

RESEARCH ARTICLE

Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective

K. Luijken¹  | R. H. H. Groenwold^{1,2}  | B. Van Calster^{2,3}  | E. W. Steyerberg^{2,4}  | M. van Smeden¹ 

¹Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

²Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

³Department of Development and Regeneration, University of Leuven, Leuven, Belgium

⁴Department of Public Health, Erasmus University Medical Center, Rotterdam, The Netherlands

Correspondence

K. Luijken, Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands. Email: K.Luijken@lumc.nl

Funding information

Netherlands Organisation for Scientific Research, Grant/Award Number: ZonMW, project 917.16.430; Patient-Centered Outcomes Research Institute (PCORI), Grant/Award Number: ME-1606-35555

It is widely acknowledged that the predictive performance of clinical prediction models should be studied in patients that were not part of the data in which the model was derived. Out-of-sample performance can be hampered when predictors are measured differently at derivation and external validation. This may occur, for instance, when predictors are measured using different measurement protocols or when tests are produced by different manufacturers. Although such heterogeneity in predictor measurement between derivation and validation data is common, the impact on the out-of-sample performance is not well studied. Using analytical and simulation approaches, we examined out-of-sample performance of prediction models under various scenarios of heterogeneous predictor measurement. These scenarios were defined and clarified using an established taxonomy of measurement error models. The results of our simulations indicate that predictor measurement heterogeneity can induce miscalibration of prediction and affects discrimination and overall predictive accuracy, to extents that the prediction model may no longer be considered clinically useful. The measurement error taxonomy was found to be helpful in identifying and predicting effects of heterogeneous predictor measurements between settings of prediction model derivation and validation. Our work indicates that homogeneity of measurement strategies across settings is of paramount importance in prediction research.

KEYWORDS

Brier score, calibration, discrimination, external validation, measurement error, measurement heterogeneity, prediction model

1 | INTRODUCTION

Prediction models have an important role in contemporary medicine by providing probabilistic predictions of diagnosis or prognosis.¹ Prediction models need to provide accurate and reliable predictions for patients that were not part of the dataset in which the model was derived (ie, derivation set).² The ability of a prediction model to predict in future patients

.....
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2019 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

TABLE 1 Possible sources of measurement heterogeneity in measurements of predictors, illustrated by examples from previously published prediction studies

Type of Predictor	Examples of Predictors	Examples of Measurement Heterogeneity
Anthropometric measurements	Height Weight Body circumference	Guidelines on imaging decisions in osteoporosis care are established using standardized measurements of height, while in clinical practice height is measured using nonstandardized techniques or self-reported values. ⁸
Physiological measurements	Blood pressure Serum cholesterol HbA1c Fasting glucose	In scientific studies, blood pressure is often measured by the average of multiple measurements performed under standardized conditions, while blood pressure measurements in practice deviate from protocol guidelines in various ways due to variability in available time and devices. ⁹
Diagnosis	Previous/current disease	The diagnosis “hypertension” can be defined as a blood pressure of $\geq 140/90$ mm Hg (without use of anti-hypertensive therapy) or as the use of anti-hypertensive drugs. ¹⁰
Treatment/ Exposure status	Type of drug used Smoking status Dietary intake	The cut-off value for an “increased length of stay in the hospital” to predict unplanned readmission may depend on the country in which the model is evaluated. ¹¹
Imaging	Presence or size of tissue on ultrasound, MRI, CT or FDG PET scans	In scientific studies, review of FDG PET scans may be protocolized or performed by a single experienced nuclear medicine physician, blinded to patient outcome. ¹² In routine practice, FDG PET scans may be reviewed under various systematics or by a multidisciplinary team. ¹³

(ie, out-of-sample) can be evaluated in an external validation study. While out-of-sample predictive performance is in general expected to be lower than performance estimated at derivation,¹ large discrepancies are often contributed to sub-optimal modeling strategies in the derivation of the model³⁻⁵ and differences between patient characteristics in derivation and validation samples.^{6,7}

Another potential source of limited out-of-sample performance is when predictors are measured differently at derivation than at (external) validation. This may occur, for instance, when predictors are categorized using different cut-off values or when predictors are based on diagnostic tests that were produced by different manufacturers (see Table 1 for examples). Although some studies have mentioned that such heterogeneity in predictor measurements might hamper out-of-sample model performance (eg,^{14,15}), effects of measurement heterogeneity in prediction studies have received little attention. Particularly, its impact on predictive performance has not been formally quantified.

In this study, we investigate the out-of-sample performance of a clinical prediction model in situations where predictor measurement strategies at the model derivation stage differed from measurement strategies at the model validation stage. The different scenarios of heterogeneous predictor measurement were defined using a well-known taxonomy of measurement error models, described by, eg, Keogh and White.¹⁶ We varied the degree of measurement error in the derivation data and validation data to recreate qualitative differences in the predictor measurement structures across settings. Hence, the measurement error perspective serves as a framework to define predictor measurement heterogeneity. We focus on logistic regression, since this model is widely applied in clinical prediction research.¹⁷

This paper is structured as follows. In Section 2, we define the measurement error models used to describe scenarios of measurement heterogeneity. In Section 3, we derive analytical expressions to identify and predict effects of measurement error on in-sample predictive performance. In Section 4, we illustrate the effects of measurement heterogeneity across settings on predictive performance in large sample simulations and contrast these to the impact of measurement error within the derivation setting. In Section 5, we present an extensive set of Monte Carlo simulations in finite samples to examine the impact of measurement heterogeneity on out-of-sample predictive performance. We end with discussing the implications of our findings in Section 6.

2 | EXPRESSING MEASUREMENT HETEROGENEITY IN TERMS OF MEASUREMENT ERROR MODELS

Consider a random sample of N independent individuals $i = 1, \dots, N$. Let Y be a binary response variable with values $y_i \in \{0, 1\}$. We define a logistic regression model for estimating the probability that $Y = 1$ given values of a set of P continuous predictor variables, $\mathbf{X} = \{X_1, \dots, X_P\}$. The probability of observing an event ($Y = 1$) given the predictors, $\pi_i = P(Y_i = 1 | \mathbf{X}_i)$, is defined as

$$\pi_i = \frac{1}{1 + \exp(-(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i))},$$

where α is an intercept (scalar), β is a P -dimensional vector of regression coefficients.

For simplicity of presentation, we consider a single vector $X \subset \mathbf{X}$. To distinguish different measurements of the same predictor, we denote an exact measurement of the predictor (eg, bodyweight measured on a scale) by X and a pragmatic measurement (eg, self-reported weight) by W . In most measurement error literature, X denotes an error-free true value and W denotes an observed error-prone version of X .¹⁸ However, for prediction purposes, it is hardly ever feasible (or even undesirable) to obtain error-free measurements in clinical practice, and hence we use the terms exact measurement for X and pragmatic measurement for W . The connection between X and W can be formally defined using measurement error models. We define a general model of measurement heterogeneity for continuous predictors in line with existing measurement error literature.^{16,18} Assuming that the relation between X and W is linear and additive, the association between W and X can be described as

$$\begin{aligned}\mathbb{E}(W|Y = y) &= \psi_{Y=y} + \theta_{Y=y}\mathbb{E}(X) + \epsilon_{Y=y}, \\ \text{Var}(W|Y = y) &= \theta_{Y=y}^2\sigma_X^2 + \sigma_{\epsilon_{Y=y}}^2,\end{aligned}\quad (1)$$

where $\epsilon_{Y=y} \sim \mathcal{N}(0, \sigma_{\epsilon_{Y=y}}^2)$ and all parameters may depend on the value of Y , indicating that measurements can differ between individuals in which the outcome is observed (cases) and individuals in which the outcome is not observed (non-cases). The parameter ψ reflects the mean difference between X and $W|Y = y$, θ indicates the linear association between measurement $W|Y = y$ and X , and σ_{ϵ}^2 reflects variance introduced by random deviations in the measurement process, where a larger σ_{ϵ}^2 indicates that the measurement W is less precise. The term *measurement error* applies to situations where both an exact measurement and a pragmatic measurement of a predictor are available within a setting (eg, the derivation set), and thus where the parameters ψ , θ and σ_{ϵ}^2 define the degree of measurement error in W with respect to X . The term *measurement heterogeneity* refers to situations where the same predictor is measured heterogeneously across settings of derivation and validation. The most precise measurement (whether available at derivation or validation) corresponds to X and the parameters ψ , θ and σ_{ϵ}^2 define the degree of heterogeneity between X and W . We now consider three types of measurement error models that are particular forms of Equation (1), based on which we specify both within-sample measurement error and measurement heterogeneity across settings.

Random measurement error model

Under $\psi = 0$ and $\theta = 1$, Equation (1) reduces to the following model:

$$\mathbb{E}(W) = \mathbb{E}(X) + \epsilon, \quad (2)$$

where $\epsilon \sim \mathcal{N}(0, \sigma_{\epsilon}^2)$ is independent of X and Y . This is referred to as the random or classical measurement error model.^{16,18} W is a mean-unbiased measurement of X , since $\mathbb{E}(W|Y) = \mathbb{E}(W) = \mathbb{E}(X)$. An example of a predictor measurement corresponding to the random measurement error model is reading body weight from the same scale. Each reading, the value may deviate slightly upwards or downwards, resulting in random deviations. Variation in the size of these deviations across settings due to precision of the available scales is an example of random measurement heterogeneity.

Systematic measurement error model

When $\psi \neq 0$ and/or $\theta \neq 1$, yet when ψ and θ have the same values for cases and noncases, predictor measurements correspond to a systematic measurement error model.¹⁶ The systematic measurement error model is defined as

$$\mathbb{E}(W) = \psi + \theta\mathbb{E}(X) + \epsilon, \quad (3)$$

where $\epsilon \sim \mathcal{N}(0, \sigma_{\epsilon}^2)$ is independent of X and Y . It follows that W is no longer a mean-unbiased measurement of X ($\mathbb{E}(W) \neq \mathbb{E}(X)$). Systematic measurement heterogeneity may occur, for example, when a blood glucose monitor is replaced by a monitor from a different manufacturer that is calibrated differently. The switch in measurement instrument may introduce a shift by a constant in the measured predictor values, ie, a change in ψ (additive systematic measurement error). Furthermore, observed values may depend on the actual value of a predictor, where θ represents linear dependencies between X and W . For instance, values of self-reported weight may be underreported, especially by individuals with a higher actual weight, ie, $\theta < 1$ (multiplicative systematic measurement error). The size of ψ and θ can differ across settings, for example when weight is measured using a scale in one setting (eg, θ might be close to 1) and as a self-reported value in another setting (eg, θ might deviate from 1), which would result in systematic measurement heterogeneity.

Differential measurement error model

In case measurement procedures differ between cases and noncases, ie, when $\psi_1 \neq \psi_0$ and/or $\theta_1 \neq \theta_0$ and/or $\sigma_{\epsilon_1}^2 \neq \sigma_{\epsilon_0}^2$, the measurements can be described by Equation (1) above, also referred to as differential measurement error.¹⁶ Differential measurement of predictors is conceivable in settings where assessment of predictors are done in an unblinded fashion, such as case-control studies.¹⁹ For example, when patient history is collected after observing the outcome event, cases may be more likely to recall health information prior to the outcome event than noncases, also known as recall bias.²⁰ This may for example lead to over-reporting in cases, ie, $\psi_1 > \psi_0$, a stronger association between reported and actual predictor values, ie, $\theta_1 > \theta_0$, or more precise predictor measurements, ie, $\sigma_{\epsilon_1}^2 < \sigma_{\epsilon_0}^2$, in cases than in noncases. Prospective differential measurement error may occur when a prediction model influences the way that predictors are measured in clinical practice. After clinical uptake of a prediction model, physicians may measure predictors differently in patients in whom they suspect the outcome of interest (potential future cases), guided by the knowledge that these particular predictors are of importance. For example, in these patients, body weight may be measured using a scale, whereas the prediction model may have been derived from self-reported measurements of body weight, introducing a difference between measurement procedures of (potential) cases and noncases (ie, differential measurement error), as well as a difference in measurement strategy between derivation and application setting (ie, differential measurement heterogeneity).

3 | PREDICTIVE PERFORMANCE UNDER WITHIN-SAMPLE MEASUREMENT ERROR

In this section, we define analytical expressions that indicate how substituting an exact predictor measurement, X , with a pragmatic predictor measurement, W , affects apparent predictive performance in the situation where both measurements X and W are available in the derivation sample of a prediction model. For brevity, we will evaluate a single-predictor model. Expressions of in-sample predictive performance under random measurement error were previously derived by Khudyakov and colleagues for a probit prediction model.²¹ The current paper extends these expressions to a logistic regression model. We measure predictive performance by the concordance-statistic (c-statistic) and Brier score, measuring discrimination and overall accuracy, respectively. Effects on calibration will be evaluated in the next sections. We will discuss expressions in terms of sample realizations, that is, realizations y_i , x_i and w_i . In the following, let $\bar{x} = \frac{1}{n} \sum_{i=1}^n (x_i | y_i)$ and s_x^2 denote the sample mean and variance of x , let $\bar{w} = \frac{1}{n} \sum_{i=1}^n (w_i | y_i)$ and s_w^2 denote the sample mean and variance of w , and let n_1 and n_0 denote the number of cases and noncases in the sample, respectively.

3.1 | C-statistic

To examine the discriminatory performance, we make use of the c-statistic, a rank-order statistic that typically ranges from 0.5 (no discrimination) to 1 (perfect discrimination) and is equal to the area under the receiver operating characteristic (ROC) curve for a binary outcome.¹⁷ Consider a data-generating model relating response variable Y to X by a logit link function, where $X|Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ (binormality). Let $\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i | y_i = 1)$ denote the sample mean of x for cases, let $\bar{x}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} (x_i | y_i = 0)$ denote the sample mean of x for noncases, and let $s_{x1}^2 + s_{x0}^2$ denote the total variance of x . Let Φ denote the cumulative distribution function of the standard normal distribution. Following Austin and Steyerberg,²² the c-statistic is approximated by

$$AUC_x = \Phi \left(\frac{\bar{x}_1 - \bar{x}_0}{\sqrt{s_{x1}^2 + s_{x0}^2}} \right).$$

Alternatively, for w , let $\bar{w}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} (w_i | y_i = 1)$ and $\bar{w}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} (w_i | y_i = 0)$ denote the sample means of w for cases and noncases, respectively, and let $s_{w1}^2 + s_{w0}^2$ denote the total variance of w . The c-statistic of a binary logistic regression model of the predictor w is then given by:

$$AUC_w = \Phi \left(\frac{\bar{w}_1 - \bar{w}_0}{\sqrt{s_{w1}^2 + s_{w0}^2}} \right). \quad (4)$$

Under the general measurement error model (Equation 1),

$$\begin{aligned}\bar{w}_0 &= \psi_0 + \theta_0 \bar{x}_0, \\ \bar{w}_1 &= \psi_1 + \theta_1 \bar{x}_1, \\ s_{w0}^2 &= s_{x0}^2 \theta_0^2 + s_{\epsilon_0}^2, \\ s_{w1}^2 &= s_{x1}^2 \theta_1^2 + s_{\epsilon_1}^2.\end{aligned}$$

The impact of measurement error on the c-statistic can now be expressed as

$$\begin{aligned}\Delta AUC &= AUC_w - AUC_x \\ &= \Phi \left(\frac{(\psi_1 + \theta_1 \bar{x}_1) - (\psi_0 + \theta_0 \bar{x}_0)}{\sqrt{s_{x1}^2 \theta_1^2 + s_{\epsilon_1}^2 + s_{x0}^2 \theta_0^2 + s_{\epsilon_0}^2}} \right) - \Phi \left(\frac{\bar{x}_1 - \bar{x}_0}{\sqrt{s_{x1}^2 + s_{x0}^2}} \right),\end{aligned}\quad (5)$$

where a $\Delta AUC < 0$ indicates that the model has less discriminatory power when w is used instead of x . Equations (4) and (5) indicate that the expected impact of substituting x by w in prediction model development has the following consequences. In case of random measurement error in w , it can be expected that the model fitted on w has a lower c-statistic and $\Delta AUC < 0$. In case of systematic measurement error in w , the c-statistic is not affected beyond random measurement error. Differential measurement error can affect model discrimination in both directions. For example, when observed measurements w are systematically shifted further from x in cases, ie, when $\psi_1 > \psi_0$ and $\theta_1 = \theta_0 = 1$, and when the difference in mean predictor values between cases and noncases in x is positive, ie, $\bar{x}_1 > \bar{x}_0$ and $AUC_x > 0.5$, the mean difference in predictor values between cases and noncases, $\bar{w}_1 - \bar{w}_0$, increases, enlarging the discriminatory power of the model, ie, $\Delta AUC > 0$. Additional random measurement error affects the c-statistic irrespective of whether the error is differential or not.

3.2 | Brier score

As a measure of overall predictive accuracy we evaluate the Brier score, which is a proper scoring rule that indicates the distance between predicted and observed outcomes. The Brier score is calculated by²³

$$BS(x) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\pi}(x_i))^2, \quad (6)$$

where $\hat{\pi}(x_i) = (1 + \exp(-(\hat{\alpha}_x + \hat{\beta}_x x_i)))^{-1}$ and a lower Brier score indicates higher accuracy of predictions. Following Blattenberger and Lad²⁴ and Spiegelhalter,²⁵ the Brier score can be decomposed into

$$BS(x) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\pi}(x_i))(1 - 2\hat{\pi}(x_i)) + \frac{1}{n} \sum_{i=1}^n \hat{\pi}(x_i)(1 - \hat{\pi}(x_i)), \quad (7)$$

resulting in a calibration component, $(y_i - \hat{\pi}(x_i))(1 - 2\hat{\pi}(x_i))$, and a refinement component, $\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))$. As Spiegelhalter already noted,²⁵ the calibration component has an expectation of 0 under the null hypothesis of perfect calibration, that is $\mathbb{E}_0(Y_i) = \hat{\pi}(x_i)$, and the expected Brier score can be expressed by the refinement term in Equation (7), that is $\mathbb{E}_0(BS(x)) = \frac{1}{n} \sum_{i=1}^n \hat{\pi}(x_i)(1 - \hat{\pi}(x_i))$. Consequently, the impact of within-sample measurement error on the Brier score of a maximum likelihood model in the derivation set can be expressed as

$$\mathbb{E}_0(\Delta BS) = \frac{1}{n} \sum_{i=1}^n \hat{\pi}(w_i)(1 - \hat{\pi}(w_i)) - \frac{1}{n} \sum_{i=1}^n \hat{\pi}(x_i)(1 - \hat{\pi}(x_i)), \quad (8)$$

where

$$\hat{\pi}(w_i) = \frac{1}{1 + \exp(-(\hat{\alpha}_w + \hat{\beta}_w(\psi_{Y=y} + x_i \theta_{Y=y} + \epsilon_{Y=y}))},$$

where a $\mathbb{E}_0(\Delta BS) > 0$ indicates that substituting x with w yields less accurate predictions. Realistically, however, a model is hardly ever perfectly calibrated (see²⁶ for an in-depth discussion of levels of calibration of prediction models). A maximum likelihood estimate of a logistic regression model attains “weak calibration” in its derivation sample by definition, meaning that no systematic overfitting or underfitting and/or overestimation or underestimation of risks occurs. In the remaining

of this paper we use the term “calibration” instead of “weak calibration” and use the term “Brier score” to refer to the decomposed empirical Brier score in Equation (7).

Expression (8) indicates that substituting x with w in a perfectly specified model has the following consequences. When the association between w and outcome y is weaker than the association between x and y , a prediction model based on w provides less extreme predicted probabilities. This results in a larger refinement term for w , ie, $\frac{1}{n} \sum_{i=1}^n \hat{\pi}(w_i)(1 - \hat{\pi}(w_i))$ is larger, and in a positive $\mathbb{E}_0(\Delta BS)$ and hence lower accuracy.

4 | MEASUREMENT ERROR VERSUS MEASUREMENT HETEROGENEITY

The expressions of predictive performance under measurement error indicate that more erroneous predictor measurements lead to less apparent discriminatory power and accuracy. However, these results cannot be generalized directly to effects of measurement error on out-of-sample performance of prediction models. We use the measurement error model taxonomy to explore how heterogeneity in measurement structures affects out-of-sample performance. Rather than distinguishing error-free and error-prone predictor measurements, the measurement error models now express deviations from homogeneity of measurements across settings.

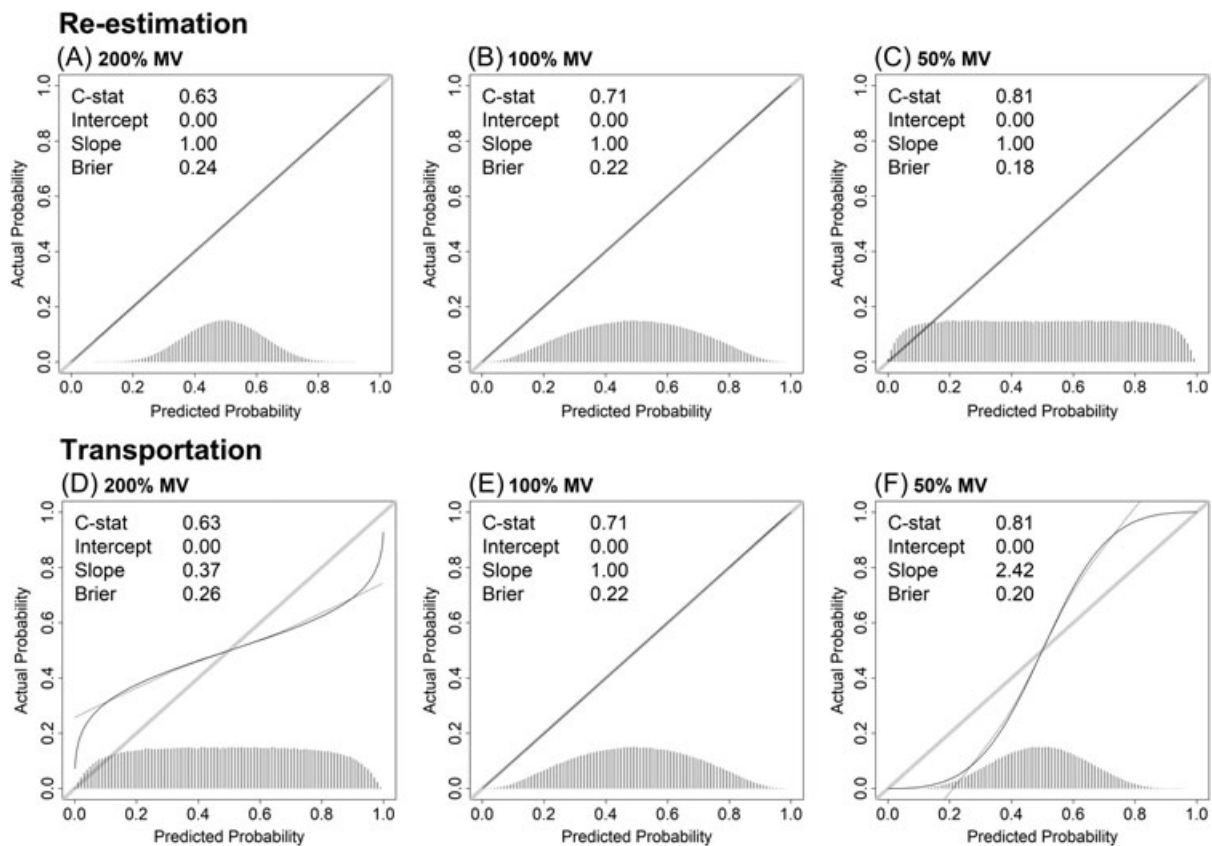


FIGURE 1 Measures of predictive performance under predictor measurement error and predictor measurement heterogeneity. The data-generating mechanism corresponded perfectly to the estimated logistic regression model. The top rows show calibration plots of a single-predictor model that is fitted using predictor measurement x and validated by re-estimating the model on the same data using w . The bottom rows show situations where the same model is transported from derivation to validation setting, specifically, the model (D) is derived using x and validated using w , (E) is derived and validated using w , and (F) is derived using w and validated using x . The calibration plots show the calibration slope (black line) and predicted probability frequencies (bottom-histograms) for situations in which the predictor measurement variance at validation equals 200% (A,D), 100% (B,E), or 50% (C,F) of the predictor measurement variance at derivation. The `val.prob` function from the `rms` package was used to compute the simulation outcome measures and to generate the calibration plots,²⁷ where we edited the legend format settings in the plot to improve readability. MV = measurement variance of the predictor measurement used for model validation relative to the predictor measurement used for derivation

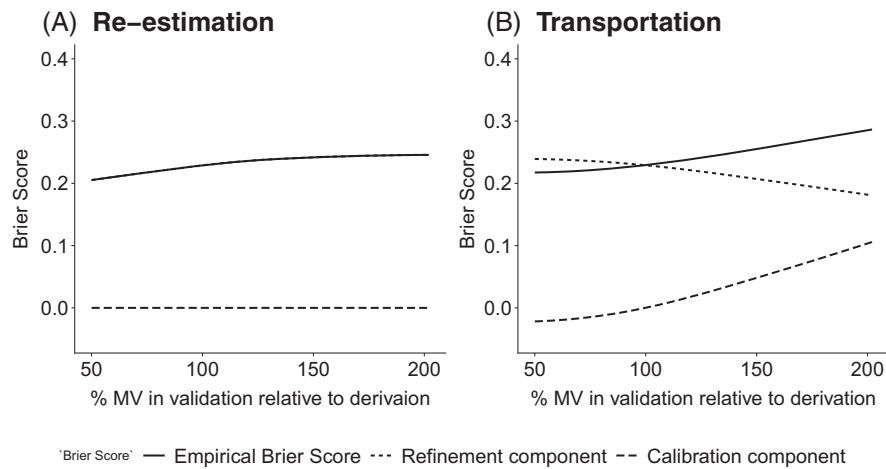


FIGURE 2 Decomposed Brier score under predictor measurement error and predictor measurement heterogeneity. The data-generating mechanism corresponded perfectly to the estimated logistic regression model. The plot displays the large sample properties of the components of the Brier score (Equation 7) under increasing random predictor measurement variance at validation, corresponding to the random measurement error model (Equation 2). The left panel shows the Brier score for a single-predictor logistic regression model that is fitted using predictor measurement x and validated by re-estimating the model on the same data using w . The right panel shows transportation from x at derivation to x at validation (up to %MV = 100) and transportation from x at derivation to w at validation (from %MV = 100 onwards). MV = measurement variance of the predictor measurement used for model validation relative to the predictor measurement used for derivation

A direct comparison of effects of measurement error and effects of measurement heterogeneity on predictive performance can be found in Figures 1 and 2, which illustrate large-sample ($N = 1\,000\,000$) properties of predictive performance measures. Effects of measurement error are illustrated by comparing in-sample predictive performance measures of a prediction model that is first estimated based on x and subsequently estimated based on w , where the latter contains increasing measurement error. Effects of measurement heterogeneity are illustrated by comparing out-of-sample predictive performance measures of a prediction model that is transported across settings with different predictor measurement structures. We explored three settings: (i) x is available at derivation and w is available at validation, (ii) w is available at both derivation and validation, and (iii) w is available at derivation and x is available at validation. In other words, this section illustrates the impact of measurement error and measurement heterogeneity as an isolated factor by evaluating the same population at both derivation and validation, and only varying the predictor measurement structures. For the purpose of demonstration, we focus on random measurement error and -heterogeneity and provide further analyses in the next section.

Additional to the c-statistic and Brier score, we evaluate calibration as a measure of predictive performance. In logistic regression, calibration can be determined using a re-calibration model, where the observed outcomes in validation data, y_V , are regressed on a linear predictor (lp).²⁸ This linear predictor is obtained by combining the regression coefficients estimated from the derivation data, $\hat{\alpha}_D$ and $\hat{\beta}_D$, with the predictor values in the validation data, x_{iV} . The recalibration model is defined as¹:

$$\text{logit}(y_V) = a + b \times \text{lp},$$

where $\text{lp} = \hat{\alpha}_D + \hat{\beta}_D x_{iV}$ and b represents the calibration slope. A calibration slope $b = 1$ indicates perfect calibration. A calibration slope $b < 1$ indicates that predicted probabilities are too extreme compared to observed probabilities,^{1,26} which is often found in situations of “statistical overfitting.” A calibration slope $b > 1$ indicates that the provided predicted probabilities are too close to the outcome incidence, also referred to as “statistical underfitting”. Additional to the calibration slope, we evaluated the difference between the average observed event rate and the mean predicted event rate (ie, calibration-in-the-large, which can be computed as the intercept of the recalibration model while using an offset for the linear predictor, ie, $a|b = 1$).¹

In situations of within-sample measurement error, ie, in the re-estimated model, all calibration plots showed a calibration slope equal to $b = 1$, indicating perfect apparent calibration (Figure 1A-C). The apparent c-statistic and Brier score improved with decreasing random measurement error. In case of measurement heterogeneity across samples, ie, in the transported model, similar changes in the c-statistic and Brier score were found. However, heterogeneous measurements led to a calibration slope $b \neq 1$, indicating that predictions were no longer valid (Figure 1D and 1F). When

measurements at validation were less precise than at derivation, the calibration slope was $b < 1$, similar to statistical overfitting. When measurements at validation were more precise than at derivation, the calibration slope was $b > 1$, similar to statistical underfitting. More elaborate illustrations of the impact of measurement heterogeneity in large sample simulations, including effects of systematic and differential measurement heterogeneity, can be found in the Supplementary File 1.

Although the total Brier score did not differ substantially between the re-estimated and transported model, the examination of the large sample properties of the decomposed Brier score (Equation 7) indicated differences in the components between the procedures (Figure 2). In the re-estimated model, the calibration term equaled zero, and the total Brier score equaled the refinement term (Figure 2A). The Brier score increased with increasing random measurement error, indicating that accuracy decreased. In the transported model, changes in the refinement term were counterbalanced by changes in the calibration term. For example, when measurements at validation were less precise than at derivation, the spread in predicted probabilities increased (refinement term in Figure 2B decreased). A decrease in the refinement term under perfect calibration would indicate that overall accuracy of the model is improving, as predicted probabilities are closer to 0 or 1. However, in the transported model this improvement was counterbalanced by a calibration term larger than zero, which indicates that predicted probabilities were too extreme compared to observed probabilities (Figure 2B).

Figures 1 and 2 illustrate that miscalibration is not introduced by measurement error per se but rather by measurement heterogeneity across settings of derivation and validation. The discrepancy in calibration between model re-estimation and model transportation can be reduced to differences in the linear predictors of the recalibration models. In case of model re-estimation, the linear predictor is expressed by

$$lp_{re-est} = \hat{\alpha}_{w(V)} + \hat{\beta}_{w(V)}w_{iV}, \quad (9)$$

indicating that the parameters $\hat{\alpha}_{w(V)}$ and $\hat{\beta}_{w(V)}$ are estimated using the predictor values measured by strategy w in the validation data. In the more realistic validation procedure in which the model is transported over different predictor measurement procedures, the linear predictor is expressed by

$$lp_{transp} = \hat{\alpha}_{x(D)} + \hat{\beta}_{x(D)}w_{iV}, \quad (10)$$

meaning that regression coefficients are estimated based on x_{iD} and that the model is validated using w_{iV} . This distinction in recalibration models sheds a different light on previous research into effects of measurement error on predictive performance. Khudyakov and colleagues derived analytically that calibration in a derivation sample is not affected by measurement error.²¹ Since their findings are based on the assumption that the linear predictor is defined as in Equation (9), previous results on the impact of measurement error on predictive performance can be interpreted as effects on in-sample predictive performance.^{21,29}

5 | PREDICTIVE PERFORMANCE UNDER MEASUREMENT HETEROGENEITY ACROSS SETTINGS

General patterns of predictive performance under measurement heterogeneity were examined in a set of Monte Carlo simulations in finite samples to evaluate their behavior under sampling variability. Simulations were performed in R version 3.3.1,³⁰ and our code is accessible online (see https://github.com/KLuijken/Prediction_Measurement_Heterogeneity_Predictor). We studied the predictive performance of a single- and a two-predictor binary logistic regression model. For the latter, we evaluated situations in which both predictors were measured heterogeneously across settings as well as situations in which one of the predictors was measured similar over settings. The data for the single-predictor model were generated from

$$\begin{aligned} \text{logit}(Y) &= \log(4)X, \\ \text{where } X &\sim \mathcal{N}(0, 1). \end{aligned}$$

The data for the two-predictor models were generated from

$$\begin{aligned} \text{logit}(Y) &= \beta^T \mathbf{X}, \\ \text{where } \mathbf{X} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{X1X2} \\ \rho_{X1X2} & 1 \end{pmatrix}\right). \end{aligned}$$

TABLE 2 Input parameters for finite sample simulations. Full-factorial simulations for the parameters ψ , θ and σ_ϵ resulted in 54 scenarios for the single-predictor model, and 162 scenarios in both the two-predictor model with and the model without a predictor that was measured homogeneously across settings. An additional 54 scenarios of differential measurement error in the single-predictor model were evaluated, resulting in a total of 432 scenarios

Factor Values		
Derivation	ψ_D	0
	θ_D	1.0
	$\sigma_{\epsilon(D)}$	0.5, 1.0, 2.0
Validation	ψ_V	0, 0.25
	θ_V	0.5, 1.0, 2.0
	$\sigma_{\epsilon(V)}$	0.5, 1.0, 2.0

The correlation between predictors, $\rho_{X_1X_2}$, varied with 0, 0.5 and 0.9. Both the β -parameters in the two-predictor models have value 2.3 in case $\rho_{X_1X_2} = 0$ or $\rho_{X_1X_2} = 0.5$, and have value 2.1 in case $\rho_{X_1X_2} = 0.9$. We varied the values of the regression coefficients in order to keep the c-statistic of the data-generating models at an approximate value of 0.80 and hence to compare predictive performance over models.²² We recreated different measurement procedures of the predictors using different specifications of the general measurement error model (Equation 1). In the derivation sample, measurements corresponded to the random measurement error model (Equation 2), while in validation various measurement structures were recreated (see Table 2 for values of input parameters). All measurements contained at least some erroneous measurement variance to generate realistic scenarios.

In total, 432 scenarios were evaluated. For each scenario, a derivation sample ($n = 2000$) and a validation sample ($n = 2000$) were generated. We did not consider smaller sample sizes, since predictive performance measures are sensitive to statistical overfitting, which would complicate the interpretation of effects of measurement heterogeneity.^{4,5} The validation procedure was repeated 10 000 times for each simulation scenario. The number of events was around 1000 in each dataset, which exceeds the minimal requirement for validation studies.^{26,31}

Simulation outcome measures

The simulation outcome measures were the average c-statistic, calibration slope, calibration-in-the-large coefficient, and Brier score. The c-statistic was computed using the `somers2` function of the `rms` package.²⁷ The calibration slope was computed by regressing the observed outcome in the validation dataset on the linear predictor, as defined in Equation (9). We evaluated calibration graphically by plotting loess calibration curves and overlaying the plots of all 10 000 resamplings.^{26,32} The calibration-in-the-large was computed as the intercept of the recalibration model, while using an offset for the linear predictor.¹ The empirical Brier score was computed using Equation (6). Additionally, we evaluated in-sample predictive performance as a reference for effects on out-of-sample performance.

5.1 | Simulation results

Identical measurement error structures at derivation and validation resulted in consistent predictive performance across settings. All out-of-sample measures of predictive performance were affected by measurement heterogeneity. Effects on predictive performance measures were largest in the single-predictor model (Table 3). The two-predictor model in which one of the predictors was measured consistently over settings (Figure 3) outperformed the model in which none of the predictors were measured consistently across settings (Figure 4). Inspection of calibration plots confirmed all patterns of miscalibration discussed below (Supplementary File 2). By and large, the impact of correlation between predictors on other parameters was minimal since the correlation structure was equal across compared settings, hence, we show combined results in the figures.

TABLE 3 Out-of-sample predictive performance measures under measurement heterogeneity in a single-predictor logistic regression model. Mean c-statistic, median calibration slope, mean calibration-in-the-large, and mean Brier score (standard deviation) at external validation of a single-predictor logistic regression model transported from a derivation set ($n = 2000$) where measurement procedures were described by the random measurement error model (Equation 2) to validation sets ($n = 2000$) with various measurement structures under Equation (1). Predictive performance measures were averaged over 10 000 repetitions. All calibration slopes in the derivation set were equal to 1.0 (0.0) and are therefore not reported

	Measurement Structure at Validation	C-statistic		Calibration Slope	Calibration-in-the-large ($\times 10$)	Brier Score	
		Derivation	Validation			Derivation	Validation
$\sigma_{\epsilon(D)}^2 < \sigma_{\epsilon(V)}^2$	$\psi = 0, \theta = 0.5$	0.745 (0.033)	0.590 (0.034)	0.247 (0.153)	-0.002 (0.006)	0.204 (0.012)	0.281 (0.033)
	$\psi = 0, \theta = 1.0$	0.745 (0.033)	0.655 (0.045)	0.380 (0.180)	0.008 (0.014)	0.204 (0.012)	0.257 (0.031)
	$\psi = 0, \theta = 2.0$	0.745 (0.033)	0.726 (0.033)	0.428 (0.125)	-0.009 (0.003)	0.204 (0.012)	0.232 (0.023)
	$\psi = 0.25, \theta = 0.5$	0.745 (0.033)	0.589 (0.034)	0.247 (0.153)	-2.202 (0.643)	0.204 (0.012)	0.283 (0.032)
	$\psi = 0.25, \theta = 1.0$	0.745 (0.033)	0.655 (0.045)	0.380 (0.180)	-2.210 (0.652)	0.204 (0.012)	0.258 (0.031)
	$\psi = 0.25, \theta = 2.0$	0.745 (0.033)	0.726 (0.033)	0.428 (0.125)	-2.205 (0.651)	0.204 (0.012)	0.233 (0.023)
$\sigma_{\epsilon(D)}^2 = \sigma_{\epsilon(V)}^2$	$\psi = 0, \theta = 0.5$	0.700 (0.068)	0.635 (0.069)	0.812 (0.291)	0.001 (0.006)	0.217 (0.020)	0.235 (0.015)
	$\psi = 0, \theta = 1.0$	0.700 (0.068)	0.700 (0.068)	1.000 (0.000)	0.001 (0.008)	0.217 (0.020)	0.218 (0.020)
	$\psi = 0, \theta = 2.0$	0.700 (0.068)	0.753 (0.042)	0.955 (0.377)	-0.002 (0.013)	0.217 (0.020)	0.204 (0.014)
	$\psi = 0.25, \theta = 0.5$	0.700 (0.068)	0.635 (0.069)	0.811 (0.293)	-1.529 (1.027)	0.217 (0.020)	0.237 (0.014)
	$\psi = 0.25, \theta = 1.0$	0.700 (0.068)	0.700 (0.068)	1.002 (0.002)	-1.530 (1.033)	0.217 (0.020)	0.219 (0.019)
	$\psi = 0.25, \theta = 2.0$	0.700 (0.068)	0.753 (0.042)	0.955 (0.377)	-1.526 (1.024)	0.217 (0.020)	0.205 (0.013)
$\sigma_{\epsilon(D)}^2 > \sigma_{\epsilon(V)}^2$	$\psi = 0, \theta = 0.5$	0.655 (0.045)	0.681 (0.045)	3.147 (1.991)	0.003 (0.007)	0.230 (0.011)	0.234 (0.009)
	$\psi = 0, \theta = 1.0$	0.655 (0.045)	0.745 (0.034)	3.106 (1.563)	0.000 (0.006)	0.230 (0.011)	0.220 (0.014)
	$\psi = 0, \theta = 2.0$	0.655 (0.045)	0.781 (0.014)	2.160 (0.969)	0.005 (0.009)	0.230 (0.011)	0.203 (0.013)
	$\psi = 0.25, \theta = 0.5$	0.655 (0.045)	0.681 (0.045)	3.156 (2.001)	-0.846 (0.528)	0.230 (0.011)	0.235 (0.008)
	$\psi = 0.25, \theta = 1.0$	0.655 (0.045)	0.745 (0.034)	3.102 (1.559)	-0.846 (0.532)	0.230 (0.011)	0.221 (0.013)
	$\psi = 0.25, \theta = 2.0$	0.655 (0.045)	0.781 (0.014)	2.159 (0.967)	-0.851 (0.535)	0.230 (0.011)	0.203 (0.013)

5.1.1 | Random measurement heterogeneity

When measurements were less precise at validation compared to derivation, ie, when $\sigma_{\epsilon(D)}^2 < \sigma_{\epsilon(V)}^2$, the c-statistic decreased and Brier score increased at validation. In the single-predictor model, the c-statistic decreased from 0.75 at derivation to 0.59 – 0.73 at validation and the Brier score increased from 0.20 at derivation to 0.23 – 0.28 at validation (Table 3, bottom rows). Furthermore, the median calibration slope at validation was smaller than 1, ranging from 0.25 – 0.43 in the single-predictor model. When measurements were more precise at validation compared to derivation, ie, when

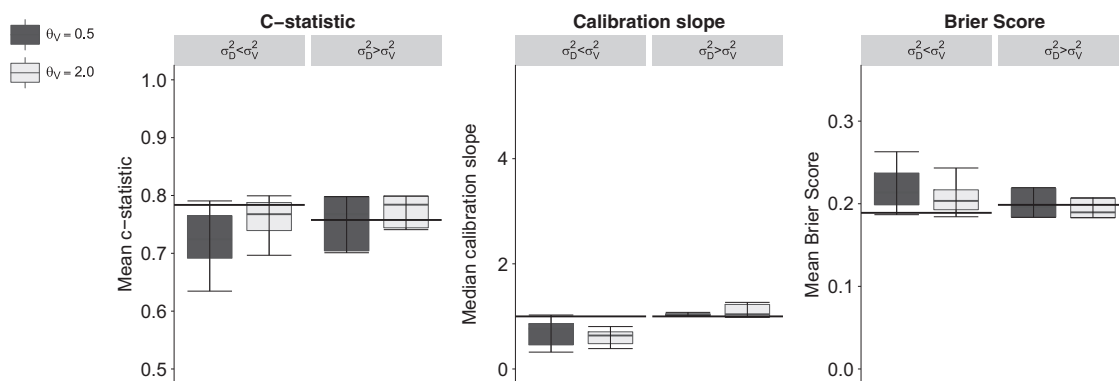


FIGURE 3 Measures of predictive performance under measurement heterogeneity in one of two predictors in finite sample simulations. Mean c-statistic, median calibration slope, and mean Brier score averaged over 10 000 repetitions with interquartile range and 95% confidence interval for a two-predictor model where one of the predictors is measured consistent across settings, whereas the other is measured heterogeneously. Horizontal bars indicate performance measures at model derivation, while boxes indicate performance at external validation. The predictor measurement structure in the derivation set ($n = 2000$) corresponds to the random measurement error model (Equation 2). In the validation set ($n = 2000$), predictor measurements consist of varying structures under Equation (1)

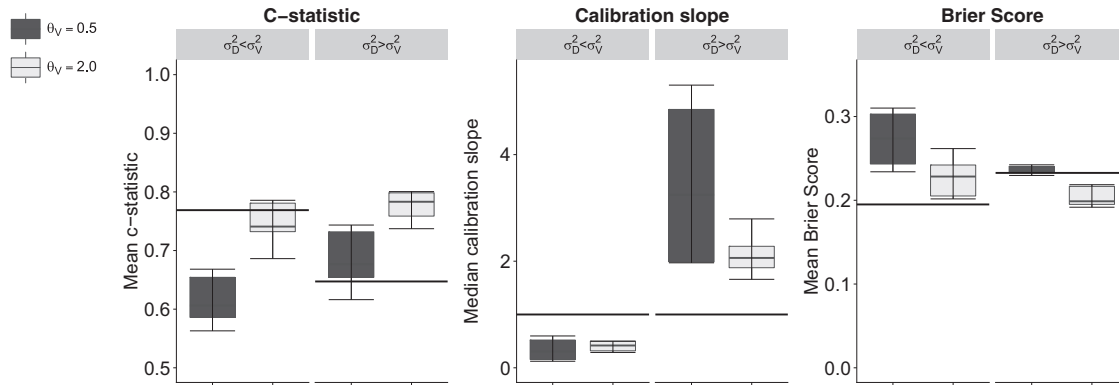


FIGURE 4 Measures of predictive performance under measurement heterogeneity in both predictors in finite sample simulations. Mean c-statistic, median calibration slope, and mean Brier score averaged over 10 000 repetitions with interquartile range and 95% confidence interval of a two-predictor logistic regression model in which both predictors are measured heterogeneously across settings. Horizontal bars indicate performance measures at model derivation, while boxes indicate performance at external validation. Measurements in the derivation set ($n = 2000$) are recreated using Equation (2), which corresponds to the random measurement error model. In the validation set ($n = 2000$), measurements correspond to various measurement structures under Equation (1)

$\sigma_{\epsilon(D)}^2 > \sigma_{\epsilon(V)}^2$, the c-statistic was increased, from 0.66 to 0.68 – 0.78 in the single-predictor model, and the Brier score was decreased, changing from 0.23 to 0.20 – 0.24 in the single-predictor model. However, the improved c-statistic and Brier score were accompanied by median calibration slopes greater than 1, ranging from 2.16 – 3.16 in the single-predictor model (Table 3, top rows). Calibration-in-the-large was not affected by random measurement heterogeneity. Similar effects on predictive performance were observed for the two-predictor models, which are presented graphically in Figures 3 and 4.

5.1.2 | Systematic measurement heterogeneity

When measurements at external validation changed by a constant compared to derivation, ie, when $\psi_D = 0$ and $\psi_V = 0.25$, the risk on observing the outcome was systematically overestimated, which is reflected in the negative value for calibration-in-the-large coefficient (Table 3). Changes in ψ had little effect on the calibration slope and Brier score, and no apparent effect on the c-statistic. Multiplicative systematic measurement heterogeneity, ie, $\theta_D \neq \theta_V$, reinforced or counterbalanced effects of random measurement heterogeneity in the direction of the systematic measurement heterogeneity. When the association between x and w was relatively weak at validation, eg, when $\theta_V = 0.5$, predictive performance deteriorated (black bars in Figures 3 and 4), whereas predictive performance improved when the association between x and w was relatively strong, eg, when $\theta_V = 2.0$ (gray bars in Figures 3 and 4).

5.1.3 | Differential measurement heterogeneity

We highlight four specific scenarios in which the single-predictor model was derived under differential random measurement error, ie, $\sigma_{\epsilon_1}^2 \neq \sigma_{\epsilon_0}^2$, and validated using nondifferential measurements, and vice versa (Table 4). Differential measurement led to miscalibration at external validation in all scenarios. The c-statistic and Brier score at validation

TABLE 4 Effects of differential measurement of predictors in events and nonevents in four scenarios. Mean c-statistic, median calibration slope, and mean Brier score (standard deviation) averaged over 10 000 repetitions for a single-predictor logistic regression model under four specific measurement error structures varying in the degree of random measurement variance under the differential measurement error model (Equation 1). By default, σ_{ϵ}^2 is set to 1.0. When $\sigma_{\epsilon_1}^2 = 0.5$, measurements are more precise in cases. When $\sigma_{\epsilon_1}^2 = 2.0$, measurements are less precise in cases

Differential Measurement Error at...		C-statistic		Calibration Slope	Brier Score	
		Derivation	Validation		Derivation	Validation
Derivation	$\sigma_{\epsilon_1}^2 = 0.5$	0.730 (0.011)	0.707 (0.012)	0.780 (0.071)	0.209 (0.004)	0.219 (0.004)
	$\sigma_{\epsilon_1}^2 = 2.0$	0.655 (0.012)	0.707 (0.012)	1.856 (0.208)	0.231 (0.003)	0.223 (0.002)
Validation	$\sigma_{\epsilon_1}^2 = 0.5$	0.706 (0.012)	0.730 (0.011)	1.293 (0.120)	0.217 (0.003)	0.211 (0.003)
	$\sigma_{\epsilon_1}^2 = 2.0$	0.706 (0.012)	0.655 (0.012)	0.547 (0.061)	0.217 (0.004)	0.237 (0.005)

TABLE 5 Key Findings. Effects of measurement heterogeneity on predictive performance in general scenarios of measurement heterogeneity. The scenarios were defined by generating different qualities of measurement across settings using the general measurement error model in Equation (1). Measurements in the derivation set corresponded to the random measurement error model (Equation 1), ie, under $\psi_D = 0$ and $\theta_D = 1.0$. Using similar logic, all patterns can be translated to differential measurement of cases and noncases (ie, when $\psi_1 \neq \psi_0$ and/or $\theta_1 \neq \theta_0$ and/or $\sigma_{\epsilon_1}^2 \neq \sigma_{\epsilon_0}^2$)

Predictor Measurements at Validation	Predictive Performance at Validation			Overall Accuracy
	Discrimination	Calibration-in-the-large	Calibration Slope	
Less precise compared to derivation;	$\sigma_{\epsilon(D)}^2 < \sigma_{\epsilon(V)}^2$	-	$b < 1$	Deteriorated
More precise compared to derivation;	$\sigma_{\epsilon(D)}^2 > \sigma_{\epsilon(V)}^2$	-	$b > 1$	Improved
Weaker association with actual predictor value, while				
- less precise compared to derivation;	$\theta_V < 1.0, \sigma_{\epsilon(D)}^2 < \sigma_{\epsilon(V)}^2$	-	Stronger $b < 1$	Stronger deterioration
- more precise compared to derivation;	$\theta_V < 1.0, \sigma_{\epsilon(D)}^2 > \sigma_{\epsilon(V)}^2$	-	Stronger $b > 1$	Less improvement
Stronger association with actual predictor value, while				
- less precise compared to derivation;	$\theta_V > 1.0, \sigma_{\epsilon(D)}^2 < \sigma_{\epsilon(V)}^2$	-	Less $b < 1$	Less deterioration
- more precise compared to derivation;	$\theta_V > 1.0, \sigma_{\epsilon(D)}^2 > \sigma_{\epsilon(V)}^2$	-	Less $b > 1$	Stronger improvement
Increased by a constant relative to derivation.	$\psi_V > 0$	$a < 0$	-	-

slightly improved when cases were measured less precise at derivation or more precise at validation. For example, when cases were measured less precise at derivation, ie, $\sigma_{\epsilon_{1(D)}}^2 > \sigma_{\epsilon_{0(D)}}^2$, the c-statistic increased from 0.66 to 0.71 at validation and the Brier score decreased from 0.23 to 0.22. However, the median calibration slope at validation was 1.86.

6 | DISCUSSION

Heterogeneity of predictor measurements across settings can have a substantial impact on the out-of-sample performance of a prediction model. When predictor measurements are more precise at derivation compared to validation, model discrimination and accuracy at validation deteriorate, and the provided predicted probabilities are too extreme, similar to when a model is overfitted with respect to the derivation data. When predictor measurements are less precise at derivation compared to validation, discrimination and accuracy at validation tend to improve, but the provided predicted probabilities are too close to the outcome prevalence, similar to statistical underfitting. These key findings of our study are summarized in Table 5. The current study emphasizes that a prediction model not only concerns the algorithm relating predictors to the outcome, but also depends on the procedures by which model input is measured, ie, qualitative differences in data collection.

Measurement error is commonly thought not to affect the validity of prediction models, based on the general idea that unbiased associations between predictor and outcome are no prerequisite in prediction studies.¹⁸ By taking the measurement error perspective, our study revealed that prediction research requires consideration of variation in measurement procedures *across* different settings of derivation and validation, rather than analyzing the amount of measurement error *within* a study. A recent systematic review by Whittle and colleagues demonstrated that measurement error was not acknowledged in many prediction studies, and pointed out the need to investigate consequences of measurement error in prediction research.³³ An important starting point for this research following from our study is that the generalizability of prediction models depends on the transportability of measurement structures.

Specification of measurement heterogeneity can help to explain discrepancies in predictive performance between derivation and validation setting in a pragmatic way. The relatedness between derivation and validation samples is generally quantified in terms of similarity in person-characteristics (also referred to as “case-mix”), and regression coefficients¹. Previously proposed measures to express sample relatedness are the mean and spread of the linear predictor⁷ or the correlation structure of predictors in both samples.³⁴ The information on sample relatedness can be incorporated in benchmark values of predictive performance to assess model transportability.⁶ While regression coefficients and case-mix distributions clearly quantify sample relatedness, it is impossible to disentangle the sources of discrepancies from these statistical measures. For example, a decrease of the regression coefficients or the spread of the linear predictor at external validation could be due to differences across settings in either person-characteristics or the means by which these characteristics were measured. Moreover, less precise predictor measurements affect both the regression coefficients and the spread of the linear predictor, meaning that measurement heterogeneity can mask similarities and differences between the individuals in a derivation and validation sample. Knowledge of substantive differences between derivation and validation setting can help researchers determining to which extent the prediction model is transportable.

In theory, measurement error correction procedures could be applied to adjust for measurement heterogeneity when data on both X and W are available.¹⁶ Alternatively, the degree of measurement heterogeneity could be quantified using the residual intraclass correlation (RICC), which expresses the clustering of measurements across physicians or centers.¹⁵ Yet, we expect that the applicability of these methods in correcting for measurement heterogeneity will be limited not only due to the fact that individual patient data of both the derivation and validation set are required, but furthermore because it is infeasible to disentangle measurement parameters from other characteristics of the data. The main contribution of the taxonomy of measurement error models rises from its aptitude to conceptualize measurement heterogeneity across settings in pragmatic terms.

The following implications for prediction studies follow from our work. Ideally, prediction models are derived from predictor measurements that resemble measurement procedures in the intended setting of application. Data collection protocols that reduce measurement error to a minimum do not necessarily benefit the performance of the model as the precision of measurements will most likely not be obtained in validation (or application) settings. Deriving a prediction model from these precise measurements could result in miscalibration similar to model overfitting and reduced discrimination and accuracy at external validation. Furthermore, researchers should bear in mind the implications of using a “readily available dataset” for model derivation or validation as data quality directly affects predictive performance of the model. For instance, validating a model in a clinical trial dataset, in which measurements typically contain minimal

measurement error, may increase measures of discrimination and accuracy, yet the model may provide predicted probabilities too close to the event rate due to miscalibration. Another example is the promising use of large routine care datasets for model validation.^{5,35,36} Predictor measurement procedures may vary greatly within such datasets or differ from the procedures used to collect the data for the derivation study, which could increase the predictor measurement variance to a level that no longer resembles the amount of measurement variance within a clinical setting. Hence, rather than analyzing data because they are available, prediction models should be derived from and validated on datasets collected with measurement procedures that are in widespread use in the intended clinical setting. Finally, it is important to clearly report which measurement procedures were used for derivation or validation of a prediction model. The influential TRIPOD Statement has drawn attention to the importance of reporting measurement procedures.¹⁴ Our findings indicate that descriptions of measurement procedures at model derivation are essential for proper external validation of the model. Likewise, the validation studies ideally contain descriptions of deviations from measurements used at derivation, as these may introduce discrepancies in predictive performance.

Our study redefines the importance of predictor measurements in the context of prediction research. We highlight heterogeneity in predictor measurement procedures across settings as an important driver of unanticipated predictive performance at external validation. Preventing measurement heterogeneity at the design phase of a prediction study, both in development and validation studies, facilitates interpretation of predictive performance and benefits the transportability of the prediction model.

ACKNOWLEDGEMENTS

The authors thank B. Rosche for his technical assistance. RG was supported by The Netherlands Organisation for Scientific Research (ZonMW, project 917.16.430). ES was supported through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1606-35555).

DISCLAIMER

All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors, or Methodology Committee.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

K. Luijken  <https://orcid.org/0000-0001-5192-8368>

R. H. H. Groenwold  <https://orcid.org/0000-0001-9238-6999>

B. Van Calster  <https://orcid.org/0000-0003-1613-7450>

E. W. Steyerberg  <https://orcid.org/0000-0002-7787-0122>

M. van Smeden  <https://orcid.org/0000-0002-5529-1541>

REFERENCES

1. Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer Science+Business Media; 2008.
2. Altman D, Royston P. What do we mean by validating a prognostic model? *Statist Med*. 2000;19(4):453-473.

3. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statist Med.* 2016;35(2):214-226.
4. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statist Med.* 2004;23(16):2567-2586.
5. Steyerberg EW, Uno H, Ioannidis JP, van Calster B. Poor performance of clinical prediction models: the harm of commonly applied methods. *J Clin Epidemiol.* 2018;98:133-143.
6. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol.* 2010;172(8):971-980.
7. Debray T, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statist Med.* 2013;32(18):3158-3180.
8. Mikula AL, Hetzel SJ, Binkley N, Anderson PA. Clinical height measurements are unreliable: a call for improvement. *Osteoporosis International.* 2016;27(10):3041-3047.
9. Drawz PE, Ix JH. BP measurement in clinical practice: time to sprint to guideline-recommended protocols. *J Am Soc Nephrol.* 2018;29(2):383-388. <https://doi.org/10.1681/ASN.2017070753>
10. Genders TS, Steyerberg EW, Hunink MM, et al. Prediction model to estimate presence of coronary artery disease: retrospective pooled analysis of existing cohorts. *Br Med J.* 2012;344:e3485.
11. Aubert CE, Folly A, Mancinetti M, Hayoz D, Donzé J. Prospective validation and adaptation of the HOSPITAL score to predict high risk of unplanned readmission of medical patients. *Swiss Med Wkly.* 2016;146:w14335.
12. Herder GJ, van Tinteren H, Golding RP, et al. Clinical prediction model to characterize pulmonary nodules: validation and added value of ¹⁸F-FDG PET. In: *The Use of 18FDG PET in NSCLC* [dissertation]. Amsterdam, The Netherlands: Vrije Universiteit Amsterdam; 2006.
13. Al-Ameri A, Malhotra P, Thygesen H, et al. Risk of malignancy in pulmonary nodules: a validation study of four prediction models. *Lung Cancer.* 2015;89(1):27-30.
14. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Medicine.* 2015;13(1).
15. Wynants L, Timmerman D, Bourne T, Van Huffel S, Van Calster B. Screening for data clustering in multicenter studies: the residual intraclass correlation. *BMC Med Res Methodol.* 2013;13(1):128.
16. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statist Med.* 2014;33(12):2137-2155.
17. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology.* 2010;21(1):128-138.
18. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective.* Boca Raton, FL: CRC Press; 2006.
19. White E. Measurement error in biomarkers: sources, assessment, and impact on studies. *IARC Sci Publ.* 2011;163:143-161.
20. Sackett DL. Bias in analytic research. In: *The Case-Control Study Consensus and Controversy.* Oxford, UK: Pergamon Press Ltd; 1979:51-63.
21. Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The impact of covariate measurement error on risk prediction. *Statist Med.* 2015;34(15):2353-2367.
22. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol.* 2012;12(1):82.
23. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.* 1950;78(1):1-3.
24. Blattenberger G, Lad F. Separating the Brier score into calibration and refinement components: a graphical exposition. *Am Stat.* 1985;39(1):26-32.
25. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statist Med.* 1986;5(5):421-433.
26. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol.* 2016;74:167-176.
27. Harrell Jr FE. rms: regression modeling strategies. R package version 4.0-0. City. 2013.
28. Cox DR. Two further applications of a model for binary regression. *Biometrika.* 1958;45(3/4):562-565.
29. Rosella LC, Corey P, Stukel TA, Mustard C, Hux J, Manuel DG. The influence of measurement error on calibration, discrimination, and overall estimation of a risk prediction model. *Popul Health Metrics.* 2012;10(1):20.
30. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
31. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol.* 2005;58(5):475-483.
32. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statist Med.* 2014;33(3):517-535.
33. Whittle R, Peat G, Belcher J, Collins GS, Riley RD. Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported. *J Clin Epidemiol.* 2018;102:38-49.
34. Kundu S, Mazumdar M, Ferket B. Impact of correlation of predictors on discrimination of risk models in development and external populations. *BMC Med Res Methodol.* 2017;17(1):63.

35. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *Br Med J*. 2016;353:i3140.
36. Cook JA, Collins GS. The rise of big clinical databases. *Br J Surg*. 2015;102(2).

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Statistics in Medicine*. 2019;38:3444–3459. <https://doi.org/10.1002/sim.8183>