Random-effects meta-analysis of the clinical utility of tests and prediction models

Wynants L.,[1] Riley R.D.,[2] Timmerman D.,[1,3] Van Calster B.[1]

1. KU Leuven, Department of Development and Regeneration, Leuven, Belgium
2. Keele University, Research Institute for Primary Care and Health Sciences, Staffordshire, ST5 5BG, U.K.
3. University Hospitals Leuven, Department of Obstetrics and Gynecology, Leuven, Belgium

Short title: Random-effects meta-analysis of clinical utility

Author responsible for proofs: Laure Wynants, laure.wynants@kuleuven.be, O&N IV Herestraat 49 - box 805, 3000 Leuven, Belgium, +32 16 32 76 70

Financial support
This study was supported by the Flemish government [Research Foundation–Flanders (FWO) projects G049312N and G0B4716N, Flanders' Agency for Innovation by Science and Technology (IWT) project IWT-TBM 070706-IOTA3] and Internal Funds KU Leuven (project C24/15/037). LW holds a post-doctoral research mandate from Interne Fondsen KU Leuven/Internal Funds KU Leuven. DT is a senior clinical investigator of FWO. The sponsors had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the work for publication. The researchers performed this work independently of the funding sources.

Acknowledgements
We thank the editors and reviewers from Statistics in Medicine for their constructive feedback and suggestions, which helped improve the article on revision.

Abstract

The use of data from multiple studies or centers for the validation of a clinical test or a multivariable prediction model allows researchers to investigate the test's/model's performance in multiple settings and populations. Recently, meta-analytic techniques have been proposed to summarize discrimination and calibration across study populations. Here, we rather consider performance in terms of Net Benefit, which is a measure of clinical utility that weighs the benefits of true positive classifications against the harms of false positives. We posit that it is important to examine clinical utility across multiple settings of interest. This requires a suitable meta-analysis method, and we propose a Bayesian trivariate random-effects meta-analysis of sensitivity, specificity, and prevalence. Across a range of chosen harm-to-benefit ratios, this provides a summary measure of Net Benefit, a prediction interval, and an estimate of the probability that the test/model is clinically useful in a new setting. In addition, the prediction interval and probability of usefulness can be calculated conditional on the known prevalence in a new setting. The proposed methods are illustrated by two case studies: one on the meta-analysis of published studies on ear thermometry to diagnose fever in children, and one on the validation of a multivariable clinical risk prediction model for the diagnosis of ovarian cancer in a multicenter dataset. Crucially, in both case studies the clinical utility of the test/model was heterogeneous across settings, limiting its usefulness in practice. This emphasizes that heterogeneity in clinical utility should be assessed before a test/model is routinely implemented.

Keywords: meta-analysis, net benefit, test accuracy, diagnostic, decision curves

1. Introduction

Clinical diagnoses are often predicted using tests and multivariable prediction models that combine various predictors.[1] Before diagnostic tests or models are introduced into clinical practice, it is of the utmost importance that their predictive performance is externally validated on new data. This preferably takes place in a setting that is independent from the test's or model's development setting, for example in new centers or by other research teams.[2] Typically, researchers assess the discriminative ability of the test/model to distinguish between patients who do and do not suffer from the suspected disease. These results are summarized using, among others, sensitivity, specificity, and the c-statistic. If the diagnostic tool yields a predicted risk that the patient suffers from the disease, measures of calibration can be used, which assess how well predicted risks correspond to observed event rates.[3]

A problem with measures of discrimination and calibration is that they do not assess the consequences of using a test/model in practice. In contrast, decision-analytic measures of clinical utility incorporate the harms of false negative and false positive classifications. A measure of clinical utility that has received broad support is the Net Benefit (NB).[4-8] Briefly, the NB quantifies the benefit of using a test/model for clinical decision making by correcting the number of true positive classifications for the number of false positive classifications using a weighting factor.[9,10] The weight reflects the ratio of the harm of a false positive and a false negative. Because the assumed harms can vary, the NB is usually calculated for a relevant range of harm ratios. A plot of the NB for various harm ratios is a decision curve.

All measures of clinical utility depend on disease prevalence,[11] which may vary across studies, centers, and regions. In addition, the predictive performance of a test/model may be heterogeneous, reflecting differences in patient case-mix or true variations in the association between predictors and the disease.[1,12,13] Recently, meta-analytic techniques have been proposed to investigate heterogeneity in predictive performance (discrimination and calibration) across populations.[14,15] The NB is influenced by an interplay of the prevalence of the outcome, discrimination, and calibration. Although heterogeneity in clinical utility may naturally be expected, methods for the meta-analysis of the clinical utility of a test/model have not been considered. Indeed, before routinely implementing a test/model in practice, it is surely essential to examine its clinical utility across multiple settings of interest. To address this, we consider a method for the meta-analysis of the NB, using a Bayesian trivariate random-effects meta-analysis of sensitivity, specificity, and prevalence.

In what follows, we introduce two motivating examples. The first concerns diagnosing fever in children, and investigates the clinical usefulness of using ear thermometry based on a systematic review of the literature. The second is a multicenter external validation of the clinical utility of the IOTA LR2 model, which is a multiple logistic regression model based on ultrasound characteristics to distinguish between benign and malignant adnexal masses. In Section 3 we introduce the NB measure of clinical utility, and explain how to calculate it from a trivariate random-effects meta-analysis of sensitivity, specificity, and prevalence. Bayesian prediction intervals may be used to predict the clinical utility in randomly selected new studies or centers. We propose to construct prediction intervals conditional on a known prevalence, if this

information is available in a new setting. We apply these ideas to our examples in Section 4. In section 5 we put the NB of the test/model into perspective by comparing it to other diagnostic strategies, such as the default strategies that classify all patients as positive or negative. We show how to graphically present the comparison using decision curves, and introduce the Bayesian probability that the test/model performs better than the default strategies in a randomly selected new setting.

2. Two motivating examples

2.1. Meta-analysis of the clinical utility of ear temperature for diagnosing fever in children

Rectal temperature measurement can be difficult in children, particularly when they are uncooperative or restless. Ear thermometry is a commonly used and attractive alternative, as the ear is easily accessible and the procedure is very quick. Craig and colleagues performed a systematic review of the accuracy of infrared ear thermometry for diagnosing fever in children.[16] They retrieved eleven studies evaluating the accuracy of ear thermometers of the 'FirstTemp' brand. All studies considered patients with an ear temperature ≥38.0 °C as test positive, and the gold standard diagnosis of fever was a rectal temperature ≥38.0 °C. The number of children with (n_1) and without (n_0) fever, and the number true positive (r_{11}) and true negative (r_{00}) classifications per study j are summarized in Table 1. The observed prevalence of fever as diagnosed by the gold standard varied between 27% and 75%, reflecting that the studies included children who were already in a hospital or an emergency unit.

A random-effects meta-analysis of the discriminative performance has already been performed, yielding a summary sensitivity of 65% and a summary specificity of 98%, and demonstrating considerable between-study heterogeneity, especially in sensitivity.[15] We may now ask the question whether using ear thermometry to diagnose fever is clinically useful to inform patient management and treatment decisions. On the one hand, one may want to avoid missing serious infectious diseases in young children, and stress the harm of a false negative classification. In this case, sensitivity is important, and the test accuracy does not appear satisfactory in most studies (Table 1). On the other hand, one may want to avoid overtreatment of fever and unnecessary hospitalization costs, when the child can be safely taken care of at home. In this case, specificity is important, and this was very good (Table 1), despite heterogeneity across settings. To make any statement regarding the likely clinical usefulness of ear thermometry in a new population, a random-effects meta-analysis of the NB is required. Such an analysis directly takes into account the relative harms of false positive and false negative classifications.

2.2. Multicenter validation of a diagnostic multivariable risk model for diagnosing ovarian cancer

The LR2 model is a logistic regression model based on ultrasound characteristics to obtain a pre-operative diagnosis of ovarian cancer, yielding a predicted probability of malignancy.[17] Testa et al. performed a multicenter validation study of the predictive performance of the LR2 model, which included 2403 patients from 18 centers. All patients were selected for surgical removal of an adnexal mass and histology was the reference standard to diagnose cancer.[17] The predicted probabilities of malignancy for cancer cases and healthy patients in the validation dataset are

shown per center in Figure 1, for the 15 largest centers. The observed prevalence of malignancy varied between 15% and 69%.

A meta-analysis in the included centers yielded a summary c-statistic of 0.92 with little between-center heterogeneity, demonstrating excellent discrimination between benign and malignant tumors.[17] However, Testa et al. showed that the LR2 model tends to underestimate the probability of malignancy in most centers, that is, there was some miscalibration of the model's predictions.[17] To conclude whether or not the LR2 model is clinically useful in new centers despite the miscalibration, it is required to perform a random-effects meta-analysis of the NB. This takes into account the harms of missing a cancer and the harms of unnecessarily referring a patient without cancer for specialized oncology care based on a predefined risk threshold.

Figure 1. Density plots of predicted probabilities of ovarian cancer from the LR2 model, per center. Plot headings indicate the center's location, number of patients and number of cancer cases.

[insert figure 1 here]

3. Investigating heterogeneity in the Net Benefit of a test/model using meta-analysis

3.1. Net Benefit

Unlike traditional measures of predictive performance, NB incorporates the consequences of using the test/model to guide clinical decision making. The method assumes that there is a risk threshold, t, at which one is uncertain about treating or not treating a patient. If P(disease)<t, the patient should forego treatment, whereas the patient should receive treatment if P(disease)≥t. The relative consequences of falsely treating a patient without disease versus falsely withholding treatment from a patient with disease are implied by t: odds(t) is the ratio of the harm associated with a false positive result and the harm associated with a false negative result.[10] For example, if a risk threshold of 0.20 is used, the harm ratio is 1:4. The harm of a false negative is four times larger than the harm of unnecessary treatment. This implies that unnecessarily treating up to four patients for each correctly treated patient is considered acceptable. Odds(t) can also be thought of as a 'harm-to-benefit' ratio. Indeed, the harm of a false negative equals the forgone benefit of being rightly treated.

NB corrects the number of true positives (r_{11}) for the number of false positives (r_{01}) weighted by the harm ratio (odds(t)), and divides the result by the total sample size (n):[10]

(1)

$$NB_t = \frac{r_{11} - \left(r_{01} \times \frac{t}{1-t}\right)}{n}.$$

NB_t quantifies benefit in terms of the net proportion of true positives at threshold t. If we evaluate a risk prediction model that yields predicted probabilities, r_{11} and r_{01} vary with the chosen risk threshold (and hence the harm ratio), as we classify patients with a predicted risk >t as positive. If we are evaluating a binary diagnostic test, r_{11} and r_{01} are constant but the harm ratio can still be varied.[18] Reasonable choices for the harm ratio reflect differences in risk aversion (the highly risk averse prefer a lower threshold) and/or health care systems. For example, when long waiting lists

for treatment are a reality, higher risk thresholds may be adopted. Note that NB_t should not be used to select the risk threshold t.[4] Rather, for a given t, reflecting a certain harm ratio, we can find out whether a test/model is clinically useful.

NB_t can also be computed for default strategies where either everyone is treated or no one is treated. In fact, for 'treat none', $r_{11}=r_{01}=0$, hence $NB_{t\ treat\ none}=0$ by definition. If NB_t of a test/model is higher than the NB_t of 'treat all' or 'treat none', the test/model is considered useful at threshold t. We will elaborate on these comparisons in section 5. In the next sections, we first consider a meta-analysis of the NB_t.

3.2. A trivariate meta-analysis model of the true positive rate, true negative rate and prevalence

3.2.1. Trivariate meta-analysis

The NB_t is a function of classification results and the risk threshold. Hence, in a meta-analysis, the summary NB_t can be computed from summary measures of prevalence, sensitivity, and specificity, at a given risk threshold.

Suppose data from J (j=1 to J) settings are available. We will use the term 'setting' throughout this work to refer to a single center in a multicenter study, or to a single study in the meta-analysis of multiple published studies. Each has n_{1j} and n_{0j} patients with and without the disease (or other outcome of interest), respectively, and $n_{1j}+n_{0j}=n_j$. In setting j, at a chosen risk threshold, the test or model under validation yields a positive classification for r_{11j} patients with the disease (true positives), and a negative classification for r_{00j} patients without the disease (true negatives). The observed sensitivity in each setting is r_{11j}/n_{1j} and the observed specificity is r_{00j}/n_{0j}. To combine this information across settings, a trivariate random-effects meta-analytic model for prevalence, sensitivity, and specificity has been proposed previously.[19] Assume a binomial distribution for the number of patients with the disease or outcome, the number of true positives, and the number of true negatives in each setting:

(2)

$$n_{1j} \sim \text{bin}\left(n_j, p_j\right)$$

$$r_{11j} \sim \text{bin}\left(n_{1j}, Se_j\right)$$

$$r_{00j} \sim \text{bin}\left(n_{0j}, Sp_j\right).$$

After applying the logit transformation, the true setting-specific prevalences (p_j), sensitivities (Se_j), and specificities (Sp_j) are assumed to be normally distributed with means γ_1, γ_2, and γ_3. The variance-covariance matrix $\boldsymbol{\Omega}$ contains the between-setting variance in the logit prevalence $\left(\tau_1^2\right)$, the logit sensitivity $\left(\tau_2^2\right)$, and the logit specificity $\left(\tau_3^2\right)$, and the covariances. Hence, we account for the heterogeneity and the correlations between the logit prevalence and the logit sensitivity (ρ_{12}), the logit prevalence and the logit specificity (ρ_{13}), and the logit sensitivity and the logit specificity (ρ_{23}):

(3)

$$\begin{pmatrix} \text{logit}\left(p_j\right) \\ \text{logit}\left(Se_j\right) \\ \text{logit}\left(Sp_j\right) \end{pmatrix} \sim N\left[\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix}, \boldsymbol{\Omega}\right], \boldsymbol{\Omega} = \begin{pmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 & \rho_{13}\tau_1\tau_3 \\ \rho_{12}\tau_1\tau_2 & \tau_2^2 & \rho_{23}\tau_2\tau_3 \\ \rho_{13}\tau_1\tau_3 & \rho_{23}\tau_2\tau_3 & \tau_3^2 \end{pmatrix}.$$

The summary NB_t can now be estimated from γ_1, the summary logit prevalence, γ_2, the summary logit sensitivity, γ_3, the summary logit specificity, and t, as follows:

(4)

$$NB_t = \frac{\exp(\gamma_2)}{1+\exp(\gamma_2)} \times \frac{\exp(\gamma_1)}{1+\exp(\gamma_1)} - \left[\left(1 - \frac{\exp(\gamma_3)}{1+\exp(\gamma_3)}\right) \times \left(1 - \frac{\exp(\gamma_1)}{1+\exp(\gamma_1)}\right) \times \frac{t}{1-t}\right].$$

The process can be repeated for each threshold (i.e., at each threshold apply a trivariate meta-analysis of sensitivity, specificity, and prevalence, and then use equation (4) to obtain the NB_t).

The trivariate model of sensitivity, specificity, and prevalence can be estimated using a frequentist or Bayesian approach. We prefer a Bayesian approach, because it yields an estimate of the posterior probability distribution of the NB_t that accounts for all parameter uncertainty and naturally enables subsequent predictions of the NB_t in new settings. However, it also requires the specification of prior distributions. We used a vague multivariable normal prior distribution for the vector of logit prevalence, logit sensitivity and logit specificity, with mean 0, variances 1000, and covariances 0. An inverse Wishart prior was used for the between-setting variance-covariance matrix $\boldsymbol{\Omega}$, with variances 10, covariances 0, and the number of degrees of freedom as small as possible (3, the number of outcomes) to reflect vague prior knowledge. However, it has been shown that seemingly vague Wishart priors may still be influential, which can affect posterior inferences for the between-setting variances, correlations, and pooled summary estimates.[20-22] Therefore, in the next section, an alternative product normal parametrization of the model is introduced, which allows specifying priors for all elements of the variance-covariance matrix separately.

3.2.2. Alternative product normal parametrization of the between-setting model.

The between-setting model (3) can be reparametrized in the product normal formulation, as proposed by Bujkiewicz et al.[21,23] This formulation models the true logit prevalence, the true logit sensitivity and the true logit specificity as conditional univariate normal distributions, using linear models for the relations between them:

(5)

$$\begin{cases} \text{logit}\left(p_j\right) \sim N\left(\eta_1, \psi_1^2\right) \\ \text{logit}\left(Se_j\right) | \text{logit}\left(p_j\right) \sim N\left(\eta_{2j}, \psi_2^2\right) \\ \eta_{2j} = \lambda_{20} + \lambda_{21} \text{logit}\left(p_j\right) \\ \text{logit}\left(Sp_j\right) | \text{logit}\left(p_j\right), \text{logit}\left(Se_j\right) \sim N\left(\eta_{3j}, \psi_3^2\right) \\ \eta_{3j} = \lambda_{30} + \lambda_{31} \text{logit}\left(p_j\right) + \lambda_{32} \text{logit}\left(Se_j\right). \end{cases}$$

Instead of specifying a prior distribution for the between-setting variance-covariance matrix **Ω** as a whole, this formulation allows us to place realistic prior distributions on separate elements of **Ω**. Hence, this formulation eases the incorporation of prior information on the between-setting standard deviations and correlations obtained from previous studies or published literature. If this information is not available, realistic weak priors can be used, such as weak Fisher priors for correlations, which restrict correlations between -1 and 1, and weak half-normal priors for variances, which are bounded by zero.[20,21] The implied prior distributions on the parameters λ_{21}, λ_{31}, and λ_{32}, and the hyper-parameters ψ_1, ψ_2, and ψ_3 are dictated by the relationships between these parameters and the elements of the between-setting variance-covariance matrix, which have been derived by Bujkiewicz and colleagues:

(6)

$$\psi_1^2 = \tau_1^2,\ \psi_2^2 = \tau_2^2 - \lambda_{21}^2 \tau_1^2,\ \psi_3^2 = \tau_3^2 - \lambda_{31}^2 \tau_1^2 - \lambda_{32}^2 \tau_2^2$$

$$\lambda_{21} = \frac{\tau_2}{\tau_1} \rho_{12},\ \ \lambda_{31} = \frac{\tau_3}{\tau_1} \rho_{13} - \lambda_{32} \lambda_{21},\ \ \ \lambda_{32} = \frac{\rho_{23}\, \tau_2 \tau_3 - \rho_{13}\, \tau_1 \tau_3 \lambda_{21}}{\tau_2^2 - \lambda_{21}^2 \tau_1^2}.$$

The remaining parameters can be given vague prior distributions, for example, $\eta_1 \sim N(0,1000)$, $\lambda_{20} \sim N(0,1000)$, $\lambda_{30} \sim N(0,1000)$. The summary logit prevalence, logit sensitivity, and logit specificity are directly linked to the re-parametrized model formulation:

(7)

$$\gamma_1 = \eta_1$$

$$\gamma_2 = \lambda_{20} + \lambda_{21} \gamma_1$$

$$\gamma_3 = \lambda_{30} + \lambda_{31} \gamma_1 + \lambda_{32} \gamma_2.$$

Hence, the summary NB_t can be computed based on γ_1, the summary logit prevalence, γ_2, the summary logit sensitivity, γ_3, the summary logit specificity, and t, the risk threshold of interest, as in equation (4).

3.3. Net Benefit for a new population

3.3.1. 95% prediction intervals

In the presence of between-study (or between-center) heterogeneity in disease prevalence or predictive performance, the summary NB_t is potentially inadequate to quantify the expected NB_t

in a new study (or center). A 95% prediction interval for the NB_t in a new setting reveals the potential impact of heterogeneity on the clinical utility in new settings. A 95% prediction interval for NB_t can be obtained in a natural way in a Bayesian framework, by sampling sensitivity, specificity, and prevalence for new studies (or centers) from their joint posterior distribution. Hence, the uncertainty in all parameters estimated in the meta-analytic model is propagated when deriving predictions for the NB_t in a new setting.

3.3.2. Prediction intervals using prior knowledge on the prevalence in the new setting

The product normal formulation of the between-setting model allows an elegant prediction of NB_t in a new setting, conditional on a known prevalence, using the posterior estimates of lambda parameters:

(8)

$$\hat{\gamma}_2 = \hat{\lambda}_{20}+\hat{\lambda}_{21}\text{logit(prevalence)}$$

$$\hat{\gamma}_3=\hat{\lambda}_{30}+\hat{\lambda}_{31}\text{logit(prevalence)}+\hat{\lambda}_{32}\hat{\gamma}_2$$

$$\widehat{\text{NB}}_t|\text{prevalence} = \frac{\exp(\hat{\gamma}_2)}{1+\exp(\hat{\gamma}_2)} \times \text{prevalence} - \left[\left(1 - \frac{\exp(\hat{\gamma}_3)}{1+\exp(\hat{\gamma}_3)}\right) \times (1 - \text{prevalence}) \times \frac{t}{1-t}\right].$$

The prevalence can be treated as known a priori, or given a distribution reflecting uncertainty in the prevalence estimate. By sampling from the posterior distribution, uncertainty in all parameters is accounted for in the prediction interval of the NB_t at a given prevalence, while the interrelations between prevalence, sensitivity, and specificity are accounted for. When a Wishart prior is used instead of the product normal formulation, the sensitivity and specificity of a new setting can be sampled and combined with the prevalence to obtain the predicted NB_t.

3.4. Implementation in WinBUGS

The models were implemented in WinBUGS. The estimates were obtained using MCMC sampling with 2 chains of 100 000 iterations, excluding a burn-in of 50 000, and using a thinning factor of 20. The convergence was checked visually by monitoring the chains for the parameters of interest (i.e., sensitivity, specificity, prevalence, NB_t, τ_1^2, τ_2^2, τ_3^2 , NB_t | prevalence, and NB_t treat all (see section 5). The NB_t, NB_t | prevalence, and NB_t treat all were monitored separately for settings in the sample and new settings sampled from the joint posterior of sensitivity, specificity and prevalence to obtain prediction intervals). The model syntax is included in Web Appendix 1 and 2. Summary estimates of NB_t are reported as posterior means with 95% credible intervals and 95% prediction intervals.

4. Application to the two case studies

In what follows, the NB_t is computed for using the in-ear thermometer to diagnose fever and for the LR2 model to diagnose ovarian cancer. The results we present are the results of the product normal formulation, using realistic priors for the between-study (between-center) variance-covariance matrix. Comparisons with the results obtained when using the Wishart distribution are presented in Appendix Table 2 and Appendix Table 3.

4.1. Meta-analysis of the clinical utility of ear temperature for diagnosing fever in children

The NB_t of using an ear thermometer was calculated at three risk thresholds: 0.2, 0.5, and 0.8. A risk threshold of 0.2 indicates that we would be willing to diagnose 4 healthy children with fever to detect one true positive case. A risk threshold of 0.5 indicates that we believe the harm of a false positive is equal to the harm of a false negative. A risk threshold of 0.8 indicates that we perceive the harm of a false positive to be 4 times larger than the harm of a false negative. We selected these risk thresholds for illustrative purposes and assume they reflect different opinions regarding the perceived relative harms of false positive and false negative diagnoses of fever clinicians may hold.

We used Fisher priors ($z \sim N(-0.20, 0.50^2)$, with $z=\log((1+\rho)/(1-\rho))$) for the correlations between logit sensitivity and logit specificity (ρ_{23}), and logit specificity and logit prevalence (ρ_{13}). The chosen prior distribution corresponds to a moderate negative correlation (Appendix Figure 1) to reflect prior knowledge on spectrum bias, the phenomenon that the performance of a test or model differs between clinical settings due to case-mix differences. Referral patterns may lead to higher sensitivities and lower specificities in high-prevalence settings.[24] Leeflang found empirical evidence for a negative correlation between specificity and prevalence, indicating that differences in prevalence may represent changes in the spectrum of people without the disease of interest.[25] People with symptoms or prior test results indicative of the disease may be referred to certain centers, yielding at the same time a higher prevalence and a group of patients without the disease that are harder to diagnose correctly. For the correlation between logit sensitivity and logit prevalence (ρ_{12}), a uniform prior (U[-0.99, 0.99]) was used, as previous empirical research could not demonstrate an overall positive or negative correlation relation between sensitivity and prevalence.[25] We used a half-normal prior distribution for the between-setting variance of logit sensitivity, logit specificity, and logit prevalence, $\tau^2_{1\,(2,3)} \sim N(0, 2^2)I(0,)$ (see Appendix Figure 1).

Arbitrary but plausible initial values were manually set for some key parameters to facilitate convergence ($\eta 1$: 0.45, $\lambda 20$: 0.7, $\lambda 30$: 0.08, $Z(\rho 12)$: 0, $Z(\rho 13)$: 0, $Z(\rho 23)$: 0, the between-setting variance of sensitivity: 0.31, the between-setting variance of specificity: 0.47, and between-setting variance of prevalence: 0.2). The initial values for remaining parameters in the model were randomly sampled from the prior distributions, using the "gen inits" function in WinBugs. In the case of vague priors, this function generates extreme initial values.

Recall that children with an ear temperature ≥38.0 °C were classified as having fever, and children with an ear temperature <38.0 °C were classified as not having fever, regardless of t. This yielded a summary sensitivity of 65%, and a summary specificity 98%. When the risk threshold was 0.2, reflecting larger perceived harms of false negatives than false positives, the summary $NB_{0.2}$ was 0.30 (95% CrI 0.19 to 0.42). This indicates that the net benefit of this diagnostic test is equivalent to the benefit of correctly classifying a net number of 30 children with fever per 100 patients, and no false positive classifications (Table 2). In section 5, we will give an interpretation that puts the magnitude of this value into perspective. The prediction interval reveals there is a 95% probability that $NB_{0.2}$ will be between 0.03 and 0.68 in a new study. For a new study with a prevalence of fever of 50%, we find a summary $NB_{0.2}$ of 0.32 and a

narrower prediction interval of 0.12 to 0.47, reflecting that we now have information on the prevalence, which restricts the likely values of NB_t in the new setting.

If we compute the NB_t at higher risk thresholds, indicating equal or lower perceived harms for false negatives than for false positives, the summary NB_t is lower. This reflects the increasing correction for the number of false positives. Results for risk thresholds 0.5 and 0.8 are given in Table 2.

Similar point estimates with generally narrower credible intervals were obtained when inverse Wishart priors were used for the between-study variance-covariance matrix, but small differences were observed, especially for NB_t at t=0.8 (see Appendix Table 2).

Sensitivity to the choice of initial values was checked by repeating the analysis with widely separated starting values for the two chains ($\eta 1$: -2.2 versus 2.2 , $\lambda 20$: -2.2 versus 2.2, $\lambda 30$: -2.2 versus 2.2, $\rho 12$: 0 versus 0.9, $Z(\rho 13)$: 0 versus -1.47, $Z(\rho 23)$: 0 versus -1.47, the between-setting variance of sensitivity: 0.1 versus 5, the between-setting variance of specificity: 0.1 versus 5, and between-setting variance of prevalence: 0.1 versus 5). All chains started mixing in the first 300 iterations and converged to summary estimates that were very similar to the ones reported.

4.2. Multicenter validation of a diagnostic multivariable risk model for ovarian cancer

The NB_t of the LR2 model for diagnosing ovarian cancer was calculated at three risk thresholds: 0.05, 0.1, and 0.5. The low risk thresholds indicate the need for a diagnostic strategy with a high sensitivity for malignancy, and perceived harms of false negatives that are at least as high as the perceived harms of false positives. In clinical reality, risk thresholds above 0.5 are not sensible because this would imply that a false positive case (unnecessarily being referred to specialized oncology care to undergo additional MRI testing) is more harmful than a false negative (an undetected cancer).

In this analysis, we used the posterior probability distributions obtained from the analysis of another validation study of the LR2 model (n=1938) as prior distributions in the current analysis.[26,27] This external dataset contained data from 19 centers (12 of which also contributed data to the current dataset), and was collected at an earlier time. The external dataset was analyzed using weak realistic priors (the same as in section 4.1). We characterized the resulting posterior distributions by their means and standard deviations (which varied with t as shown in Appendix Table 1), and chose parametric distributions to match their shapes. We used normal distributions for η_1, λ_{20}, and λ_{30}, lognormal distributions for the between-center variance of logit sensitivity, logit specificity, and logit prevalence, and normal distributions for the Fisher transformations of the correlations. The resulting distributions were used as prior distributions in the current analyses.

To facilitate convergence, the means of the posteriors obtained in the external datasets were used to set initial values for some key parameters ($\eta 1$, $\lambda 20$, $\lambda 30$, $Z(\rho 12)$, $Z(\rho 13)$, $Z(\rho 23)$, and the between-setting variance of sensitivity, specificity, and prevalence). The initial values for remaining parameters in the model were randomly sampled using the "gen inits" function in WinBugs.

Because the number of patients classified as test positive decreases as the adopted risk threshold increases, the summary sensitivity decreases from 0.95 at t=0.05 to 0.63 at t=0.5, and the summary specificity increases from 0.68 at t=0.05 to 0.95 at t=0.5 (Table 3). At a risk threshold of 0.05, the summary $NB_{0.05}$ is 0.27 (95% CrI 0.21 to 0.34). Hence, the net benefit of the model is equivalent to the benefit of a strategy that correctly detects a net number of 27 cancer cases per 100 patients, without false positive classifications. A more complete interpretation is given in section 5. The heterogeneity in $NB_{0.05}$ between centers is quite large, as reflected by the 95% prediction interval: 0.05 to 0.66. A more precise prediction of the $NB_{0.05}$ in a new center may be obtained by conditioning on a known prevalence. Regional centers typically have a lower cancer prevalence than university hospitals with a specialized gynecological oncology unit. For new center with a known prevalence of 15%, the $NB_{0.05}$ lies between 0.12 and 0.14 with 95% certainty. For a new center with a known prevalence of 35%, the $NB_{0.05}$ lies between 0.29 and 0.33 with 95% certainty.

With higher risk thresholds, the NB_t is lower, reflecting that by adopting these perceived harm ratios gradually more weight is given to false positive classifications (Table 3).

The point estimates, 95% credible intervals and 95% prediction intervals were influenced by the choice of prior for the variance-covariance matrix, but differences nearly disappeared when NB_t was estimated conditional on a known prevalence in a new setting (see Appendix Table 3). Credible intervals were often wider when the Wishart prior was used.

5. Putting the Net Benefit into perspective: is the test/model clinically useful?

5.1. The probability that a test/model is clinically useful in a new setting

The NB_t of a test/model is usually compared to the NB_t of 'treat all' and 'treat none'. The NB_t of treat all equals $p-[(1-p)\times t/(1-t)]$, with p the prevalence of the disease, while the NB_t of treat none equals 0, irrespective of t. Similar to NB_t of a test/model, NB_t of 'treat all' may vary between settings, because it is dependent on the prevalence. By sampling from the posterior distribution of the prevalence, the summary NB_t of treat all and a 95% prediction interval can be obtained in the Bayesian random-effects meta-analytic framework outlined above. By using a priori knowledge on the prevalence, we can also calculate the NB_t of treat all in a new setting with the given prevalence.

If the NB_t of the test/model of interest is below that of treat all or treat none, the test/model is harmful, because decisions made without use of the test/model have higher clinical utility.[9,10] In contrast, if the NB_t of the test/model of interest higher than that of treat all and treat none, the test/model is clinically useful:

(9)

$$P(\text{useful}) = P[NB_t > \max(NB_{t\ \text{treat all}}, 0)]\,.$$

In the Bayesian framework, we can sample from the joint posterior distribution of sensitivity, specificity and prevalence, and evaluate in each sample whether the NB_t of the test/model in a new setting is larger than the NB_t of treat all and treat none. Hence, we obtain the probability that the test/model is useful in any new setting. By using a priori knowledge on the prevalence, we

can also calculate the probability that the test/model is clinically useful in a new setting with the given prevalence.

At a risk threshold of 0.2, treat all is the best default strategy for fever, with a $NB_{0.2}$ of 0.33 (95% CrI 0.20 to 0.46). At the 0.5 and 0.8 risk thresholds, the summary NB_t of treat all is <0, making treat none the best default strategy (Table 4). Nonetheless, there is considerable heterogeneity in the NB_t of treat all in new studies, reflecting heterogeneity in prevalence. For example, at t=0.5, the 95% prediction interval is -0.65 to 0.57. The probability that using an in-ear thermometer is better than the two default strategies of diagnosing fever in all or none of the children depends on the adopted risk threshold. At t=0.2 (i.e., a harm ratio of 1:4) there is a 44% chance that using an ear thermometer is clinically useful in any new setting, and a 34% chance that this strategy is clinically useful in a setting with a known prevalence of fever of 0.50. However, if we assume that false positive and false negative diagnoses are equally bad (t=0.5), there is a 97% chance that using an in-ear thermometer is clinically useful in any new setting, and a 99.9% chance that this strategy is clinically useful in a setting with a known prevalence of fever of 0.50. At t=0.8, the probability of usefulness in any new setting is 95%, and the probability of usefulness in a new setting with a prevalence of 0.50 is 99%.

The best default strategy to diagnose ovarian malignancy at risk thresholds 0.05 and 0.1 is treat all ($NB_{0.05}$ 0.26, 95% CrI 0.20 to 0.34; $NB_{0.1}$ 0.22, 95% CrI 0.15 to 0.30), while it is treat none at t=0.5 (Table 5). Here too, the between-center heterogeneity in the NB_t of treat all is large, with the 95% prediction interval at t=0.05 ranging from 0.02 to 0.69. If the perceived harms of false negatives are 19 times larger than the perceived harms of false positives (t=0.05), there is a 69% chance that LR2 is useful in any new center. The probability of usefulness is 99.9 % in a center with a malignancy rate of 15%. Interestingly, if the malignancy rate is 35%, the probability that LR2 is useful is lower: 75%. This is likely due to the miscalibration of LR2, which seemed to be especially pronounced in centers with a high prevalence.[17] The probabilities that the LR2 model is useful to diagnose ovarian cancer in a new center when risk thresholds 0.1 or 0.5 are adopted are given in Table 5.

The summary estimates, 95% credible intervals and 95% prediction intervals of the NB_t for treat all were influenced by the choice of prior in both case studies (see Appendix Table 2 and Appendix Table 3). The probability of usefulness was also influenced by the choice of prior; the biggest difference was observed in the fever case study with t=0.2 (0.31 with the Wishart prior versus 0.44 with the realistic weak informative priors).

A straightforward extension of the proposed methods to quantify the probability of clinical utility in new settings is to include a routinely suggested competitor test/model in the comparison. One may want to quantify the probability that the model/test of interest performs better than the routinely used test/model and both default strategies: $P[NB_t > \max(NB_{t\ \text{routinely used model/test}}, NB_{t\ \text{treat all}}, 0)]$. Routinely used competitor models/tests may be the temperature taken in the mouth or under the arm for the first example, and the RMI algorithm for ovarian cancer in the second example.[17,26,27]

5.2. Decision curves

A decision curve is a plot of the NB_t of a test/model and the NB_t of default strategies for a range of relevant risk thresholds, and allows an easy graphical comparison of diagnostic strategies.[10] In the Bayesian framework outlined above, the NB_t, pointwise credible intervals and pointwise prediction intervals can be computed for each risk threshold of interest, and subsequently plotted and connected, as in Figure 2 below for the fever example and Figure 3 below for the ovarian cancer example.

In the first example, it is immediately clear that the summary NB_t of infrared ear thermometry (blue curve) is below the summary NB_t of treating all patients (purple curve) up to risk threshold 0.23. For higher risk thresholds, the summary NB_t of using an ear thermometer is higher than the summary NB_t of both default strategies. The large between-study heterogeneity in NB_t is reflected by the broad prediction intervals around the summary curves (light transparent bands).

The large heterogeneity makes it impossible to visually assess whether ear thermometry will have clinical utility to diagnose fever in a new setting. Therefore, we added a plot of the probability that ear thermometry is useful in a new population, that is, the probability that the NB_t of ear thermometry is higher than the NB_t of treat all or treat none in a new setting. The probability that using an ear thermometer for diagnosing fever is useful in a new setting is above 90% for risk thresholds between 0.39 and 0.86. Hence, if the perceived harm of a false negative is between 1.7 and 0.2 times as large as the harm of a false positive, using an in-ear thermometer is a good diagnostic strategy. In contrast, if the perceived harm of a false negative is at least 9 times larger than the harm of a false positive ($t \leq 0.1$), reflecting that is very important not to miss cases of fever, the probability that using an in-ear thermometer is clinically useful in a new setting is less than 10%. In this case, it is better to err on the safe side and assume all patients have fever, or to use a rectal thermometer when possible.

In the second example, the summary curve for the LR2 model (blue line) is higher than the summary curve of treat all (purple line) and treat none (black horizontal line) at all considered risk thresholds. However, the heterogeneity in the NB_t of LR2 and treat all is very large, as reflected by the large prediction intervals. This is mainly caused by the large heterogeneity in the prevalence of malignancy across centers. The probability that LR2 has clinical utility to make treatment decisions in new centers is above 90% for risk threshold between 0.20 and 0.50. Hence, if the clinician's judgement is that a false negative is between four and one times as harmful as a false positive, the LR2 model is a good tool to pre-operatively diagnose ovarian cancer. If the clinician believes that a false negative is 19 times as harmful as a false positive, the probability that the LR2 model is clinically useful in a new setting is still 70%.

Extensions of the plot are straightforward: one may add the NB_t curve of another diagnostic test, for example, temperature taken in the mouth or under the arm for the first example, and other diagnostic models for ovarian cancer in the second example. In addition, one may create a plot depicting NB_t curves and the probability that the test/model is useful for a known prevalence in a new setting. Note that in this case, there will be no heterogeneity in the NB_t curve for treat all, and the heterogeneity in the NB_t of the test/model will only reflect heterogeneity in diagnostic accuracy (not heterogeneity in prevalence).

Figure 2. Decision curve for diagnosing fever using an in-ear thermometer. The bottom panel shows the probability that using an in-ear thermometer is clinically useful (i.e., has a higher NB_t than treat all and treat none) in a randomly chosen new study.

[insert figure 2 here]

Figure 3. Decision curve for diagnosing ovarian malignancy using the LR2 model. The bottom panel shows the probability that using LR2 is clinically useful (i.e., has a higher NB_t than treat all and treat none) in a randomly chosen new center.

[insert figure 3 here]

6. Discussion

When heterogeneity in disease prevalence or diagnostic accuracy exists, the clinical usefulness of diagnostic tests and models differs between populations. We proposed to evaluate the Net Benefit (NB_t) of a test or model across different studies or centers through a trivariate random-effects meta-analysis of sensitivity, specificity, and prevalence. This approach directly models the binomial distributions of true positive counts, true negative counts, and the number of diseased patients, and estimates the summary Net Benefit in a second step. Differences between settings are modelled through a product normal formulation of the between-study (or between-center) model of the logit sensitivity, logit specificity, and logit prevalence. This allows the user to specify realistic prior distributions for all elements of the variance-covariance matrix separately.

Probability statements on the likely NB_t in new settings provide crucial insights into the heterogeneity of the clinical utility. Using a Bayesian analysis, we sampled values for a new setting from the joint posterior distribution of sensitivity, specificity, and prevalence, to formulate statements on the likely NB_t in a new setting. Examples include 95% prediction intervals for the NB_t, 95% prediction intervals for the NB_t for settings with a known prevalence, and the probability that a test/model is useful in a new setting (i.e., a higher NB_t for the test/model than for default strategies of treating all or treating none). Heterogeneity in the clinical utility of the diagnostic test/model was demonstrated in two case studies.

Heterogeneity in diagnostic accuracy is omnipresent: a systematic review estimated that heterogeneity affects 70% of published meta-analyses.[28] However, even if the heterogeneity in diagnostic accuracy is negligible, heterogeneity in disease prevalence may still render the NB_t of the test/model under investigation highly heterogeneous, as demonstrated in our case study on the diagnosis of ovarian cancer. The NB_t of treating all patients without using a test/model is also heterogeneous across settings, whenever there is heterogeneity in the prevalence of the disease. This makes it very difficult to assess when a test/model is clinically useful, that is, superior to treat all or treat none: prediction intervals are very wide and swamp graphical depictions of the decision curves. Therefore, we proposed to calculate the posterior probability that the test/model is clinically useful in a new setting.

Our case studies have demonstrated that the probability of clinical usefulness in a new setting may be far from 0 or 1 at certain risk thresholds, effectively illustrating variability in the clinical usefulness of tests/models. This finding highlights once more the necessity of validating

diagnostic tests and risk prediction models in multiple relevant care settings to investigate heterogeneity in predictive performance and clinical utility, and assess generalizability and transportability.[1,2] A practical approach is proposed in Box 1. If a risk prediction model has no clinical utility in certain settings, it may be worthwhile to update the model for these settings by recalibrating the model intercept or adjusting regression coefficients.[29,30] It has been shown that perfectly calibrated models are never harmful.[13] However, in practice, it may be difficult to obtain perfect calibration in all settings.[3]

Box 1. Practical recommendations for validation.

[insert Box 1 here]

To get a more precise estimate of the clinical utility in a new setting, we suggest using prior knowledge on the prevalence in the target setting, if available. A similar idea was developed by Willis and colleagues,[31,32] who proposed to use a priori information on the test positive rate and prevalence to exclude studies from the meta-analysis and obtain tailored estimates of test accuracy. Rather than excluding studies/centers from the meta-analysis, our approach uses all available information to explicitly model the correlations between prevalence and diagnostic accuracy. By subsequently conditioning on the known prevalence, we obtain a relevant summary NB_t, a prediction interval, and the probability that the test/model will be useful in a new setting with the given prevalence. It is straightforward to reformulate our between-setting model of sensitivity, specificity, and prevalence in terms of test positive rates, positive predictive values, and negative predictive values. Such an alternative formulation may be of practical use, allowing assessments of clinical usefulness conditional on the test positive rate. Centers that wish to use the test/model under investigation may be more likely to know the test positive rate in their setting than the prevalence.

As an alternative to the proposed random-effects meta-analysis of the NB_t, you may consider combining the individual patient data from all centers or studies. One then computes the NB_t based on the pooled dataset as if all observations came from the same setting. This approach is equivalent to computing a weighted average of the study or center-specific NB_ts, using the number of observations in each study or center as weights. As this approach fails to acknowledge any potential heterogeneity in clinical utility across settings, we do not recommend it. Kerr and colleagues proposed to calculate separate NB_t-values for subpopulations, especially when diagnostic models have different predicted risk distributions.[4] Although Kerr did not focus on subpopulations defined by different care settings, but rather on patient subgroups within one population (e.g., men and women), the idea may be applied to centers and studies as well. This approach recognizes heterogeneity in clinical utility, but a random-effects meta-analytic approach has the advantages of providing an overall summary estimate of NB_t, and borrowing of strength for small populations. Indeed, estimates of the NB_t may be unreliable for small centers or studies.

Instead of performing a two-step approach (i.e. first performing a multivariable meta-analysis of diagnostic accuracy and prevalence and then, post-estimation, calculating the NB_t) one could rather perform a random-effects meta-analysis of the study- or center-specific NB_ts directly. However, a number of difficulties are associated with this approach. First, it requires estimates of

the sampling variance of the NB_t, for which no closed formula exists.[9,33] Obtaining bootstrap standard errors for each study/center at each risk threshold is computationally intensive. Second, a between-study (or between-center) model of the NB_t would need to be specified. Often, a normal distribution is chosen, but this may not be appropriate. At low thresholds asymmetric distributions may be expected, especially when the disease prevalence is high, as the NB_t is bounded by 1 and the maximum attainable NB_t equals the prevalence. At high risk thresholds, one may also expect an asymmetric distribution, one that is peaked near zero, with many studies/centers with a NB_t well below zero. Especially in settings in which a diagnostic model discriminates well but is badly calibrated such that it overestimates the risk of an event, NB_ts below zero may occur at high risk thresholds.[13] Lastly, the NB_t estimates are not independent from their variance estimates. Typical variance-stabilizing functions, such as the logit and arcsine transformations, are not appropriate for the analysis of the NB_t, which can take on any value between $-\infty$ and 1. For these reasons, we prefer the proposed two-step approach over a direct meta-analysis of the NB_t.

The most important advantage of studying the clinical utility of a test/model, is that the NB_t transcends traditional measures of discrimination and calibration, to incorporate consequences for clinical decision-making into the evaluation. The strength of our meta-analytic approach is that we explicitly address heterogeneity in clinical utility. In practice, a test/model may be widely recommended if the meta-analysis demonstrates clinical utility in a broad range of care settings. This generalist approach may be more feasible than a particularistic approach in which studies are conducted for each setting separately, and the use of the test/model is restricted to those centers in which utility has been demonstrated. A particularistic approach is often characterized by small sample sizes and high uncertainty in individual studies.

This study has a number of limitations. For one, we merely investigate the presence of heterogeneity, but do not address the sources of heterogeneity. In the future, meta-regressions may be undertaken to incorporate study or center characteristics in the between-setting model as potential sources of heterogeneity. We assumed normality of the logit sensitivity, logit specificity, and logit prevalence. These are common assumptions,[15,19,34] but future simulation studies could investigate whether our proposed approach is robust against violations of this assumption. Our analyses have shown that our method is sensitive to the choice of priors for the between-setting variance-covariance matrix, which is in accordance with previous studies.[20-22] Hence, we recommend picking priors based on available knowledge, or realistic weak priors, which is facilitated through the product normal formulation of the between-setting model.[21,23] In case no prior knowledge is available, we suggest to use the priors specified in section 4.1. The presented case studies have only focused on the difference in the NB_t between one test/model and the default strategies (treat all and treat none). The applications can easily be extended to compare the clinical utility of competing diagnostic tests or risk prediction models.

A hindrance to the uptake of our method in practice may be that individual patient-data from all studies (or centers) is needed when a risk model is evaluated, as the classification then depends on the adopted risk threshold. One needs to evaluate for each risk threshold whether patients' predicted risks fall below or above it. For binary tests, individual patient data is not needed, as

long as the numbers of diseased patients, true positives and false positives can be calculated from the reported outcomes.

In summary, we have demonstrated a method to calculate the NB_t based on a trivariate random-effects meta-analysis of diagnostic accuracy and disease prevalence. The findings on our case studies suggest that the heterogeneity in clinical utility of a test/model across care settings should be quantified, before it is routinely implemented in practice.

References

1. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* New York, NY: Springer US; 2009.
2. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol.* 2015;68(1):25-34.
3. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol.* 2016;74:167-176.
4. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *J Clin Oncol.* 2016;34(21):2534-2540.
5. Localio AR, Goodman S. Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. *Ann Intern Med.* 2012;157(4):294-295.
6. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. *Lancet Oncol.* 2015;16(4):e173-e180.
7. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. *JAMA.* 2015;313(4):409-410.
8. Baker SG, Schuit E, Steyerberg EW, et al. How to interpret a small increase in AUC with an additional risk prediction marker: decision analysis comes through. *Stat Med.* 2014;33(22):3946-3959.
9. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ.* 2016;352:i6.
10. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making.* 2006;26(6):565-574.
11. Hilden J. Prevalence-free utility-respecting summary indices of diagnostic power do not exist. *Stat Med.* 2000;19(4):431-440.
12. Vergouwe Y, Moons KGM, Steyerberg EW. External Validity of Risk Models: Use of Benchmark Values to Disentangle a Case-Mix Effect From Incorrect Coefficients. *Am J Epidemiol.* 2010;172(8):971-980.
13. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making.* 2015;35(2):162-169.
14. Snell KI, Hua H, Debray TP, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol.* 2016;69:40-50.
15. Riley RD, Ahmed I, Debray TP, et al. Summarising and validating test accuracy results across multiple studies for use in clinical practice. *Stat Med.* 2015;34(13):2081-2103.
16. Craig JV, Lancaster GA, Taylor S, Williamson PR, Smyth RL. Infrared ear thermometry compared with rectal thermometry in children: a systematic review. *Lancet.* 2002;360(9333):603-609.
17. Testa A, Kaijser J, Wynants L, et al. Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicentre international IOTA study. *Br J Cancer.* 2014;111(4):680-688.
18. Vickers AJ, Cronin AM, Gonen M. A simple decision analytic solution to the comparison of two binary diagnostic tests. *Stat Med.* 2013;32(11):1865-1876.
19. 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(18):2384-2399.
20. Burke DL, Bujkiewicz S, Riley RD. Bayesian bivariate meta-analysis of correlated effects: Impact of the prior distributions on the between-study correlation, borrowing of strength, and joint inferences. *Stat Methods Med Res.* 2016:0962280216631361.

21. Bujkiewicz S, Thompson JR, Sutton AJ, et al. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Stat Med.* 2013;32(22):3926-3943.
22. Wei Y, Higgins JP. Bayesian multivariate meta-analysis with multiple outcomes. *Stat Med.* 2013;32(17):2911-2934.
23. Bujkiewicz S, Thompson JR, Riley RD, Abrams KR. Bayesian meta-analytical methods to incorporate multiple surrogate endpoints in drug development process. *Stat Med.* 2016;35(7):1063-1089.
24. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol.* 1992;45(10):1143-1154.
25. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ.* 2013;185(11):E537-544.
26. Van Holsbeke C, Van Calster B, Bourne T, et al. External Validation of Diagnostic Models to Estimate the Risk of Malignancy in Adnexal Masses. *Clin Cancer Res.* 2012;18(3):815-825.
27. Timmerman D, Van Calster B, Testa AC, et al. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. *Ultrasound Obstet Gynecol.* 2010;36(2):226-234.
28. Willis BH, Quigley M. The assessment of the quality of reporting of meta-analyses in diagnostic research: a systematic review. *BMC Med Res Methodol.* 2011;11(1):163.
29. Kappen TH, Vergouwe Y, van Klei WA, van Wolfswinkel L, Kalkman CJ, Moons KGM. Adaptation of Clinical Prediction Models for Application in Local Settings. *Med Decis Making.* 2012;32(3):E1-E10.
30. Janssen KJM, Kalkman CJ, Grobbee DE, Bonsel GJ, Moons KGM, Vergouwe Y. Updating prediction rules: Simple methods give promising results. *Eur J Epidemiol.* 2006;21:49-49.
31. Willis BH, Hyde CJ. What is the test's accuracy in my practice population? Tailored meta-analysis provides a plausible estimate. *J Clin Epidemiol.* 2015;68(8):847-854.
32. Willis BH, Hyde CJ. Estimating a test's accuracy using tailored meta-analysis; How setting-specific data may aid study selection. *J Clin Epidemiol.* 2014;67(5):538-546.
33. Vickers A, Cronin A, Elkin E, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak.* 2008;8(1):53.
34. 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(12):1331-1332; author reply 1332-1333.

Table 1. True positives (r_{11}), number with fever (n_1), sensitivity, true negatives (r_{00}), number without fever (n_0), and specificity, for each temperature study included in the meta-analysis.

First author	$\mathbf{r_{11j}}$	$\mathbf{n_{1j}}$	**Sensitivity**	$\mathbf{r_{00j}}$	$\mathbf{n_{0j}}$	**Specificity**
Brennan	150	203	0.74	155	167	0.93
Davis	9	18	0.50	46	48	0.96
Green	8	9	0.89	12	12	1.00
Greenes	53	109	0.49	193	195	0.99
Hoffman	30	42	0.71	56	58	0.97
Hooker	10	15	0.67	24	24	1.00
Lanham	53	103	0.51	74	75	0.99
Muma	48	87	0.55	136	136	1.00
Nypaver	282	425	0.66	445	453	0.98
Rhoads	7	27	0.26	38	38	1.00
Stewart	57	59	0.97	20	20	1.00

Table 2. Summary sensitivity, specificity, and Net Benefit of ear thermometry to diagnose fever.

	t=0.2	t=0.5	t=0.8
Sensitivity	0.65	0.65	0.65
Specificity	0.98	0.98	0.98
NB_t ear	0.30	0.29	0.27
95% CrI	0.19 to 0.42	0.17 to 0.42	0.15 to 0.40
95% PI	0.03 to 0.68	0.02 to 0.66	-0.05 to 0.66
NB_t ear | prev=.5	0.32	0.32	0.29
95% CrI	0.24 to 0.40	0.23 to 0.39	0.21 to 0.37
95% PI	0.12 to 0.47	0.10 to 0.46	0.04 to 0.44

NB_t: Net Benefit; CrI: credible interval; PI: prediction interval; t: risk threshold.

Table 3. Summary sensitivity, specificity, and Net Benefit of the LR2 model to diagnose ovarian malignancies pre-operatively.

	t=0.05	t=0.1	t=0.5
Sensitivity	0.95	0.90	0.63
Specificity	0.68	0.80	0.95
NB_t LR2	0.27	0.25	0.16
95% CrI	0.21 to 0.34	0.20 to 0.31	0.11 to 0.21
95% PI	0.05 to 0.66	0.05 to 0.63	-0.01 to 0.52
NB_t LR2 | prev=0.15	0.13	0.12	0.06
95% CrI	0.12 to 0.13	0.11 to 12	0.04 to 0.07
95% PI	0.12 to 0.14	0.10 to 0.13	-0.01 to 0.08
NB_t LR2 | prev=0.35	0.32	0.30	0.19
95% CrI	0.31 to 0.33	0.29 to 0.31	0.17 to 0.22
95% PI	0.29 to 0.33	0.26 to 0.32	0.11 to 0.25

NB_t: Net Benefit; CrI: credible interval; PI: prediction interval; t: risk threshold

Table 4. Summary Net Benefit of treat all and the probability that ear thermometry will be useful in a new study.

	t=0.2	t=0.5	t=0.8
NB_t treat all	0.33	-0.08	-1.69
95% CrI	0.20 to 0.46	-0.29 to 0.13	-2.21 to -1.16
95% PI	-0.04 to 0.73	-0.65 to 0.57	-3.14 to -0.10
NB_t treat all | prev=0.5	0.38	0	-1.5
P(ear useful)	0.44	0.97	0.95
P(ear useful | prev=0.5)	0.34	1.00	0.99

NB_t: Net Benefit; CrI: credible interval; PI: prediction interval; t: risk threshold.

Table 5. Summary Net Benefit of treat all and the probability that LR2 will be useful in a new center.

	t=0.05	t=0.1	t=0.5
NB_t treat all	0.26	0.22	-0.40
95% CrI	0.20 to 0.34	0.15 to 0.30	-0.53 to -0.26
95% PI	0.02 to 0.69	-0.03 to 0.66	-0.88 to 0.43
NB_t treat all | prev=0.15	0.11	0.06	-0.70
NB_t treat all | prev=0.35	0.32	0.28	-0.30
P(LR2 useful)	0.69	0.78	0.95
P(LR2 useful | prev=0.15)	1.00	1.00	0.93
P(LR2 useful | prev=0.35)	0.75	0.91	1.00

NB_t: Net Benefit; CrI: credible interval; PI: prediction interval; t: risk threshold.

Box 1. Practical recommendations for validation.

Recommendations for the validation of the predictive performance of a test/model:

- Collect data from a broad range of clinical care settings in which the test/model is intended to be used.[1]
- Assess the predictive performance of the test/model in terms of calibration and discrimination,[1] preferably using summary statistics from random-effects models.[14,15]
- Assess the heterogeneity in predictive performance across settings, for example by using a random effects meta-analysis, and by providing 95% prediction intervals and probabilistic statements for the calibration and discrimination in new settings.[14,15]
- Consider model recalibration or updating in settings where calibration is problematic.[29]

Recommendations for the validation in terms of clinical utility:

- Conduct a decision curve analysis and compare (for relevant harm to benefit ratios) the Net Benefit of the test/model to competing strategies (e.g., treat all, treat none),[9] preferably using summary statistics from a random-effects model.
- Assess the heterogeneity in clinical utility, by conducting a random-effects meta-analysis of Net Benefit and calculating 95% prediction intervals and probabilistic statements about the Net Benefit in new settings.
    - P(useful) is close to zero: advise against the test/model. Serious calibration issues are likely.
    - P(useful) is close to one: recommend the test/model.
    - P(useful) is close to 0.5: clinical utility is too variable to make a general recommendation; then
        - Investigate clinical usefulness in specific settings by conditioning on prevalence. Recommend the model settings in where utility is demonstrated.
        - Consider model recalibration or updating in settings where utility is not demonstrated.[29]
        - Caution is advised when the test/model has not been validated in a setting similar to yours. Findings may not generalize to your setting.[29]

Appendix

Appendix Figure 1. Fisher prior for correlations (upper panel) and half-normal prior for between-setting variances (lower panel).

[insert appendix figure 1 here]

Appendix Table 1. Normal priors for η_1, λ_{20}, and λ_{30}, lognormal priors for between-setting variances, and Fisher priors for correlations, for the NB_t analysis of LR2, based on posterior distributions obtained from the analysis of an external dataset.

Parameter	distribution	mean	sd
t=0.05			
η_1	Normal	-1.04	0.23
λ_{20}	Normal	2.99	0.36
λ_{30}	Normal	1.37	1.73
τ_1^2	Lognormal	-0.21	0.38
τ_2^2	Lognormal	-1.08	1.09
τ_3^2	Lognormal	-1.49	0.62
Z (ρ_{12})	Normal	0.07	0.49
Z (ρ_{13})	Normal	-0.63	0.33
Z (ρ_{23})	Normal	-0.41	0.41
t=0.10			
η_1	Normal	-1.04	0.22
λ_{20}	Normal	2.31	0.31
λ_{30}	Normal	2.12	2.26
τ_1^2	Lognormal	-0.28	0.38
τ_2^2	Lognormal	-1.35	1.23
τ_3^2	Lognormal	-1.21	0.68
Z (ρ_{12})	Normal	-0.16	0.42
Z (ρ_{13})	Normal	-0.74	0.33
Z (ρ_{23})	Normal	-0.23	0.38
t=0.50			
η_1	Normal	-1.05	0.24
λ_{20}	Normal	0.87	0.26
λ_{30}	Normal	4.12	0.65
τ_1^2	Lognormal	-0.09	0.39
τ_2^2	Lognormal	-0.99	0.69
τ_3^2	Lognormal	-0.37	0.66
Z (ρ_{12})	Normal	0.50	0.33
Z (ρ_{13})	Normal	-0.15	0.32
Z (ρ_{23})	Normal	-0.55	0.35

Appendix Table 2. Sensitivity to choice of priors for the variance-covariance matrix for the ear thermometry case study.

	t=0.2		t=0.5		t=0.8	
	Inverse Wishart	Weak realistic	Inverse Wishart	Weak realistic	Inverse Wishart	Weak realistic
Sensitivity	0.64	0.65	0.64	0.65	0.64	0.65
Specificity	0.98	0.98	0.98	0.98	0.98	0.98
NB_t ear	0.30	0.30	0.29	0.29	0.25	0.27
95% CrI	0.20 to 0.42	0.19 to 0.42	0.19 to 0.39	0.17 to 0.42	0.15 to 0.36	0.15 to 0.40
95% PI	0.05 to 0.65	0.03 to 0.68	0.04 to 0.63	0.02 to 0.66	0.01 to 0.59	-0.05 to 0.66
NB_t ear | prev=.5	0.32	0.32	0.31	0.32	0.28	0.29
95% CrI	0.25 to 0.38	0.24 to 0.40	0.24 to 0.37	0.23 to 0.39	0.21 to 0.34	0.21 to 0.37
95% PI	0.09 to 0.46	0.12 to 0.47	0.10 to 0.44	0.10 to 0.46	0.06 to 0.40	0.04 to 0.44
NB_t treat all	0.33	0.33	-0.07	-0.08	-1.67	-1.69
95% CrI	0.23 to 0.44	0.20 to 0.46	-0.24 to 0.10	-0.29 to 0.13	-2.09 to -1.25	-2.21 to -1.16
95% PI	0.02 to 0.66	-0.04 to 0.73	-0.55 to 0.44	-0.65 to 0.57	-2.91 to -0.35	-3.14 to -0.10
P(ear useful)	0.31	0.44	1.00	0.97	0.98	0.95
P(ear useful | prev=0.5)	0.27	0.34	1.00	1.0	0.99	0.99

Appendix Table 3. Sensitivity to choice of priors for the variance-covariance matrix for the IOTA case study.

	t=0.05		t=0.10		t=0.50	
	Inverse Wishart	Weak realistic	Inverse Wishart	Weak realistic	Inverse Wishart	Weak realistic
Sensitivity	0.95	0.95	0.90	0.90	0.64	0.63
Specificity	0.66	0.68	0.79	0.80	0.95	0.95
NB_t LR2	0.31	0.27	0.29	0.25	0.18	0.16
95% CrI	0.22 to 0.41	0.21 to 0.34	0.21 to 0.38	0.20 to 0.31	0.11 to 0.26	0.11 to 0.21
95% PI	0.07 to 0.68	0.05 to 0.66	0.06 to 0.67	0.05 to 0.63	0.02 to 0.52	-0.01 to 0.52
NB_t LR2 | prev=.15	0.13	0.13	0.11	0.12	0.05	0.06
95% CrI	0.12 to 0.13	0.12 to 0.13	0.11 to 0.12	0.11 to 12	0.04 to 0.07	0.04 to 0.07
95% PI	0.12 to 0.14	0.12 to 0.14	0.10 to 0.13	0.10 to 0.13	0.003 to 0.08	-0.01 to 0.08
NB_t LR2 | prev=.35	0.32	0.32	0.30	0.30	0.19	0.19
95% CrI	0.31 to 0.33	0.31 to 0.33	0.29 to 0.31	0.29 to 0.31	0.17 to 0.21	0.17 to 0.22
95% PI	0.30 to 0.33	0.29 to 0.33	0.26 to 0.31	0.26 to 0.32	0.11 to 0.24	0.11 to 0.25
NB_t treat all	0.30	0.26	0.27	0.22	-0.32	-0.40
95% CrI	0.21 to 0.41	0.20 to 0.34	0.17 to 0.38	0.15 to 0.30	-0.51 to -0.13	-0.53 to -0.26
95% PI	0.04 to 0.71	0.02 to 0.69	-0.02 to 0.70	-0.03 to 0.66	-0.81 to 0.45	-0.88 to 0.43
P(LR2 useful)	0.64	0.69	0.77	0.78	0.98	0.95
P(LR2 useful | prev=0.15)	1.00	1.00	1.00	1.00	0.98	0.97
P(LR2 useful | prev=0.35)	0.78	0.75	0.93	0.91	1.00	1.00