Avenues for Further Research 20

Yulun Liu and Yong Chen

Acronyms

GLMMs	Generalized linear mixed models
HSROC	Hierarchical summary receiver operating characteristic
IPD	Individual patient-level data
MRI	Magnetic resonance imaging
NPV	Negative predictive value
PPV	Positive predictive value
ROP	Retinopathy of prematurity
SROC	Summary receiver operating characteristic

20.1 Review of Existing Statistical Work on Diagnostic Meta-analysis

Systematic review of test performance is a rigorous approach for synthesizing evidence in the evaluation of diagnostic/screening tests performance. Previous chapters have been focusing on guiding the progress of diagnostic test assessments and discussing the major challenges during systematic reviews, such as small study

Y. Liu · Y. Chen (✉)
Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: yulunliu@pennmedicine.upenn.edu; ychen123@mail.med.upenn.edu, ychen123@upenn.edu, ychen123@pennmedicine.upenn.edu

G. Biondi-Zoccai (ed.), *Diagnostic Meta-Analysis*,
https://doi.org/10.1007/978-3-319-78966-8_20

effects, appraising inconsistency, and moderators. When the included studies meet the prespecified quality criteria, the results can be quantitatively summarized by a meta-analysis, providing the estimates for quantities of key interest while accounting for the possible heterogeneity.

To date, a variety of statistical methods for diagnostic meta-analysis have been developed in the presence and absence of a gold standard. Assume that the performance of a candidate test has been measured against a gold standard. The simplest method is to apply univariate fixed-effect or random-effects meta-analysis to estimate sensitivity and specificity separately, ignoring any correlations that may exist between the two measures. However, sensitivity and specificity are often negatively correlated across studies [1] due to the fact that different thresholds may have been used to define positive and negative test results. The current methods essentially can be classified into two categories. The first category includes the summary receiver operating characteristic (SROC) curve approach (or Moses-Littenberg model) [2, 3] and a hierarchical summary receiver operating characteristic (HSROC) model [2–5], which were based on modeling of accuracy and scale parameters while accounting for between-study heterogeneity. The second category includes models based on sensitivity and specificity, including the bivariate general mixed-effects models and bivariate generalized linear mixed models (GLMMs) [1, 5–9]. Interestingly, Harbord et al. [10] found that the bivariate GLMMs and HSROC models are closely related and even equivalent in the absence of covariates.

Despite that various statistical methods have been developed and available as guidance for investigators, it is time to consider future directions of diagnostic tests in meta-analysis. In fact, there remain many interesting and important topics in diagnostic meta-analysis that need to be investigated.

20.2 Advanced Methods of Diagnostic Meta-analysis

This subsection is an incomplete collection of topics that we believe are important for future research on meta-analysis of diagnostic test accuracy studies. These include (a) the robustness of model misspecifications and (b) the identifiability of models and the assumption of conditional independence for multiple diagnostic tests in the absence of a gold standard.

20.2.1 Model Robustness

Although the bivariate GLMMs and HSROC models take into consideration the correlation between sensitivity and specificity across studies, the standard likelihood-based inference sometimes suffers from computational issues, such as non-convergence or sensitivity to the choice of initial values due to the complexity of likelihood and the small number of studies; see Chen et al. [11]. To circumvent these difficulties, composite likelihood [12]-based inference of meta-analysis of diagnostic tests has been developed [13]. Such a procedure not only avoids the computational issues but also offers robustness to misspecification of joint distributions of sensitivity and specificity. In practice, many of diagnostic test accuracy studies involve not only case-control studies but also cohort studies. The bivariate GLMMs and HSROC models focus only on sensitivity and specificity and ignore the information on disease prevalence that is contained in cohort studies. As a consequence, such methods cannot provide estimates of measures related to disease prevalence, including positive and negative predictive values (PPV and NPV), which reflect the clinical utility of a diagnostic test. Additionally, due to possible clinical variability or artifactual variation, sensitivity and specificity may vary with disease prevalence [14, 15]. Chu et al. [16] proposed a trivariate model to jointly analyze sensitivity, specificity, and disease prevalence. Chen et al. [11] proposed a general framework of jointly analyzing case-control and cohort studies while producing robust inference on positive and negative predictive values. They also applied their method to the surveillance of melanoma patients where the goal was to detect the recurrence of melanoma in regional lymph nodes and/or distant sites at a point when it remains treatable. This method not only provided robust estimates of diagnostic accuracy for the four modern diagnostic imaging modalities but also produced patient-specific estimates of positive/negative predictive value of the recurrence of melanoma under various clinical settings, which directly supports clinical decision-making [11]. Ma et al. [17] developed Bayesian inference of this model. Although the composite likelihood-based inference can address the computational issues in standard likelihood-based inference and is robust to the misspecifications of correlations among sensitivity, specificity, and disease prevalence, more robust models are still warranted. For example, van Houwelingen et al. [6, 7] have relaxed the normality assumption of random effects to mixture distributions. Chen et al. [18] have developed beta-binomial distributions as an alternative to allow heavy-tailed distributions. More work along this line toward robust inference is needed.

20.2.2 Absence of Gold Standard Test: Identifiability and Conditional Dependence

In diagnostic meta-analysis, a common problem occurs when the selected reference test may not be a gold standard due to measurement error, high cost, or nonexistence [19]. Failure to account for the errors in reference test can lead to substantial bias in the evaluation of candidate test accuracy [20]. Several statistical methods have been proposed for dealing with such a situation in the literature. Among them, two models have been developed to account for an imperfect reference test, namely, a multivariate generalized linear mixed model [21] and a hierarchical summary receiver operating characteristic model [22]. In practice, investigators may have to choose between one of these two models. In order to provide a useful guideline for modeling with diagnostic meta-analysis, Liu et al. [23] provided a unification of these models and showed that these two models, although with very different formulations, are closely related and are mathematically equivalent in the absence of study-level covariates. Moreover, they have provided the exact relations between the parameters of these two models and assumptions under which two models can be reduced to equivalent sub-models. In other settings, studies may rely on two or more imperfect reference tests to verify the results of a candidate test, or studies may have multiple candidate tests with an imperfect reference. In the former case, the composite reference standard was developed by Alonzo and Pepe [24]; this method combines information from several imperfect reference tests to obtain a "pseudo-gold standard." Such a method is appealing because it provides a simple fixed rule to assign a final diagnosis to each subject in a study population, reducing the effect of misclassification of disease status [25]. For the latter case, the latent class models have been developed for estimating diagnostic accuracy [26, 27], among others. Nevertheless, some possible limitations of latent class approach have been discussed in the literature [28, 29].

It is worth noting that two important issues need to be carefully considered during the evaluating the accuracy of multiple candidate tests in the absence of a gold standard, namely, model identifiability and dependence of diagnostic tests. First, when two or more candidate tests in the absence of a gold standard are simultaneously applied to each subject of a population, the lack of identifiability may occur. For example, if two imperfect diagnostic tests are considered and the data is summarized as a 2 × 2 table with at most three degrees of freedom; yet, in fact, there are five unknown parameters (one disease prevalence, two sensitivities, and two specificities) in the probability distribution that characterizes these data. To overcome such non-identifiability, the Bayesian approach was conducted through the knowledge of unknown test characteristics as prior information [19]. Gustafson et al. [30] proposed to use nested models, i.e., model expansion and model contraction, to alleviate the identifiable issue, and concluded that non-identifiable models with moderate amount of prior information often outperform simpler but identifiable models. The second issue is the assumption of conditional independence. Some models and inferences for multiple tests rely critically on the assumption that the tests are independent conditional on disease status; see Hui and Walter [31], Pepe and Janes [32], and Chu et al. [21]. However, it is not always satisfied in practice. Dendukuri and Joseph [33] considered the conditional dependence between two tests by allowing pairwise correlation between two tests and random-effects model for correlation between more than two tests. In summary, the issue of model identifiability and conditional independence remains challenging, and further work in this direction is in great need.

20.3 Future Work and Direction

Traditional meta-analyses provide the results based on aggregated data (or study-level data) from published studies. Over the past few decades, although statistical methods relying on aggregated data have been well-studied, these procedures may be highly susceptible to ecological fallacy bias in the literature [34–37]. In contrast, individual patient-level data (IPD) meta-analysis, which synthesizes the evidence from patient-level data, is regarded as a gold standard. IPD meta-analysis offers several advantages compared with the traditional meta-analysis, including bias reduction, the ability to undertake updated analyses (e.g., follow-up data), and subgroup analyses [38]. More specifically, since IPD meta-analysis allows the results that are derived directly from each study, it has potential to substantially reduce the effects of publication and reporting biases [38]. Moreover, IPD meta-analysis collects more detailed information on individual-level characteristics/covariates; it therefore can increase statistical power to carry out subgroup analyses through meta-regression [34]. In particular, when the heterogeneity is present, the interpretation of overall summary results (e.g., study-level covariates) can be misleading, whereas IPD meta-analysis allows investigation on individual characteristic as potential sources of heterogeneity between studies [39]. Despite these benefits, however, IPD may not be always available from all relevant studies due to high cost or logistic reasons [38]. Additionally, in some situations, those studies with availability of IPD may represent a biased subset of the available studies [38, 40, 41].

Recently, incorporating IPD, if available, into aggregated data has received increasing attention, which offers opportunities to inform personalized medical decisions based on patient-level characteristics and produces results tailored to the individual patients or clinically relevant subgroups [42, 43]. In the following two subsections, we will discuss the future work efforts needed to address a set of statistical challenges in combining both IPD and aggregated data, development of diagnostic prediction research, and assessment of prediction models for further aiding of clinical decision-making. In addition, we will also discuss the opportunities and potential challenges when IPD is used alone.

20.3.1 Combination of Aggregated Data and Individual Patient-Level Data

IPD may be unavailable for all studies; the circumstance arises when IPD are accessible for a subset of studies and aggregated data alone are available for the remaining studies. To utilize all available data, several methods have been proposed to combine both IPD and aggregated data using treatment interventions or diagnostic studies [43–45]. Among them, only few published work focuses on how to synthesize both data from diagnostic tests, as well as to evaluate accuracy-by-covariate interactions; for example, see Riley et al. [45], where they have extended the standard bivariate random-effects meta-analysis.

When there is more than one diagnostic test simultaneously used to evaluate their accuracy, it is essential for patients and clinicians to select the most effective diagnostic test. In such case, the network meta-analysis, which is an extension of traditional pairwise meta-analysis, has been applied to compare multiple interventions for a combination of IPD and aggregated data. To our best knowledge, very few statistical methods on the synthesis of IPD and aggregated data for multiple diagnostic accuracy studies have been developed. Further research is needed on this topic. Additionally, for either pairwise meta-analyses or network meta-analyses, it is important to consider the case when there is no gold standard.

In clinical practice, patients and care providers often face decisional dilemmas when multiple diagnostic tests are available, and therefore, prediction models are essential tools in aiding decision-making. The diagnostic prediction model is useful to convert combinations of multiple predictors, such as individual characteristics (e.g., age and smoking status), test results, and biomarkers, with preassigned weights to an estimated absolute risk or probability of disease [46, 47]. By modeling these predictors, a commonly used statistical method is through the multivariable regression framework, such as logistic or Cox regression [48]. In fact, many prediction models are constructed from a single dataset. However, with the availability of IPD, the prediction models based on IPD has become increasingly appealing for improving the development and validation of prediction models [49]. For example, several authors [50–52] incorporated previously published univariable predictor-outcome association to construct a novel prediction model through univariate meta-analysis. When the multivariable associations are available from the literature, it will be difficult to incorporate them due to inclusion of different predictors, model overfitting, and other practical factors. These potential challenges have been discussed in Debray et al. [53]. Before implementing a diagnostic prediction model in clinical practice, model validation is also required, particularly for two major factors—discrimination and calibration [54, 55]. Debray et al. [56] focused on investigating the generalizability of prediction model through the internal-external cross validation to combine model development with validation. A principle on IPD meta-analysis for prediction modeling can be found in Debray et al. [57]. Riley et al. [48] highlighted the importance of external validation of prediction modeling (e.g., discrimination and calibration) on IPD meta-analysis. Nevertheless, several important issues remain open, including novel methods of model development and validation, particularly for the case in the absence of a gold standard, combination of tests, missing predictors, and between-studies heterogeneity in predictor effects.

20.3.2 Partial Verification Bias/No Gold Standard for Individual Patient-Level Data

Despite IPD method offers many opportunities, it still poses many methodological challenges, such as partial verification bias and no gold standard. Next we give two case studies to illustrate the potential challenges using IPD alone.

Case study 1: An example on the issue of verification bias is the study of endometrial carcinoma reported by Rockall et al. [58]. The histology test is considered as a gold standard, but an invasive method, for the diagnosis of the myometrial and cervical invasion in endometrial carcinoma. As an alternative, the magnetic resonance imaging (MRI) with gadolinium enhancement has been used as a surrogate; it is a noninvasive, highly accurate, and less expensive diagnostic test for detecting lymph node metastases [59, 60]. This study includes 96 patients with endometrial carcinoma who had a MRI test performed between May 1995 and November 2004. Out of 96 patients, 68 had a negative MRI test and 28 had positive MRI. For those patients with positive results, 18% of them have been evaluated by the gold standard test of the endometrial carcinoma. For those patients with negative results, 66% of them have been evaluated by the gold standard test following the MRI testing. This design, only partially verifies the subjects with gold standard, is more cost-effective compared to the standard design where all subjects are evaluated by both tests.

Case study 2: An example on the imperfect reference test is the study of retinopathy of prematurity (ROP), which is an eye disease that occurs in premature infants. It is a leading cause of avoidable blindness in children worldwide [61]. When infants with ROP are diagnosed in early stage, they can often be effectively treated with laser retinal ablative surgery or other treatments [62, 63]. In this ROP study, the enrolled infants have undergone a sequential screening examinations on their paired eyes by study-certified ophthalmologists (hereafter referred as the ophthalmology test), which is often treated as a gold standard. Such screening process essentially tends to be time-intensive for the ophthalmologists, stressful for the infants, and related to medicolegal liability concerns [64–66]. The telemedicine-based digital retinal imaging test (hereafter referred as the imaging test) has been widely used in practice. In this ROP study, the preliminary findings suggest that the prevalence rates of ROP significantly differ among subpopulations; specifically, the prevalence rates of female and male groups are 21% and 31%, respectively. The sensitivity and specificity of both diagnostic tests (i.e., the ophthalmology test and the imaging tests) are approximately the same across subpopulations.

In case study 1, since the subjects were evaluated by the gold standard selectively, i.e., subjects with positive results from the candidate test were less likely to be evaluated by the gold standard compared to the subject with negative result from the candidate test, ignoring such selective verification can lead to bias in the estimate of diagnostic accuracy. Such a problem has been recognized by researchers [67, 68], and this type of bias is known as the partial verification bias. Statistical methods have been proposed to correct for the potential partial verification bias when using IPD data alone [68–72]. For multiple studies, Ma et al. [17] recently proposed a hybrid GLMM to correct bias in diagnostic meta-analyses. However, little work has been done in the setting of correlated data or longitudinal studies.

In case study 2, the evaluation from study-certified ophthalmologists is also error-prone. In fact, previous studies have suggested that the agreement between two independent ophthalmologists is poor, suggesting that the reference test is not a gold standard. This problem is related to the Hui-Walter framework [31]. Specifically, Hui and Walter proposed the model to estimate the accuracy of diagnostic tests when the accuracy of the gold standard is unknown [31]. In particular, their proposed approach requires that (1) two diagnostic tests are both applied to two populations with different disease prevalence rates and (2) the results of one diagnostic test are assumed to be independent of the other ones within the disease subpopulation and the disease-free subpopulation. Additionally, the accuracy of both diagnostic tests is assumed to be consistent among two different subpopulations. Compared to the Hui-Walter framework, the key difference is that the ROP study involves the correlated and clustered data. Such correlated or clustered data are common collected in medical research. Further work is required to deal with such problem.

In conclusion, significant efforts are underway to enhance statistical methods for diagnostic test accuracy studies. This chapter aims to provide an overview of the recent statistical advances on meta-analysis of diagnostic tests and suggest a few directions for future research. We believe that more advances in this important topic will have direct impacts to better clinical decision-making and more effective screening of diseases.

References

1. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58:982–90.
2. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Mak. 1993;13:313–21.
3. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med. 1993;12:1293–316.
4. Walter S. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med. 2002;21:1237–56.
5. Arends L, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Med Decis Mak. 2008;28:621–38.
6. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med. 2002;21:589–624.
7. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. Stat Med. 1993;12:2273–84.
8. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol. 2006;59:1331–2.
9. Hamza TH, van Houwelingen HC, Stijnen T. The binomial distribution of meta-analysis was preferred to model within-study variability. J Clin Epidemiol. 2008;61:41–51.
10. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8:239–51.
11. Chen Y, Liu Y, Ning J, Cormier J, Chu H. A hybrid model for combining case–control and cohort studies in systematic reviews of diagnostic tests. J R Stat Soc Ser C Appl Stat. 2015;64:469–89.
12. Lindsay BG. Composite likelihood methods. Contemp Math. 1988;80:221–39.
13. Chen Y, Liu Y, Ning J, Nie L, Zhu H, Chu H. A composite likelihood method for bivariate meta-analysis in diagnostic systematic reviews. Stat Methods Med Res. 2017;26:914–30.
14. Feinstein A. Misguided efforts and future challenges for research on "diagnostic tests". J Epidemiol Community Health. 2002;56:330–2.
15. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. Can Med Assoc J. 2013;185:E537–44.
16. Chu H, Nie L, Cole SR, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. Stat Med. 2009;28:2384–99.
17. Ma X, Chen Y, Cole SR, Chu H. A hybrid Bayesian hierarchical model combining cohort and case–control studies for meta-analysis of diagnostic tests: accounting for partial verification bias. Stat Methods Med Res. 2016;25:3015–37.

18. Chen Y, Liu Y, Chu H, Ting Lee ML, Schmid CH. A simple and robust method for multivariate meta-analysis of diagnostic test accuracy. Stat Med. 2017;36:105–21.
19. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol. 1995;141:263–72.
20. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. Can Med Assoc J. 2006;174:469–76.
21. Chu H, Chen S, Louis TA. Random effects models in a meta-analysis of the accuracy of two diagnostic tests without a gold standard. J Am Stat Assoc. 2009;104:512–23.
22. Dendukuri N, Schiller I, Joseph L, Pai M. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. Biometrics. 2012;68:1285–93.
23. Liu Y, Chen Y, Chu H. A unification of models for meta-analysis of diagnostic accuracy studies without a gold standard. Biometrics. 2015;71:538–47.
24. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Stat Med. 1999;18:2987–3003.
25. Naaktgeboren CA, Bertens LC, van Smeden M, de Groot JA, Moons KG, Reitsma JB. Value of composite reference standards in diagnostic research. BMJ. 2013;347:f5605.
26. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. Biometrics. 1996;52:797–810.
27. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. Stat Methods Med Res. 1998;7:354–70.
28. Pepe MS, Alonzo TA. Comparing disease screening tests when true disease status is ascertained only for screen positives. Biostatistics. 2001;2:249–60.
29. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics. 2004;60:427–35.
30. Gustafson P, et al. On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables [with comments and rejoinder]. Stat Sci. 2005;20:111–40.
31. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics. 1980;36:167–71.
32. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. Biostatistics. 2006;8:474–84.
33. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics. 2001;57:158–67.
34. Lambert PC, et al. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. J Clin Epidemiol. 2002;55:86–94.
35. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI, Anti-Lymphocyte Antibody Induction Therapy Study Group. Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Stat Med. 2002;21:371–87.
36. Thompson SG, Higgins J. How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002;21:1559–73.
37. Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol. 2004;57:683–97.
38. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010;340:c221.
39. Smith CT, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. Stat Med. 2005;24:1307–19.
40. Steinberg K, Smith SJ, Stroup DF, Olkin I, Lee NC, Williamson GD, Thacker SB. Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies. Am J Epidemiol. 1997;145:917–25.
41. Higgins JP, Green S. Cochrane handbook for systematic reviews of interventions, vol. 4. Chichester: John Wiley & Sons; 2011.
42. Thompson SG, Higgins JP. Can meta-analysis help target interventions at individuals most likely to benefit? Lancet. 2005;365:341–6.

43. Riley RD, Steyerberg EW. Meta-analysis of a binary outcome using individual participant data and aggregate data. Res Synth Methods. 2010;1:2–19.
44. Sutton AJ, Kendrick D, Coupland CA. Meta-analysis of individual-and aggregate-level data. Stat Med. 2008;27:651–69.
45. Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. Stat Med. 2008;27:6111–36.
46. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, Maas AI. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med. 2008;5:e165.
47. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med. 2015;13:1.
48. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140.
49. Ahmed I, Debray TP, Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol. 2014;14:3.
50. Steyerberg EW, Eijkemans MJ, Van Houwelingen JC, Lee KL, Habbema JD. Prognostic models based on literature and individual patient data in logistic regression analysis. Stat Med. 2000;19:141–60.
51. Debray TP, Koffijberg H, Lu D, Vergouwe Y, Steyerberg EW, Moons KG. Incorporating published univariable associations in diagnostic and prognostic modeling. BMC Med Res Methodol. 2012;12:121.
52. Greenland S. Quantitative methods in the review of epidemiologic literature. Epidemiol Rev. 1987;9:1–30.
53. Debray T, Koffijberg H, Vergouwe Y, Moons KG, Steyerberg EW. Aggregating published prediction models with individual participant data: a comparison of different approaches. Stat Med. 2012;31:2697–712.
54. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115:928–35.
55. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010;21:128–38.
56. Debray T, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32:3158–80.
57. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG, Cochrane IPD Meta-analysis Methods Group. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Med. 2015;12:e1001886.
58. Rockall A, Meroni R, Sohaib SA, Reynolds K, Alexander-Sefre F, Shepherd JH, Jacobs I, Reznek RH. Evaluation of endometrial carcinoma on magnetic resonance imaging. Int J Gynecol Cancer. 2007;17:188–96.
59. Saez F, Urresola A, Larena JA, Martín JI, Pijuán JI, Schneider J, Ibáñez E. Endometrial carcinoma: assessment of myometrial invasion with plain and gadolinium-enhanced MR imaging. J Magn Reson Imaging. 2000;12:460–6.
60. Nakao Y, Yokoyama M, Hara K, Koyamatsu Y, Yasunaga M, Araki Y, Watanabe Y, Iwasaka T. MR imaging in endometrial carcinoma as a diagnostic tool for the absence of myometrial invasion. Gynecol Oncol. 2006;102:343–7.
61. Gilbert C. Retinopathy of prematurity: a global perspective of the epidemics, population of babies at risk and implications for control. Early Hum Dev. 2008;84:77–82.
62. Schaffer DB, Palmer EA, Plotsky DF, Metz HS, Flynn JT, Tung B, Hardy RJ. Prognostic factors in the natural course of retinopathy of prematurity. The Cryotherapy for Retinopathy of Prematurity Cooperative Group. Ophthalmology. 1993;100:230–7.

63. Good WV, Hardy RJ, E.M.S. Group. The multicenter study of early treatment for retinopathy of prematurity (ETROP). New York: Elsevier; 2001.
64. Yen KG, Hess D, Burke B, Johnson RA, Feuer WJ, Flynn JT. The optimum time to employ telephotoscreening to detect retinopathy of prematurity. Trans Am Ophthalmol Soc. 2000;98:145.
65. Richter GM, Williams SL, Starren J, Flynn JT, Chiang MF. Telemedicine for retinopathy of prematurity diagnosis: evaluation and challenges. Surv Ophthalmol. 2009;54:671–85.
66. Ying G-S, Quinn GE, Wade KC, Repka MX, Baumritter A, Daniel E, e-ROP Cooperative Group. Predictors for the development of referral-warranted retinopathy of prematurity in the telemedicine approaches to evaluating acute-phase retinopathy of prematurity (e-ROP) study. JAMA Ophthalmol. 2015;133:304–11.
67. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926–30.
68. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983;39:207–15.
69. Zhou X-H. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. Commun Stat Theory Methods. 1993;22:3177–98.
70. Zhou X-H. Correcting for verification bias in studies of a diagnostic test's accuracy. Stat Methods Med Res. 1998;7:337–53.
71. Harel O, Zhou XH. Multiple imputation for correcting verification bias. Stat Med. 2006;25:3769–86.
72. De Groot J, Janssen KJ, Zwinderman AH, Moons KG, Reitsma JB. Multiple imputation to correct for partial verification bias revisited. Stat Med. 2008;27:5880–9.