Risk of bias assessments in individual participant data meta-analyses of test accuracy and prediction models

Levis, Brooke; Snell, Kym I E; Damen, Johanna A A; Hattle, Miriam; Ensor, Joie; Dhiman, Paula; Andaur Navarro, Constanza L; Takwoingi, Yemisi; Whiting, Penny F; Debray, Thomas P A; Reitsma, Johannes B; Moons, Karel G M; Collins, Gary S; Riley, Richard D

DOI:
10.1016/j.jclinepi.2023.10.022

Journal of Clinical Epidemiology 165 (2024) 111206

Journal of Clinical Epidemiology

REVIEW ARTICLE

Risk of bias assessments in individual participant data meta-analyses of test accuracy and prediction models: a review shows improvements are needed

Brooke Levis[a,b,*], Kym I.E. Snell[c,d], Johanna A.A. Damen[e], Miriam Hattle[c,d], Joie Ensor[c,d], Paula Dhiman[f], Constanza L. Andaur Navarro[e], Yemisi Takwoingi[c,d], Penny F. Whiting[g], Thomas P.A. Debray[e], Johannes B. Reitsma[e], Karel G.M. Moons[e], Gary S. Collins[f], Richard D. Riley[c,d,**]

[a]*Centre for Prognosis Research, School of Medicine, Keele University, Keele, Staffordshire, UK*
[b]*Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada*
[c]*Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK*
[d]*National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, Birmingham, UK*
[e]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[f]*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK*
[g]*School of Social and Community Medicine, University of Bristol, Bristol, UK*

Accepted 30 October 2023; Published online 2 November 2023

Abstract

Objectives: Risk of bias assessments are important in meta-analyses of both aggregate and individual participant data (IPD). There is limited evidence on whether and how risk of bias of included studies or datasets in IPD meta-analyses (IPDMAs) is assessed. We review how risk of bias is currently assessed, reported, and incorporated in IPDMAs of test accuracy and clinical prediction model studies and provide recommendations for improvement.

Study Design and Setting: We searched PubMed (January 2018−May 2020) to identify IPDMAs of test accuracy and prediction models, then elicited whether each IPDMA assessed risk of bias of included studies and, if so, how assessments were reported and subsequently incorporated into the IPDMAs.

Results: Forty-nine IPDMAs were included. Nineteen of 27 (70%) test accuracy IPDMAs assessed risk of bias, compared to 5 of 22 (23%) prediction model IPDMAs. Seventeen of 19 (89%) test accuracy IPDMAs used Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2), but no tool was used consistently among prediction model IPDMAs. Of IPDMAs assessing risk of bias, 7 (37%) test accuracy IPDMAs and 1 (20%) prediction model IPDMA provided details on the information sources (e.g., the original manuscript, IPD, primary investigators) used to inform judgments, and 4 (21%) test accuracy IPDMAs and 1 (20%) prediction model IPDMA provided information or whether assessments were done before or after obtaining the IPD of the included studies or datasets. Of all included IPDMAs, only seven test accuracy IPDMAs (26%) and one prediction model IPDMA (5%) incorporated risk of bias assessments into their meta-analyses. For future IPDMA projects, we provide guidance on how to adapt tools such as Prediction model Risk Of Bias ASsessment Tool (for prediction models) and QUADAS-2 (for test accuracy) to assess risk of bias of included primary studies and their IPD.

Funding: Brooke Levis was supported by a Fonds de Recherche du Québec−Santé Postdoctoral Training Fellowship. Kym IE Snell, Joie Ensor, Gary S Collins, Miriam Hattle, and Richard D Riley were supported by funding from the MRC-NIHR Better Methods Better Research panel (grant reference: MR/V038168/1). Gary S Collins was also supported by Cancer Research UK (programme grant: C49297/A27294). Paula Dhiman was supported by the NIHR Biomedical Research Centre, Oxford. Yemisi Takwoingi was funded by a National Institute for Health Research (NIHR) Postdoctoral Fellowship. Yemisi Takwoingi, Richard D Riley, Kym IE Snell, Joie Ensor, and Miriam Hattle were supported by the NIHR Birmingham Biomedical Research Center. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

* Corresponding author. Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Jewish General Hospital, 3755 Cote Ste-Catherine, Montreal, Quebec H3T 1E2, Canada. Tel.: +1-514-340-8222x28389; fax: +1-514-340-7564.

** Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham B15 2TT, UK. Tel.: +44-0-121-414-3344; fax: +44-0-121-414-3971.

E-mail addresses: brooke.levis@mail.mcgill.ca (B. Levis); r.d.riley@bham.ac.uk (R.D. Riley).

https://doi.org/10.1016/j.jclinepi.2023.10.022

Conclusion: Risk of bias assessments and their reporting need to be improved in IPDMAs of test accuracy and, especially, prediction model studies. Using recommended tools, both before and after IPD are obtained, will address this.

Keywords: Risk of bias; Individual participant data meta-analysis; Test accuracy; Prediction models; Applicability; Quality; QUADAS-2; PROBAST

1. Introduction

Individual participant data meta-analyses (IPDMAs) are increasingly common [1,2]. They involve obtaining, checking, harmonizing, and synthesizing participant-level data from multiple studies, rather than pooling published or reported study results (aggregate data). IPDMAs differ from aggregate data meta-analyses in that participant eligibility criteria for an IPDMA may differ from eligibility criteria in the primary studies, and IPDMA can lead to improvements through, for example, better standardization of variable definitions (e.g., index tests, predictors, reference standards, outcomes), and improved analysis methods both within and across included datasets or studies. Furthermore, collaborating investigators may be able to provide additional information about the original studies.

As for aggregate data meta-analyses, assessing risk of bias (RoB) and applicability of included studies (and their IPD) should be a critical part of any IPDMA project. RoB relates to the internal validity of an included study (e.g., *does it avoid bias in study results?*) whereas applicability relates to external validity (e.g., *does it match the population or setting of interest?*).

The Preferred Reporting Items for Systematic Review and Meta-Analysis of Individual Participant Data reporting guideline [3], which is mainly focused on intervention studies, includes items for assessing and reporting RoB within and across studies. It states that review authors should describe "*how findings of IPD checking were used to inform the assessment*" and "*if and how risk of bias assessment was used in any data synthesis.*" The authors provide brief guidance, but do not recommend specific RoB assessment tools, describe how existing tools might be tailored, or discuss how RoB judgments might be incorporated into analyses.

A recently published textbook provides preliminary guidance on how to undertake RoB assessments in IPDMAs [2]. The textbook emphasizes that RoB of included datasets or studies should be examined at multiple stages (in particular, before and after IPD collection) and might be done by using and adapting elements of existing tools (e.g., the Cochrane Risk of Bias-2 [ROB-2] tool [4]). Despite the importance of RoB assessment, a recent review of 323 IPDMAs of intervention effects found that only 43% used a satisfactory technique to assess RoB of included trials, and only 40% accounted for RoB when interpreting results [5].

IPDMAs can also be conducted to summarize test accuracy and to develop or validate clinical prediction models. Test accuracy studies evaluate the performance of an index test (or the comparative accuracy of two or more index tests) against a reference standard (e.g., in terms of sensitivity, specificity, positive and negative predictive value, area under the receiver operating characteristic curve). Prediction model studies typically (i) develop a multivariable model for predicting an outcome (prognosis) or detecting a particular condition (diagnosis) in individuals, or (ii) evaluate the performance of one or more existing models (e.g., in terms of their calibration and discrimination performance).

In systematic reviews and meta-analyses of aggregate data, QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies-2; www.bristol.ac.uk/population-health-sciences/projects/quadas/quadas-2/) [6] and PROBAST (Prediction model Risk Of Bias ASsessment Tool; www.probast.org) [7,8] can be used to examine the methodological quality of test accuracy and prediction model studies, respectively. These are the only tools designed specifically for these study types, and they are recommended by the Cochrane test accuracy and prognosis groups [9,10]. QUADAS-2 signaling items are categorized into domains of *patient selection*, *index test*, *reference standard*, and *flow and timing*; an extension for assessing the RoB in comparative accuracy studies, QUADAS-C, was published in 2021 [11]. PROBAST signaling items are categorized into domains of *participants*, *predictors*, *outcome*, and *analysis*. However, it is not currently clear whether and how RoB is assessed in IPDMAs for test accuracy or prediction model research. QUADAS-2 and PROBAST also include domain-level items to assess applicability, but it is not clear whether IPDMAs do this.

This article aimed to review how RoB and applicability concerns are assessed, reported, and incorporated in recent IPDMAs of test accuracy and prediction modeling studies. Based on the findings, we also provide guidance on how researchers might undertake and improve RoB assessments in future IPDMAs of such studies.

2. Materials and methods

2.1. Part 1: review

2.1.1. IPDMA eligibility criteria

For a published IPDMA project to be eligible for our review, it must (1) aim to conduct an IPDMA; (2) use IPD from multiple studies or data sources to (i) examine or compare test accuracy or (ii) develop or validate a multivariable prediction

What is new?

Key findings

- Our review suggests risk of bias is rarely assessed in individual participant data meta-analyses (IPDMAs) for prediction model research but is often assessed for IPDMAs of test accuracy using the QUADAS-2 tool. Even when risk of bias is considered, better reporting of risk of bias results is still needed in IPDMAs of both test accuracy and prediction model studies.

What this adds to what was known?

- A recent review of 323 IPDMAs of intervention effects found that only 43% used a satisfactory technique to assess the risk of bias of the included trials, and only 40% accounted for risk of bias when interpreting results. Our research shows that similar concerns also hold for IPDMAs of test accuracy and prediction model studies.

What is the implication and what should change now?

- As with meta-analyses of aggregate data, risk of bias should be routinely assessed, reported, and incorporated in IPDMAs of test accuracy and prediction model studies. We provide guidance on how to do this, both before and after IPD are obtained.

model; (3) include human data (i.e., not animal data); and (4) have a medical focus (i.e., trying to answer a medical question), and not a methodological focus. We excluded IPDMAs (1) using IPD from a single multisite or multicenter study (e.g., all part of the same overarching study but combining different research centers); (2) using a network meta-analysis approach; (3) using simulated or reconstructed data (e.g., IPD reconstructed from published 2 $\times$ 2 tables); or (4) using machine learning or artificial intelligence as the primary analyses. The latter require substantial additional considerations that are outside the scope of this review. Protocols for planned IPDMAs (without results) and abstracts without an associated full text (e.g., conference abstracts) were also excluded.

2.1.2. Database search and study selection

To identify a set of recent IPDMAs, we searched PubMed (on May 7, 2020) via DistillerSR (Evidence Partners, Ottawa, Canada) from January 1, 2018, to May 7, 2020, using a search strategy that included elements for (1) IPD, (2) meta-analysis, and (3) test accuracy or prediction models. The complete search strategy is provided in Appendix 1. Titles and abstracts were reviewed by one

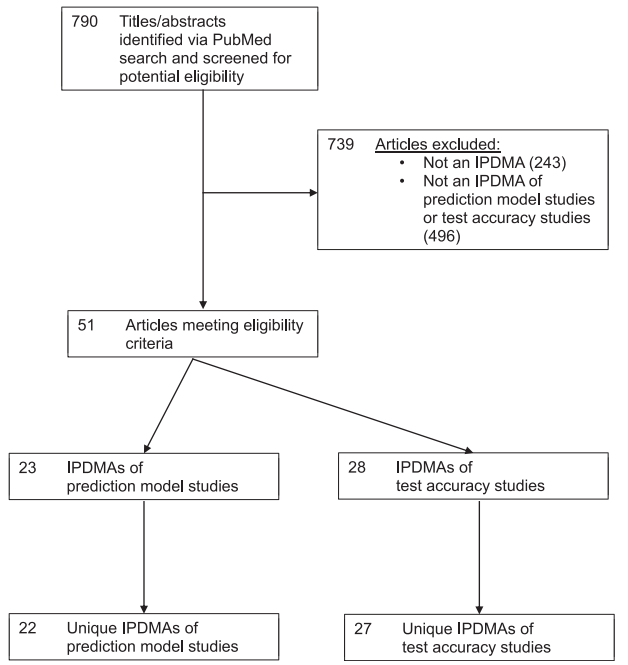

Fig. 1. Flow chart. IPDMAs, individual participant data meta-analyses.

investigator (B.L.) with substantial experience of IPDMA projects, consulting with another experienced investigator (R.D.R.) as necessary. If a decision regarding eligibility could not be made based on the abstract alone, the full text was retrieved and evaluated for eligibility. Eligibility based on the full text was confirmed during the data extraction phase by two investigators, including the investigator who reviewed all titles and abstracts. Duplicate entries were excluded manually. In addition, in two instances where the same IPD project team published two IPDMAs with the same datasets, we included the publication that addressed the primary research question of the IPDMA collaboration and excluded the publication that addressed a secondary research question.

2.1.3. Undertaking data extraction

For each included IPDMA article, we extracted:

- the PubMed ID;
- the first author's surname, the journal, and year of publication;
- the type of IPDMA (test accuracy or prediction model);
- the objective, including whether a single test or multiple tests were evaluated or compared (test accuracy IPDMAs only), and whether prediction models were developed, validated, or both (prediction model IPDMAs only);
- whether a formal RoB and/or applicability assessment was conducted.

Table 1. Risk of bias and applicability assessment in IPDMAs of test accuracy studies

Study	Objective (to evaluate/compare single vs. multiple tests)	Whether risk of bias and/or applicability was assessed	Tool used	Levels assessed (study and/or participant level)	Sources used to inform assessments
Adderley, 2020 [12]	Single	Yes	QUADAS-2	Study only	Paper + unclear for others
Al-Rubaie, 2018 [13]	Multiple	No	–	–	–
Bima, 2020 [14]	Single	Yes	QUADAS-2	Study only	Unclear
Broger, 2020 [15]	Multiple	No	–	–	–
Gupta, 2020 [16]	Multiple	Yes	Newcastle Ottawa Scale	Study and participant	IPD + authors + unclear for others
Haase, 2019 [17]	Single	Yes	QUADAS-2	Study only	Paper
Herrmann, 2018 [18]	Multiple	Yes	Single item on patient selection (applicability)	Study and participant	IPD
Hsu, 2019 [19]	Multiple	Yes	QUADAS	Study only	Unclear
Kalafat, 2018 [20]	Single	Yes	QUADAS-2	Study only	Unclear
Karlas, 2018 [21]	Single	No	–	–	–
Kazakos, 2020 [22]	Multiple	Yes	QUADAS-2	Study only	Paper
Klein Nulent, 2018 [23]	Single	Yes	QUADAS-2	Study only	Paper + unclear for others
Kubo, 2018 [24]	Single	Yes	QUADAS-2	Study only	Unclear
Lee, 2019 [25]	Single	No	–	–	–
Levis, 2019 [26]	Single	Yes	QUADAS-2	Study and participant	Paper + IPD + unclear for others
Ley, 2019 [27]	Single	Yes	QUADAS-2	Study only	Unclear
Nguyen-Khac, 2018 [28]	Single	Yes	QUADAS-2	Study only	Unclear
Parpia, 2020 [29]	Multiple	No	–	–	–
Pavlovic, 2019 [30]	Multiple	No	–	–	–
Raskovalova, 2021 [31]	Multiple	Yes	QUADAS-2	Study only	Unclear
Suh, 2018 [32]	Single	Yes	QUADAS-2	Study only	Unclear
Suh, 2019 [33]	Single	Yes	QUADAS-2	Study only	Unclear
Thiele, 2020 [34]	Single	Yes	QUADAS-2	Study only	Unclear
van Doorn, 2018 [35]	Single	No	–	–	–
Westra, 2019 [36]	Multiple	Yes	QUADAS-2	Study only	Unclear
Whiting, 2018 [37]	Single	Yes	QUADAS-2	Study only	Unclear
Yoshida, 2021 [38]	Single	Yes	QUADAS-2	Study only	Paper + unclear for others

If RoB and/or applicability was assessed, we extracted information on

- tools or items used (including whether [and what] adaptations were made);
- whether assessments were done at the study or participant level (i.e., whether RoB was assessed for the study as a whole, or whether RoB was assessed separately for each included participant based on the individual-level characteristics, such as diagnostic instrument used or number of days between assessments);
- what sources were used to inform assessments (e.g., the original manuscript, the study protocol, the IPD, correspondence with primary investigators);

Timing of assessment (before or after receiving IPD)	Presentation of assessments	Incorporation of risk of bias assessments into analyses
Unclear	Study by study + in aggregate; domain scores only	Not done
–	–	–
Unclear	Study by study + in aggregate; domain scores only	Not done
–	–	–
After + unclear before	Study by study only; all item scores	Not done
Unclear	Study by study + in aggregate; domain scores only	Not done
After	Not reported	Additional patients added in sensitivity analyses
Unclear	Study by study only; all item scores	Not applicable (all studies at low risk of bias)
Before + unclear after	Study by study only; domain scores only	One study excluded due to risk of selection bias
–	–	–
Unclear	Study by study + in aggregate; domain scores + signaling items	Not done
Before only	Study by study + in aggregate; domain scores only	Studies with high risk of bias were excluded from analyses
Unclear	Study by study + in aggregate; domain scores only	Not done
–	–	–
After + unclear before	Study by study only; domain scores + signaling items	Subgroup analyses were conducted based on QUADAS-2 item scores
Unclear	Study by study + in aggregate; domain scores only	Studies with high risk of bias were excluded in sensitivity analyses
Unclear	Study by study only; domain scores only	Not done
–	–	–
–	–	–
Unclear	Study by study + in aggregate; domain scores only	Subgroup analyses were conducted to compare studies with low risk of bias to studies with high risk of bias based on a QUADAS-2 score of 5+.
Unclear	Study by study + in aggregate; domain scores only	Additional analyses were conducted among studies with consecutive enrollment
Unclear	Study by study + in aggregate; domain scores only	Meta-regression including a variable for blinding was conducted
Unclear	Study by study + in aggregate; domain scores + signaling items	Not done
–	–	–
Unclear	Study by study + in aggregate; domain scores only	Not done
Unclear	In aggregate only; domain scores only	Not done
Unclear	Study by study + in aggregate; domain scores + signaling items	Not done

Abbreviations: QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies-2; IPDMAs, individual participant data meta-analyses.

- whether assessments were done before and/or after receiving the data;
- how assessments were presented in the report (e.g., the level of detail), and;
- whether assessments were incorporated into analyses (e.g., sensitivity analyses excluding studies with a high RoB).

The data extraction sheet can be found in Appendix 2. IPDMAs were assigned to 13 investigators for data extraction and verification. Data were extracted by one investigator and checked by a second investigator. Disagreements were resolved by consensus, consulting the senior investigator (R.D.R.) as necessary. For each IPDMA, we considered information provided in (1) the main

publication, (2) any appendices, and (3) any protocol or registration that was mentioned in the main publication.

2.1.4. Summarizing data extracted on RoB and applicability

Data extraction results were summarized separately for each IPDMA article type: test accuracy or prediction model. For each type, we determined the percentage of IPDMA articles that included an assessment of RoB and applicability. Among those that completed an assessment, we summarized how the assessments were made (tools used, including any adaptations; sources of information; timing of assessment), how assessments were reported (level of detail), and whether assessments were incorporated into analyses (e.g., subgroup analyses). We calculated 95% confidence intervals (CIs) for proportions using the Clopper and Pearson 'exact' approach.

2.2. Part 2: developing guidance for future IPDMA projects

Based on the findings of the review, we produced guidance for undertaking RoB and applicability assessments in IPDMAs of test accuracy and prediction models. We followed the framework recommended in Chapters 15 and 17 of Riley et al. [2], to utilize items within existing tools proposed for non-IPD reviews of test accuracy and prediction models.

3. Results of the review

3.1. Included articles and characteristics

A total of 790 titles and abstracts were retrieved from the search, of which 27 unique IPDMAs of test accuracy studies [12−38] and 22 unique IPDMAs of prediction model studies met inclusion criteria and were included [39−60] (Fig. 1). Of the test accuracy IPDMAs, 17 aimed to evaluate the accuracy of a single diagnostic index test (63%), and 10 aimed to evaluate or compare multiple index tests (37%). Of the prediction model IPDMAs, 11 aimed to develop a prediction model (including internal validation, assessment of the added value of a particular predictor to an existing model, and creation of risk scores/groups) (50%), nine aimed to externally validate an existing prediction model (including model updating [e.g., recalibration] and data splits by different independent studies or sources) (41%), and two aimed to both develop and externally validate a model (9%).

3.2. RoB findings for test accuracy IPDMAs

Nineteen of 27 (70%, 95% CI: 50−86%) test accuracy IPDMAs assessed RoB of their included primary studies (or sources), of which 17 used the QUADAS-2 tool [12,14,17,20,22−24,26−28,31−34,36−38], one used the original QUADAS tool [19], and one used the Newcastle Ottawa Scale [16]. A summary of findings is presented in Table 1, and specific examples are described in Box 1.

The sources used to inform RoB assessments were not clearly reported; seven IPDMAs reported some information on sources (26%, 95% CI: 11−46%), with six using the original manuscript [12,17,22,23,26,38], two using the IPD itself [16,26], and one contacting study authors [16]. Timing of RoB assessments, that is, at what stage of the IPDMA itself, was also not clearly reported; four IPDMAs provided some information on timing (15%, 95% CI: 4−34%). One IPDMA stated that RoB assessments were only done prior to obtaining the IPD [23] and three other IPDMAs gave partial information, with one reporting carrying out assessments before receiving the IPD [20] and two reporting the assessments being performed after obtaining IPD [16,26].

Two IPDMAs (7%, 95% CI: 1−25%) assessed RoB at the participant level. One IPDMA noted that participants' RNA sequence data informed the response for various items on the Newcastle Ottawa Scale [16]. Another IPDMA assessed a QUADAS-2 signaling item (Item 4.1: *Was there an appropriate interval between index test(s) and reference standard?*) at the individual participant level, allowing different participants to have a different RoB, and deemed that another QUADAS-2 item (Item 2.2: *If a threshold was used, was it prespecified?*) was not applicable in the IPD context, given that the availability of IPD allows for examining accuracy at any threshold [26]. The other 17 IPDMAs that assessed RoB only made assessments at the study level.

In terms of presenting RoB results, 18 IPDMAs (67%, 95% CI: 46−83%) presented results separately for each included primary test accuracy study [12,14,16,17,19,20,22−24,26−28,31−34,36,38], while one (4%, 95% CI: 0−19%) presented results at the aggregate level (i.e., summarized across all studies) [37]. Of the 17 IPDMAs using QUADAS-2, most (13 of 17, 76%, 95% CI: 50−93%) reported domain judgments, while four reported answers for all signaling questions and domain-level judgments separately.

Seven (26%, 95% CI: 11−46%) IPDMAs incorporated RoB assessments into analyses, including four that excluded studies in main or sensitivity analyses [20,23,27,32], two that compared subgroups based on answers to signaling questions or overall study level RoB judgment [26,31], and one that included a QUADAS-2 signaling question (blinding) as part of a meta-regression [33]. In one IPDMA, incorporation of assessments was not applicable as all included studies were judged to have a low RoB [19].

3.3. Applicability findings for test accuracy IPDMAs

Of the 17 IPDMAs that used QUADAS-2 (including its applicability domains), one assessed the QUADAS-2 applicability domain on patient selection (*Are there concerns that the included patients do not match the review question?*) at

Box 1 Examples of IPDMAs assessing and incorporating risk of bias and applicability.

Study	Aspect	Example
Haase, 2019 [17]	• Considering participants with inconclusive test results	• In an IPDMA of computed tomography angiography (CTA) for obstructive coronary artery disease diagnosis in patients with stable chest pain, the IPDMA team included all participants in their primary analysis, regardless of whether they had evaluable or unevaluable CTA examinations. The IPDMA team applied a worst-case scenario in which unevaluable CTA results were considered false positive if coronary angiography was negative and false negative if coronary angiography was positive.
Herrmann, 2018 [18]	• Removing vs. including participants from analyses to address applicability	• In an IPDMA of two-dimensional shear wave elastography for evaluation of liver fibrosis, some primary studies did not perform liver biopsy in patients with known liver cirrhosis based on the clinical histories. The IPDMA team did not include these patients in their main analyses but noted that excluding them may have led to a focus on less severe cirrhosis patients. To assess a potential bias, they performed an additional analysis where they reincluded these patients.
Levis, 2019 [26]	• Using the IPD to get different risk of bias and applicability classifications than would be possible based on the published reports alone • Assessing risk of bias at the participant level	• In an IPDMA on the accuracy of the Patient Health Questionnaire-9 for screening to detect major depression, primary studies often used a wide range of time intervals between the index test assessment (depression screening tool) and reference standard assessment (diagnostic interview), raising concerns about risk of bias and lack of applicability. • The IPDMA team used the IPD from such studies to identify and include the subset of participants with reasonable time intervals, thereby alleviating these concerns. • Moreover, within the subset with ''reasonable'' time intervals, the IPDMA team assessed QUADAS-2 signaling item 4.1 (*Was there an appropriate interval between index test(s) and reference standard?*) at the individual participant level, allowing different participants to have a different risk of bias based on the length of the interval.
Levis, 2019 [26]	• Modifying QUADAS-2 to suit the IPDMA context	• The IPDMA team deemed that QUADAS-2 item 2.2 (*If a threshold was used, was it prespecified?*) was not applicable in the IPD context, given that the availability of IPD allowed for the examination of accuracy at all thresholds.
Levis, 2019 [26]	• Incorporating risk of bias assessments into analyses: subgroup analyses	• The IPDMA team performed subgroup analyses based on QUADAS-2 item scores, including for items that were assessed at the participant level.
Suh, 2019 [33]	• Incorporating risk of bias assessments into analyses: meta-regression	• In an IPDMA of 2-hydroxyglutarate magnetic resonance spectroscopy for prediction of isocitrate dehydrogenase mutant glioma, the IPDMA team conducted a meta-regression to explain the effects of study heterogeneity. In the meta-regression, they included a variable on the blinding of the index test assessor to the reference standard result.

the participant level, allowing different participants to have a different applicability concern and compared subgroups based on participant-level judgments [26].

In addition to the IPDMAs above that used QUADAS or QUADAS-2 (which assess *both* RoB and applicability concerns), one IPDMA excluded some participants in main

Table 2. Risk of bias and applicability assessment in IPDMAs of prediction model studies

Study	Objective (to develop vs. validate a model)	Whether risk of bias and/or applicability was assessed	Tool used	Levels assessed (study and/or participant level)	Sources used to inform assessments
Al-Shahi Salman, 2018 [39]	Development + validation	No	–	–	–
Antonopoulos, 2020 [40]	Development + validation	No	–	–	–
Cao, 2020 [41]	Development	No	–	–	–
Condoluci, 2020 [42]	Development + validation	No	–	–	–
Crawford, 2018 [43]	Development + validation	Yes	Checklist defined by IPDMA authors[a]	Study only	Unclear
Depmann, 2018 [44]	Development	No	–	–	–
Ediebah, 2018 [45]	Development + validation	No	–	–	–
Hopkins, 2019 [46]	Development + validation	No	–	–	–
Hudda, 2019 [47]	Development + validation	No	–	–	–
Jaja, 2018 [48]	Development + validation	No	–	–	–
Jonkman, 2019 [49]	Development	No	–	–	–
Kievit, 2018 [50]	Development	No	–	–	–
Lee, 2019 [51]	Development	No	–	–	–
Malda, 2019 [52]	Development	Yes	Newcastle Ottawa Scale	Study only	IPD + unclear for others
Pennells, 2019 [53]	Validation	No	–	–	–
Phillips, 2020 [54]	Validation	No	–	–	–
Saczkowski, 2018 [55]	Development	Yes	Down's and Black quality score	Study only	Unclear
Shinohara, 2019 [56]	Development	No	–	–	–
Spronk, 2020 [57]	Development	No	–	–	–
Verma, 2019 [58]	Development	Yes	Adapted QUADAS-2[b]	Study only	Unclear
Vickers, 2018 [59]	Development	No	–	–	–
Vollgraff Heidweiller-Schreurs, 2020 [60]	Development	Yes	Adapted QUADAS-2[c]	Study only	Unclear

analyses due to applicability concerns but included them in sensitivity analyses [18].

3.4. RoB findings for prediction model IPDMAs

Five of 22 (23%, 95% CI: 8−45%) IPDMAs of prediction model studies assessed RoB of the primary studies, of which one used an author-defined checklist [43], one used the Newcastle Ottawa Scale [52], one used the Down's and Black quality score [55], and two used adapted versions of QUADAS-2 [58,60]. See Table 2. All five assessed RoB at the study-level only. The sources used to inform assessments and the timing of assessments were not clearly reported; only one IPDMA reported at least some information, noting that some of the assessments were based on study-level information (e.g., follow-up rate) derived from the IPD, although other potential sources were unclear [52].

All five prediction model IPDMAs that assessed RoB presented results separately for each included primary study. The two prediction model IPDMA studies using QUADAS-2 reported domain-level judgments for each study, with the exception of one IPDMA that reported on the blinding signaling question separately [60]. Of the three IPDMAs using other quality assessment tools, one presented results for each item of the tool separately, for each study separately [43], one only provided a single overall score per study [55], and one presented results for each item separately and provided an overall score per study [52].

Only one (5%, 95% CI: 0−23%) IPDMA incorporated RoB assessments into their meta-analyses, considering the total Down's and Black quality score as a candidate predictor for the prediction model being developed [55]. In one other IPDMA, incorporation of assessments was reported to not be possible due to a lack of data for the main outcome among the relevant studies [60].

Timing of assessment (before or after receiving IPD)	Presentation of assessments	Incorporation of risk of bias assessments into analyses
–	–	–
–	–	–
–	–	–
–	–	–
Unclear	Study by study + in aggregate; item by item	Not done
–	–	–
–	–	–
–	–	–
–	–	–
–	–	–
–	–	–
–	–	–
–	–	–
After + unclear before	Study by study + in aggregate; item by item + total score	Not done
–	–	–
–	–	–
Unclear	Study by study + in aggregate; one value (0–32) per study	Down's and Black score was a candidate predictor for the model
–	–	–
–	–	–
Unclear	Study by study + in aggregate; domain scores only	Not done
–	–	–
Unclear	Study by study + in aggregate; domain scores only, except for blinding	Not done; described as not possible

Abbreviations: QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies-2; IPDMAs, individual participant data meta-analyses.

[a] Checklist included five yes/no items: consecutive sample, sufficient follow-up length for outcome to develop, possibility of replication based on published report, blinding of outcome assessors to index test, and whether sample size was justified.

[b] The reference standard domain was omitted.

[c] In domain 4 ('flow and timing'), the time interval between test and delivery was considered not applicable.

3.5. Applicability findings for prediction model IPDMA projects

The Newcastle Ottawa Scale, Down's and Black quality score, and QUADAS-2 tool, which were used in four of the above IPDMAs [52,55,58,60] include some items related to applicability, but none of the IPDMAs using these tools incorporated the applicability judgments into analyses, except for the IPDMA that included the total Down's and Black quality score as a candidate predictor [55].

4. Guidance for examining RoB and applicability in future IPDMAs

The findings of the review suggest that improvements in RoB and applicability assessments in IPDMAs are needed. We now provide recommendations for examining RoB and applicability in IPDMAs for test accuracy and prediction models, adapting guidance provided in the textbook of Riley et al. [2] (See Box 2 for a summary and Box 1 for specific examples).

Authors of IPDMAs of test accuracy and prediction model studies should assess the methodological quality of each study providing IPD by using information from the study's published report(s), appendices, protocols, and the IPD itself. This should be done using items from QUADAS-2 (see Box 3) and PROBAST (see Box 4), for test accuracy and prediction models, respectively, considering both RoB and applicability concerns.

Ideally, assessments should be undertaken in two stages: First, authors can assess the methodological quality of potential datasets *before* seeking the IPD (this may lead to IPD not being sought from studies deemed at a high RoB,

Box 2 Summary of recommendations for examining risk of bias and applicability in IPDMAs for test accuracy and prediction models.

Topic	Guidance
Sources of information to inform judgments for each study or dataset	• Published report(s) that used the datasets, appendices, protocols, and the IPD itself
Items to consider in assessment	• Items from QUADAS-2 (or QUADAS-C) for test accuracy IPDMAs • Items from PROBAST for prediction models IPDMAs • See Box 3 and Box 4 for specific recommendations regarding each item from each tool
Stages and timing of assessment	• First, examine risk of bias before seeking the IPD (this may lead to IPD not being sought from studies deemed at a high risk of bias) • Then, examine risk of bias after the IPD is obtained, cleaned, and checked (also using any extra information available from the IPD itself or study authors)
Tailoring of signaling questions	• Tailor the signaling questions based on the clinical context (e.g., adding items related to clinical or educational qualifications of personnel making classifications)
Study vs. participant-level assessments	• Consider what items might be applicable at the participant level (e.g., timing of assessments) as opposed to the study level (e.g., consecutive recruitment) • Pay special attention to participants with inconclusive results on the index test, reference standard, predictor(s), or outcome
Reporting of assessments	• Report results of the risk of bias and applicability assessment for each included study/dataset and overall. • Consider reporting judgments from individual signaling questions in addition to domain-level judgments
Use and interpretation of assessments	• Use results of the risk of bias and applicability assessments to inform analyses and interpretation of the IPDMA's results (e.g., sensitivity analyses reinstating participants or excluding studies and participants at high risk of bias)

especially if obtaining IPD would not resolve high RoB concerns). Then, once the IPD is obtained, cleaned, and checked, the authors can update the quality assessments using any extra information available from the IPD itself (or from study authors collaborating on the IPD project).

As QUADAS-2 and PROBAST (www.probast.org) are generic tools designed for all medical domains and areas, the IPDMA team may need to tailor the signaling questions based on the clinical context. For example, additional signaling questions may be necessary, and some signaling questions may no longer be applicable in the IPD context. For instance, one IPDMA on depression screening tool accuracy added an item to the QUADAS-2 reference standard domain related to the clinical qualification of the assessor [26]. In addition, the QUADAS-2 signaling question "Item 2.2: *If a threshold was used, was it prespecified?*" is irrelevant if provided IPD allow the IPDMA team to select their own thresholds (or evaluate all possible thresholds). Similarly, the analysis domain of PROBAST may be redundant, given that the IPDMA team has the freedom to alter and improve the analyses conducted.

Authors of IPDMAs of test accuracy studies should pay special attention to participants in the dataset who have inconclusive index test or reference standard results. While the original study may have excluded such participants from analyses, the IPDMA team can reinstate them in either main or sensitivity analyses, as appropriate, thus reducing RoB related to included participants. The availability of IPD also allows application of multiple imputation methods to impute missing index test or reference standard results [61,62].

Authors should consider the incorporation of participant-level assessments when relevant (e.g., timing of assessments) into their judgments. Results of the RoB and applicability assessment for each included dataset (and overall) can be summarized in a table, graphically, or both. The results should not only be reported but also incorporated into the discussion and conclusions of the IPDMA report. When appropriate, results should also be used to inform sensitivity analyses.

Box 3 QUADAS-2. Domains and signaling questions from the QUADAS-2 tool [6], which may be used to examine the methodological quality of IPD from each study or dataset contributing to the IPDMA project for test accuracy research.

Domain and signaling questions	Guidance
Domain 1: Patient selection	
1.1 Was a consecutive or random sample of patients enrolled?	Judge at the study/dataset level, in the same way as for aggregate data meta-analyses
1.2 Was a case-control design avoided?	Judge at the study/dataset level, in the same way as for aggregate data meta-analyses
1.3 Did the study avoid inappropriate exclusions?	May be possible to judge at the participant level. In addition, if the IPD allows for the reinstation of participants excluded for inappropriate reasons, this concern can be avoided. Note, however, that IPD cannot overcome the inappropriate exclusion of participants from a study's original sampling frame
Applicability: Are there concerns that the included patients and setting do not match the review question?	May be possible to judge at the participant level. In addition, if additional inclusion/exclusion criteria can be applied to each dataset, this concern can be avoided
Domain 2: Index test	
2.1 Were the index test results interpreted without knowledge of the results of the reference standard?	May be possible to judge at the participant level
2.2 If a threshold was used, was it prespecified?	Judge at the study/dataset level. However, if the IPD allows for all possible thresholds to be evaluated, then this item is not applicable and can be omitted
Applicability: Are there concerns that the index test, its conduct, or interpretation differ from the review question?	May be possible to judge at the participant level. In addition, if additional inclusion/exclusion criteria can be applied to each dataset, this concern can be avoided
Domain 3: Reference standard	
3.1 Is the reference standard likely to correctly classify the target condition?	May be possible to judge at the participant level (if multiple reference standards are used)
3.2 Were the reference standard results interpreted without knowledge of the results of the index test?	May be possible to judge at the participant level
Applicability: Are there concerns that the target condition as defined by the reference standard does not match the review question?	Judge at the study/dataset level, in the same way as for aggregate data meta-analyses. In addition, if additional inclusion/exclusion criteria can be applied to each dataset, this concern can be avoided
Domain 4: Index test	
4.1 Was there an appropriate interval between index test and reference standard?	May be possible to judge at the participant level. In addition, if additional inclusion/exclusion criteria can be applied to each dataset, this concern can be avoided
4.2 Did all patients receive a reference standard?	Judge at the study/dataset level, in the same way as for aggregate data meta-analyses
4.3 Did all patients receive the same reference standard?	Judge at the study/dataset level, in the same way as for aggregate data meta-analyses
4.4 Were all patients included in the analysis?	Judge at the study/dataset level. However, if the IPD allows for the reinstation of previously excluded participants (e.g., those this inconclusive index test or reference standard results), then this concern can be avoided

Source: The first column of Box 3 presents QUADAS-2 domains and signaling questions, freely available at https://www.bristol.ac.uk/population-health-sciences/projects/quadas/quadas-2/. QUADAS-2 was originally published by Whiting et al. [6]

Box 4 PROBAST. Domains and signaling questions within the first three domains of the PROBAST tool (Prediction model study Risk Of Bias ASsessment Tool) [7,8], which may be used to examine the methodological quality of IPD from each study or dataset contributing to the IPDMA project for prediction model research.

Domain and signaling questions	Guidance
Domain 1: Participant selection	
1.1 Were appropriate data sources used, e.g., cohort or randomized controlled trial for prognostic prediction model research, or cross-sectional study for diagnostic prediction model research?	Judge at the study/dataset level, in the same way as for aggregate data meta-analyses
1.2 Were all inclusions and exclusions of participants appropriate?	May be possible to judge at the participant level. In addition, if the IPD allows for the reinstation of participants excluded for inappropriate reasons, this concern can be avoided. Note, however, that IPD cannot overcome the inappropriate exclusion of participants from a study's original sampling frame
Applicability: Concern that the included participants and setting do not match the review question	May be possible to judge at the participant level. In addition, if additional inclusion/exclusion criteria can be applied to each dataset, this concern can be avoided
Domain 2: Predictors	
2.1 Were predictors defined and assessed in a similar way for all participants?	If the availability of IPD allows the IPDMA team to redefine variables objectively, concerns about definitions can be avoided
2.2 Were predictor assessments made without knowledge of outcome data?	May be possible to judge at the participant level
2.3 Are all predictors available at the time the model is intended to be used?	May be possible to judge at the participant level
Applicability: Concern that the definition, assessment, or timing of predictors in the model do not match the review question	May be possible to judge at the participant level. In addition, if additional inclusion/exclusion criteria can be applied to each dataset, this concern can be avoided
Domain 3: Outcome	
3.1 Was the outcome determined appropriately?	May be possible to judge at the participant level. In addition, if the availability of IPD allows the IPDMA team to redefine variables objectively, this concern can be avoided. For example, component participant-level data can be used to construct a new composite outcome
3.2 Was a prespecified or standard outcome definition used?	If the availability of IPD allows the IPDMA team to redefine variables, this concern can be avoided
3.3 Were predictors excluded from the outcome definition?	If the availability of IPD allows the IPDMA team to redefine variables, this concern can be avoided
3.4 Was the outcome defined and determined in a similar way for all participants?	If the availability of IPD allows the IPDMA team to redefine variables objectively, concerns about definitions can be avoided
3.5 Was the outcome determined without knowledge of predictor information?	May be possible to judge at the participant level
3.6 Was the time interval between predictor assessment and outcome determination appropriate?	May be possible to judge at the participant level
Applicability: Concern that the outcome, its definition, timing, or determination do not match the review question	May be possible to judge at the participant level. In addition, if additional inclusion/exclusion criteria can be applied to each dataset, this concern can be avoided
Domain 4: Analysis	
4.1. Were there a reasonable number of participants with the outcome?	With the exception of 4.1, this domain can be omitted, given that the IPDMA team has the freedom to alter and improve the analyses conducted, including the application of multiple imputation methods to address missing data
4.2. Were continuous and categorical predictors handled appropriately?	
4.3. Were all enrolled participants included in the analysis?	
4.4. Were participants with missing data handled appropriately?	
4.5. Was selection of predictors based on univariable analysis avoided?	
4.6. Were complexities in the data (e.g., censoring, competing risks, and sampling of controls) accounted for appropriately?	
4.7. Were relevant model performance measures evaluated appropriately?	
4.8. Were model overfitting and optimism in model performance accounted for?	
4.9. Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?	

Source: The first column of Box 4 presents the PROBAST domains and signaling items, freely available at http://www.probast.org/. The PROBAST domains were originally published by Wolff et al. and Moons et al. [7,8],

5. Discussion

The number of IPDMA studies is increasing, yet they have unique challenges and issues compared to traditional systematic reviews and meta-analyses based on aggregate data [2]. An important aspect is to examine the RoB and applicability of included studies and the IPD provided, but the results of our review reveal that improvements are needed in this regard for test accuracy and prediction model IPDMAs. This echoes a similar finding for IPDMAs of randomized trials [5].

A key finding is that RoB was rarely assessed in IPDMAs of prediction model research, unlike for IPDMAs of test accuracy where QUADAS-2 was mainly used. However, even when RoB was considered, better reporting of RoB assessments is needed, including the specific questions/items/domains assessed and the tool used to do so. Going forward, we recommended IPDMA researchers use and adapt tools such as QUADAS-2 (or QUADAS-C when relevant) and PROBAST, as outlined in the previous section, for test accuracy and prediction models, respectively. The RoB assessments can be summarized graphically or in a table, and the findings incorporated into the results, discussion, and conclusions of the IPDMA. In their IPDMA projects, Levis et al. [26] provide supplementary tables summarizing the judgments for each study, while Haase et al. [17] provide summary tables and figures across all studies in addition to the results of individual studies.

In the IPDMA context, it is important to assess the methodological quality of the provided IPD (and not just the quality of the available reports). Provision of data does not guarantee quality, and in fact the IPD may elucidate concerns about the quality of the data that were not apparent based on the published reports alone [2]. Conversely, a big advantage in IPDMA projects is that the IPD can be used to reduce RoB and improve applicability. For instance, availability of IPD may allow reinclusion of participants previously excluded from a study's original analysis; it may allow a subset of participants to be identified that (compared to the full dataset) more closely match the target population; and it may allow the analysis team to apply more appropriate analytical methods. However, IPD cannot overcome any inappropriate exclusion of participants from a study's original sampling frame, nor can it overcome the use of imperfect reference standards.

A major (but not well-recognized) advantage of IPDMA projects is being able to refine and update RoB classifications after receiving the IPD itself and through discussion with IPD providers. This was not emphasized by the large majority of IPDMA projects we reviewed. IPD may lead to a different RoB and applicability classification than initially considered when using the reported information from that study. For instance, regarding flow and timing, in the IPDMA of Levis et al., relevant primary studies often used a wide range of time intervals between the index test (Patient Health Questionnaire-9 assessment) and the reference standard (diagnostic interview), with potential for bias; however, when using the IPD from such studies the subset of participants with appropriate time intervals could be selected, thereby alleviating these concerns.

There are some limitations to consider from our review. First, our review was not a 'systematic' review as we only sought to obtain a representative sample of relevant articles. Only PubMed was searched, and only one investigator (B.L.) assessed articles for eligibility, with the support of a second investigator (R.D.R.). The search string was not developed by a librarian or peer-reviewed using PRESS, but it was approved by all authors and achieved its goal of identifying a relevant sample of IPDMAs. Our protocol was not registered, but it was finalized prior to commencing the review. In addition, our search was conducted in 2020; the write-up was delayed due to COVID-19-related issues, but we do not expect a fundamental shift in the last few years of how RoB is assessed in IPDMAs. QUADAS-C was published after our search [11] and citation checking identified one IPDMA by one of its authors [63]. PROBAST was published in 2019 [7,8], only 1 year before the end of our search. Thus, it is possible that more recently published IPDMAs of prediction model studies have incorporated PROBAST. Nonetheless, the original PROBAST publications did not provide specific guidance for using PROBAST in IPDMA projects, as we have now done here. Finally, the recommendations provided here are based on consensus among a small set of experts in IPDMAs of test accuracy and prediction model research. Additional research may be needed to extend the initial guidance offered here.

In summary, RoB and applicability assessments need to be improved in test accuracy and, in particular, prediction model IPDMA projects. The use of QUADAS-2 (and QUADAS-C if applicable) and PROBAST, both before and after IPD are obtained, can address this. Alongside our recommendations in Boxes 2–4, further development and dissemination of tailored tools will also help improve assessments of RoB and applicability in IPDMAs.

CRediT authorship contribution statement

Brooke Levis: Conceptualization, Methodology, Investigation, Validation, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Kym I.E. Snell:** Investigation, Validation, Writing – review & editing. **Johanna A.A. Damen:** Investigation, Validation, Writing – review & editing. **Miriam Hattle:** Investigation, Validation, Writing – review & editing. **Joie Ensor:** Investigation, Validation, Writing – review & editing. **Paula Dhiman:** Investigation, Validation, Writing – review & editing. **Constanza L. Andaur Navarro:** Investigation, Validation, Writing – review & editing. **Yemisi Takwoingi:** Conceptualization, Methodology,

Investigation, Validation, Writing − review & editing. **Penny F. Whiting:** Conceptualization, Methodology, Writing − review & editing. **Thomas P.A. Debray:** Conceptualization, Methodology, Investigation, Validation, Writing − review & editing. **Johannes B. Reitsma:** Conceptualization, Methodology, Investigation, Validation, Writing − review & editing. **Karel G.M. Moons:** Conceptualization, Methodology, Investigation, Validation, Writing − review & editing. **Gary S. Collins:** Conceptualization, Methodology, Investigation, Validation, Writing − review & editing. **Richard D. Riley:** Conceptualization, Methodology, Investigation, Validation, Formal analysis, Writing − original draft, Writing − review & editing, Supervision.

Data availability

Data will be made available on request.

Declaration of competing interest

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure and declare no competing interests, except Prof Riley receives royalties from sales of his book 'Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research'.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2023.10.022.

References

[1] Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010;340:c221.
[2] Riley RD, Stewart LA, Tierney JF. Individual participant data meta-analysis: a handbook for healthcare research. Chichester: Wiley; 2021.
[3] Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, et al. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD statement. JAMA 2015;313:1657−65.
[4] Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898.
[5] Wang H, Chen Y, Lin Y, Abesig J, Wu IX, Tam W. The methodological quality of individual participant data meta-analysis on intervention effects: systematic review. BMJ 2021;373:n736.
[6] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529−36.
[7] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51−8.
[8] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1−33.
[9] Reitsma JB, Rutjes A, Whiting P, Yang B, Leeflang MM, Bossuyt PM, Deeks JJ. Chapter 8: assessing risk of bias and applicability. In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, editors. Cochrane handbook for systematic reviews of diagnostic test accuracy. Version 2.0 (updated July 2023). Cochrane; 2023. Available at https://training.cochrane.org/handbook-diagnostic-test-accuracy/current. Accessed December 13, 2023.
[10] Prognosis Methods Group. Tools. Available at https://methods.cochrane.org/prognosis/tools. Accessed March 3, 2023.
[11] Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, et al. QUADAS-C: a tool for assessing risk of bias in comparative diagnostic accuracy studies. Ann Intern Med 2021;174:1592−9.
[12] Adderley N, Humphreys CJ, Barnes H, Ley B, Premji ZA, Johannson KA. Bronchoalveolar lavage fluid lymphocytosis in chronic hypersensitivity pneumonitis: a systematic review and meta-analysis. Eur Respir J 2020;56(2):2000206.
[13] Al-Rubaie ZTA, Askie LM, Hudson HM, Ray JG, Jenkins G, Lord SJ. Assessment of NICE and USPSTF guidelines for identifying women at high risk of pre-eclampsia for tailoring aspirin prophylaxis in pregnancy: an individual participant data meta-analysis. Eur J Obstet Gynecol Reprod Biol 2018;229:159−66.
[14] Bima P, Pivetta E, Nazerian P, Toyofuku M, Gorla R, Bossone E, et al. Systematic review of aortic dissection detection risk score plus D-dimer for diagnostic rule-out of suspected acute aortic syndromes. Acad Emerg Med 2020;27(10):1013−27.
[15] Broger T, Nicol MP, Székely R, Bjerrum S, Sossen B, Schutz C, et al. Diagnostic accuracy of a novel tuberculosis point-of-care urine lipoarabinomannan assay for people living with HIV: a meta-analysis of individual in- and outpatient data. PLoS Med 2020;17(5):e1003113.
[16] Gupta RK, Turner CT, Venturini C, Esmail H, Rangaka MX, Copas A, et al. Concise whole blood transcriptional signatures for incipient tuberculosis: a systematic review and patient-level pooled meta-analysis. Lancet Respir Med 2020;8(4):395−406.
[17] Haase R, Schlattmann P, Gueret P, Andreini D, Pontone G, Alkadhi H, et al. Diagnosis of obstructive coronary artery disease using computed tomography angiography in patients with stable chest pain depending on clinical probability and in clinically important subgroups: meta-analysis of individual patient data. BMJ 2019;365:l1945.
[18] Herrmann E, de Lédinghen V, Cassinotto C, Chu WCW, Leung VYF, Ferraioli G, et al. Assessment of biopsy-proven liver fibrosis by two-dimensional shear wave elastography: an individual patient data-based meta-analysis. Hepatology 2018;67(1):260−72.
[19] Hsu C, Caussy C, Imajo K, Chen J, Singh S, Kaulback K, et al. Magnetic resonance vs transient elastography analysis of patients with nonalcoholic fatty liver disease: a systematic review and pooled analysis of individual participants. Clin Gastroenterol Hepatol 2019;17(4):630−637.e8.
[20] Kalafat E, Laoreti A, Khalil A, Da Silva Costa F, Thilaganathan B. Ophthalmic artery Doppler for prediction of pre-eclampsia: systematic review and meta-analysis. Ultrasound Obstet Gynecol 2018;51(6):731−7.
[21] Karlas T, Petroff D, Sasso M, Fan JG, Mi YQ, de Lédinghen V, et al. Impact of controlled attenuation parameter on detecting fibrosis using liver stiffness measurement. Aliment Pharmacol Ther 2018;47(7):989−1000.
[22] Kazakos CT, Karageorgiou V. Retinal changes in schizophrenia: a systematic review and meta-analysis based on individual participant data. Schizophr Bull 2020;46(1):27−42.
[23] Klein Nulent TJW, Noorlag R, Van Cann EM, Pameijer FA, Willems SM, Yesuratnam A, et al. Intraoral ultrasonography to

measure tumor thickness of oral cancer: a systematic review and meta-analysis. Oral Oncol 2018;77:29−36.
[24] Kubo T, Furuta T, Sakuda T, Ochi M, Adachi N. Conventional 99mTc-(hydroxy) methylene diphosphate remains useful to predict osteosarcoma response to neoadjuvant chemotherapy: individual patient data and aggregate data meta-analyses. Medicine (Baltimore) 2018;97(51):e13308.
[25] Lee CH, Kang DY, Han M, Hur SH, Rha SW, Her SH, et al. Differential cutoff points and clinical impact of stent parameters of various drug-eluting stents for predicting major adverse clinical events: an individual patient data pooled analysis of seven stent-specific registries and 17,068 patients. Int J Cardiol 2019;282:17−23.
[26] Levis B, Benedetti A, Thombs BD. DEPRESsion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ 2019;365:l1476.
[27] Ley B, Winasti Satyagraha A, Rahmat H, von Fricken ME, Douglas NM, Pfeffer DA, et al. Performance of the Access Bio/CareStart rapid diagnostic test for the detection of glucose-6-phosphate dehydrogenase deficiency: a systematic review and meta-analysis. PLoS Med 2019;16(12):e1002992.
[28] Nguyen-Khac E, Thiele M, Voican C, Nahon P, Moreno C, Boursier J, et al. Non-invasive diagnosis of liver fibrosis in patients with alcohol-related liver disease by transient elastography: an individual patient data meta-analysis. Lancet Gastroenterol Hepatol 2018;3(9):614−25.
[29] Parpia S, Takach Lapner S, Schutgens R, Elf J, Geersing GJ, Kearon C. Clinical pre-test probability adjusted versus age-adjusted D-dimer interpretation strategy for DVT diagnosis: a diagnostic individual patient data meta-analysis. J Thromb Haemost 2020;18:669−75.
[30] Pavlovic V, Yang L, Chan HL, Hou J, Janssen HL, Kao JH, et al. Peginterferon alfa-2a (40 kD) stopping rules in chronic hepatitis B: a systematic review and meta-analysis of individual participant data. Antivir Ther 2019;24:133−40.
[31] Raskovalova T, Deegan PB, Mistry PK, Pavlova E, Yang R, Zimran A, et al. Accuracy of chitotriosidase activity and CCL18 concentration in assessing type I Gaucher disease severity: a systematic review with meta-analysis of individual participant data. Haematologica 2021;106(2):437−45.
[32] Suh CH, Kim HS, Jung SC, Choi CG, Kim SJ. 2-Hydroxyglutarate MR spectroscopy for prediction of isocitrate dehydrogenase mutant glioma: a systemic review and meta-analysis using individual patient data. Neuro Oncol 2018;20(12):1573−83.
[33] Suh CH, Kim SJ, Jung SC, Choi CG, Kim HS. The "Central Vein Sign" on T2*-weighted images as a diagnostic tool in multiple sclerosis: a systematic review and meta-analysis using individual patient data. Sci Rep 2019;9(1):18188.
[34] Thiele M, Hugger MB, Kim Y, Rautou PE, Elkrief L, Jansen C, et al. 2D shear wave liver elastography by Aixplorer to detect portal hypertension in cirrhosis: an individual patient data meta-analysis. Liver Int 2020;40(6):1435−46.
[35] van Doorn S, Geersing GJ, Kievit RF, van Mourik Y, Bertens LC, van Riet EES, et al. Opportunistic screening for heart failure with natriuretic peptides in patients with atrial fibrillation: a meta-analysis of individual participant data of four screening studies. Heart 2018;104:1236−7.
[36] Westra J, Tu S, Campo G, Qiao S, Matsuo H, Qu X, et al. Diagnostic performance of quantitative flow ratio in prospectively enrolled patients: an individual patient-data meta-analysis. Catheter Cardiovasc Interv 2019;94(5):693−701.
[37] Whiting P, Birnie K, Sterne JAC, Jameson C, Skinner R, Phillips B, et al. Accuracy of cystatin C for the detection of abnormal renal function in children undergoing chemotherapy for malignancy: a systematic review using individual patient data. Support Care Cancer 2018;26(5):1635−44.
[38] Yoshida K, Desbiolles A, Feldman SF, Ahn SH, Alidjinou EK, Atsukawa M, et al. Hepatitis B core-related antigen to indicate high viral load: systematic review and meta-analysis of 10,397 individual participants. Clin Gastroenterol Hepatol 2021;19(1):46−60.e8.
[39] Al-Shahi Salman R, Frantzias J, Lee RJ, Lyden PD, Battey TWK, Ayres AM, et al. ICH Growth Individual Patient Data Meta-analysis Collaborators. Absolute risk and predictors of the growth of acute spontaneous intracerebral haemorrhage: a systematic review and meta-analysis of individual patient data. Lancet Neurol 2018;17(10):885−94.
[40] Antonopoulos AS, Odutayo A, Oikonomou EK, Trivella M, Petrou M, Collins GS, et al. SAFINOUS-CABG (Saphenous Vein Graft Failure—an Outcomes Study in Coronary Artery Bypass Grafting) group. Development of a risk score for early saphenous vein graft failure: an individual patient data meta-analysis. J Thorac Cardiovasc Surg 2020;160:116−127.e4.
[41] Cao X, Ganti AK, Stinchcombe T, Wong ML, Ho JC, Shen C, et al. Predicting risk of chemotherapy-induced severe neutropenia: a pooled analysis in individual patients data with advanced lung cancer. Lung Cancer 2020;141:14−20.
[42] Condoluci A, Terzi di Bergamo L, Langerbeins P, Hoechstetter MA, Herling CD, De Paoli L, et al. International prognostic score for asymptomatic early-stage chronic lymphocytic leukemia. Blood 2020;135:1859−69.
[43] Crawford F, Cezard G, Chappell FM, PODUS Group. The development and validation of a multivariable prognostic model to predict foot ulceration in diabetes using a systematic review and individual patient data meta-analyses. Diabet Med 2018;35:1480−93.
[44] Depmann M, Eijkemans MJC, Broer SL, Tehrani FR, Solaymani-Dodaran M, Azizi F, et al. Does AMH relate to timing of menopause? Results of an individual patient data meta- analysis. J Clin Endocrinol Metab 2018;103:3593−600.
[45] Ediebah DE, Quinten C, Coens C, Ringash J, Dancey J, Zikos E, et al. Canadian Cancer Trials Group and the European Organization for Research and Treatment of Cancer. Quality of life as a prognostic indicator of survival: a pooled analysis of individual patient data from Canadian Cancer Trials Group clinical trials. Cancer 2018;124(16):3409−16.
[46] Hopkins AM, Shahnam A, Zhang S, Karapetis CS, Rowland A, Sorich MJ. Prognostic model of survival outcomes in non-small cell lung cancer patients initiated on afatinib: pooled analysis of clinical trial data. Cancer Biol Med 2019;16(2):341−9.
[47] Hudda MT, Fewtrell MS, Haroun D, Lum S, Williams JE, Wells JCK, et al. Development and validation of a prediction model for fat mass in children and adolescents: meta-analysis using individual participant data. BMJ 2019;366:l4293.
[48] Jaja BNR, Saposnik G, Lingsma HF, Macdonald E, Thorpe KE, Mamdani M, et al. SAHIT collaboration. Development and validation of outcome prediction models for aneurysmal subarachnoid haemorrhage: the SAHIT multinational cohort study. BMJ 2018;360:j5745.
[49] Jonkman NH, Colpo M, Klenk J, Todd C, Hoekstra T, Del Panta V, et al. Development of a clinical prediction model for the onset of functional decline in people aged 65-75 years: pooled analysis of four European cohort studies. BMC Geriatr 2019;19:179.
[50] Kievit RF, Gohar A, Hoes AW, Bots ML, van Riet EE, van Mourik Y, et al. Queen of Hearts and RECONNECT consortium. Efficient selective screening for heart failure in elderly men and women from the community: a diagnostic individual participant data meta-analysis. Eur J Prev Cardiol 2018;25(4):437−46.
[51] Lee CMY, Colagiuri S, Woodward M, Gregg EW, Adams R, Azizi F, et al. Comparing different definitions of prediabetes with subsequent risk of diabetes: an individual participant data meta-analysis involving 76 513 individuals and 8208 cases of incident diabetes. BMJ Open Diabetes Res Care 2019;7(1):e000794.
[52] Malda A, Boonstra N, Barf H, de Jong S, Aleman A, Addington J, et al. Individualized prediction of transition to psychosis in 1,676 Individuals at clinical high risk: development and validation of a

multivariable prediction model based on individual patient data meta-analysis. Front Psychiatry 2019;10:345.
[53] Pennells L, Kaptoge S, Wood A, Sweeting M, Zhao X, White I, et al. Emerging Risk Factors Collaboration. Equalization of four cardiovascular risk algorithms after systematic recalibration: individual-participant meta-analysis of 86 prospective studies. Eur Heart J 2019;40:621−31.
[54] Phillips B, Morgan JE, Haeusler GM, Riley RD, PICNICC Collaborative. Individual participant data validation of the PICNICC prediction model for febrile neutropenia. Arch Dis Child 2020;105:439−45.
[55] Saczkowski RS, Brown DJA, Abu-Laban RB, Fradet G, Schulze CJ, Kuzak ND. Prediction and risk stratification of survival in accidental hypothermia requiring extracorporeal life support: an individual patient data meta-analysis. Resuscitation 2018;127:51−7.
[56] Shinohara K, Tanaka S, Imai H, Noma H, Maruo K, Cipriani A, et al. Development and validation of a prediction model for the probability of responding to placebo in antidepressant trials: a pooled analysis of individual patient data. Evid Based Ment Health 2019;22(1):10−6.
[57] Spronk I, Van Loey NEE, Sewalt C, Nieboer D, Renneberg B, Moi AL, et al. Quality of life study group. Recovery of health-related quality of life after burn injuries: an individual participant data meta-analysis. PLoS One 2020;15:e0226653.
[58] Verma R, Chiang J, Qian H, Amin R. Maximal static respiratory and sniff pressures in healthy children. a systematic review and meta-analysis. Ann Am Thorac Soc 2019;16(4):478−87.
[59] Vickers A, Vertosick EA, Sjoberg DD, Hamdy F, Neal D, Bjartell A, et al. Value of intact prostate specific antigen and Human Kallikrein 2 in the 4 Kallikrein predictive model: an individual patient data meta-analysis. J Urol 2018;199:1470−4.
[60] Vollgraff Heidweiller-Schreurs CA, van Osch IR, Heymans MW, Ganzevoort W, Schoonmade LJ, Bax CJ, et al. Cerebroplacental ratio in predicting adverse perinatal outcome: a meta-analysis of individual participant data. BJOG 2021;128:226−35.
[61] Albert PS. Imputation approaches for estimating diagnostic accuracy for multiple tests from partially verified designs. Biometrics 2007;63:947−57.
[62] Gad AM, Ali AAM, Mohamed RH. A multiple imputation approach to evaluate the accuracy of diagnostic tests in presence of missing values. Commun Math Biol Neurosci 2022;21.
[63] Boyd LNC, Ali M, Leeflang MMG, Treglia G, de Vries R, Le Large TYS, et al. Diagnostic accuracy and added value of blood-based protein biomarkers for pancreatic cancer: a meta-analysis of aggregate and individual participant data. EClinicalMedicine 2022;55:101747.