Account of methodological development

Corresponding Author:
Ben Van Calster, KU Leuven Department of Development and Regeneration, Herestraat 49 Box 7003, Leuven 3000, Belgium
Email: ben.vancalster@med.kuleuven.be

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

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

[1] KU Leuven Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, Box 2446, Leuven 3001, Belgium
Email: laure.wynants@esat.kuleuven.be, sabine.vanhuffel@esat.kuleuven.be
[2] KU Leuven iMinds Department Medical Information Technologies, Kasteelpark Arenberg 10, Box 2446, Leuven 3001, Belgium
[3] Center for Medical Decision Sciences, Department of Public Health, Erasmus Medical Center, Wytemaweg 80, 3015 CN Rotterdam, The Netherlands.
Email: y.vergouwe@erasmusmc.nl
[4] KU Leuven Department of Development and Regeneration, Herestraat 49 Box 7003, Leuven 3000, Belgium
Email: ben.vancalster@med.kuleuven.be, dirk.timmerman@uzleuven.be.


## Abstract

Clinical risk prediction models are increasingly being developed and validated on multicenter datasets. In this paper, 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

## 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 is 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]

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) increased. Pavlou et al. recently pointed out that calibrated results can be obtained with the mixed effects logistic regression model if marginal predictions are used.[8] 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 paper, 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 deduct 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.

## 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 non-event, $X_{kij}$ the $k^{th}$ 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:

**(1)**

$$Y_{ij} \sim \text{bin}(1, p_{ij})$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_m+\sum_{k=1}^{K}\beta_{k\,m}X_{kij}\,.$$

The intercept $\alpha_m$ 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:

**(2)**

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

**(3)**

$$\hat{p}_{LR\,ij}=\frac{1}{1+\exp(-LP_{LR\,ij})}.$$

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 $\tau^2$. The random center intercepts $a_j$ are assumed to be normally distributed with mean zero.

**(4)**

$$Y_{ij} \sim \text{bin}(1, p_{ij})$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_c + a_j + \sum_{k=1}^{K} \beta_{k\,c} X_{kij}$$

$$a_j \sim N(0, \tau^2)$$

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 linear predictor given the random intercept for the $j^{th}$ center is

**(5)**

$$LP_{MLR\,c\,ij} = \hat{\alpha}_c + \hat{a}_j + \sum_{k=1}^{K} \hat{\beta}_{k\,c} X_{kij}\,.$$

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 $\tau^2$ is lower (i.e., uniform shrinkage for all centers in homogeneous populations).[4] Conditional predicted probabilities $\hat{p}_{MLR\,c\,ij}$ are obtained by taking the inverse logit of the conditional linear predictor.

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, ($\hat{a}_j = 0$).

**(6)**

$$LP_{MLR\,a\,ij} = \hat{\alpha}_c + 0 + \sum_{k=1}^{K} \hat{\beta}_{k\,c} X_{kij}\,.$$

Predicted probabilities assuming an average random center intercept (0), $\hat{p}_{MLR\,a\,ij}$, are obtained by taking the inverse logit of the linear predictor. 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:

**(7)**

$$\hat{p}_{MLR\ m\ ij} = \int_{-\infty}^{\infty} \frac{1}{1+\exp(-LP_{MLR\ cond\ ij})} f(\hat{a}_j) d\hat{a}_j$$

$$\hat{p}_{MLR\ m\ ij} = \int_{-\infty}^{\infty} \frac{1}{1+\exp[-(\hat{\alpha}_c + \hat{a}_j + \sum_{k=1}^{K} \hat{\beta}_{k\ c} X_{kij})]} f(\hat{a}_j)\, d\hat{a}_j,$$

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 linear predictor of the mixed effects model $LP_{MLR\ m\ ij}$ 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 Web 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 non-events. The C-index expresses the probability that for a randomly selected pair of an event and a non-event, the event has a higher predicted probability.[12] For the computation of the standard C-index, pairs of events and non-events belonging to different clusters are compared, as well as pairs from the same cluster. It is estimated by

**(8)**

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

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

**(9)**

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

This corresponds to the average center-specific C-index, weighted by the number of pairs of events and non-events 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 linear predictor LP, obtained by applying formula (2). To perform logistic calibration one fits the following model to a validation dataset:

**(10)**

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

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.[15-18] We elaborate on the effect of sample size in 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

**(11)**

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_{j\,cal}+\beta_{cal\,w}LP_{ij}+b_{j\,cal}LP_{ij}$$

$$\begin{pmatrix}a_{j\,cal}\\ b_{j\,cal}\end{pmatrix}\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\tau_a^2 & \tau_{ab}\\ \tau_{ab} & \tau_b^2\end{bmatrix}\right).$$

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

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.[19] 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 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. Web Appendix 3 presents an overview of the formulas for the C-index and logistic calibration in this comprehensive framework.

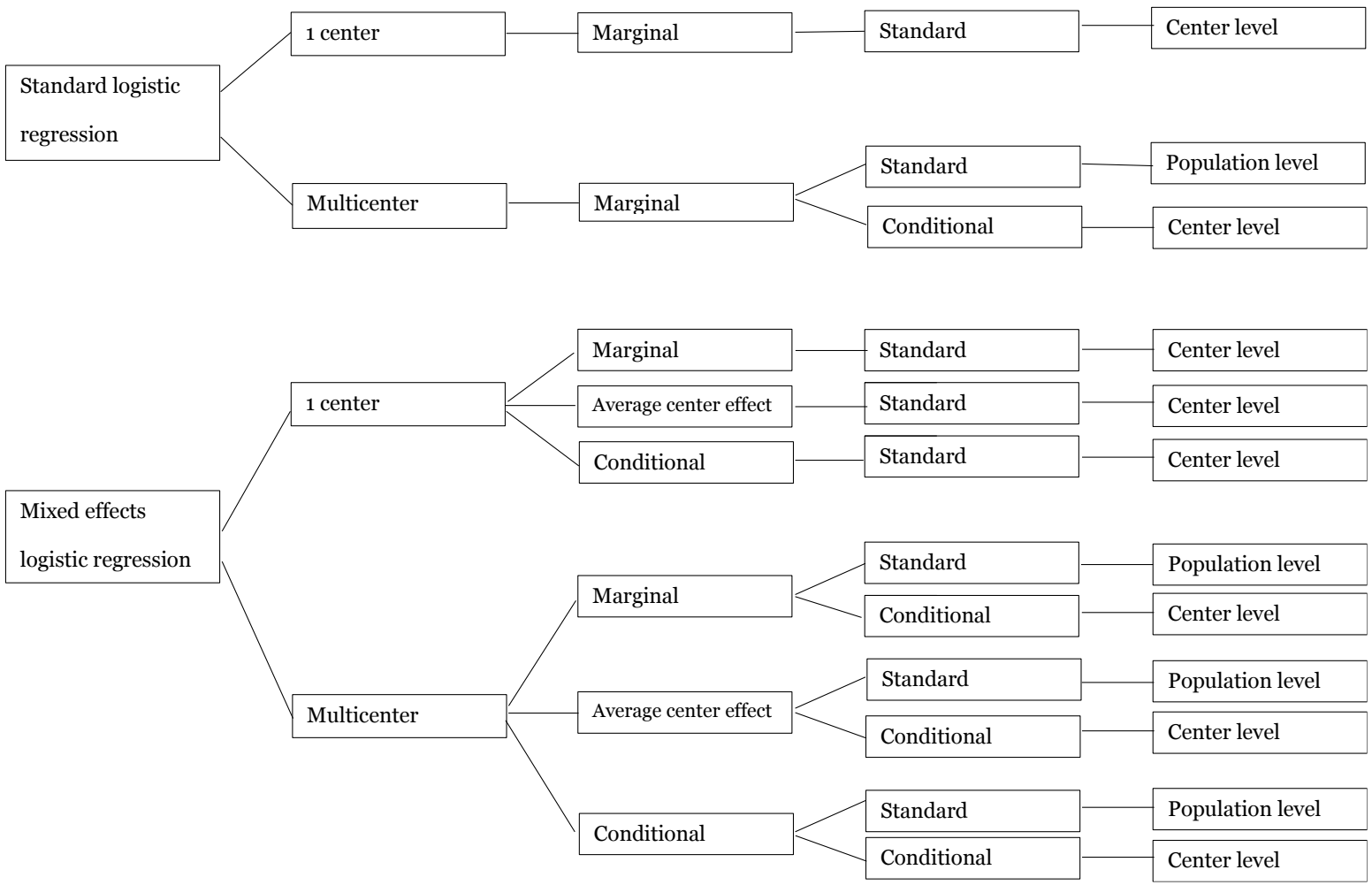


Figure 1. A comprehensive framework of options for model validation, subject to the type of prediction model that is being evaluated (standard or mixed effects logistic regression) and the available validation dataset (one center or multicenter)

## 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).[20-22] Using a cumulative Gaussian approximation to the logistic function leads to the following approximation:[20]

**(12)**

$$\alpha_m \approx \alpha_c / f$$

$$\beta_m \approx \beta_c / f,$$

$$\text{with } f = \sqrt{1+\tau^2 c^2}$$

$$\text{and } c = \frac{16\sqrt{3}}{15\pi}.$$

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

**(13)**

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

$$= \alpha_{cal} + \beta_{cal}(\hat{\alpha}_m + \sum_{k=1}^{l} \hat{\beta}_{k\,m} X_{ki})$$

$$\approx \alpha_{cal} + \frac{\beta_{cal}}{f}(\hat{\alpha}_c + \sum_{k=1}^{l} \hat{\beta}_{k\,c} X_{ki}).$$

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

**(14)**

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_{cal\,w} + a_{j\,cal} + \beta_{cal\,w} LP_{MLR\,a\,ij} + b_{j\,cal} LP_{MLR\,a\,ij}$$

$$= \alpha_{cal\,w} + \beta_{cal\,w}(\hat{\alpha}_c + \sum_{k=1}^{l} \hat{\beta}_{k\,c} X_{kij})$$

$$\approx \alpha_{cal} + \beta_{cal\,w} f(\hat{\alpha}_m + \sum_{k=1}^{l} \hat{\beta}_{k\,m} X_{kij}).$$

This demonstrates how the calibration slope may deviate from one due to the choice of modelling 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 $\tau^2$ will often be estimated with error. As shown in Web 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

## Simulation study

### *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 (intraclass correlation (ICC)=20%), and one with little clustering (ICC=5%).[23] We fixed the number of centers (J) at 20.[24]

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 $\alpha$ 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:

**(16)**

$$Z_{ij} \sim \text{unif}(0,1)$$

$$Y_{ij}=\begin{cases}1 & \text{if } z_{ij}\leq p_{ij}\\ 0 & \text{if } z_{ij}>p_{ij}\end{cases}.$$

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 (8 parameters for the regression coefficients plus one extra parameter for the random intercept variance). The required number of non-events 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 eigth 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 to 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 non-converging 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 linear predictor for the random intercept model: the conditional linear predictor, including the center-specific intercept estimates (equation 5), and the linear predictor assuming an average random intercept (equation 6). The marginal linear predictor 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).[25] The lmer function from the lme4 package[26] 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 Web Appendix 5.

## *Results*

Calibration

Severe clustering (ICC 20%, 100 events per variable)

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 Web 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 2A, triangles), while the predictions from the standard model were only calibrated at the population level (Figure 2B, circles). The center-level calibration slopes tended to be larger than one for the predictions of the standard model (Figure 2A, circles) and the population-level calibration slopes were smaller than one for the predictions assuming an average random intercept (Figure 2B, 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 can be seen in 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 $b_{cal\,j}$ (not shown) reflect center-specific differences from this slope. The average estimated variance of the $b_{cal\,j}$ was <0.0005 for the three types of predictions.

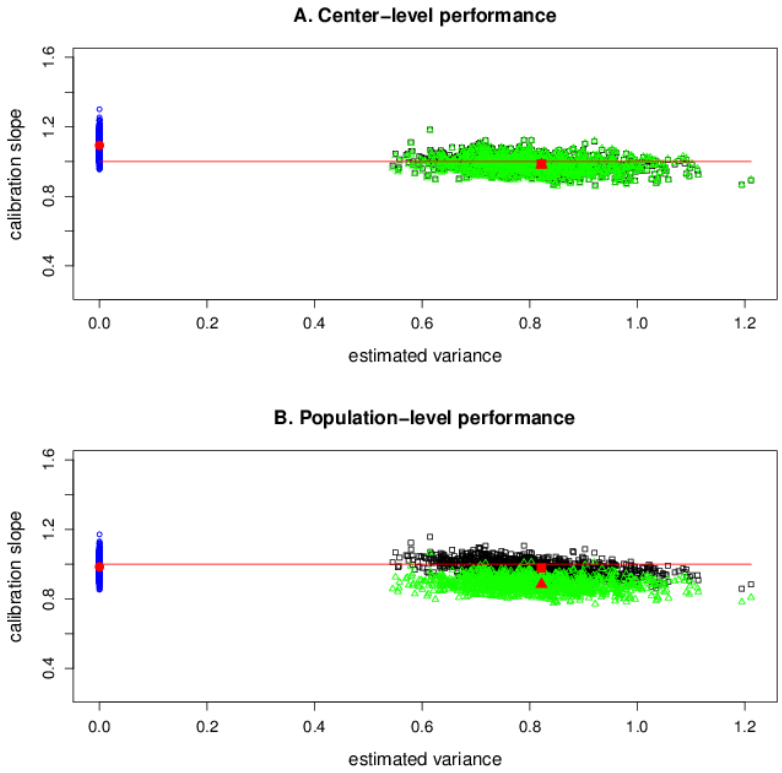


Figure 2. Center-level (panel A) and population-level (panel B) calibration slopes of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor of the random intercept model assuming an average random intercept (triangles), by estimated

random intercept variance in samples with 100 events per variable 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.

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 (Web 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 are 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 Web Appendix 2.

Mild clustering (ICC 5%, 100 events per variable)

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 Web Appendix 7, Figure A7). The predictions from the standard model yielded within-center calibration slopes slightly above one (Figure 3A, circles), while the predictions assuming an average random intercept yielded population-level calibration slopes slightly below one (Figure 3 B, triangles).

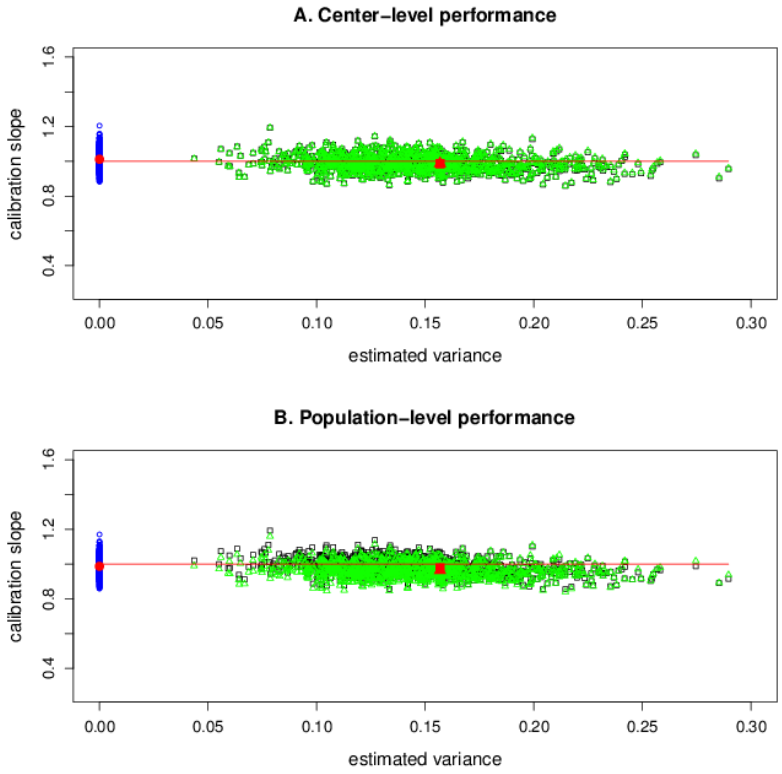


Figure 3. Center-level (panel A) and population-level (panel B) calibration slopes of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable 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.

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 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 4A).

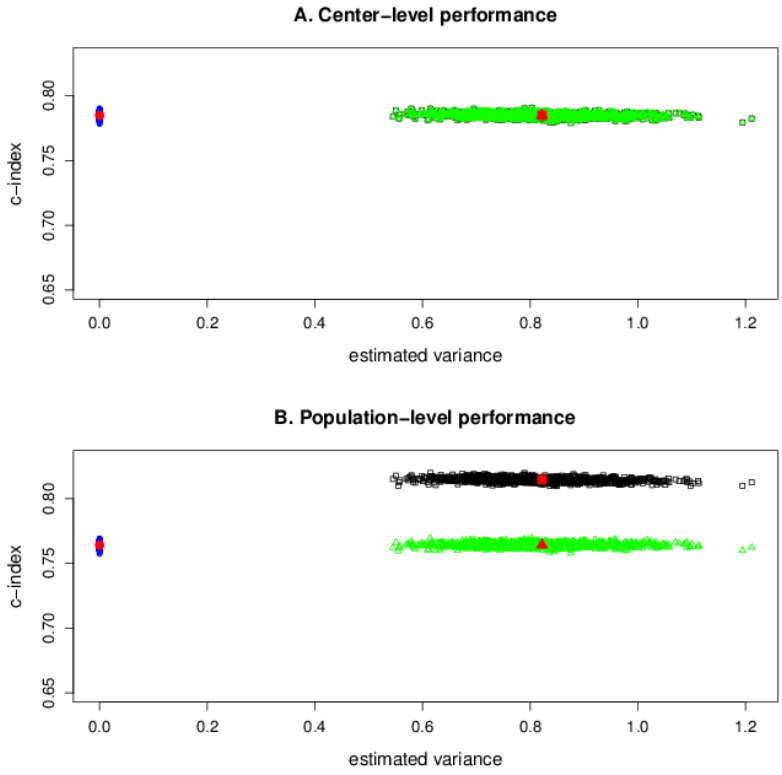

Figure 4. Center-level (panel A) and population-level (panel B) C-indexes of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable 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.

The population-level C-indexes for the predictions assuming an average random intercept (Figure 4B), 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 (Web Appendix 8, Figure A8B, squares).

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

## 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 ten 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 (Web Appendix 9, Table A3). Standard errors were larger in the mixed effects model.

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 linear predictor 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) (see Web 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 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.

## 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).

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

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.

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 consequence of the much-described finding that marginal regression coefficients are typically closer to zero than conditional regression coefficients.[20-22] 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 a 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.[27-29] One suggested guideline is to collect data from at least fifty clusters,[29] although this may be hard to obtain in practice. A sufficiently large number of events per variable 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.[24]

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.[30] 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, 24, 27-31] 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 [32, 33, 34] 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, 35]

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.[36] 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, 37]. If centers are large, the center-specific performance can be studied. An example on preoperative tumor diagnosis was recently published by the IOTA consortium.[19]

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.[38] 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, 39-41]

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.

## Funding

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 (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 paper 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 ElaborationThe TRIPOD Statement: 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: 80-6.
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, Klei W, Moons K and Vergouwe Y. 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, Koffijberg H and Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med*. 2013; 32: 3158-80.
7. Skrondal A and Rabe-Hesketh S. Prediction in multilevel generalized linear models. *J R Stat Soc Ser A*. 2009; 172: 659-87.
8. Pavlou M, Ambler G, Seaman S and Omar RZ. 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 and Vergouwe Y. 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. *Computational Statistics & Data Analysis*. 2012; 56: 1966-80.
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: 562-5.
14. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ and Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016.
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: 411-21.
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: 935-42.

18. Peduzzi P, Concato J, Kemper E, Holford TR and Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996; 49: 1373-9.
19. 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.
20. Zeger SL, Liang K-Y and Albert PS. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*. 1988; 44: 1049-60.
21. Neuhaus JM, Kalbfleisch JD and Hauck WW. A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data. *International Statistical Review / Revue Internationale de Statistique*. 1991; 59: 25-35.
22. Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. *Stat Methods Med Res*. 1992; 1: 249-73.
23. Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S and Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol*. 2004; 57: 785-94.
24. 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: 1504-11.
25. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2011.
26. Bates D, Maechler M and Bolker B. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42. 2011.

27. Maas CJM and Hox JJ. Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. 2005; 1: 86-92.
28. Paccagnella O. Sample Size and Accuracy of Estimates in Multilevel Models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*. 2011; 7: 111-20.
29. Moineddin R, Matheson FI and Glazier RH. A simulation study of sample size for multilevel logistic regression models. *BMC Med Res Methodol*. 2007; 7.
30. Molenberghs G and Verbeke G. *Models for discrete longitudinal data*. 2005.
31. 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.
32. 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: 2419-29.
33. 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: 1136-49.
34. Maas CJ and Hox JJ. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*. 2004; 46: 427-40.
35. 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.
36. van Klaveren D, Götz HM, Op de Coul EL, Steyerberg EW and Vergouwe Y. 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.

37. 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: 2081-103.
38. 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: 971-80.
39. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE and Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008; 61: 76-86.
40. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ and Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. 2004; 23: 2567-86.
41. Van Houwelingen HC and Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat Med*. 1995; 14: 1999-2008.
42. Vittinghoff E and McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol*. 2007; 165: 710-8.
43. Peduzzi P, Concato J, Feinstein AR and Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol*. 1995; 48: 1503-10.
44. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A and Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *J Clin Epidemiol*. 2011; 64: 993-1000.

45. 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: 361-87.
46. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr. and Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med*. 2000; 19: 1059-79.

# Appendix 1. Correspondence between predictions from the standard logistic regression model and marginalized predictions from the mixed effects logistic regression model

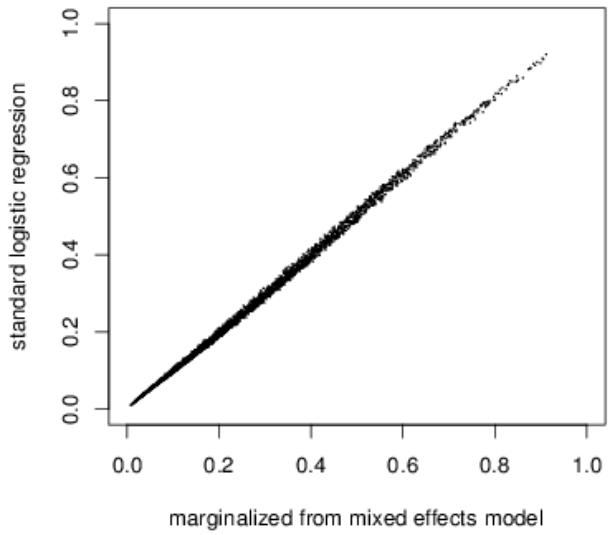


Figure A1. Marginal predictions from a standard logistic regression model versus marginalized predictions from a mixed effects logistic regression model.

The logistic regression model and the mixed effects logistic regression model were developed in a sample with 100 EPV, drawn from a population with clustered data (ICC 20%). The predictions depicted are the marginalized predicted probabilities of the mixed effects logistic regression model and marginal predicted probabilities obtained by the standard logistic regression model.

# Appendix 2. The influence of the amount of events per variable

## *Calibration and sample size*

The most common reason for calibration slopes deviating from one is an inadequate sample size for model development. The importance of the number of events per variable (EPV) has been shown in many simulation studies.[16-18, 42-44] If the ratio of the number of events to the number of parameters to estimate is small, for instance, less than ten, overfitting will likely occur. The model captures idiosyncratic characteristics of the sample at hand,[15] and fails to maintain its predictive ability when tested in new data. Calibration slopes of overfitted models tested in new data are typically smaller than one. The effects of the choice of the modelling technique on the calibration slope will in practice often be difficult to observe, because of the simultaneous effect of overfitting.

## *Simulation study*

*Design*

The simulation study was designed as described in the main article. Samples of EPV 5 were only drawn from heavily clustered data (ICC=20%) to facilitate the

interpretation of the results, since we expect that the choice of the modelling technique on the calibration slope will be difficult to observe in weakly clustered data, because of the simultaneous effect of overfitting.

*Results*

Calibration

31 of 1000 samples (3%) were removed because they did not include patients from all centers, making it impossible to give conditional predictions in some centers. 1 sample (0.1%) was removed because of a change of more than $10^{-5}$ in the deviance of the last two iterations, and 1 sample (0.1%) was removed because it was outlying in a plot of estimated regression coefficients versus standard errors. The performance of non-converged models was similar to the performance of converged models.

The random intercept variance was often underestimated (median 0.66) and the estimates were very variable across samples, due to the small sample size (Figure A2). For the same reason, calibration slopes were highly variable.

The low center-level calibration slopes of the random intercept model indicated overfitting (at the correctly estimated $\tau^2$, mean $\hat{\beta}_{\text{cal within}}$=0.74 when conditional predictions (big square) were used and mean $\hat{\beta}_{\text{cal within}}$=0.74 when

predictions assuming an average random intercept (big triangle) were used, Figure A2, panel A). The predictions from the standard model gave within-center calibration slopes closer to one (mean $\hat{\beta}_{\text{cal within}}$=0.83, Figure A2, panel A, big circle). This is the result of two opposite effects: small sample sizes lead to overfitting and decreased calibration slopes, while using marginal predictions leads to increased calibration slopes at the center level.

The standard model was overfitted at the population level (mean $\hat{\beta}_{\text{cal}}$=0.75, Figure A2, panel B, big circle). The conditional predictions had a slightly lower population-level calibration slope (at the correctly estimated $\tau$, mean $\hat{\beta}_{\text{cal}}$=0.70, Figure A2, panel B, big square). The predictions assuming an average random intercept gave the lowest population-level calibration slopes (mean $\hat{\beta}_{\text{cal}}$=0.67, Figure A2, panel B, big triangle).

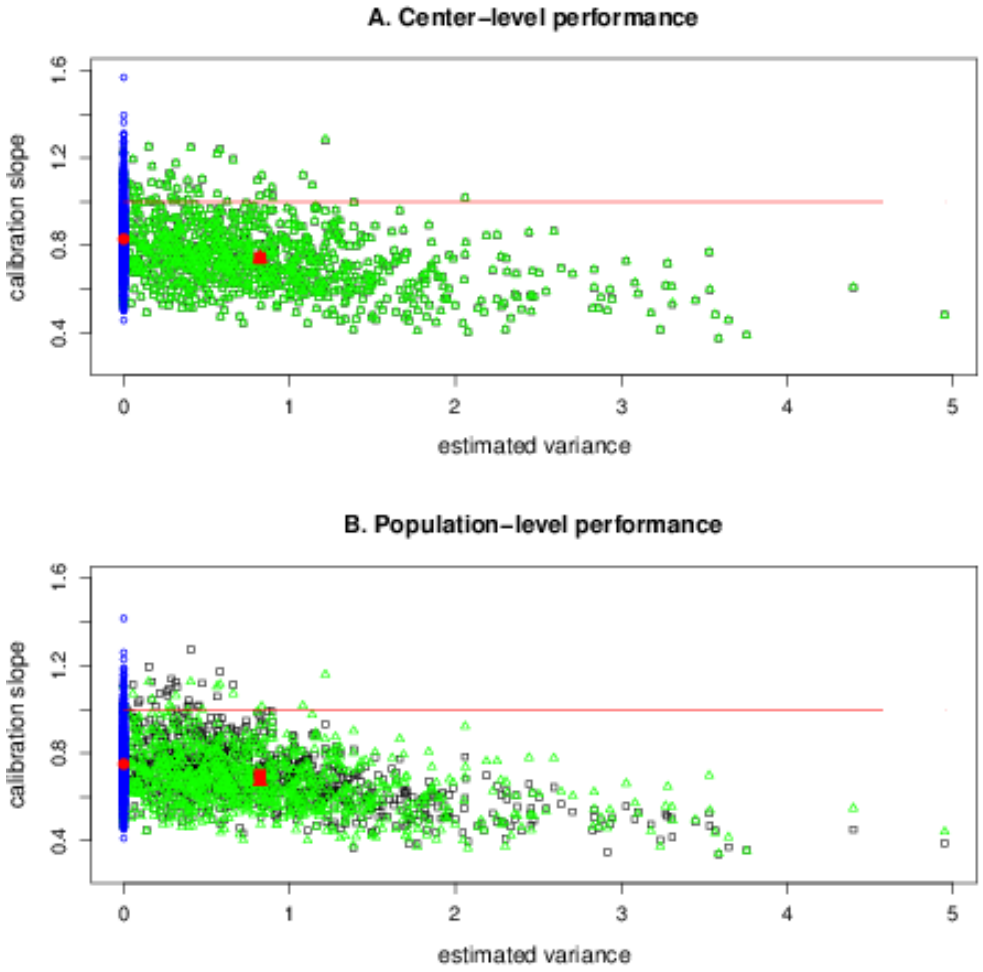


Figure A2. Center-level (panel A) and population-level (panel B) calibration slopes of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 5 events per variable 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 within-center calibration intercepts were smaller than zero for the standard model, indicating that on average, predicted probabilities were too high. The population-level calibration intercepts were larger than zero for the predictions assuming an average random intercept, indicating that on average, predicted probabilities were too low (figure A3).

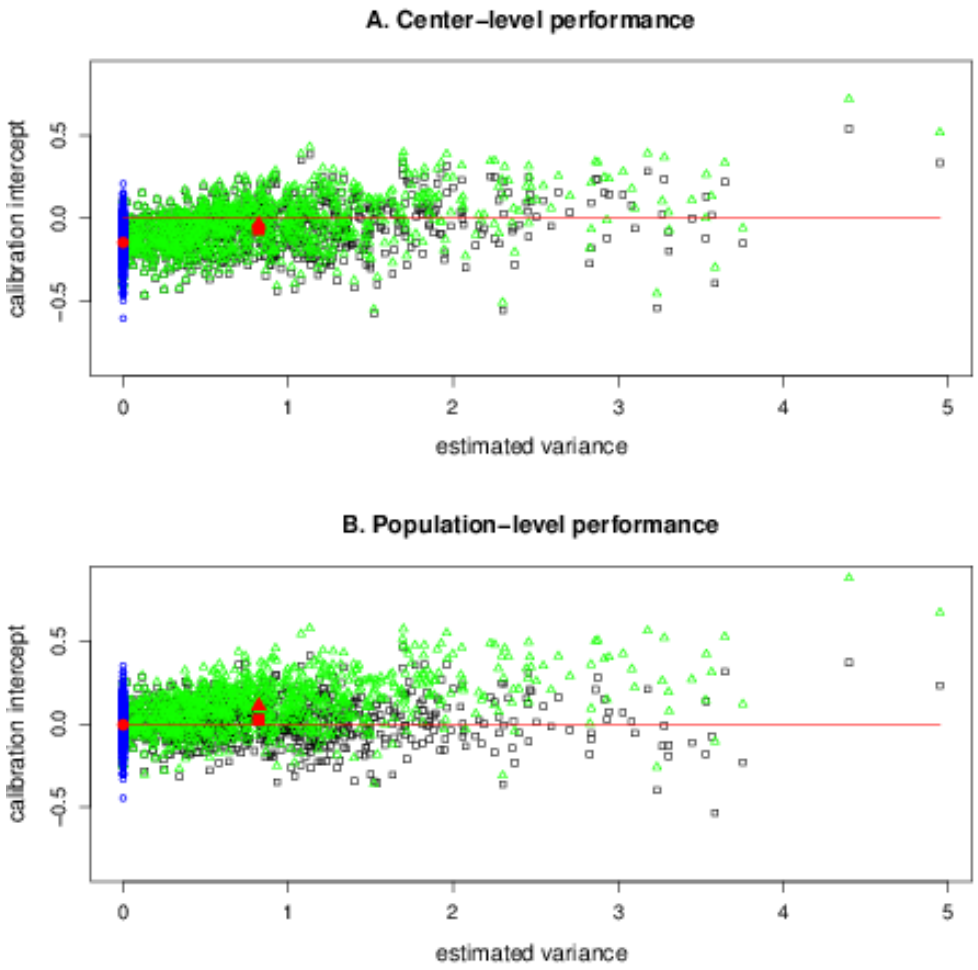


Figure A3. Center-level (panel A) and population-level (panel B) calibration intercepts of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model from the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 5 events per variable 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 intercept.

Discrimination

In small samples (EPV 5), the mean C-index was lower (0.758 and 0.758 for the mixed and standard model, respectively) than in large samples, due to overfitting. Higher population-level C-indexes were obtained with the conditional predictions than with the predictions from the standard model and the predictions assuming an average random intercept. The random intercept variance was sometimes estimated close to zero in small samples (EPV 5). In these cases, the C-index dropped to the level of the C-index of the standard model.

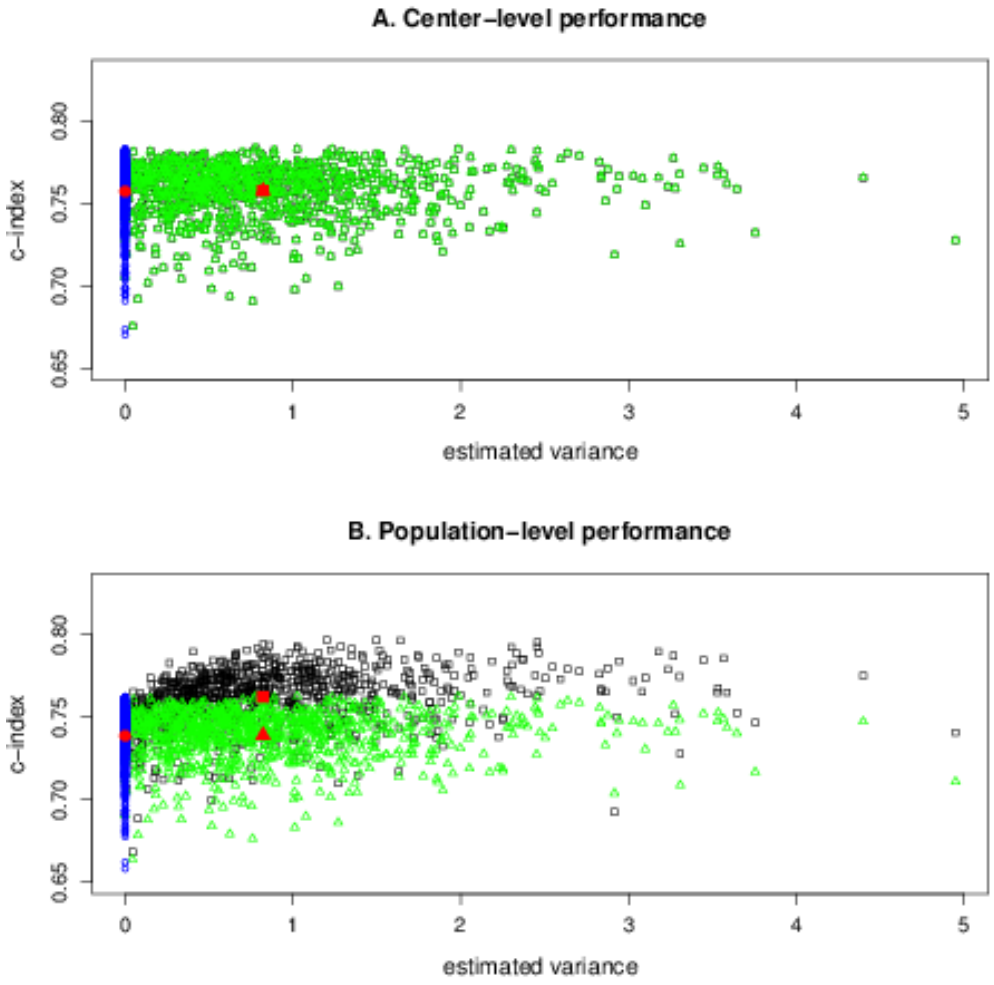


Figure A4. Population-level (panel B) and center-level (panel A) C-indexes of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 5 events per variable 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.

### *Discussion*

When the dataset was small (EPV=5), both the standard model and the mixed effects model yielded poorly calibrated results. The regression models were heavily overfitted. The poor calibration was worse in the mixed effects model, because the problem of overfitting in small datasets, yielding beta coefficients that are too far from zero, was enhanced by the effect that conditional regression coefficients are generally more extreme than marginal regression coefficients. Although the standard model was seemingly better calibrated, fitting a model that does not reflect the data structure is not an adequate way to deal with problems caused by small sample sizes. A more appropriate solution would be to collect larger datasets for model development. Guidelines suggest to collect at least ten events per parameter that needs to be estimated.[16, 18, 45] Additional shrinkage or penalization of the regression coefficients may be needed when the number of events per variable is smaller than twenty.[40, 46] When statistical variable selection needs to be performed, up to fifty events per candidate predictor are recommended.[16, 17] We did not investigate a scenario with weak clustering and a low EPV. However, given that the performance results for the different types of predictions were very similar in the simulations with weak clustering and a high EPV, we expect the overfitting

caused by the low sample size would have rendered the effects of ignoring clustering on model performance invisible in the simulation results.

# Appendix 3. Formulas for the C-index and logistic calibration in a comprehensive framework of the standard and within-center validation of marginal predictions, conditional predictions and predictions assuming an average random intercept

| | **Marginal predictions** | **Conditional predictions** | **Predictions assuming an average random intercept** |
|---|---|---|---|
| **Standard validation** | | | |
| C-index | $\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(\hat{p}_{LR\,ij}>\hat{p}_{LR\,i'j'} \text{ and } y_{ij}=1 \text{ and } y_{i'j'}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(y_{ij}=1 \text{ and } y_{i'j'}=0\right)}$ | $\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(\hat{p}_{MLR\,c\,ij}>\hat{p}_{MLR\,c\,i'j'} \text{ and } y_{ij}=1 \text{ and } y_{i'j'}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(y_{ij}=1 \text{ and } y_{i'j'}=0\right)}$ | $\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(\hat{p}_{MLR\,a\,ij}>\hat{p}_{MLR\,a\,i'j'} \text{ and } y_{ij}=1 \text{ and } y_{i'j'}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{j'=1}^{J}\sum_{i'=1}^{n_{j'}} I\left(y_{ij}=1 \text{ and } y_{i'j'}=0\right)}$ |
| Calibration slope $\beta_{cal}$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+\beta_{cal}LP_{LR\,ij}$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+\beta_{cal}LP_{MLR\,cond\,ij}$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+\beta_{cal}LP_{MLR\,a\,ij}$ |
| Calibration intercept $\alpha_{cal}$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+LP_{LR\,ij}$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+LP_{MLR\,cond\,ij}$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal}+LP_{MLR\,a\,ij}$ |
| **Conditional validation** | | | |
| C-index | $\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(\hat{p}_{LR\,ij}>\hat{p}_{LR\,i'j} \text{ and } y_{ij}=1 \text{ and } y_{i'j}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(y_{ij}=1 \text{ and } y_{i'j}=0\right)}$ | $\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(\hat{p}_{MLR\,c\,ij}>\hat{p}_{MLR\,c\,i'j} \text{ and } y_{ij}=1 \text{ and } y_{i'j}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(y_{ij}=1 \text{ and } y_{i'j}=0\right)}$ | $\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(\hat{p}_{MLR\,a\,ij}>\hat{p}_{MLR\,a\,i'j} \text{ and } y_{ij}=1 \text{ and } y_{i'j}=0\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} I\left(y_{ij}=1 \text{ and } y_{i'j}=0\right)}$ |
| Calibration slope $\beta_{cal\,w}$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+\beta_{cal\,w}LP_{LR\,ij}+b_jLP_{LR\,ij}$ <br> $\begin{pmatrix}a_j\\b_j\end{pmatrix}\sim N\begin{pmatrix}0 & \tau_a^2\\0' & \tau_b^2\end{pmatrix}$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+\beta_{cal\,w}LP_{MLR\,c\,ij}+b_jLP_{MLR\,c\,ij}$ <br> $\begin{pmatrix}a_j\\b_j\end{pmatrix}\sim N\begin{pmatrix}0 & \tau_a^2\\0' & \tau_b^2\end{pmatrix}$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+\beta_{cal\,w}LP_{MLR\,a\,ij}+b_jLP_{MLR\,a\,ij}$ <br> $\begin{pmatrix}a_j\\b_j\end{pmatrix}\sim N\begin{pmatrix}0 & \tau_a^2\\0' & \tau_b^2\end{pmatrix}$ |
| Calibration slope $\alpha_{cal\,w}$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+LP_{LR\,ij}$ <br> $(a_j)\sim N(0,\tau_a^2)$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+LP_{MLR\,c\,ij}$ <br> $(a_j)\sim N(0,\tau_a^2)$ | $\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\alpha_{cal\,w}+a_j+LP_{MLR\,a\,ij}$ <br> $(a_j)\sim N(0,\tau_a^2)$ |

Table A1. Formulas for the C-index and logistic calibration in a comprehensive framework of the standard and within-center validation of marginal predictions, conditional predictions and predictions assuming an average random intercept.

## Appendix 4. Calibration with a biased estimate of the between-center variance

When we have two mixed effects models fitted on the same dataset, but one has a biased estimate of the between-center variance $\hat{\tau}'^2$ (e.g., underestimated in a sample with very few centers), the following equations hold:

$$\frac{\hat{\beta}_c}{\sqrt{1+\tau^2 c^2}} \approx \frac{\hat{\beta}_{c'}}{\sqrt{1+\hat{\tau}'^2 c^2}}$$

$$\hat{\beta}_{c'} \approx \hat{\beta}_c \times \sqrt{\frac{1+\hat{\tau}'^2 c^2}{1+\tau^2 c^2}},$$

with c defined as above. Hence, the within-center calibration slopes of these two models will differ by factor $\sqrt{\frac{1+\tau^2 c^2}{1+\hat{\tau}'^2 c^2}}$. If the model with the correct estimate of the between-center variance is perfectly calibrated, the model with the biased estimate will yield calibration slopes smaller than one if $\tau^2$ is overestimated and calibration slopes larger than one if $\tau^2$ is underestimated.

This is illustrated in the simulation study (see Figure A5). If the random intercept variance was estimated correctly (0.822), the predictions from the random intercept model assuming average random intercepts were well calibrated at the center level (mean $\hat{\beta}_{\text{cal w}}$=0.98, large triangle). As we expected, the

calibration slope was lower when the random intercept variance was overestimated and higher when the random intercept variance was underestimated. In the extreme case where the random intercept variance was set to zero by using a standard logistic regression model for prediction (circles), the within-center calibration slope was 1.09 on average (large dot). The association between the estimated random intercept variance and the within-center calibration slope was negative and close to the theoretical approximation.

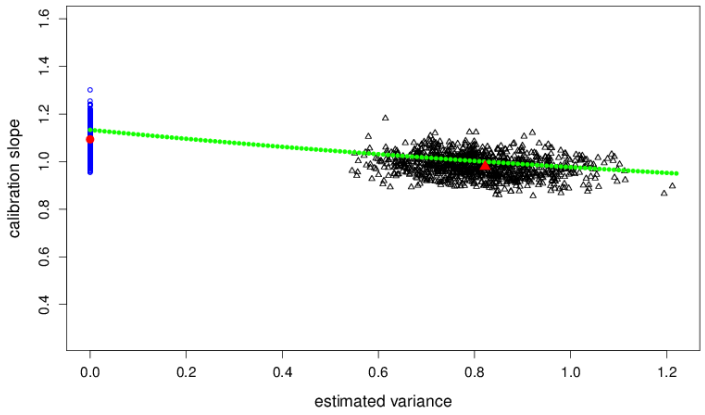


Figure A5. Center-level calibration slopes of the standard logistic regression model (circles) and the random intercept model assumingaverage random intercepts (triangles), by estimated random intercept variance. Small symbols indicate calibration slopes observed in the samples, large filled symbols indicate observed average calibration slopes at estimated variance=0 for the standard logistic regression model and at the correctly estimated variance (0.822, ICC=20%) for the random intercept model. The dotted line represents the expected relationship.

## Appendix 5. R code for simulations

```
#####1. Generation of source populations
rm(list=ls(all=TRUE))
library (rms)
library (lme4)

set.seed(1986)

#Step 1. Fix the number of clusters (nb) and draw the number of observations per
#cluster (n2) from a poisson distribution.
nb <- 20
mean <- 6.82
sd <- 0.3
L1<-rnorm(nb, mean, sd)
L<-round(exp(L1))
n2 <- rpois(nb, L)
cent <-seq(1:nb)
center <-rep(cent, n2)
n <- length(center) #number of observations in the source population
pat_nr <- seq(1:n)

#Step 2. Generate random cluster intercepts (randeff).
set.seed(2022)
randeff        <- rnorm(nb, 0, 0.9068997)  # this is to obtain an ICC of 20%
randeffect <-rep(randeff, n2)

## Step 3. Generate predictors x1 to x8
x1      <- rnorm(n, 0, 1)
x2      <- rnorm(n, 0, 0.6)
x3      <- rnorm(n, 0, 0.4)
x4      <- rnorm(n, 0, 0.2)
x5      <- rbinom(n, 1, 0.20)
x6      <- rbinom(n, 1, 0.30)
x7      <- rbinom(n, 1, 0.30)
x8      <- rbinom(n, 1, 0.40)
```

```
# Step 4: Set beta coefficients and obtain the logit.
b1      <- 0.8
b2      <- 0.8
b3      <- 0.8
b4      <- 0.8
b5      <- 0.8
b6      <- 0.8
b7      <- 0.8
b8      <- 0.8
logit   <- -
2.1+b1*x1+b2*x2+b3*x3+b4*x4+b5*x5+b6*x6+b7*x7+b8*x8+randeffect

#Step 5: Calculate the predicted probabilities of experiencing an event
p       <- plogis(logit)

#Step 6: Obtain dichotomous outcome y
y       <- ifelse(runif(n)<=p, 1, 0)
data <- as.data.frame(cbind(pat_nr, x1, x2, x3, x4, x5, x6, x7, x8, randeffect, center,
y), dimnames = list(c(1:n), Cs(pat_nr, x1, x2, x3, x4, x5, x6, x7, x8, randeffect,
center, y)))

####2.Sampling from the source population and model building within samples
#This code corresponds to sampling with 100 events per variable. The code can be
#adjusted for other conditions with minor changes.
rm(list=ls(all=TRUE))
library (Hmisc)
library (rms)
library (lme4)
library(arm)
Nsample <- 1000  #specify the number of samples to be drawn

#Step 1. Specify the number of clusters
Nsamplev2 <-20

#Step 2. Specify the number of events to be drawn
NEVENTS <- 900
```

```
NNONEVENTS <- round((NEVENTS/mean(data$y))-NEVENTS)

#Matrices to store results
Ncenter_sample <- matrix(NA, nrow=Nsample, ncol=1)
center_sample <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
Nevents_sample <- matrix(NA, nrow=Nsample, ncol=1)
Npats_sample <- matrix(NA, nrow=Nsample, ncol=1)
Pevents_sample <- matrix(NA, nrow=Nsample, ncol=1)
Npats_center_sample <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)

mat_sd_ranef0 <- matrix(NA, nrow=Nsample, ncol=1)
mat_sd_ranef1 <- matrix(NA, nrow=Nsample, ncol=1)
mat_regcof <- matrix(NA, nrow=Nsample, ncol=9)
mat_SEregcof <- matrix(NA, nrow=Nsample, ncol=9)
mat_ranef <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
mat_SEranef <- matrix(NA, nrow=Nsample, ncol=Nsamplev2)
Niterm1 <- matrix(NA, nrow=Nsample, ncol=1)
Equaldeviance <- matrix(NA, nrow=Nsample, ncol=1)

mat_regcofu <- matrix(NA, nrow=Nsample, ncol=9)
mat_SEregcofu <- matrix(NA, nrow=Nsample, ncol=9)
Niterm1u <- matrix(NA, nrow=Nsample, ncol=1)
Equaldevianceu <- matrix(NA, nrow=Nsample, ncol=1)

realvalresultB <- matrix(NA,nrow=Nsample, ncol=8) #Validation results for
predictions average random intercept
dimnames(realvalresultB) <- list(c(1:Nsample), Cs(C,calib_slope, calib_interc,
Cw,ML_calib_slope,var_ML_calib_slope,
ML_calib_interc,var_ML_calib_interc))
realvalresultBcond <- matrix(NA,nrow=Nsample, ncol=8) ) #Validation results for
conditional predictions
dimnames(realvalresultBcond) <- list(c(1:Nsample), Cs(C,calib_slope,
calib_interc, Cw,ML_calib_slope,var_ML_calib_slope,
ML_calib_interc,var_ML_calib_interc))
realvalresultBu <- matrix(NA,nrow=Nsample, ncol=8) ) #Validation results for
marginal predictions
```

```
dimnames(realvalresultBu) <- list(c(1:Nsample), Cs(C,calib_slope, calib_interc,
Cw,ML_calib_slope,var_ML_calib_slope,
ML_calib_interc,var_ML_calib_interc))logliko <-
matrix(NA,nrow=Nsample,ncol=1)

logliko <- matrix(NA,nrow=Nsample,ncol=1)
loglik1 <- matrix(NA,nrow=Nsample,ncol=1)
logliko u <- matrix(NA,nrow=Nsample,ncol=1)
loglik1u <- matrix(NA,nrow=Nsample,ncol=1

#Start sampling from the domain
tmpA <- aggregate(data$pat_nr, list(center = data$center), FUN=length)
pt_center <- as.vector(tmpA[,2])
names(pt_center) <- tmpA[,1]

for (r in 1:Nsample){
set.seed(r)

#Step 3. Sample clusters (samp_center) from the source population
samp_center <- names(pt_center)

#Get the data from sampled clusters (dsample_lev2)
loc.pt_center <- pt_center[samp_center]
indices.cutoffs <- c(0,cumsum(loc.pt_center))
ptnrs_samp2 <- matrix(NA,nrow=sum(loc.pt_center),ncol=1)
for (i in 1:Nsamplev2){
ptnrs_samp2[(indices.cutoffs[i]+1):indices.cutoffs[i+1],] <-
data[data$center==names(loc.pt_center)[i],1]
}
dsample_lev2 <- data[ptnrs_samp2,]
dsample_lev2<-dsample_lev2[order(dsample_lev2$center), ]

#Step 4. Stratified sampling of events (sample_lev1_y1) and non-events
(sample_lev1_y0).
sdatay1 <- dsample_lev2[dsample_lev2$y==1,]
pt_nr_y1 <- as.vector(sdatay1$pat_nr)
sample_lev1_y1 <- sample(pt_nr_y1, NEVENTS , replace = FALSE)
```

```
sdatay0 <- dsample_lev2[dsample_lev2$y==0,]
pt_nr_y0 <- as.vector(sdatay0$pat_nr)
sample_lev1_y0 <- sample(pt_nr_y0, NNONEVENTS, replace = FALSE)
sample_lev1 <- c(sample_lev1_y0, sample_lev1_y1)
dsample <- data[sample_lev1,]

#Step 5. Fit a random intercept model in the sample: full model (m1) and null
#model (m0)
        #mixed effects model
iter_m1<-capture.output(m1<-lmer(y~x1+x2+x3+x4+x5+x6+x7+x8+(1|center),
family = "binomial", data=dsample, verbose=T ))
m0<-lmer(y~(1|center), family = "binomial", data=dsample)
Last2it<-iter_m1[c(length(iter_m1)-1):c(length(iter_m1))]
Last2itsplit<-strsplit(Last2it,":")
Equal<- Last2itsplit [[1]][2]== Last2itsplit [[2]][2]#check convergence
        #standard model
m1u<-glm(y~x1+x2+x3+x4+x5+x6+x7+x8,family=binomial ,data=dsample,
maxit=50) #deviance en fixed parameter schattingen voor alle iteraties tijdens het
fitten
m0u<-glm(y~ 1,family=binomial, data=dsample, maxit=50)#nulmodel
Equalu<- m1u$converge

# Set Npats_center, center and ranef to missing when we sampled <1 observation
per cluster
center_s <- length(unique(dsample$center))
N_missing_center <- Nsamplev2-Ncenter_s
add_center <- rep(x=-99, times=N_missing_center)
if (Ncenter_s < Nsamplev2){
center_s <- c(unique(dsample$center), add_center)
Npats_center_s <- c((aggregate(dsample$pat_nr, list(center = dsample$center),
FUN=length)[,2]),add_center)
clustereff_s <- c(attr(m1, "ranef"),add_center)
clustereff_s_SE <- c(as.numeric(se.ranef(m1)$center), add_center)
}
if (Ncenter_s == Nsamplev2) {
Npats_center_s <- aggregate(dsample$pat_nr, list(center = dsample$center),
FUN=length)[,2]
```

```
center_s <- unique(dsample$center)
clustereff_s <- attr(m1, "ranef")
clustereff_s_SE <- as.numeric(se.ranef(m1)$center)
}

#save sample summaries
Ncenter_sample[r,] <- Ncenter_s
center_sample[r,] <- center_s
Npats_center_sample[r,] <- Npats_center_s
Nevents_sample[r,] <- sum(dsample$y)
Npats_sample[r,] <- length(dsample[,1])
Pevents_sample[r,] <- mean(dsample$y)

#save model parameters
mat_sd_ranef0[r,]<- sigma.hat(m0)$sigma$center
mat_sd_ranef1[r,]<- sigma.hat(m1)$sigma$center
mat_regcof[r,]<- as.numeric(fixef(m1))
mat_SEregcof[r,]<- se.coef(m1)$fixef
mat_ranef[r,] <- clustereff_s
mat_SEranef[r,] <- clustereff_s_SE
mat_dims[r,] <- as.numeric(attr(m1,"dims"))
Niterm1[r,] <- length(iter_m1)-1
Equaldeviance[r,] <- Equal
mat_regcofu[r,]<- m1u$coef
mat_SEregcofu[r,]<- summary(m1u)$coefficients[, 2]
Niterm1u[r,] <- m1u$iter
Equaldevianceu[r,] <- Equalu

#save log likelihood of full and null model
loglik0[r,]<-logLik(m0)
loglik1[r,]<-logLik(m1)
loglik0u[r,]<-logLik(m0u)
loglik1u[r,]<-logLik(m1u)

#Step 6. Evaluate model performance in the test set
        #Predictions assuming average random intercept
```

```
datatest<-data.frame(cbind(data[-sample_lev1,2:9], data$y[-sample_lev1],
data$center[-sample_lev1]))
datatest$lpB <- cbind(1, as.matrix(data[-sample_lev1,2:9])) %*% fixef(m1)
names(datatest)<-c("x1","x2", "x3", "x4","x5","x6","x7","x8", "y","center", "lpB")
perfm_m1B<- lrm(datatest$y~datatest$lpB)
perfm_m1Bi<- lrm(datatest$y~offset(datatest$lpB))
realvalresultB[r,1] <- perfm_m1B$stats[6] # C
realvalresultB[r,2] <- perfm_m1B$coefficients[2] #calib slope
realvalresultB[r,3] <- perfm_m1Bi$coefficients[1] #calib intercept
C.w.index <- rep(NA, length(unique(data$center)))
n.comp <- rep(NA, length(unique(data$center)))
for(p in 1:length(unique(data$center))){
data_p <- datatest[datatest$center==p,]
prediction <- plogis(data_p$lpB)
outcome <- data_p$y
result <- concordance.prob.logistic(outcome, prediction)
C.w.index[p] <- result$C
n.comp[p] <- result$n.comp
}
realvalresultB[r,4] <- sum(C.w.index*n.comp/sum(n.comp, na.rm = TRUE), na.rm
= TRUE) # Cwithin
calibslope_mlB <- lmer(y~lpB+(lpB|center), family = "binomial", data=datatest)
realvalresultB[r,5] <- as.numeric(fixef(calibslope_mlB)[2]) #within calib slope
realvalresultB[r,6] <- (sigma.hat(calibslope_mlB)$sigma$center[2])^2 #var
within calib slope
calibint_mlB <- lmer(y~offset(lpB)+(1|center), family = "binomial", data=datatest)
realvalresultB[r,7] <- as.numeric(fixef(calibint_mlB)[1]) # within calib slope
realvalresultB[r,8] <- (sigma.hat(calibint_mlB)$sigma$center[1])^2 #var within
calib slope

        #Conditional predictions
ri<-ifelse(datatest$center==1,clustereff_s[1],
        ifelse(datatest$center==2,clustereff_s[2],
        ifelse (datatest$center==3, clustereff_s[3],
        ifelse (datatest$center==4, clustereff_s[4],
        ifelse (datatest$center==5, clustereff_s[5],
        ifelse (datatest$center==6, clustereff_s[6],
```

```
        ifelse (datatest$center==7, clustereff_s[7],
        ifelse (datatest$center==8, clustereff_s[8],
        ifelse (datatest$center==9, clustereff_s[9],
        ifelse (datatest$center==10, clustereff_s[10],
        ifelse (datatest$center==11, clustereff_s[11],
        ifelse (datatest$center==12, clustereff_s[12],
        ifelse (datatest$center==13, clustereff_s[13],
        ifelse (datatest$center==14, clustereff_s[14],
        ifelse (datatest$center==15, clustereff_s[15],
        ifelse (datatest$center==16, clustereff_s[16],
        ifelse (datatest$center==17, clustereff_s[17],
        ifelse (datatest$center==18, clustereff_s[18],
        ifelse (datatest$center==19, clustereff_s[19],
        ifelse (datatest$center==20, clustereff_s[20],0
        ))))))))))))))))))))
lpc<-datatest$lpB+ri
cond<-as.data.frame(cbind(datatest$center,datatest$y,lpc))
colnames(cond)<-c("center","y","lp")
perfm_m1B<- lrm(cond$y~cond$lp)
realvalresultBcond[r,1] <- perfm_m1B$stats[6] # C
realvalresultBcond[r,2] <- perfm_m1B$coefficients[2] #calib slope
perfm_m1Bi<- lrm(cond$y~offset(cond$lp))
realvalresultBcond[r,3] <- perfm_m1Bi$coefficients[1] #calib intercept
C.w.index <- rep(NA, length(unique(cond$center)))
n.comp <- rep(NA, length(unique(cond$center)))
for(p in 1:length(unique(cond$center))){
data_p <- cond[cond$center==p,]
prediction <- plogis(data_p$lp)
outcome <- data_p$y
result <- concordance.prob.logistic(outcome, prediction)
C.w.index[p] <- result$C
n.comp[p] <- result$n.comp
}
realvalresultBcond[r,4] <- sum(C.w.index*n.comp/sum(n.comp, na.rm = TRUE),
na.rm = TRUE) # Cwithin
calibslope_mlB <- lmer(y~lp+(lp|center), family = "binomial", data=cond)
```

```
realvalresultBcond[r,5] <- as.numeric(fixef(calibslope_mlB)[2]) #within calib
slope
realvalresultBcond[r,6] <- (sigma.hat(calibslope_mlB)$sigma$center[2])^2 #var
within calib slope
calibinter_mlB <- lmer(y~offset(lp)+(1|center), family = "binomial", data=cond)
realvalresultBcond[r,7] <- as.numeric(fixef(calibinter_mlB)[1]) #within calib
intercept
realvalresultBcond[r,8] <- (sigma.hat(calibinter_mlB)$sigma$center[1])^2 #var
within calib intercept

#Marginal predictions
datatest$lpBu <- cbind(1, as.matrix(data[-sample_lev1,2:9])) %*% m1u$coef
names(datatest)<-c("x1","x2", "x3", "x4","x5","x6","x7","x8", "y","center", "lpB",
"lpBu")
perfm_m1Bu<- lrm(datatest$y~datatest$lpBu)
realvalresultBu[r,1] <- perfm_m1Bu$stats[6] # C
realvalresultBu[r,2] <- perfm_m1Bu$coefficients[2] #calib slope
perfm_m1Bui<- lrm(datatest$y~offset(datatest$lpBu))
realvalresultBu[r,3] <- perfm_m1Bui$coefficients[1] #calib slope
C.w.index <- rep(NA, length(unique(data$center)))
n.comp <- rep(NA, length(unique(data$center)))
for(p in 1:length(unique(data$center))){
data_p <- datatest[datatest$center==p,]
prediction <- plogis(data_p$lpBu)
outcome <- data_p$y
result <- concordance.prob.logistic(outcome, prediction)
C.w.index[p] <- result$C
n.comp[p] <- result$n.comp
}
realvalresultBu[r,4] <- sum(C.w.index*n.comp/sum(n.comp, na.rm = TRUE),
na.rm = TRUE) # Cwithin
calibslope_mlBu <- lmer(y~lpBu+(lpBu|center), family = "binomial",
data=datatest)
realvalresultBu[r,5] <- as.numeric(fixef(calibslope_mlBu)[2]) #within calib slope
realvalresultBu[r,6] <- (sigma.hat(calibslope_mlBu)$sigma$center[2])^2 #var
within calib slope
```

```
calibinter_mlBu <- lmer(y~offset(lpBu)+(1|center), family = "binomial",
data=datatest)
realvalresultBu[r,7] <- as.numeric(fixef(calibinter_mlBu)[1]) # within calib
intercept
realvalresultBu[r,8] <- (sigma.hat(calibinter_mlBu)$sigma$center[1])^2 #var
within calib intercept

#write convergence results
write.table(cbind(r, Niterm1[r,], iter_m1), file="path\\iter", col.names=F,
row.names=F, append=T)
write.table(x=Equaldeviance, file=" path \\EPV 100 ICC 20%\\Equaldeviance",
append=F, na="NA", col.names=F)
write.table(cbind(r, Niterm1u[r,], iter_m1), file=" path \\iteru", col.names=F,
row.names=F, append=T)
write.table(x=Equaldevianceu, file=" path \\Equaldevianceu", append=F,
na="NA", col.names=F)

} #end of the simulation loop

#Save all results
```

# Appendix 6. Detailed simulation results: calibration

| Level of performance evaluation | Calibration measure | Type of prediction | | |
|---|---|---|---|---|
| | | Marginal predictions | Predictions assuming an average random center intercept | Conditional predictions |
| Conditional (center level) | Calibration slope (1 is ideal) | 1.09 (0.05) [<0.0005] | 0.98 (0.05) [<0.0005] | 0.99 (0.05) [<0.0005] |
| | Calibration intercept (0 is ideal) | -0.13 (0.02) [0.83] | -0.01 (0.04) [0.89] | -0.01 (0.03) [0.05] |
| Standard (population level) | Calibration slope (1 is ideal) | 0.99 (0.05) | 0.88 (0.04) | 0.98 (0.04) |
| | Calibration intercept (0 is ideal) | -0.00 (0.02) | 0.12 (0.03) | -0.00 (0.03) |

Table A2. Calibration slopes and intercepts at the center and population level for marginal predictions from the standard logistic regression model and predictions assuming an average random center intercept and conditional predictions from the random intercept logistic regression model. Results are presented as mean (sd) [mean between-center variance of the random calibration intercept or slope].

## Appendix 7. Simulation results: calibration intercepts

The within-center calibration intercepts were below zero for the predictions from the standard model, indicating that on average, predicted probabilities were too high. The population-level calibration intercepts were above zero for the predictions assuming an average random intercept, indicating that on average, predicted probabilities were too low. Note that in our simulations, the "true" prediction model had a negative intercept. In the case of a positive intercept, within-center calibration intercepts larger than zero may be expected for the predictions of the standard model, and population-level calibration intercepts smaller than zero may be expected for the predictions assuming an average random intercept (results not shown).

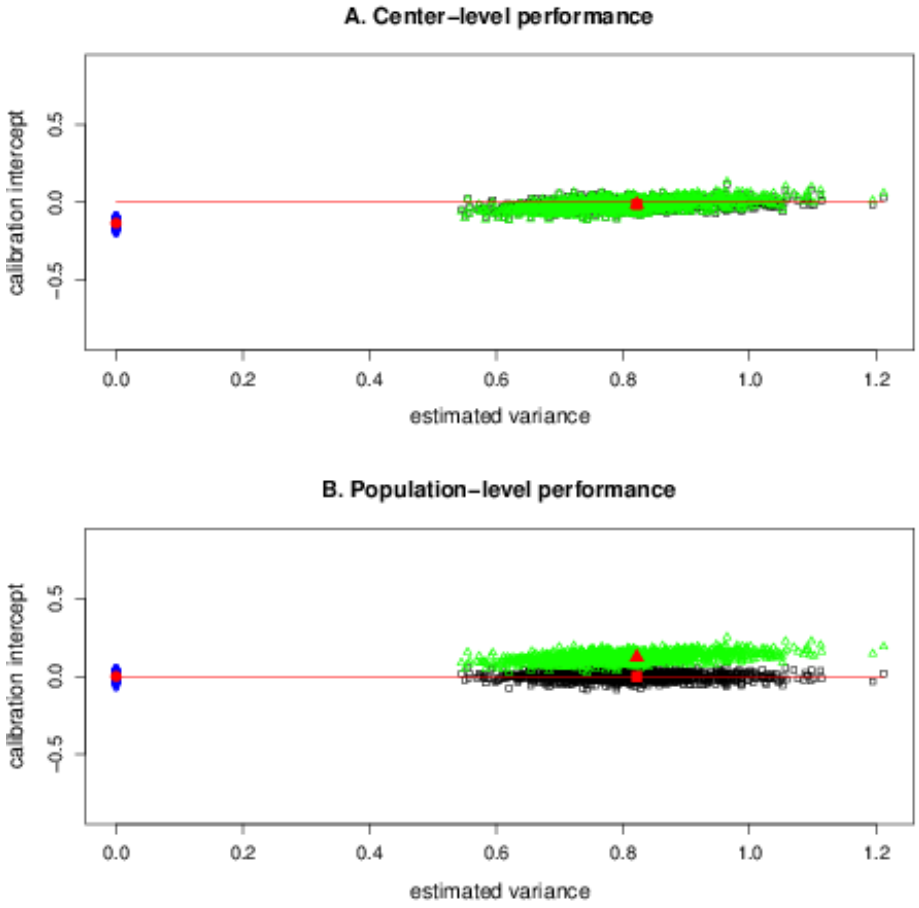


Figure A6. Center-level (panel A) and population-level (panel B) calibration intercepts of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model from the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable 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 intercept.

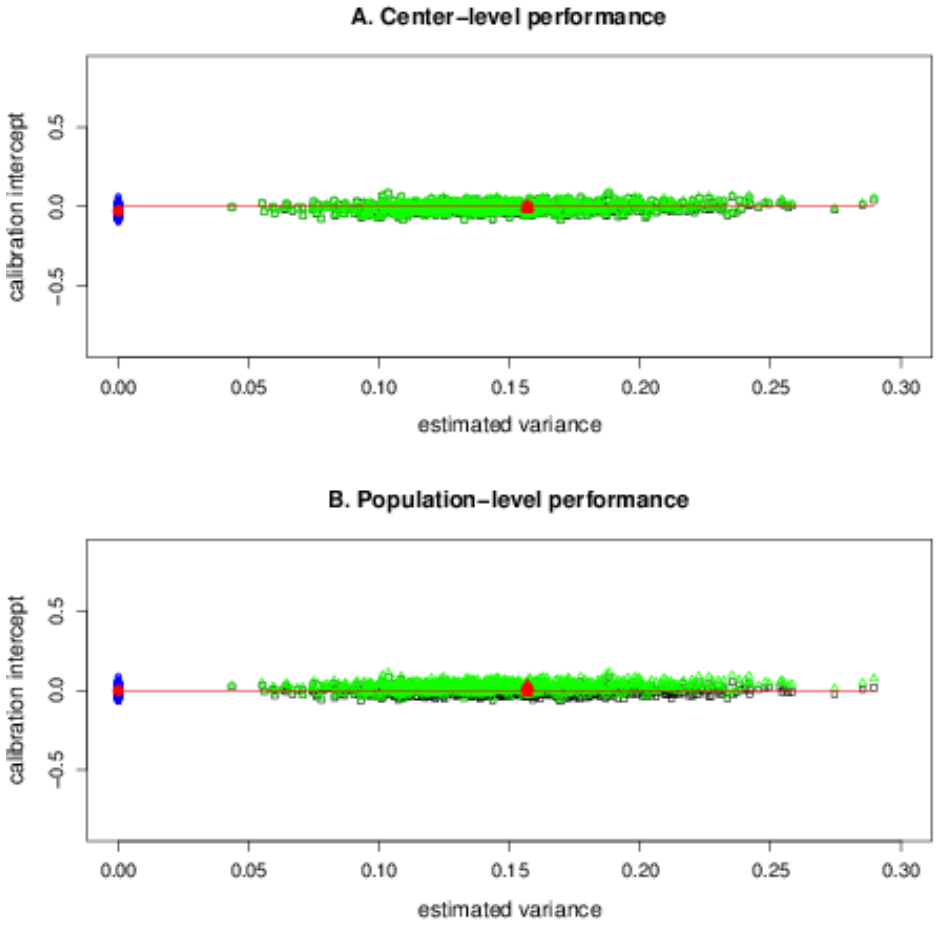


Figure A7. Center-level (panel A) and population-level (panel B) calibration intercepts of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model from the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable 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 intercept.

## Appendix 8. Simulation results: C-indexes

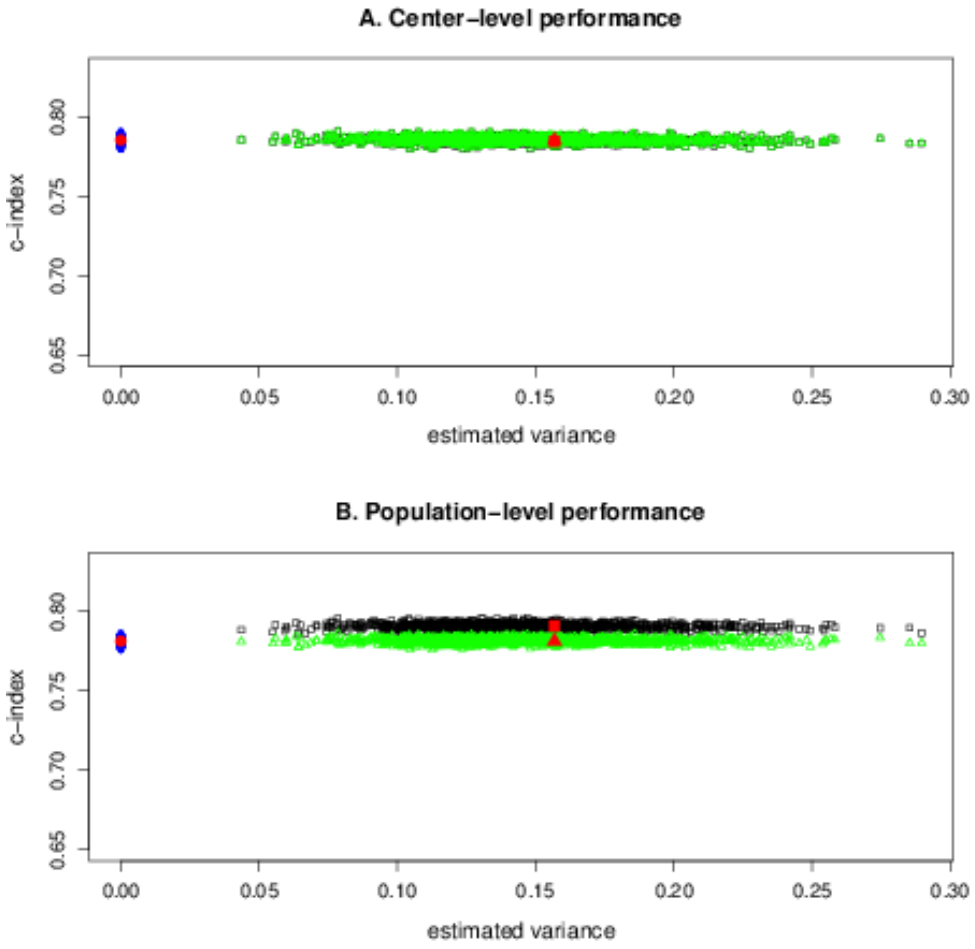


Figure A8. Center-level (panel A) and population-level (panel B) C-indexes of the standard logistic regression model (circles), the conditional linear predictor of the random intercept model (squares) and the linear predictor assuming an average random intercept (triangles), by estimated random intercept variance in samples with 100 events per variable and true random effects variance=0.159 (ICC=5%). 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 (0.159) for the random intercept model.

# Appendix 9. Case study results

| | Mixed effects model $\widehat{\beta}$ (se) | Standard model $\widehat{\beta}$ (se) |
|---|---|---|
| Intercept | -4.729 (0.280) | -4.520 (0.203) |
| Age (in years) | 0.036 (0.004) | 0.035 (0.003) |
| Proportion of solid tissue | 3.655 (0.189) | 3.614 (0.181) |
| More than 10 locules (yes/no) | 1.858 (0.191) | 1.728 (0.183) |
| Number of papillations (0,1,2,3,>3) | 0.633 (0.045) | 0.592 (0.043) |
| Presence of acoustic shadows (yes/no) | -2.151 (0.242) | -2.161 (0.232) |
| Presence of ascites (yes/no) | 2.975 (0.203) | 2.848 (0.193) |

Table A3. Estimated regression coefficients from a mixed effects logistic regression model and a standard logistic regression model to predict the risk of ovarian mass malignancy.

| | Population-level C | Population-level calibration slope | Population-level calibration intercept | Within-center C | Within-center calibration slope (variance) | Within-center calibration intercept (variance) |
|---|---|---|---|---|---|---|
| Marginal | 0.90 | 0.95 | 0.62 | 0.88 | 0.97 (0.014) | 0.41 (0.554) |
| Average cluster effect | 0.90 | 0.91 | 0.70 | 0.88 | 0.94 (0.015) | 0.48 (0.582) |
| Marginalized from mixed effects | 0.90 | 0.99 | 0.60 | 0.88 | 1.02 (0.017) | 0.39 (0.529) |
| Conditional | 0.91 | 0.88 | 0.29 | 0.88 | 0.93 (0.009) | 0.27 (0.248) |

Table A4. Population-level and center-level discrimination and calibration statistics for the marginal predicted risks from a standard logistic regression model and for the predicted risks assuming an average random intercept, the marginalized

predicted risks and the conditional predicted risks from a mixed effects standard logistic regression model for ovarian tumor malignancy.