SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2018, Vol. 27(6) 1723–1736
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280216668555
journals.sagepub.com/home/smm
SAGE

Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study

L Wynants,[1,2] Y Vergouwe,[3] S Van Huffel,[1,2] D Timmerman[4] and B Van Calster[3,4]

Abstract

Clinical risk prediction models are increasingly being developed and validated on multicenter datasets. In this article, we present a comprehensive framework for the evaluation of the predictive performance of prediction models at the center level and the population level, considering population-averaged predictions, center-specific predictions, and predictions assuming an average random center effect. We demonstrated in a simulation study that calibration slopes do not only deviate from one because of over- or underfitting of patterns in the development dataset, but also as a result of the choice of the model (standard versus mixed effects logistic regression), the type of predictions (marginal versus conditional versus assuming an average random effect), and the level of model validation (center versus population). In particular, when data is heavily clustered (ICC 20%), center-specific predictions offer the best predictive performance at the population level and the center level. We recommend that models should reflect the data structure, while the level of model validation should reflect the research question.

Keywords

Mixed model, logistic regression, clinical prediction model, calibration, discrimination, predictive performance, bias

1 Introduction

Clinical risk prediction models estimate the probability that an individual experiences a certain event (diagnostic model), or will experience it in the future (prognostic model).[1,2] They can be used as tools for clinical decision support in the context of evidence-based medicine, and to discuss risks and treatment options with patients. Risk models are often built using regression techniques, such as logistic regression (for diagnosis) and Cox regression (for prognosis).

Increasingly, multicenter data are collected to construct or validate risk prediction models. The main advantages of collecting data at multiple sites are the increased generalizability of the results and reduced recruitment times.[3] Despite these advantages, the clustered nature of multicenter data poses additional methodological challenges.[4] Since patients from one center may be more similar than patients from different centers, patients can no longer be assumed to be independent. Mixed effects models (also known as hierarchical or multilevel models) can be used to analyze the clustered data properly.[4] In the context of prediction, a mixed effects model with center-specific intercepts (random intercept model) and possibly also center-specific slopes (random slope model), has the additional advantage of yielding conditional predictions, tailored to the center a patient belongs to.[5–7]

[1]KU Leuven Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Leuven, Belgium
[2]KU Leuven iMinds Department Medical Information Technologies, Leuven, Belgium
[3]Center for Medical Decision Sciences, Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands
[4]KU Leuven Department of Development and Regeneration, Leuven, Belgium

Corresponding author:
B Van Calster, KU Leuven Department of Development and Regeneration, Herestraat 49 Box 805, Leuven 3000, Belgium.
Email: ben.vancalster@med.kuleuven.be

An important aspect of a clinical prediction model is its performance in new individuals. The literature available to date has not yet provided evidence that a mixed effects model's predictive performance is superior to a standard regression model's, nor is it clear how exactly predictions for individuals from new centers should be obtained from mixed effect models. In previous research comparing a standard logistic regression model and a mixed effects logistic regression model, the random intercept was substituted with zero in order to make predictions for new centers, which were not included in the dataset used for model development.[5] In this way, predictions for new individuals assume an average random center effect. The mixed effects model produced miscalibrated results at the population level, that is, the predicted probabilities of experiencing the event did not reflect the observed probabilities. Calibration slopes deviated from one, and the miscalibration was worse when the degree of clustering (the intraclass correlation (ICC)) increased. Pavlou et al.[8] recently pointed out that calibrated results can be obtained with the mixed effects logistic regression model if marginal predictions are used. These are obtained by integrating over the estimated random effects distribution, rather than substituting the random intercept by zero.[7] However, Pavlou focused solely on the predictive performance of the model at the population level, while others have distinguished between performance at the population level and at the center level, and stressed the relevance of the latter.[9,10]

In this article, we investigate whether a mixed effects logistic regression model has a better predictive performance in terms of calibration and discrimination than a standard logistic regression model in clustered data. In the first section, we present a generalized framework for performance evaluation at the population level and the center level, which incorporates marginal predictions, predictions assuming an average random effect, and conditional predictions with a known center effect. In the second section, we review what is known about the difference between marginal and conditional regression coefficients, and deduce what this implies for model calibration. In the third section, we present a simulation study, in which we investigate the performance of mixed effect logistic regression models and standard logistic regression models within the framework proposed in the first section. In the fourth section, we present an example on the prediction of the risk of tumor malignancy, using clinical data from the International Ovarian Tumor Analysis Group.[11] Finally, we discuss the implications of our findings and formulate recommendations for practice with respect to the development and validation of clinical risk prediction models in clustered data using logistic regression analysis.

2 A framework of performance evaluation of prediction models in clustered data

In this section, we first review how to obtain predictions from the standard logistic regression model and the mixed effects logistic regression model. Then, we review population-level and center-level measures of predictive performance. Finally, we present a framework of the different options to evaluate predictive performance in multicenter data.

Logistic regression is a common technique for estimating risk prediction models for diagnosis. Let Y_{ij} be the event indicator for individual i ($i = 1, \ldots, n_j$) from center j ($j = 1, \ldots J$) with a value of 1 for an event and a value of 0 for a nonevent, X_{kij} the kth predictor ($k = 1, \ldots, K$), and $p_{ij} = P(Y_{ij} = 1)$ the probability that the individual experiences the event of interest. The logistic regression model expresses p_{ij} as a linear combination of predictors X_{kij}, using the logit as a link function:

$$\begin{aligned} &Y_{ij} \sim \text{bin}(1, p_{ij}) \\ &\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_m + \sum_{k=1}^{K} \beta_{k\,m} X_{kij} \end{aligned} \tag{1}$$

The intercept α_m and regression coefficients $\beta_{k\,m}$ are estimated using maximum likelihood. The standard logistic regression model is fitted on patients from different centers without taking clustering into account. It is a population-averaged or marginal model: its regression coefficients $\beta_{k\ m}$ represent the average effects in the population, and the predicted probability for an individual patient reflects the average probability of patients with the same observed values of predictors, ignoring the centers the patients came from. The predicted probability of an event is computed by taking the inverse logit of the linear predictor (LP) of the estimated model

$$LP_{LR\,ij} = \hat{\alpha}_m + \sum_{k=1}^{K} \hat{\beta}_{k\,m} X_{kij} \tag{2}$$

$$\hat{p}_{\text{LR ij}} = \frac{1}{1+\exp(-\text{LP}_{\text{LR ij}})} \tag{3}$$

In clustered data, a mixed effects logistic regression model can be used for model development.[4–6] The simplest version is a random intercept model, which models heterogeneity of the event rate across centers by allowing the intercepts to vary. In this case, one extra parameter needs to be estimated, alongside the beta coefficients of fixed effect predictors and the overall intercept: the random intercept variance τ^2. The random center intercepts a_j are assumed to be normally distributed with mean zero.

$$\begin{gathered} Y_{ij} \sim \text{bin}(1,\ p_{ij}) \\ \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_c + a_j + \sum_{k=1}^{K} \beta_{k\ c} X_{kij} \\ a_j \sim N(0, \tau^2) \end{gathered} \tag{4}$$

The mixed effects model is a center-specific model and the regression coefficients $\beta_{k\ c}$ reflect the predictor effects within a center. The conditional LP given the random intercept for the jth center is

$$\text{LP}_{\text{MLR c ij}} = \hat{\alpha}_c + \hat{a}_j + \sum_{k=1}^{K} \hat{\beta}_{k\ c} X_{kij} \tag{5}$$

The $\hat{a}_j$ are typically estimated using empirical Bayes estimation, which shrinks them to zero. The degree of shrinkage is higher if less center-level information is available (e.g., stronger shrinkage for small centers) or if the between-center variance τ^2 is lower (i.e., uniform shrinkage for all centers in homogeneous populations).[4] Conditional predicted probabilities $\hat{p}_{\text{MLR c ij}}$ are obtained by taking the inverse logit of the conditional LP.

To obtain a prediction for a patient of a center not included in the development set, one can replace the random intercept by the average random intercept ($\hat{a}_j$ =0).

$$\text{LP}_{\text{MLR a ij}} = \hat{\alpha}_c + 0 + \sum_{k=1}^{K} \hat{\beta}_{k\ c} X_{kij} \tag{6}$$

Predicted probabilities assuming an average random center intercept (0), $\hat{p}_{\text{MLR a ij}}$, are obtained by taking the inverse logit of the LP. This will yield the prediction for an individual from a center with an average intercept. Due to the nonlinearity of the logit transformation, this does not correspond to the average but to the median probability of patients with the same observed values of predictors across centers.

Although the mixed effects logistic regression model is a center-specific model, one can obtain marginal predictions by integrating over the distribution of the random effects:

$$\begin{aligned} \hat{p}_{\text{MLR m ij}} &= \int_{-\infty}^{\infty} \frac{1}{1+\exp(-\text{LP}_{\text{MLR cond ij}})} f(\hat{a}_j)d\hat{a}_j \\ \hat{p}_{\text{MLR m ij}} &= \int_{-\infty}^{\infty} \frac{1}{1+\exp[-(\hat{\alpha}_c + \hat{a}_j + \sum_{k=1}^{K} \hat{\beta}_{k\ c} X_{kij})]} f(\hat{a}_j)d\hat{a}_j \end{aligned} \tag{7}$$

where $f(\hat{a}_j)$ is the density function of a normal distribution with mean zero and variance $\hat{\tau}^2$. The integral often cannot be solved analytically and must be evaluated by numerical averaging after sampling a large number of random effects from their fitted distribution. The marginalized LP of the mixed effects model $\text{LP}_{\text{MLR m ij}}$ can be obtained by performing a logit transformation on the marginal predicted probabilities.[8] These predictions are very similar to the marginal predictions obtained by the standard logistic regression model, as shown in online Appendix 1. In summary, the mixed effects model yields three types of predictions: conditional predictions, predictions for an individual in a center with an average random intercept, and marginal predictions.[7]

The predictive performance of a model is crucial and needs extensive evaluation, preferably using data from new clinical settings. Key aspects of predictive performance are discrimination and calibration, with or without considering the clustered nature of multicenter data. Discrimination refers to the ability of the model to distinguish between events and nonevents. The C-index expresses the probability that for a randomly selected pair of an event and a nonevent, the event has a higher predicted probability.[12] For the computation of the standard C-index, pairs of events and nonevents belonging to different clusters are compared, as well as pairs from the same cluster. It is estimated by

$$\hat{C} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(\hat{p}_{ij} > \hat{p}_{i'j'} \text{ and } y_{ij} = 1 \text{ and } y_{i'j'} = 0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(y_{ij} = 1 \text{ and } y_{i'j'} = 0\right)} \tag{8}$$

In multicenter data, the within-center C-index is computed by only comparing pairs of events and nonevents within the J centers[10]

$$\hat{C}_w = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(\hat{p}_{ij} > \hat{p}_{i'j} \text{ and } y_{ij} = 1 \text{ and } y_{i'j} = 0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(y_{ij} = 1 \text{ and } y_{i'j} = 0\right)} \tag{9}$$

This corresponds to the average center-specific C-index, weighted by the number of pairs of events and nonevents per center. Other weights may be used as well.[9]

Calibration refers to the ability of the model to provide accurate risk estimates for individual patients. This can be checked with logistic calibration.[2,13,14] Consider, for the standard logistic regression model, a LP, obtained by applying formula (2). To perform logistic calibration one fits the following model to a validation dataset:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_{cal} + \beta_{cal} LP_{ij} \tag{10}$$

The estimated calibration slope deviates from one if the predicted probabilities are too extreme (too close to zero or one) ($\hat{\beta}_{cal} < 1$) or not extreme enough ($\hat{\beta}_{cal} > 1$). A calibration slope smaller than one typically indicates overfitting, which often occurs when models are fitted in small datasets.[15–23] We elaborate on the effect of sample size in online Appendix 2. Calibration-in-the-large assesses whether predicted probabilities are correct on average and is checked by including LP_{ij} as an offset in equation (10), instead of estimating its effect.[2,13,14] The calibration intercept deviates from zero if the predicted probabilities are on average overestimated ($\hat{\alpha}_{cal}|(\beta_{cal} = 1) < 0$) or underestimated ($\hat{\alpha}_{cal}|(\beta_{cal} = 1) > 0$).

Mixed effects logistic calibration evaluates the predictions conditionally, reflecting differences in model calibration between centers,[5] using

$$\begin{gathered}\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_{cal\,w} + a_{j\,cal} + \beta_{cal\,w} LP_{ij} + b_{j\,cal} LP_{ij} \\ \begin{pmatrix} a_{j\,cal} \\ b_{j\,cal} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_a^2 & \tau_{ab} \\ \tau_{ab} & \tau_b^2 \end{bmatrix}\right)\end{gathered} \tag{11}$$

$\beta_{cal\,w}$ now is the average within-center calibration slope. The random effects $a_{j\,cal}$ and $b_{j\,cal}$ follow a bivariate normal distribution, with τ_b^2 the variance of the within-center calibration slopes $b_{j\,cal}$ and τ_{ab} the covariance between calibration intercepts and calibration slopes. Calibration-in-the-large is assessed by fixing $\beta_{cal\,w}$ to one, τ_b^2 to zero, and estimating $\alpha_{cal\,w}$ and the variance of the random calibration intercepts τ_a^2.

Figure 1 shows the different options to evaluate predictive performance in a comprehensive framework. Prediction models can be developed with standard or mixed effects regression analysis; validation data can be obtained from a single center or from multiple centers. Conditional predictions and predictions assuming an average random intercept are only available for mixed effects models, while marginal predictions can be derived from both types of models. It may seem natural to use conditional (within-center) measures only for conditional predictions and standard (population level) performance measures for marginal predictions. However, the choice of performance measure in multicenter validation data should depend on the use of the prediction model and the research question. The conditional performance measures should be used to assess the performance within centers. Consider a model predicting the risk that an ovarian mass in a patient is malignant.[24] The treatment decision is made in the center the patient is treated in, requiring adequate conditional performance of the prediction model. Conditional performance measures are not useful when the validation dataset contains data from a single center. For this reason, we will not focus on that situation the remainder of this work, although Figure 1 includes this option for completeness. When a model is validated in a single center, the validation results may not be generalizable to other centers. When multicenter data is available, standard performance measures will quantify how well the model

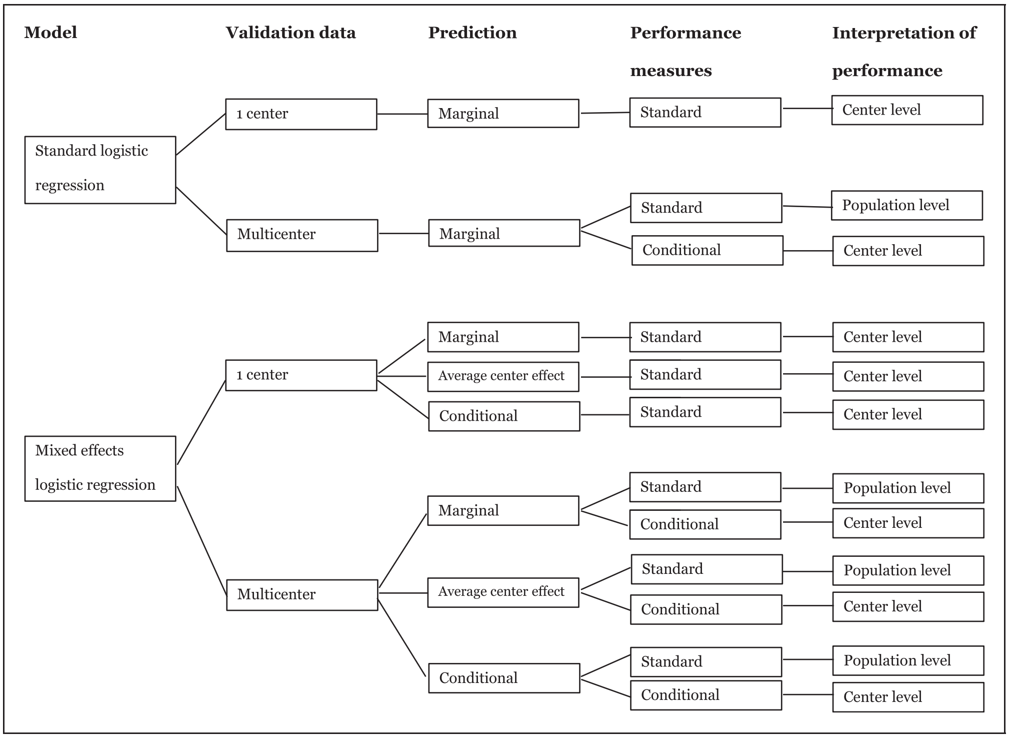

Figure 1. A comprehensive framework of options for model validation, subject to the type of prediction model that is being evaluated (standard or mixed effects logistic regression) and the available validation dataset (one center or multicenter).

performs in the entire population of individuals, as an overall measure of performance. This is useful, for example, for the recommendation of a prediction model in national guidelines. Online Appendix 3 presents an overview of the formulas for the C-index and logistic calibration in this comprehensive framework.

3 Calibration slopes for marginal and center-specific logistic regression models

Marginal effect estimates (denoted by subscript m) are typically closer to zero than conditional effect estimates (denoted by subscript c).[25–27] Using a cumulative Gaussian approximation to the logistic function leads to the following approximation[25]

$$\begin{aligned} \alpha_m &\approx \alpha_c/f \\ \beta_m &\approx \beta_c/f, \\ \text{with } f &= \sqrt{1+\tau^2 c^2} \\ \text{and } c &= \frac{16\sqrt{3}}{15\pi} \end{aligned} \tag{12}$$

This implies that, when a standard logistic regression model has an overall calibration slope β_{cal}, the overall calibration slope of the corresponding mixed effects model using the average random intercept could be approximated by β_{cal}/f

$$\begin{aligned} \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \alpha_{cal} + \beta_{cal} LP_{LR\ ij} \\ &= \alpha_{cal} + \beta_{cal}(\hat{\alpha}_m + \sum_{k=1}^{1} \hat{\beta}_{k\ m} X_{ki}) \\ &\approx \alpha_{cal} + \frac{\beta_{cal}}{f}(\hat{\alpha}_c + \sum_{k=1}^{1} \hat{\beta}_{k\ c} X_{ki}) \end{aligned} \tag{13}$$

Likewise, when a mixed effects model has within-center calibration slope $\beta_{\mathrm{cal\ w}}$ assuming an average random center effect ($a_{\mathrm{j\ cal}} = b_{\mathrm{j\ cal}} = a_j \ = b_j \ = 0$), the calibration slope of the corresponding standard model would be approximated by $\beta_{\mathrm{cal\ w}} \times f$

$$\begin{aligned}\log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \alpha_{\mathrm{cal\ w}} + \ a_{\mathrm{j\ cal}} + \beta_{\mathrm{cal\ w}}\mathrm{LP}_{\mathrm{MLR\ a\ ij}} + b_{\mathrm{j\ cal}}\mathrm{LP}_{\mathrm{MLR\ a\ ij}} \\ &= \alpha_{\mathrm{cal\ w}} + \beta_{\mathrm{cal\ w}}(\hat{\alpha}_c + \sum_{k=1}^{1}\hat{\beta}_{\mathrm{k\ c}}X_{kij}) \\ &\approx \alpha_{\mathrm{cal}} + \beta_{\mathrm{cal\ w}}f\,(\hat{\alpha}_m + \sum_{k=1}^{1}\hat{\beta}_{\mathrm{k\ m}}\,X_{kij})\end{aligned} \tag{14}$$

This demonstrates how the calibration slope may deviate from one due to the choice of modeling technique. For example, if a prediction model was fitted using mixed effects logistic regression, and this model was perfectly calibrated in a center with an average random effect, the corresponding standard model would have a within-center calibration slope larger than one in a center within an average random effect.

In practice, the random effect variance τ^2 will often be estimated with error. As shown in online Appendix 4, overestimation will decrease the estimated calibration slope, while underestimation has the opposite effect. Fitting a standard model can be seen as an extreme case of the latter, setting the estimated between-center variance to zero.

4 Simulation study

4.1 Design

In this simulation study, we compare the performance of mixed effect and standard logistic regression models in multicenter validation data. We first created source populations, from which samples with different sizes were drawn. We fitted a random intercept model and a standard logistic regression model in each sample and tested them in the remaining part of the source population, within the framework for performance evaluation presented in the first section.

We generated two source populations of approximately 20,000 patients: one population with heavily clustered data (ICC = 20%), and one with little clustering (ICC = 5%).[28] We fixed the number of centers (J) at 20.[29] The number of patients per center (n_j) was drawn from a Poisson distribution with a separate, randomly generated lambda for each center. This yielded center sizes ranging from approximately 600 to 2000.

We generated the data for the source populations according to a predefined true random intercept model. Each center was assigned a random center intercept a_j, generated from a normal distribution of which the variance was determined by the desired ICC. The true model included four normally distributed continuous predictors and four dichotomous predictors, each with a beta coefficient of 0.8. X_1 through X_4 were continuous with mean 0 and standard deviations 1, 0.6, 0.4, and 0.2, respectively. X_5 through X_8 were dummy variables with prevalence 0.2, 0.3, 0.3, and 0.4, respectively. We set the overall intercept α equal to -2.1 to obtain an event rate of the outcome Y_{ij} of 0.30. For each patient, we computed the probability of an event (p_{ij}) from the generated predictors and random intercepts, using equation (5) and applying the inverse logit transformation. We generated Y_{ij} by comparing p_{ij} to a randomly drawn value from a uniform distribution

$$\begin{aligned} &Z_{ij} \sim \mathrm{unif}(0,1) \\ &Y_{ij} = \begin{cases} 1 & \text{if} \quad z_{ij} \leq p_{ij} \\ 0 & \text{if}\ z_{ij} > p_{ij} \end{cases}\end{aligned} \tag{15}$$

We drew samples from the source population with either 100 (for ICC = 5% and ICC = 20%) or 5 (only for ICC = 20%) events per variable (EPV). The number of events to be sampled was calculated by multiplying the preset EPV value by nine (eight parameters for the regression coefficients plus one extra parameter for the random intercept variance). The required number of nonevents to be sampled was computed such that the event rate in the source population (0.3) was preserved. We sampled patients without replacement from all centers, without stratification for center. Each simulation was based on 1000 samples.

We built a random intercept logistic regression model and a standard logistic regression model containing all eight predictors in each sample. We used the following convergence criteria for the mixed effects model: a change of less than 10^{-5} in deviances of the models fitted in the last two iterations, 10–100 iterations to fit the model, and no outlying estimated regression coefficients and standard errors (visual inspection). For the standard model, we used a positive convergence tolerance of 10^{-9}, a maximum of 50 iterations and a visual check of estimated regression coefficients and standard errors as convergence criteria. Samples with nonconverging models were removed from the analysis.

We tested the models in the part of the source population that was not used for model development. Hence, the development set and the validation set are from the same population. We used two versions of the LP for the random intercept model: the conditional LP, including the center-specific intercept estimates (equation (5)), and the LP assuming an average random intercept (equation (6)). The marginal LP was obtained from the standard logistic regression model (equation (2)). We computed the standard (equation (10)) and within-center (equation (11)) calibration slopes and intercepts, and the standard (equation (8)) and within-center C-index (equation (9)) for all predictions.

All simulations and calculations were performed in R version 2.14.0 (Vienna, Austria).[30] The lmer function from the lme4 package[31] was used to fit mixed effect logistic regression models using Laplace approximation, and the rms package was used for model evaluation.[12] The R code is provided in online Appendix 5.

5 Results

5.1 Calibration

5.1.1 Severe clustering (ICC 20%, 100 EPV)

The conditional predictions from the random intercept model were well calibrated at the center level and the population level (Figure 2, squares; estimates are tabulated in online Appendix 6). The average calibration slopes close to one indicate that there was hardly any overfitting. The predictions from the random intercept model assuming average random intercepts were only calibrated at the center level (Figure 2(a), triangles), while the predictions from the standard model were only calibrated at the population level (Figure 2(b), circles). The center-level calibration slopes tended to be larger than one for the predictions of the standard model (Figure 2(a), circles) and the population-level calibration slopes were smaller than one for the predictions assuming an average random intercept (Figure 2(b), triangles).

The association between the estimated random intercept variance and the within-center calibration slope is slightly negative and close to the theoretical approximation, as shown in online Appendix 4. Note that the within-center calibration slopes plotted in Figure 2 reflect the calibration slopes in the center with an average calibration slope, while the estimated $b_{\text{cal j}}$ (not shown) reflect center-specific differences from this slope. The average estimated variance of the $b_{\text{cal j}}$ was <0.0005 for the three types of predictions.

Calibration-in-the-large was also satisfactory for conditional predictions at the population and the center level, while the predictions assuming average random intercepts were only calibrated at the center level and the predictions from the standard model were only calibrated at the population level (online Appendices 6 and 7, Figure A6). The average estimated variance of the center-specific calibration intercepts was 0.83 for the predictions assuming an average random intercept and 0.89 for the predictions from the standard model. The conditional predictions yielded a much lower average estimated variance of center-specific calibration intercepts (0.05), indicating that most of the between-center differences in the event rates were accounted for by using random intercepts in the prediction.

The results from the simulation with severe clustering and small samples (EPV 5) are presented in online Appendix 2.

5.1.2 Mild clustering (ICC 5%, 100 EPV)

The results are similar to the results of the simulation with severe clustering, although differences in calibration between the three types of predictions are smaller due to the lower between-center variance (Figure 3 and online Appendix 7, Figure A7). The predictions from the standard model yielded within-center calibration slopes slightly above one (Figure 3(a), circles), while the predictions assuming an average random intercept yielded population-level calibration slopes slightly below one (Figure 3(b), triangles).

5.2 Discrimination

The empirical Bayes estimates are constant within each center and therefore do not influence the estimated within-center C-indexes. Hence, the obtained within-center C-indexes of the conditional predictions and the predictions

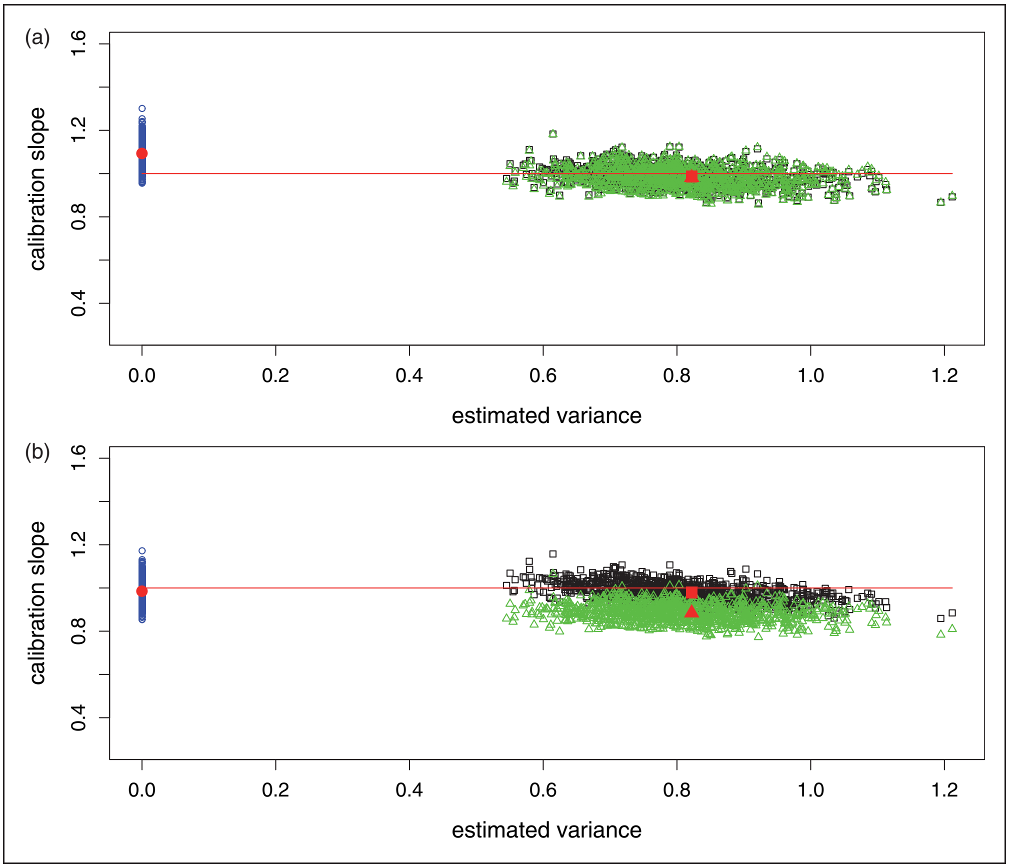

Figure 2. Center-level (panel a) and population-level (panel b) calibration slopes of the standard logistic regression model (circles), the conditional LP of the random intercept model (squares) and the LP of the random intercept model assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 EPV and true random effects variance = 0.822 (ICC = 20%). Small symbols indicate calibration slopes in the samples, large filled symbols indicate average calibration slopes at estimated variance = 0 for the standard logistic regression model and at the correctly estimated variance (0.822) for the random intercept model. The horizontal line represents the ideal calibration slope.

for an individual from an average center are by definition the same, and they are very similar to the within-center C-index of the standard model (Figure 4(a)).

The population-level C-indexes for the predictions assuming an average random intercept (Figure 4(b)), were very similar to the population-level C-indexes for the predictions from the standard model. Higher population-level C-indexes were obtained with the conditional predictions. This effect was even present in datasets with a low ICC (online Appendix 8, Figure A8B, squares).

The results from the simulation with small samples (EPV 5) and strong clustering (ICC = 20%) are shown in online Appendix 2.

6 Empirical example

To illustrate our findings on real data, we developed and evaluated models to pre-operatively diagnose ovarian cancer. The development dataset consisted of 3506 women with ovarian masses (949, 27% with malignancies), collected by the International Ovarian Tumor Analysis (IOTA) consortium between 1997 and 2007 in 21 international centers. We used six clinical and ultrasound predictors: age, the proportion of solid tissue, the presence of more than 10 locules, the number of papillary structures (0, 1, 2, 3, >3, linear effect), the presence of acoustic shadows, and the presence of ascites. This yielded an EPV of 136 for the random intercept model. The ICC was 15% ($\hat{\tau}^2 = 0.59$), accounting for the predictors. The regression coefficients of the standard model tended to be closer to zero than those of the mixed effects model, apart from the coefficient of acoustic shadows (online Appendix 9, Table A3). Standard errors were larger in the mixed effects model.

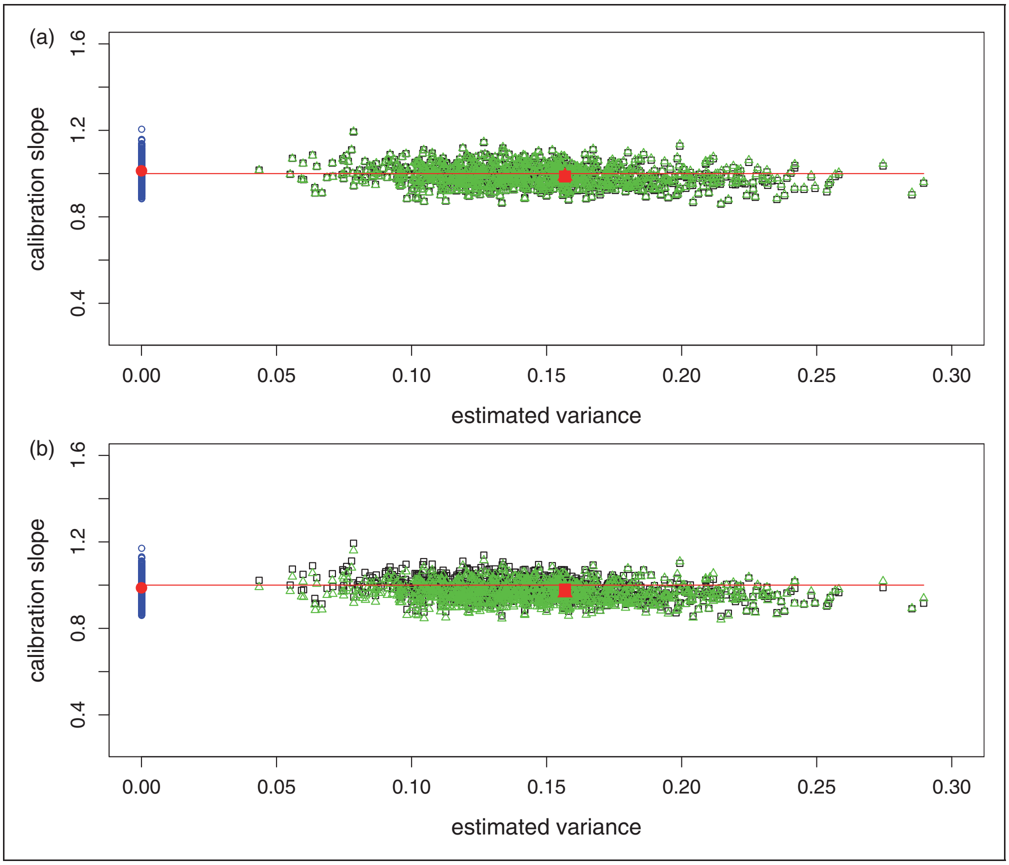

Figure 3. Center-level (panel a) and population-level (panel b) calibration slopes of the standard logistic regression model (circles), the conditional LP of the random intercept model (squares) and the LP assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 EPV and true random effects variance = 0.157 (ICC = 5%). Small symbols indicate calibration slopes in the samples, large filled symbols indicate average calibration slopes at estimated variance = 0 for the standard logistic regression model and at the correctly estimated variance (0.157) for the random intercept model. The horizontal line represents the ideal calibration slope.

All predictions (marginal, with average random intercept and conditional) were validated using conditional and standard performance measures (Figure 1), in a dataset of 2224 women (915 (41%) with malignancies), collected between 2009 and 2012 in 15 of the 21 centers of the development set. The ICC was 14% ($\hat{\tau}^2 = 0.53$) after accounting for the LP of the mixed effects model assuming an average random intercept.

The calibration slope at the population level was close to one for the marginalized predictions from the random effects model (0.99), and slightly lower for the marginal predictions from the standard model (0.95). As expected, the calibration slope was lower for the predictions assuming an average random intercept (0.91). Surprisingly, the calibration slope of the conditional predictions was also lower (0.88). It is likely that this is due to differences in the true random center intercepts between the development and validation datasets.

The within-center calibration slopes for the predictions assuming an average random intercept and for the conditional predictions were slightly below 1 (0.94 and 0.93) (online Appendix 9, Table A4). The within-cluster calibration slope for the standard model was higher (0.97) than the within-center calibration slope for predictions assuming an average random center intercept, which is typical. The within-center calibration slope for the marginalized predictions from the mixed effect model was 1.02. The random variance of the center-specific calibration slopes was nearly half as large for the conditional predictions, as for all other predictions. This indicates that the center-specific calibration was more stable when conditional predictions were used.

The population-level calibration intercept was 0.29 for the conditional predictions, 0.62 for the marginal predictions from the standard model, 0.60 for the marginalized predictions from the mixed effects model, and 0.70 for the predictions assuming an average random intercept. This is explained by the changed event rates

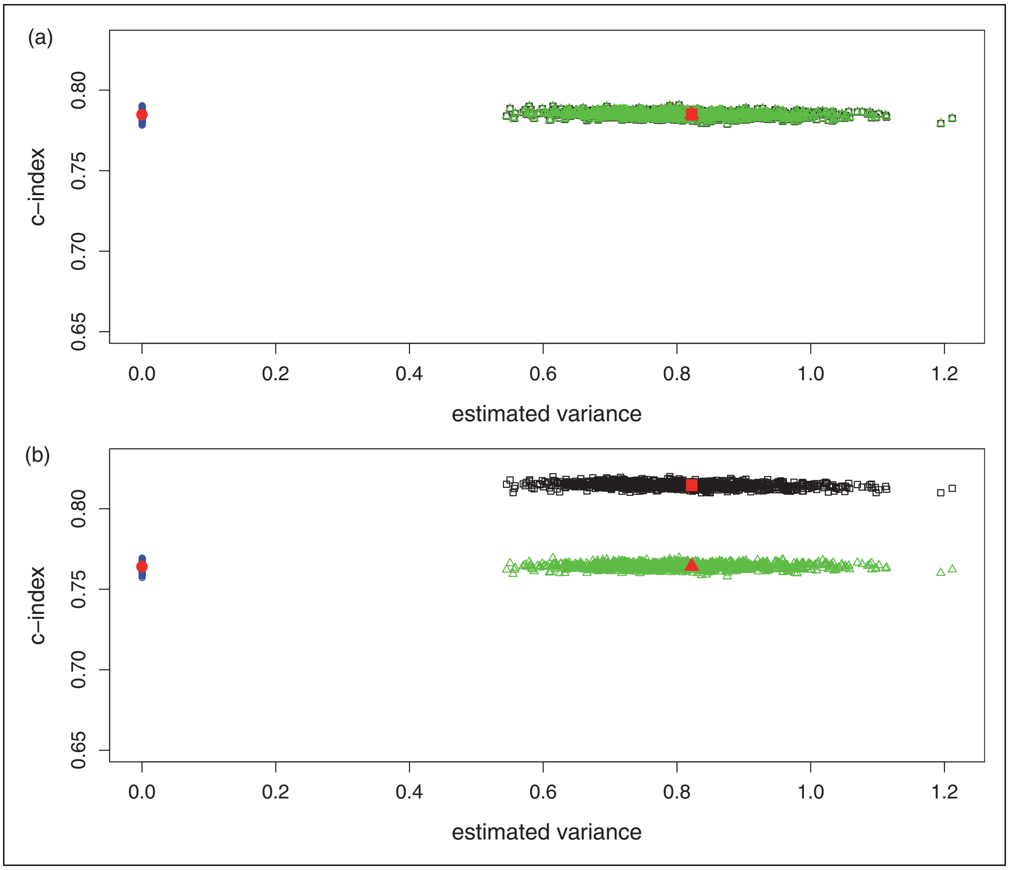

Figure 4. Center-level (panel a) and population-level (panel b) C-indexes of the standard logistic regression model (circles), the conditional LP of the random intercept model (squares) and the LP assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 EPV and true random effects variance = 0.822 (ICC = 20%). Small symbols indicate C-indexes in the samples, large filled symbols indicate average C-indexes at estimated variance = 0 for the standard logistic regression model and at the correctly estimated variance (0.822) for the random intercept model.

within the centers: in 12 out of the 15 centers in the validation dataset, the event rate was higher than in the development set.

The within-center calibration intercept was 0.27 for the conditional predictions, 0.41 for the marginal predictions from the standard model, 0.39 for the marginalized predictions from the mixed effects model, and 0.48 for the predictions assuming an average random intercept. The conditional predictions yielded the within-center calibration intercept closest to zero and the estimated variances of the center-specific calibration intercepts were half as large for conditional predictions as for all other types of predictions.

At the center level, all predictions yielded a very similar C-index (0.88). At the population level, the discrimination of the conditional predictions was superior (0.91) to the other predictions (0.90). The discrepancy might have been higher, if the random center intercepts in the validation data were more like the ones in the development data.

7 Discussion

We investigated whether ignoring clustering in multicenter data influences the predictive performance of a risk prediction model, comparing standard to mixed effects logistic regression. Our results have shown that it does, but the consequences of ignoring clustering are dependent on the level at which the model is evaluated (the population or the center level), and on the aspect of predictive performance that is evaluated (calibration or discrimination) (Table 1).

Predictions from mixed effects models assuming an average random intercept are poorly calibrated at the population level, while marginal predictions are poorly calibrated at the center level. We showed that this is a

Table 1. Schematic overview of the effect of the type of prediction on the conditional and standard performance measures, in the absence of overfitting and assuming a representative development dataset.

		Marginal predictions	Predictions assuming an average random center intercept	Conditional predictions
Calibration	Conditional (center level)	Calibration slope > 1 Calibration intercept ≠ 0	Well calibrated	Well calibrated
	Standard (population level)	Well calibrated	Calibration slope < 1 Calibration intercept ≠ 0	Well calibrated
Discrimination	Conditional (center level)	Good discrimination	Good discrimination	Good discrimination
	Standard (population level)	Good discrimination	Good discrimination	Superior discrimination

consequence of the much-described finding that marginal regression coefficients are typically closer to zero than conditional regression coefficients.[25–27] The consequence, from a calibration perspective, is that predicted probabilities from a standard logistic regression model are too close to the event rate in the population to reflect the event rates within centers.[2,13] For instance, within an average center, more than 80% of patients with a predicted risk of 0.8 will experience the event, while of all patients in that center with a predicted risk of 0.2, less than 20% will experience the event. In contrast, conditional predictions from the mixed effects model (that include center-specific effects) were well calibrated at both the population level and the center level (Table 1). This is in line with earlier research showing that conditional predictions from mixed effect models yield better calibration-in-the-large at the center level.[5] Hence, we advise to use a mixed effect model to obtain better within-center calibration.

Nonetheless, we must note that the degree of clustering in typical outcomes of prediction models is generally small. Our simulations in a source population with weak clustering (ICC = 5%) have shown that the calibration results of the standard logistic regression model and the mixed effects logistic regression model are very similar.

We showed that the calibration of mixed effects models depends on the estimation of the between-center variance in heavily clustered data. Research has shown that a large number of clusters is needed to obtain good estimates of the between-cluster variance.[32–34] One suggested guideline is to collect data from at least 50 clusters,[34] although this may be hard to obtain in practice. A sufficiently large number of EPV also contributes to a good estimation of the between-cluster variance.[16] When data from very few centers (e.g., five) is available, it would be preferable to use a fixed effects regression model, containing dummy variables for centers.[29]

Additional simulations in small samples (EPV = 5) showed that overfitting yields poorly calibrated results, both for standard and mixed effects logistic regression models. Calibration was poorer for the mixed effects model, because the problem of overfitting in small datasets was worsened by the fact that conditional regression coefficients are generally more extreme than marginal regression coefficients. Although the standard model was seemingly better calibrated, ignoring clustering is not an adequate solution for problems caused by small sample sizes.

Discrimination at the population level was better for the conditional predictions obtained by mixed effects logistic regression than for the other predictions (Table 1). This was even observable when the degree of clustering was low. Center-specific intercept estimates contain additional information when comparing predicted probabilities for patients from different centers, enhancing discrimination.

Our study has the following limitations. We only considered mixed effects logistic regression to account for clustering. Other methods are available, such as fixed effects logistic regression with dummy variables for centers. Like the mixed effects logistic regression model, it offers center-specific predictions and the regression coefficients have a conditional interpretation.[35] Hence, it may perform similarly to the mixed effects regression model in terms of discrimination and calibration. The optimal choice most likely depends on the number of centers, with mixed effects models being more appropriate if the number of clusters is large.[4,29,32–36] Further, we assumed that the assumptions underlying the regression models hold. For example, we assumed that the random intercepts were normally distributed. This may not always be the case in practice, but evidence to date[37–39] suggests that random effects models are quite robust against violations of this assumption. Random slopes were beyond the scope of this research. More research on how random slopes can be included in the development and external validation of prediction models is needed.

Based on our findings, we advise researchers to use a modeling technique that reflects the structure of the data, and to collect sufficiently large datasets to avoid overfitting. The need for center-specific models may be alleviated if we manage to include patient or center characteristics that explain the differences between centers in the prediction model.

To make predictions for new individuals we suggest to use conditional predictions. Center-specific random intercepts are required for conditional predictions, but they are not available for new centers. In the absence of data from the new center, numerical integration of the predictions over the estimated random effects distribution may be used. However, these marginal predicted probabilities will not be well calibrated at the center level. Another option is to substitute the center-specific random effect by zero. These predictions are easy to obtain, and will be well calibrated in centers with an average effect. Alternative options are to estimate the center-specific intercept from the outcome prevalence of the new center, or to use the intercept of a similar center from the model development set.[6,40]

Whether an investigator should evaluate the prediction model at the population level or the center level, depends on the situation. If the goal is to implement a prediction model nationwide, for example, based on national guidelines, the discrimination at the national level can be considered. When risk models are used to support decision-making within centers, they should perform well at the center level. Sometimes, the prediction model is used for decision support at a higher level than the cluster level. For example, a recently developed prediction model to screen for *Chlamydia trachomatis* infection was developed in a dataset that was clustered within neighborhoods.[41] Given that the intended user of the model screens individuals from various neighborhoods, population-level performance measures were of interest.

Besides investigating performance in the average center, it is also useful to investigate potential heterogeneity in model performance across centers, for example, by studying the variance of random effects, prediction intervals of performance statistics, or empirical Bayes estimates of center-specific calibration intercepts and slopes.[5,9,42] If centers are large, the center-specific performance can be studied. An example on preoperative tumor diagnosis was recently published by the IOTA consortium.[24]

When interpreting validation results, it should be kept in mind that differences in model performance can be the result of many causes, including case-mix differences in the center populations.[43] Obtaining good discrimination and calibration in every single center may be difficult. If a model does not perform well in specific centers, the local performance can be improved by using conditional predictions from a mixed effects logistic regression model, or by applying model updating techniques.[2,5,6,44–46]

Clustering in multicenter data should be accounted for when developing prediction models. At the same time, the level at which prediction models are used, determines how the performance should be assessed. It is important to understand that the choice of the model (standard or mixed effects logistic regression), the predictions used (marginal, assuming an average random intercept or conditional), and the level of performance evaluation (population or center level) will have an impact on the estimated predictive performance. We recommend the use of conditional predictions, when available, given their good performance at both the population and the center level.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: LW is a doctoral fellow of the Flanders' Agency for Innovation by Science and Technology (IWT). DT is a senior clinical investigator of the Research Foundation Flanders (FWO). YV is supported by the Netherlands Organization for Scientific Research (Grant 917.11.383). The research was further supported the Research Foundation Flanders (FWO) (Grant G0B4716N), the KU Leuven Research Fund (grant C24/15/037), Bijzonder Onderzoeksfonds KU Leuven (BOF) Center of Excellence (CoE) (#: PFV/10/002, OPTEC), the Belgian Federal Science Policy Office (IUAP #P7/19/,DYSCO, 'Dynamical systems, control and optimization', 2012–2017), and the European Research Council (ERC Advanced Grant, #339804 BIOTENSORS). This article reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

References

1. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; **162**: W1–W73.
2. Steyerberg EW. *Clinical prediction models: A practical approach to development, validation, and updating*. New York, NY: Springer US, 2009.
3. Sprague S, Matta JM, Bhandari M, et al. Multicenter collaboration in observational research: improving generalizability and efficiency. *J Bone Joint Surg Am* 2009; **91**(Suppl 3): 80–86.
4. Snijders TAB and Bosker RJ. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*, 2nd ed. London: Sage, 2012.
5. Bouwmeester W, Twisk J, Kappen T, et al. Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Med Res Methodol* 2013; **13**.
6. Debray TPA, Moons KGM, Ahmed I, et al. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013; **32**: 3158–3180.
7. Skrondal A and Rabe-Hesketh S. Prediction in multilevel generalized linear models. *J R Stat Soc Ser A* 2009; **172**: 659–687.
8. Pavlou M, Ambler G, Seaman S, et al. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. *BMC Med Res Methodol* 2015; **15**: 59.
9. van Klaveren D, Steyerberg E, Perel P, et al. Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol* 2014; **14**.
10. Van Oirbeek R and Lesaffre E. Assessing the predictive ability of a multilevel binary regression model. *Comput Stat Data Anal* 2012; **56**: 1966–1980.
11. Kaijser J, Bourne T, Valentin L, et al. Improving strategies for diagnosing ovarian cancer: a summary of the International Ovarian Tumor Analysis (IOTA) studies. *Ultrasound Obstet Gynecol* 2013; **41**: 9.
12. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer, 2001.
13. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; **45**: 562–565.
14. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016; **74**: 167–176.
15. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004; **66**: 411–421.
16. Wynants L, Bouwmeester W, Moons KG, et al. A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *J Clin Epidemiol* 2015; **68**: 8.
17. Steyerberg EW, Eijkemans MJ and Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol* 1999; **52**: 935–942.
18. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; **49**: 1373–1379.
19. Vittinghoff E and McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol* 2007; **165**: 710–718.
20. Peduzzi P, Concato J, Feinstein AR, et al. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995; **48**: 1503–1510.
21. Courvoisier DS, Combescure C, Agoritsas T, et al. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *J Clin Epidemiol* 2011; **64**: 993–1000.
22. Harrell FE, Lee KL and Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**: 361–387.
23. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, et al. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 2000; **19**: 1059–1079.
24. 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.
25. Zeger SL, Liang K-Y and Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; **44**: 1049–1060.
26. Neuhaus JM, Kalbfleisch JD and Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int Stat Rev* 1991; **59**: 25–35.
27. Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. *Stat Methods Med Res* 1992; **1**: 249–273.
28. Adams G, Gulliford MC, Ukoumunne OC, et al. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol* 2004; **57**: 785–794.
29. Kahan BC and Harhay MO. Many multicenter trials had few events per center, requiring analysis via random-effects models or GEEs. *J Clin Epidemiol* 2015; **68**: 1504–1511.

30. R Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2011.
31. Bates D, Maechler M and Bolker B. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42. 2011.
32. Maas CJM and Hox JJ. Sufficient sample sizes for multilevel modeling. *Methodology* 2005; **1**: 86–92.
33. Paccagnella O. Sample size and accuracy of estimates in multilevel models. *Methodology* 2011; **7**: 111–120.
34. Moineddin R, Matheson FI and Glazier RH. A simulation study of sample size for multilevel logistic regression models. *BMC Med Res Methodol* 2007; **7**.
35. Molenberghs G and Verbeke G. *Models for discrete longitudinal data*. New York: Springer, 2005.
36. Kahan BC. Accounting for centre-effects in multicentre trials with a binary outcome – when, why, and how? *BMC Med Res Methodol* 2014; **14**: 1–11.
37. Neuhaus JM, McCulloch CE and Boylan R. Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. *Stat Med* 2013; **32**: 2419–2429.
38. Kahan BC and Morris TP. Analysis of multicentre trials with continuous outcomes: when and how should we account for centre effects? *Stat Med* 2013; **32**: 1136–1149.
39. Maas CJ and Hox JJ. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Comput Stat Data Anal* 2004; **46**: 427–440.
40. 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.
41. van Klaveren D, Götz HM, Op de Coul EL, et al. Prediction of *Chlamydia trachomatis* infection to facilitate selective screening on population and individual level: a cross-sectional study of a population-based screening programme. *Sex Transm Infect* 2016; **92**: 433–440.
42. 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**: 2081–2103.
43. Vergouwe Y, Moons KGM and 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**: 971–980.
44. Janssen KJM, Moons KGM, Kalkman CJ, et al. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008; **61**: 76–86.
45. Steyerberg EW, Borsboom GJ, van Houwelingen HC, et al. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004; **23**: 2567–2586.
46. Van Houwelingen HC and Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat Med* 1995; **14**: 1999–2008.