# Protocol Template for a Systematic review of measurement properties

## Introduction

Core Outcome Sets are a minimum set of outcomes designed to be fielded in all trials or clinical research in a field to ensure consistency and relevance of outcomes in clinical research. Their adoption allows clinical practice guideline developers and systematic review teams to be able to make optimal use of all trials and research studies by using these standardized sets of outcomes 1,2. For over 30 yrs, OMERACT has led efforts to developing Core Outcome Sets for use in rheumatic diseases and musculoskeletal disorders. The impact of this work is clearly demonstrated in rheumatoid arthritis (RA), where an estimated 80% of pharmacological trials now use the RA Core Outcome Set 3. This widespread adoption enables more direct, consistent, and accurate comparisons across studies, strengthening the evidence base for treatment decisions.

OMERACT has developed guidelines for Core Outcome Set development including the stages of domain and instrument selection 4,5. The latter determines if an instrument has passed the OMERACT Filter of having Truth, Discrimination and Feasibility in a defined population and setting. A key component of instrument selection is the systematic review of evidence on measurement properties within the target population 4,6–8.

The purpose of this protocol is to describe the process that will be followed in conducting a systematic review of the literature to identify evidence of the measurement property performance of the [INSTRUMENT NAME AND VERSION] in a population of persons living with [DISEASE].

## Methods

This systematic review of the measurement properties of [INSTRUMENT NAME AND VERSION] is designed based on the OMERACT Instrument selection process 5,9 and following the principles of conducting a systematic review 10–13. This protocol will be registered with PROSPERO and will follow PRISMA-P reporting of protocol standards 14.

**Eligibility criteria**

Population

Studies evaluating measurement properties conducted in adult populations of persons with [DISEASE] will be included.

Outcome Instrument(s)

The following [INSTRUMENT NAME AND VERSION] will be assessed in this review.

Measurement properties of interest

The aim of this review is to evaluate the measurement properties considered essential according to the OMERACT Filter pillars of Truth, Discrimination and Feasibility 4,6,15–17. These include construct validity (known-groups and correlational), test-retest reliability, longitudinal construct validity, clinical trial discrimination and establishing thresholds of meaning (benchmarking scores or thresholds for evaluating change) 4,5,7,8.

Selection criteria

Existing studies on measurement properties of the instruments are included if they meet the following criteria:

* Full-text original article offering primary data
* The purpose of the article is to evaluate one or more measurement properties of one or more of the instruments of interest (e.g., the study does not just apply the instruments as part of a clinical study)
* Study population of adult individuals (age >18 years) with [DISEASE]. Studies covering children and adults will only be considered if the measurement properties are reported separately for the adults. Only the data for adults will be used.

Setting

There is no restriction on setting for eligibility in this study.

Language

At the search level, no restrictions were placed on language.

Search strategy

Relevant studies will be identified through an electronic search of the following databases: MEDLINE and MEDLINE In-process (Ovid), EMBASE (Ovid) and CINHAL (Ebsco).

The terms used for the search strategy will make use of the Boolean logic suggested in the OMERACT Handbook 5 however for this review we wished to find all evidence for these outcome instruments and did not limit based on disease at this step. The search terms therefore will combine text terms capturing all ways an instrument has been described in the literature (full name or acronyms for example) with the search terms for the measurement properties and the overlap will be considered a hit. The measurement property component of the search strategy will be based on the recommendations of the Consensus-based Standards for the selection of health Measurement Instruments Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) 18 group in their handbook and earlier publication 19–22. A detailed search strategy is provided in Appendix 1.

## Data Management

The results of the literature search will be collected in an [LIBRARY NAME], where duplicate studies will be flagged and removed. The search results will then be imported into [SOFTWARE NAME] to facilitate management of the selection and identification of the measurement properties evaluated in each selected paper. Tracking of progress and results of the evaluation of the selected paper will be done on the OMERACT knowledge transfer tool called the Summary of Measurement Properties table (SOMP) 9. Figure 1 provides an example of a completed SOMP. Quality assessment (risk of bias) will be recorded on a pretested excel spreadsheet and details of the measurement property evaluations extracted onto standardized word tables developed at OMERACT 5,23.

A document with numbers and text

AI-generated content may be incorrect.

**Figure 1:** Summary of Measurement Properties (SOMP) table. The SOMP documents the studies used in the review, the risk of bias assessment (cell colour), whether the Measurement Property findings were above standard thresholds (+ or - symbol) and at the bottom the steps and results in synthesis of findings phase.

## Selection process with Artificial Intelligence (AI) (OPTIONAL SECTION)

We will use the AI tools embedded in DistillerSR software to assist in the screening of titles and abstracts. We will screen a random set of 200 records in duplicate by two human reviewers independently to act as a seed 24–26. Inter-rater agreement of at least 90% will be reached before allowing AI to take part in the screening. Disagreements will be resolved by discussion of the two reviewers (REVIEWER INITIALS) or by involving a third, senior reviewer (REVIEWER INITIALS). The remaining title and abstracts will be screened by a single human reviewer (REVIEWER INITIALS) and AI until 95% recall is obtained, calculated as [True Positives / (True Positives + False Negatives)], according to the guidance described by Hamel and colleagues 27. Once this threshold is achieved, the AI tool will screen the remaining studies. The AI estimates the probability that the article should be included (the threshold score). Literature suggests that conservative range of threshold scores should be used 28; we will set inclusion at a threshold greater than 0.7 and exclusion at less than or equal to 0.7. A single reviewer will verify a random sample of 10% of the records excluded by the AML tool to ensure they were true exclusions. We will utilize DistillerSR’s background priority screening function throughout this whole stage 27.

All included records will be retrieved for full article review. Two reviewers (REVIEWER INITIALS) will screen the full text articles for 50 articles in duplicate and resolve disagreements by discussion or with a third reviewer (REVIEWER INITIALS). If adequate agreement is achieved, one of the pair will complete the remainder of the full text screening. If any systematic reviews of measurement properties of any of our instruments are identified, the references cited in the review will be checked against our search results to capture any additional relevant articles.

The results of the screening process will be documented in a PRISMA-COSMIN-OMI recommended flow diagram 11,29 including an additional feature of the number of articles for each measurement property for each of our target instruments at the bottom. This number will be compared to the total on the SOMP when it is populated.

The final set of selected articles will be added to the appropriate SOMP Table (one SOMP table for each instrument), and the measurement properties addressed in that article marked with an X on the table 5,9 to record the measurement properties studied in each published article (see Figure 1 for sample of SOMP). The version of questionnaire and the study population will also be recorded on the SOMP for quick reference. This provides a full view of the pool of evidence being used in this review for each version of the instruments and for each population. The numeric total for each property for each tool will be compared (and should be the same as) the total at the bottom of the PRISMA flow diagram.

## Data extraction process

Initial extraction having been completed with the placement of the X’s in the OMERACT Summary of Measurement Properties (SOMP), we will now move into more detailed extraction in critical appraisal and description of the study results. The SOMP will track the evolution of the review at each step.

Risk of bias in individual studies

We will assess the risk of bias in the results of the measurement property studies by evaluating the quality of the methods using the COSMIN-OMERACT Good Methods checklist for each property 4,5. The focus of its items is on identifying critical flaws in the methods, flaws which could lead to the wrong value for reliability or validity evidence. Studies that avoid these errors are deemed to have “good enough methods” to have their evidence used in the synthesis stage. There are different checklists for each measurement property. Once this checklist is completed, the reviewer decides if, based on these responses, this evidence should be considered in the synthesis with full confidence (green), with some caution (amber) or it should not be used in this review (red). While we recognize this is often due to under-reporting in the published report, we will focus on the reported methods rather than actual methods in this review. The risk of bias items for each measurement property are included in Appendix 2.

Two reviewers will conduct the Good Methods check independently on all included studies and will reach agreement on their view of the quality of the methods used. They will assign a rating of Green (go ahead, no indicators of risk of bias), Amber (proceed but use results with some caution, some potential risk of bias), or Red (stop, an indication of a high risk of bias) for each piece of evidence for each measurement property. This colour coding is added to the corresponding cell on the Summary of Measurement Property Table.

Detailed description of the study methods and results.

We will use a pretested standardized data extraction form developed by OMERACT 5 to extract information on the methods used and results for each measurement property. Detailed description will only be carried out on studies rated as a Green or Amber on the good methods check. Measurement property evidence rated as a Red (high risk of bias) will not be used as evidence. A single reviewer will extract data for each included study; the second reviewer will examine the extracted data for agreement and make notes of any disagreement. Any disagreements will be resolved through discussion between reviewers or with a senior author if necessary.

Data items

For each article, we will extract a general description of the article including author, year of publication, and instrument version of interest that were tested, measurement property(ies) assess(ed), study design/methods, and sample description (e.g., age, sex and/or gender, length of follow-up).

For each measurement property assessed, we will capture the following on standardized reporting tables designed with input from existing literature on this property and reporting recommendations 30,31:

* **Construct validity/known groups:** brief sample description, mean to identify subgroups to be compared, a priori hypotheses of differences between the subgroups, observed results, hypothesis confirmed (yes/no) \*
* **Construct validity/correlation:** brief sample description, construct used for comparison, a priori hypotheses, results, hypothesis confirmed (yes/no).\*
  + - \*It is important to note that in many studies of construct validity, a priori hypotheses are often not described. This is a risk for bias, but once acknowledged as a risk of bias, the results could provide some useable information. We therefore needed a way to interpret the results of the study. We developed a decision rule to guide how we would use these results (See Table 1). In the absence of hypotheses we would look to implied hypotheses often found in methods or discussion. In absence of those we look to predetermined values for correlations for common comparators (see Table 2). Our review team identified these at the beginning of our review and the full table is available on request.
* **Inter-method reliability:** type of reliability assessment, brief characteristics of the sample, characteristics of the testing situation, sample recruited, scores for each rater/method/machine, statistic(s) used, results, minimal detectable difference with 95% confidence interval
* **Test-retest reliability:** characteristics of testing situation, sample recruited and sample considered stable for analysis, scores at baseline and retest, statistic used, results, minimal detectable change at 95% confidence level
* **Longitudinal construct validity (responsiveness):** sample description, study methods, groups being contrasted (e.g., within-person change over time or between-person differences), hypothesis of change and anchors used, anticipated results, statistic used, results observed
* **Discrimination supporting between group difference in change (differences in differences) for application in a clinical trial setting:** intervention/comparator (and sample size), scores at baseline and primary time point, effect sizes at primary time point per trial arm, effect size and p-value of change scores of treatment arm compared to placebo arm. In absence of clinical trial data, data from parallel comparisons of differences in change score from longitudinal data can be used 5.
* **Thresholds of meaning:** brief sample description, description of study methods, threshold assessed (e.g., MID, MCID, PASS, LDA), method (anchor or distributional), definition of threshold of meaning using this approach, threshold of meaning and AUC if available, % of sample meeting or exceeding this threshold. Empirical cumulative distribution functions are created to depict different thresholds and how they would describe the samples.

|  |  |  |
| --- | --- | --- |
| **1.A priori available** | **2.No a priori hypotheses found in report** | **3.No a priori, no implied hypotheses found** |
| Use authors’ a priori hypotheses | Look in methods, discussion for implied hypothesis or relationship | Use set of predetermined hypotheses made by review team at beginning of review |

**Table 1:** Decision rule for identification of a priori hypotheses for construct validity and responsiveness (longitudinal construct validity). This guideline was established as the a priori hypothesis is often missing in reports of validity evidence.

|  |  |  |
| --- | --- | --- |
| **Hypotheses related to Physical Functioning (PF) as target domain** | | |
| **Comparative construct** | **Rationale** | **Review teams assigned a priori Agreed upon July 23, 2024.** |
| PF versus Pain intensity | In MSK disorders, pain and function are often highly correlated. In arthritis, function may depart from this relationship when residual joint damage limits function in the absence of pain. | Moderate or better,  r>0.5 |
| PF versus Pain interference | As above, with the interference having greater correlation because it is about the impact of pain on lived life, likely including physical function | Moderate, but higher than observed with pain intensity. r>0.60 |
| PF versus other measures of PF (i.e., disease specific for back) | Same construct, different instruments, one might have different coverage of domain than another | High,  r>0.75 |
| PF versus Generic health status (such as SF-36 PCS, or EQ5D | Generic encompasses many other conditions as well as current MSK disorder, will likely be less specific to just the MSK disorder. | Moderate  r>0.5 |

**Table 2.** Sample of predetermined hypotheses that will be used in situations when no hypothesis were articulated by authors of primary studies.

Comparison of results to published standards

The adequacy of the results of the measurement property evaluations will be judged against the OMERACT provisional standards for adequate performance 5 (Table 3). These standards were developed by reviewing the standards applied in other systematic reviews of measurement properties in the literature, as well as foundational literature for that property. Over 70 standards were found leading to multiple standards for each property (i.e., 44 for construct validity). These were consolidated by collapsing similar standards together, and then a decision made as to the best one to use at OMERACT. This was called the provisional standards. The X in the SOMP cell will be changed to a ‘+’ if the standard is met or exceeded, a ‘–‘ if the performance of the instrument in this study does not meet the standard, or a ‘±’ if the findings are equivocal or ambivalent.

|  |  |
| --- | --- |
| **Measurement property** | **Provisional standards for adequate performance** |
| Construct validity | Pre-specified hypotheses are met. Should be shown with similar constructs, dissimilar constructs and across known groups to show both presence and absence of a relationship as appropriate for each domain of interest. |
| Inter-method reliability | Intra-class correlation coefficient (ICC); weighted Kappa coefficient (Kw)  Excellent > 0.90.  Good >0.75 (considered the threshold for adequate performance and for a positive rating)  Excellent needed for measurement if done for individual clinical decision making. |
| Test-retest reliability | Intra-class correlation coefficient (ICC); weighted Kappa coefficient (Kw)  Excellent > 0.90.  Good >0.75 (considered adequate for a positive rating)  Excellent needed for measurement if done for individual clinical decision making. |
| Longitudinal construct validity | Consistency with a priori hypothesis of the magnitude and direction of change that should be seen in that situation of change.  If a priori hypothesis suggests a large effect should be observed, one should see an effect size or standardized response mean of >0.80. If moderate is expected, look for 0.5-0.79, small effect 0.2-0.5. Consistency with a priori theory of direction and magnitude of change would be given a positive finding. Findings outside the anticipated range should be considered a negative finding. |
| Clinical trial discrimination (Sensitivity in clinical trials) | **Gold**: Randomized groups demonstrate change in their scores congruent with anticipated effect of the study.  **Silver**: Two group comparison (not randomized) are compared and differences in their change scores are congruent with anticipated results.  **Bronze**: Longitudinal data are provided for the groups that have changed and separately for groups that have remained stable or had a different amount of change compared to the first group.  SRM/ES/T test is greater in change group than in stable group, or group expected to have smaller change. This relative difference is aligned with expected difference in the change experienced in each arm/group.  Ratio of effect size statistics squared is also a way of articulating the relative responsiveness of one measure over another. (ESgroup12/ESgroup22).  If reporting on % exceeding a threshold of meaning (i.e., response criteria), should report proportions for each group.  Results should show a logical, significant relationship to the a priori hypotheses and expectations for the relative difference in the change experienced in the two groups. |
| Thresholds of meaning | There are no “standards” for the value of a calculated threshold. We expect that reporting and context be as clear as possible for users 32 and matched to the intended application.   * ‘+’ = the chosen method was clearly described and results reported and is in a similar context of use (population, setting). * ‘±’ means results were derived only from distribution-based methods (ie, 1 SEM or ½ SD) but in an otherwise similar context.   Points to consider:   * Thresholds are dependent on the anchors used and should be reported and interpreted in that context (i.e., threshold for identifying levels of disease activity), and with sensitivity and specificity of the cut point provided. * For change thresholds, describe relation of both minimal important difference (MID) and minimal detectable change (MDC) and guide interpretation accordingly. Both must be exceeded to be confident in the threshold of change. * Congruence across multiple anchors will bring confidence in the meaning of a threshold score. Difference between results from multiple anchors can be shown using empirical cumulative distribution functions. |

**Table 3:** OMERACT provisional standards for at least adequate performance of an instrument in measurement property evaluation.

## Data Synthesis

Synthesis takes place at two levels using a best-evidence synthesis approach 33 combining information on the quality of the study, the results of the study, and the consistency of the results across multiple studies. The first synthesis gathers the evidence that has been compiled for each measurement property (i.e., the columns of the SOMP, and narrative reporting tables). The number of studies will be combined with the quality of their methods along with the consistency and appraisal of the results are combined and given a final stoplight rating using the criteria shown in Table 4. Our goal is to achieve green rating which is the presence of two or more good quality studies showing favourable findings and in the absence of conflicting results from others studies judged to be of good methods (amber or green good methods check).

|  |  |
| --- | --- |
| **Measurement Property Rating** | **Description** |
| **Green** | 2 or more studies with good enough methods (exclude all high risk of bias ratings for this property) showing consistent positive support (“+”) for the measurement property. With the majority of the remaining studies showing similar + or ± findings as either green or amber quality of methods (no conflicting results). |
| **Amber** | 1 study using good enough methods with + or ambivalent (+/-) support OR…  2+ studies showing some risk of bias with + or +/- results OR…  Inconclusive, inconsistent results from studies with good enough methods |
| **Red** | Multiple studies with good enough methods, but results are showing instrument did not reach performance standards (-) |
| **White** | No studies with good enough methods to assess this measurement property |

**Table 4**. Description of the criteria for synthesis of primary evidence available for each measurement property. Each property is assigned a colour to represent this synthesis.

This provides a profile of the evidence for the instrument performance across the measurement properties as seen in the second last row in the SOMP table (Figure 1).

The final synthesis step is to decide if, based on this profile, the instrument is good enough to represent this domain in a core outcome set. This will be done using an algorithm based on the instrument’s profile across the measurement properties and knowledge that this instrument had already passed content validity and feasibility at an earlier stage (recorded at top of SOMP). An instrument can be endorsed fully, provisionally or not yet endorsed for use in a core outcome set in this field. Briefly, if all the individual measurement properties (columns of the SOMP table) have a synthesis rating of Green, the instrument meets criteria for a full endorsement for inclusion in the core outcome set to represent the target domain in the defined population. If any of the measurement properties are rated as a red (evidence against the instrument) or white (no evidence available), this indicates that the instrument has not passed the OMERACT Filter for inclusion in a COS at this time. Remaining profiles with a mixture of green and amber are provisionally endorsed, identified as Amber in the final synthesis. Amber ratings for an instrument are accompanied by a statement of the work that needs to be done to bring this up to a full endorsement. It is worth noting that many of the amber ratings could be as the result of reporting issues rather than the actual methods used in the studies. More recently there have been guidelines developed to improve the reporting of primary studies of measurement properties, and these are currently being updated 31. These standards will, as they are adopted, improve our ability to be confident in relying on methods and results as reported in published papers.

The focus of this protocol has been on the systematic review of measurement properties, one important step in the OMERACT process for instrument selection. Following this review the compiled evidence will be presented to the Technical Advisory Group and then to the broader OMERACT community for ratification on the working group’s recommendation of this instrument’s ability to represent the target domain in a core outcome set based on this review and the earlier appraisals of domain match and feasibility. Our work has been focused on the [INSTRUMENT NAME AND VERSION] and their ability to represent the target domains of [INSERT TARGET DOMAIN].

# References

1. Clarke M. Standardising outcomes for clinical trials and systematic reviews. Curr Control Trials Cardiovasc Med. 2007;8(1):39-39. doi:10.1186/1745-6215-8-39

2. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete or unusable reports of biomedical research. The Lancet (British edition). 2014;383(9913):267-276. doi:10.1016/S0140-6736(13)62228-X

3. Kirkham JJ, Bracken M, Hind L, Pennington K, Clarke M, Williamson PR. Industry funding was associated with increased use of core outcome sets. J Clin Epidemiol. 2019;115:90-97. doi:10.1016/j.jclinepi.2019.07.007

4. Beaton DE, Maxwell LJ, Shea BJ, et al. Instrument selection using the OMERACT filter 2.1: The OMERACT methodology. Journal of rheumatology. 2019;46(8):1028-1035. doi:10.3899/jrheum.181218

5. Beaton D, Maxwell L, Grosskleg S, et al. The OMERACT Handbook. 2nd ed. (Beaton D, Boers M, Bingham C, et al., eds.).; 2021.

6. Boers M, Kirwan JR, Wells G, et al. Developing Core Outcome Measurement Sets for Clinical Trials: OMERACT Filter 2.0. J Clin Epidemiol. 2014;67(7):745-753. doi:10.1016/j.jclinepi.2013.11.013

7. D’Agostino MA, Beaton DE, Maxwell LJ, et al. Improving domain definition and outcome instrument selection: Lessons learned for OMERACT from imaging. Semin Arthritis Rheum. 2021;51(5):1125-1133. doi:10.1016/j.semarthrit.2021.08.004

8. Maxwell LJ, Beaton DE, Boers M, et al. The evolution of instrument selection for inclusion in core outcome sets at OMERACT: Filter 2.2. Semin Arthritis Rheum. 2021;51(6):1320-1330. doi:10.1016/j.semarthrit.2021.08.011

9. Beaton D, Boers M, Professor of Clinical Epidemiology E, et al. Summary of Findings Tables for Measurement Property Reviews: The Evolution and Application of OMERACT’s Summary of Measurement Properties (SOMP) Table. https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement

10. Covidence. Protocol Development for Systematic Reviews A Practical Guide.; 2024. Accessed January 26, 2025. https://www.covidence.org/wp-content/uploads/2024/10/A\_practical\_guide\_Protocol\_Development\_for\_Systematic\_Reviews.pdf

11. Elsman EBM, Mokkink LB, Terwee CB, et al. Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. J Clin Epidemiol. Published online 2024:111422. doi:10.1016/j.jclinepi.2024.111422

12. Furlan AD, Irvin E. Conducting a Systematic Review and Meta-analysis in Rehabilitation. Am J Phys Med Rehabil. 2022;101(10):965-974. doi:10.1097/PHM.0000000000001933

13. Johnston BC, Patrick DL, Devji T, et al. Patient‐reported outcomes. In: Cochrane Handbook for Systematic Reviews of Interventions. John Wiley & Sons, Ltd; 2019:479-492. doi:10.1002/9781119536604.ch18

14. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015;4(1):1-1. doi:10.1186/2046-4053-4-1

15. Boers M, Brooks P, Strand C V, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. J Rheumatol. 1998;25(2):198-199.

16. Kirwan JR, Boers M, Tugwell P. Updating the omeract filter at omeract 11. Journal of rheumatology. 2014;41(5):975-977. doi:10.3899/jrheum.131306

17. Maxwell LJ, Beaton DE, Boers M, et al. The evolution of instrument selection for inclusion in core outcome sets at OMERACT: Filter 2.2. Semin Arthritis Rheum. 2021;51(6):1320-1330. doi:10.1016/j.semarthrit.2021.08.011

18. COSMIN. Homepage. Accessed May 7, 2024. https://www.cosmin.nl/

19. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Quality of life research. 2018;27(5):1147-1157. doi:10.1007/s11136-018-1798-3

20. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Quality of life research. 2018;27(5):1171-1179. doi:10.1007/s11136-017-1765-4

21. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Med Res Methodol. 2010;10(1):22-22. doi:10.1186/1471-2288-10-22

22. Terwee CB, Jansma EP, Riphagen II, de Vet HCW. Development of a Methodological PubMed Search Filter for Finding Studies on Measurement Properties of Measurement Instruments. Quality of Life Research. 2009;18(8):1115-1123. doi:10.1007/s11136-009-9528-5

23. Beaton D, Maxwell L, Grosskleg S, et al. Instrument Selection for Core Outcome Measurement Sets. 2.1.; 2021.

24. O’Blenis P. How To Train Your Robot: Best Practices For Automated Screening. February 23, 2023. Accessed January 26, 2025. https://www.distillersr.com/resources/blog/how-to-train-your-robot-best-practices-for-automated-screening

25. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: A retrospective evaluation of the Abstrackr machine learning tool. Syst Rev. 2018;7(1):45-45. doi:10.1186/s13643-018-0707-8

26. Gates A, Guitard S, Pillay J, et al. Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. Syst Rev. 2019;8(1):278-278. doi:10.1186/s13643-019-1222-2

27. Hamel C, Hersi M, Kelly SE, et al. Guidance for using artificial intelligence for title and abstract screening while conducting knowledge syntheses. BMC Med Res Methodol. 2021;21(1):285-285. doi:10.1186/s12874-021-01451-2

28. Burns JK, Etherington C, Cheng-Boivin O, Boet S. Using an artificial intelligence tool can be as accurate as human assessors in level one screening for a systematic review. Health Info Libr J. 2024;41(2):136-148. doi:10.1111/hir.12413

29. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Rev. 2021;10(1):89-89. doi:10.1186/s13643-021-01626-4

30. OMERACT. Downloadable Forms. Accessed January 26, 2025. https://omeract.org/instrument-selection/downloadable-forms/

31. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. Quality of life research. 2021;30(8):2197-2218. doi:10.1007/s11136-021-02822-4

32. Devji T, Carrasco-Labra A, Qasim A, et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ. 2020;369:m1714-m1714. doi:10.1136/bmj.m1714

33. SLAVIN R. BEST EVIDENCE SYNTHESIS - AN INTELLIGENT ALTERNATIVE TO METAANALYSIS. J Clin Epidemiol. 1995;48(1):9-18. doi:10.1016/0895-4356(94)00097-A

# Appendix. 1 - Detailed Search Strategy

# Appendix. 2 - Risk of Bias Items for each Measurement Property