevant MeSH headings and text words in the titles and abstracts of all relevant PubMed records in the gold standard PubMed records; (2) we searched for relevant text words in 100 systematic reviews of measurement properties of health status measurement instruments that we collected for a review of these studies [5]; (3) we screened the search strategies of these 100 systematic reviews for relevant search terms; (4) we used the MeSH database to identify additional relevant MeSH headings; and (5) we added a few terms based on our own expertise in developing measurement instruments and assessing measurement properties.

Phase 2b: Search term combination

In order to develop the search filter for finding measurement properties, relevant search terms were combined based on sensitivity and precision. Sensitivity is the number of relevant records in the gold standard retrieved by the search filter as a proportion of the total number of relevant records in the gold standard. Precision (or positive predictive value) is the number of relevant records retrieved by the search filter as a proportion of the total number of records retrieved (Appendix 1). First, we determined sensitivity and precision of all terms individually (univariate) in the gold standard set of records (all 10,000 records). Next, we determined sensitivity and precision of combinations of related terms, e.g. all terms for reliability, to identify the most sensitive term for each measurement property. Then, we combined search terms in sequence of their sensitivity, starting with MeSH headings and the most sensitive terms for each measurement property. We took precision into account by giving priority to terms with a higher precision. For each additional term, we determined sensitivity and precision of the whole filter to see how the filter improved.

During the development and testing of the filter, we combined the search terms for measurement properties with the exclusion filter by applying Boolean NOT, in order to mimic the future use of the filter.

We developed two filters: first, we developed a sensitive filter. We aimed at 98% sensitivity, based on the performance of currently available search filters in other fields (http://hiru.mcmaster.ca/hiru/HIRU_Hedges_MEDLINE_Strategies.aspx). The best search filters have a sensitivity around 99%. We expected the sensitivity of our filter to be slightly lower because of the large variety of terminology used in the field of measurement.

Secondly, we developed a more precise filter to be used for a less extensive search. We aimed at a precision of 10%, which is comparable to the best search filters for finding clinical trials and diagnostic studies. We aimed to keep the sensitivity of the precise filter around 95%. The precise filter was developed by removing search terms from the first filter one by one with a relatively low contribution to the sensitivity and a high false positive rate (i.e. retrieving many irrelevant records). After removing each term, we determined sensitivity and precision of the whole filter to examine how the performance changed.

Phase 3: Search filter evaluation (internal validity)

In the third phase, the performance of the filters (combined with the exclusion filter) were tested in the gold standard by calculating sensitivity, precision, specificity, and number needed to read (NNR), which is the number of records that need to be read to identify one relevant record (Appendix 1).

Phase 4: Search filter validation (external validity)

In the final phase, the search filters (combined with the exclusion filter) were validated against two existing PubMed searches that were previously hand searched by two independent researchers. The first validation set was a systematic search of all studies on the measurement properties of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). This is a self-report questionnaire for the measurement of pain and physical functioning of osteoarthritis patients [6]. This search was performed on February 13, 2008 by one of the authors [CBT] on the entire PubMed

database, using the terms WOMAC[tw] OR ("western ontario"[tw] AND ("McMaster Universities"[tw] OR "McMaster University"[tw])). The search consisted of 824 records, containing 100 studies on measurement properties. The second validation set was a systematic review of physical activity questionnaires. This search was performed on September 24, 2007 (on the entire database) using the terms ((exercise[mesh] OR "physical activity"[tiab] OR motor activity[mesh]) AND (questionnaire[mesh] OR questionnaire*[tiab])). The search consisted of 8,837 records, containing 242 studies on measurement properties [7]. We calculated sensitivity, specificity, precision, and number needed to read of both filters in both validation sets.

Results

The exclusion filter (Appendix 2) identified 3,587 irrelevant records in the 10,000 records of the gold standard. The remaining 6,413 records were hand searched. We indentified 116 studies on measurement properties in the gold standard, which gives a prevalence of 1.16%.

About 200 possible search terms were identified and tested individually against the gold standard set of records. The most sensitive and most precise search terms are presented in Table 1.

The most sensitive search term was "reproducib*[tw]" with a sensitivity of 43.1% and a precision of 30.9%. The most precise search term was "internal consistency[tiab]" with a precision of 100% but a low sensitivity of 6.9%. While the most sensitive search terms also had a reasonable precision, all terms with a high precision had a low sensitivity.

Table 1 Most sensitive and most precise terms (univariate)

Term	Sensitivity (%)	Precision (%)
Most sensitive terms		
reproducib*[tw]	43.1	30.9
methods[sh]	42.2	5.2
valid*[tiab]	39.7	31.5
reproducibility of results[MeSH]	38.8	33.3
reliab*[tiab]	37.9	32.8
Most precise terms		
internal consistency[tiab]	6.9	100
ceiling effect[tiab]	0.9	100
coefficient of variation[tiab]	5.2	66.7
observer variation[MeSH]	11.2	44.8
psychometrics[MeSH]	9.5	42.3
validation Studies[pt]	8.6	35.7
discriminative[tiab]	0.9	33.3
precision[tw]	7.8	31.0

Only terms with at least 30% sensitivity or 30% precision are presented

The final two search filters and the exclusion filter are presented in Appendix 2. The main differences between the filters are: first, the sensitive filter does contain some terms that the precise filter does not contain (e.g. "outcome assessment (health care)"[MeSH] OR outcome assessment[tiab] OR outcome measure*[tw] OR "Health Status Indicators"[Mesh]). Second, in the sensitive search filter, some terms are combined while these are separated in the precise filter (e.g. (multitrait[tiab] AND scaling[tiab] AND (analysis[tiab] OR analyses[tiab])) versus ("multitrait scaling analysis"[tiab] OR "multitrait scaling analyses"[tiab])).

A guide for using the filters is presented in Fig. 1. The filters should be used in combination with search terms for the construct of interest, the kind of measurement instruments of interest, and the population of interest. These terms should be defined by the users, preferably with the help of an information specialist.

Search filter evaluation (internal validity) and search filter validation (external validity)

The performance of the filters (combined with the exclusion filter) in the gold standard is presented in Tables 2 and 3. The sensitive search filter was able to retrieve 113 of the 116 studies on measurement properties, which gives a sensitivity of 113/116 = 97.4%. This filter retrieved 2,594 records, which gives a precision of 113/2594 = 4.4%. The number needed to read is 23. The precise search filter was able to retrieve 108 of the 116 studies on measurement properties, which gives a sensitivity of 93.1%. This filter retrieved 1,150 records, which gives a precision of 9.4% and a number needed to read of 11.

The performance of the filters (combined with the exclusion filter) in the two validation sets are also presented in Table 3. Sensitivity of the sensitive search filter in the two validation sets was 98.0 and 94.6% and precision was 13.2 and 5.6%, respectively. Sensitivity of the precise filter was 94.0 and 89.7% and precision was 25.3 and 11.0%, respectively.

Discussion

We developed a highly sensitive search filter for finding studies on measurement properties in PubMed. This filter was able to retrieve 97.4% of the relevant records in the gold standard. We also developed a more precise search filter. This filter reduced the number of records that need to be read to identify one study on measurement properties from 87 (10,000/116) without using a filter to 11 (1,150/108) when using the filter.

Fig. 1 A guide for using the search filters

Search strategy

#1 construct search
#2 population search
#3 instrument search
#4 #1 AND #2 AND #3 AND filter for measurement properties
#5 #4 NOT exclusion filter

The construct search should be defined by the user. It includes search terms for the construct to be measured. For example: quality of life, physical activity, etc.
The population search should also be defined by the user. It includes search terms for the population of interest. For example: children, diabetes, etc.
The instrument search is optional and should also be defined by the user. It includes search terms for the instruments of interest. For example: questionnaires, performance-based tests, etc.

These searches should then be combined with the search filter for measurement properties to find all studies on the measurement properties of the instruments of interest that measure the construct of interest in the population of interest. One can choose to use the sensitive filter for a comprehensive search or the precise filter for a less extensivesearch. The exclusion filter is meant to remove irrelevant records from the search, such as case reports and animal studies.

It is important to use the exclusion filter exactly as indicated above. One should not run the exclusion filter separately and then link with NOT.

Table 2 Performance of both filters in the gold standard

	Gold standard		
	Relevant study	Nonrelevant study	Total
Sensitive search filter			
Search filter			
Retrieved	113	2,469	2,582
Not retrieved	3	7,415	7,418
Total	116	9,884	10,000
Precise search filter			
Search filter			
Retrieved	108	1,032	1,140
Not retrieved	8	8,852	8,860
Total	116	9,884	10,000

The sensitivity of both filters was slightly lower than we aimed for (97.4% instead of 98% and 93.1% instead of 95%). This means that there is a slight risk of missing relevant studies when using these filters. However, the sensitivities are still quite acceptable, when compared to other search filters (http://hiru.mcmaster.ca/hiru/HIRU_Hedges_MEDLINE_Strategies.aspx).

The performance of both filters was very good in the two validation sets. The performance was better in the set of records of studies on measurement properties of the WOMAC questionnaire than in the set of records of studies on measurement properties of the physical activity questionnaires. This might be due to the fact that studies on the same questionnaire are often performed according to a similar methodology, which might lead to more consistent use of terminology. Another reason might be that the

Table 3 Performance of the filters in the gold standard and validation sets

Filter	Set of records	Sensitivity (%)	Precision (%)	Specificity (%)	NNR
Search filter evaluation (internal validity)					
1 (sensitive)	Gold standard	97.4	4.4	75.0	23
2 (more precise)	Gold standard	93.1	9.5	89.6	11
Search filter validation (external validity)					
1 (sensitive)	WOMAC	98.0	13.2	11.0	8
1 (sensitive)	Physical activity	94.6	5.6	51.1	18
2 (more precise)	WOMAC	94.0	25.3	61.6	4
2 (more precise)	Physical activity	89.7	11.0	77.7	10

WOMAC Western Ontario and McMaster Universities Osteoarthritis Index, *NNR* number needed to read

methodology and reporting of studies on measurement properties have received more attention and have been further developed in the field of health status and quality of life measurement (e.g. WOMAC) than in other fields (e.g. physical activity).

Although we tested the performance of the filters always in combination with the exclusion filter, we decided to present the exclusion filter as a separate filter, because this enables users to choose to use the filter for measurement properties without the exclusion filter if they want to retrieve all publication types, or human and animal studies. Moreover, information specialists recommend using exclusions (Boolean NOT) always at the end of the search strategy.

This study has several methodological strengths: first, the gold standard, a random sample of PubMed, is representative for the literature in which the filters are going to be used. This will increase the likelihood of a good performance of the filters in future studies. We did not only include high quality or recent studies (or high quality journals) in our gold standard, but also poor and older studies because the filters should also be able to find these studies. For the same reason, we also included records that were not yet indexed by the NLM. Many published search filters, like those developed by Haynes and Wilczynski et al. [8–11], are tested against recent high quality studies. The sensitivity of these filters in the "real world" is likely to be overestimated.

Second, we analyzed the performance of the filters in a way that mimics the real use of the filters, e.g. in a systematic review. Therefore, we calculated sensitivity based on screening of the abstracts, not on screening of the full-text articles. Three abstracts were missed by our sensitive filter (PMID 11681521, 10747220, and 9650947) because they did not contain any terms for measurement properties in the abstracts. They were selected by hand search because of statements like "The results obtained using these techniques are compared" or "A comparison of organism recoveries and morphologies was undertaken with both … (WT) and (ES)". When we read the full-text articles of these three abstracts, it appeared that only two of them included some information on measurement properties. However, we still counted all three abstracts as false negatives because we would have selected these abstracts in a real situation, e.g. when screening abstracts for a systematic review. Therefore, we wanted the filter to retrieve them. If we would have calculated sensitivity based on the full-text articles, as has been done in many other studies [8–11], we would have overestimated the real sensitivity of the filter, because in that case, the one study that did not include information on measurement properties that we missed would not have been counted as false negative.

Third, our filters have been validated in two very different settings, i.e. one set of records from a search for finding studies on measurement properties of a disease-specific health questionnaire and one set of records from a search for finding studies on measurement properties of physical activity questionnaires. The performance of the filters in these two settings is promising. Nevertheless, it would be worthwhile to validate the filters in new validation sets, especially in the field of (health-related) quality of life research, where there are many instruments available to measure the same construct, with different measurement properties. It would also be worthwhile to analyze whether the performance of the filters is different e.g. for disease-specific versus generic instruments or for different medical fields.

This study also has some limitations: first, we did not hand search all records in the gold standard because we used an exclusion filter. We might have missed studies on measurement properties by using this exclusion filter. If that was the case, the performance of the measurement properties filter might have been either overestimated or underestimated, depending on whether the filter would have retrieved these missed records.

Secondly, the gold standard contained only 116 studies on measurement properties, and therefore the initial performance of the filter was based only on 116 studies. However, the validation sets contained 100 and 242 studies on measurement properties, respectively, which means that in total the filter has been tested on 458 studies on measurement properties.

The performance of our sensitive filter is higher than that of many other filters. For example, our filter has a higher sensitivity than 23 available search filters for finding diagnostic studies (highest sensitivity 86.9%) [12, 13]. This might be the result of the generalizability of our gold standard set of records, of using multiple sources for search term selection, and the inclusion of over 150 search terms in the filter. Large search filters are easy to use in PubMed because the filter can be copied and pasted at once into the search box.

The performance of the filters can be improved in the future when records on measurement properties are properly indexed or when indexation is corrected. This can be facilitated by reaching consensus among researchers on terminology of measurement properties. For example, the search terms "reproducib* [tw]" and "reliab*[tiab]" retrieved almost a similar amount of studies. In the COSMIN Delphi study, international consensus was reached on using the term "reliability" [14]. Such efforts will facilitate indexing by the NLM and improve retrieval of studies. In addition, standards for reporting studies on measurement properties should be developed. Such standards do not yet exist. For randomized clinical trials, this, with considerable effort of the Cochrane Collaboration, has resulted in increased performance of search filters up to over 99% [15].

Practical recommendations for using the filters

For using the search filters, a computer with internet access is required. PubMed is freely available all over the world. Users of the filters should make a choice of the filter they want to use. This depends on the aim of their search. The sensitive search filter is especially suitable for researchers to use in systematic reviews of studies on measurement properties. The precise filter can be used by researchers or clinicians for a less extensive search, e.g. to obtain an overview of the measurement properties of one specific measurement instrument to be used as an outcome measure in a particular study or in clinical practice. In both cases, the filter should be used in combination with search terms for the construct of interest, search terms for the kind of measurement instruments of interest, and search terms for the population of interest. These terms should be defined by the users, preferably with help of an information specialist. The exclusion filter could be used to exclude irrelevant study types. If users want to retrieve all publication types, or they want to include human and animal studies, they should not use the exclusion filter.

If users of the filters think that the performance of the filters might improve by adding additional terms, they are free to test and validate this. Adding additional terms might improve the sensitivity, but at the cost of lowering the precision because new terms will also yield new irrelevant studies.

Conclusion

We developed a highly sensitive search filter and a more precise search filter for finding studies on measurement properties in PubMed, using a strong methodology. The performance of both filters is very good, as demonstrated in the gold standard as well as in two validation sets. The performance of the filters can be improved even more in the future by improved indexing of studies on measurement properties by the NLM and by improved reporting of these studies by the authors. The use of these search filters will contribute to evidence-based instrument selection and improved quality of measurement in all medical fields.

Acknowledgments We acknowledge Wieneke Mokkink for her help with hand searching the records of the validation set on the measurement properties of the WOMAC.

Conflict of interest statement None.

Appendix 1: Search filter definitions

	Gold standard	
	Relevant study	Nonrelevant study
Search filter		
Retrieved	a	b
Nonretrieved	c	d
	a + c	b + d

Sensitivity: The number of relevant records in the gold standard retrieved by the search filter as a proportion of the total number of relevant records in the gold standard

$$\frac{\text{number of relevant records retrieved by search filter}}{\text{total number of relevant records in the gold standard}} \times 100$$

$$\frac{a}{a+c} \times 100.$$

Precision: The number of relevant records retrieved as a proportion of the total number of records retrieved

$$\frac{\text{number of relevant records retrieved by search filter}}{\text{total number of records retrieved by the search filter}} \times 100$$

$$\frac{a}{a+b} \times 100.$$

Specificity: The number of records that are not relevant and are not retrieved as a proportion of the total number of records that are not relevant

$$\frac{\text{total number of records not relevant and not retreived}}{\text{total number of records not relevant}} \times 100$$

$$\frac{d}{b+d} \times 100.$$

Number needed to read: The number of records that need to be read to identify one relevant record

$$\frac{1}{\text{precision}}.$$

Appendix 2: Search filters for finding studies on measurement properties

Filter 1: Sensitive search filter for measurement properties

(instrumentation[sh] OR methods[sh] OR Validation Studies[pt] OR Comparative Study[pt] OR "psychometrics" [MeSH] OR psychometr*[tiab] OR clinimetr*[tw] OR clinometr*[tw] OR "outcome assessment (health care)"[MeSH] OR outcome assessment[tiab] OR outcome measure*[tw] OR

"observer variation"[MeSH] OR observer variation[tiab] OR "Health Status Indicators"[Mesh] OR "reproducibility of results"[MeSH] OR reproducib*[tiab] OR "discriminant analysis"[MeSH] OR reliab*[tiab] OR unreliab*[tiab] OR valid*[tiab] OR coefficient[tiab] OR homogeneity[tiab] OR homogeneous[tiab] OR "internal consistency"[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR (item[tiab] AND (correlation*[tiab] OR selection*[tiab] OR reduction*[tiab])) OR agreement[tiab] OR precision[tiab] OR imprecision[tiab] OR "precise values"[tiab] OR test–retest[tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab])) OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater[tiab] OR intertester[tiab] OR inter-tester[tiab] OR intratester[tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer[tiab] OR intraobserver[tiab] OR intra-observer[tiab] OR intertechnician[tiab] OR inter-technician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer[tiab] OR intra-examiner[tiab] OR interassay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR interparticipant[tiab] OR inter-participant[tiab] OR intraparticipant[tiab] OR intra-participant[tiab] OR kappa[tiab] OR kappa's[tiab] OR kappas[tiab] OR repeatab*[tiab] OR ((replicab*[tiab] OR repeated[tiab]) AND (measure[tiab] OR measures[tiab] OR findings[tiab] OR result[tiab] OR results[tiab] OR test[tiab] OR tests[tiab])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR "known group"[tiab] OR factor analysis[tiab] OR factor analyses[tiab] OR dimension*[tiab] OR subscale*[tiab] OR (multitrait[tiab] AND scaling[tiab] AND (analysis[tiab] OR analyses[tiab])) OR item discriminant[tiab] OR interscale correlation*[tiab] OR error[tiab] OR errors[tiab] OR "individual variability"[tiab] OR (variability[tiab] AND (analysis[tiab] OR values[tiab])) OR (uncertainty[tiab] AND (measurement[tiab] OR measuring[tiab])) OR "standard error of measurement"[tiab] OR sensitiv*[tiab] OR responsive*[tiab] OR ((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically[tiab]) AND (important[tiab] OR significant[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR meaningful change [tiab] OR "ceiling effect"[tiab] OR "floor effect"[tiab] OR "Item response model"[tiab] OR IRT[tiab] OR Rasch[tiab] OR "Differential item functioning"[tiab] OR DIF[tiab] OR "computer adaptive testing"[tiab] OR "item bank"[tiab] OR "cross-cultural equivalence"[tiab])

Filter 2: Precise search filter for measurement properties

(instrumentation[sh] OR Validation Studies[pt] OR "reproducibility of results"[MeSH Terms] OR reproducib*[tiab] OR "psychometrics"[MeSH] OR psychometr*[tiab] OR clinimetr*[tiab] OR clinometr*[tiab] OR "observer variation"[MeSH] OR observer variation[tiab] OR "discriminant analysis"[MeSH] OR reliab*[tiab] OR valid*[tiab] OR coefficient[tiab] OR "internal consistency"[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR "item correlation"[tiab] OR "item correlations"[tiab] OR "item selection"[tiab] OR "item selections"[tiab] OR "item reduction"[tiab] OR "item reductions"[tiab] OR agreement[tw] OR precision[tw] OR imprecision[tw] OR "precise values"[tw] OR test–retest [tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab])) OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater[tiab] OR intertester[tiab] OR inter-tester[tiab] OR intratester[tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer[tiab] OR intraobserver[tiab] OR intra-observer[tiab] OR intertechnician[tiab] OR inter-technician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer[tiab] OR intra-examiner[tiab] OR inter-assay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR interparticipant[tiab] OR inter-participant[tiab] OR intraparticipant[tiab] OR intra-participant[tiab] OR kappa[tiab] OR kappa's[tiab] OR kappas[tiab] OR "coefficient of variation"[tiab] OR repeatab*[tw] OR ((replicab*[tw] OR repeated[tw]) AND (measure[tw] OR measures[tw] OR findings[tw] OR result[tw] OR results[tw] OR test[tw] OR tests[tw])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR "known group" [tiab] OR "factor analysis"[tiab] OR "factor analyses"[tiab] OR "factor structure"[tiab] OR "factor structures"[tiab] OR dimensionality[tiab] OR subscale*[tiab] OR "multitrait scaling analysis"[tiab] OR "multitrait scaling analyses"[tiab] OR "item discriminant"[tiab]OR "interscale correlation"[tiab] OR "interscale correlations"[tiab] OR ((error[tiab] OR errors[tiab]) AND (measure*[tiab] OR correlat*[tiab] OR evaluat*[tiab] OR accuracy[tiab] OR accurate[tiab] OR precision[tiab] OR mean[tiab])) OR "individual variability"[tiab] OR "interval variability"[tiab] OR "rate variability"[tiab] OR "variability analysis"[tiab] OR (uncertainty[tiab] AND (measurement[tiab]

OR measuring[tiab])) OR "standard error of measurement"[tiab] OR sensitiv*[tiab] OR responsive*[tiab] OR (limit[tiab] AND detection[tiab]) OR "minimal detectable concentration"[tiab] OR interpretab*[tiab] OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR "meaningful change"[tiab] OR "minimal important change"[tiab] OR "minimal important difference"[tiab] OR "minimally important change"[tiab] OR "minimally important difference"[tiab] OR "minimal detectable change"[tiab] OR "minimal detectable difference"[tiab] OR "minimally detectable change"[tiab] OR "minimally detectable difference"[tiab] OR "minimal real change"[tiab] OR "minimal real difference"[tiab] OR "minimally real change"[tiab] OR "minimally real difference"[tiab] OR "ceiling effect"[tiab] OR "floor effect"[tiab] OR "Item response model"[tiab] OR IRT[tiab] OR Rasch[tiab] OR "Differential item functioning"[tiab] OR DIF[tiab] OR "computer adaptive testing"[tiab] OR "item bank"[tiab] OR "cross-cultural equivalence"[tiab])

Exclusion filter

("addresses"[Publication Type] OR "biography"[Publication Type] OR "case reports"[Publication Type] OR "comment"[Publication Type] OR "directory"[Publication Type] OR "editorial"[Publication Type] OR "festschrift"[Publication Type] OR "interview"[Publication Type] OR "lectures"[Publication Type] OR "legal cases"[Publication Type] OR "legislation"[Publication Type] OR "letter"[Publication Type] OR "news"[Publication Type] OR "newspaper article"[Publication Type] OR "patient education handout"[Publication Type] OR "popular works"[Publication Type] OR "congresses"[Publication Type] OR "consensus development conference"[Publication Type] OR "consensus development conference, nih"[Publication Type] OR "practice guideline"[Publication Type]) NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms])

References

1. Bot, S. D., Terwee, C. B., van der Windt, D. A., Bouter, L. M., Dekker, J., & de Vet, H. C. (2004). Clinimetric evaluation of shoulder disability questionnaires: A systematic review of the literature. *Annals of the Rheumatic Diseases, 63*, 335–341.
2. Brettle, A. J., Long, A. F., Grant, M. J., & Greenhalgh, J. (1998). Searching for information on outcomes: Do you need to be comprehensive? *Quality in Health Care, 7*, 163–167.
3. Jenkins, M. (2004). Evaluation of methodological search filters—A review. *Health Information and Libraries Journal, 21*, 148–163.
4. Glanville, J., Bayliss, S., Booth, A., Dundar, Y., Fernandes, H., Fleeman, N. D., et al. (2008). So many filters, so little time: The development of a search filter appraisal checklist. *Journal of the Medical Library Association, 96*, 356–361.
5. Mokkink, L. B., Terwee, C. B., Stratford, P., Alonso, J., Patrick, D. L., Riphagen, I., et al. (2009). Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Quality of Life Research, 18*, 313–333.
6. Bellamy, N., Buchanan, W. W., Goldsmith, C. H., Campbell, J., & Stitt, L. W. (1998). Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes in antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *Journal of Rheumatology, 15*, 1833–1840.
7. van Poppel, M. N. N., Chinapaw, M., Mokkink, L. B., & Terwee, C. B. (2008). Can physical activity be measured using questionnaires? A systematic review of validity, reliability and responsiveness of physical activity questionnaires for adults (submitted).
8. Haynes, R. B., & Wilczynski, N. L. (2004). Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: Analytical survey. *BMJ, 328*, 1040.
9. Haynes, R. B., McKibbon, K. A., Wilczynski, N. L., Walter, S. D., & Werre, S. R. (2005). Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: Analytical survey. *BMJ, 330*, 1179.
10. Wilczynski, N. L., & Haynes, R. B. (2003). Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. AMIA Annual Symposium Proceedings, 719–723.
11. Wilczynski, N. L., & Haynes, R. B. (2004). Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: An analytic survey. *BMC Medicine, 2*, 23.
12. Ritchie, G., Glanville, J., & Lefebvre, C. (2007). Do published search filters to identify diagnostic test accuracy studies perform adequately? *Health Information and Libraries Journal, 24*, 188–192.
13. Leeflang, M. M., Scholten, R. J., Rutjes, A. W., Reitsma, J. B., & Bossuyt, P. M. (2006). Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *Journal of Clinical Epidemiology, 59*, 234–240.
14. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. J., et al (submitted). International consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes: Results of the COSMIN study.
15. Glanville, J. M., Lefebvre, C., Miles, J. N., & Camosso-Stefinovic, J. (2006). How to identify randomized controlled trials in MEDLINE: Ten years on. *Journal of the Medical Library Association, 94*, 130–136.