Universiteit
Leiden
The Netherlands

The impact of epidemiologic methods on findings in studies of causal effects and prediction modelling

Luijken, K.

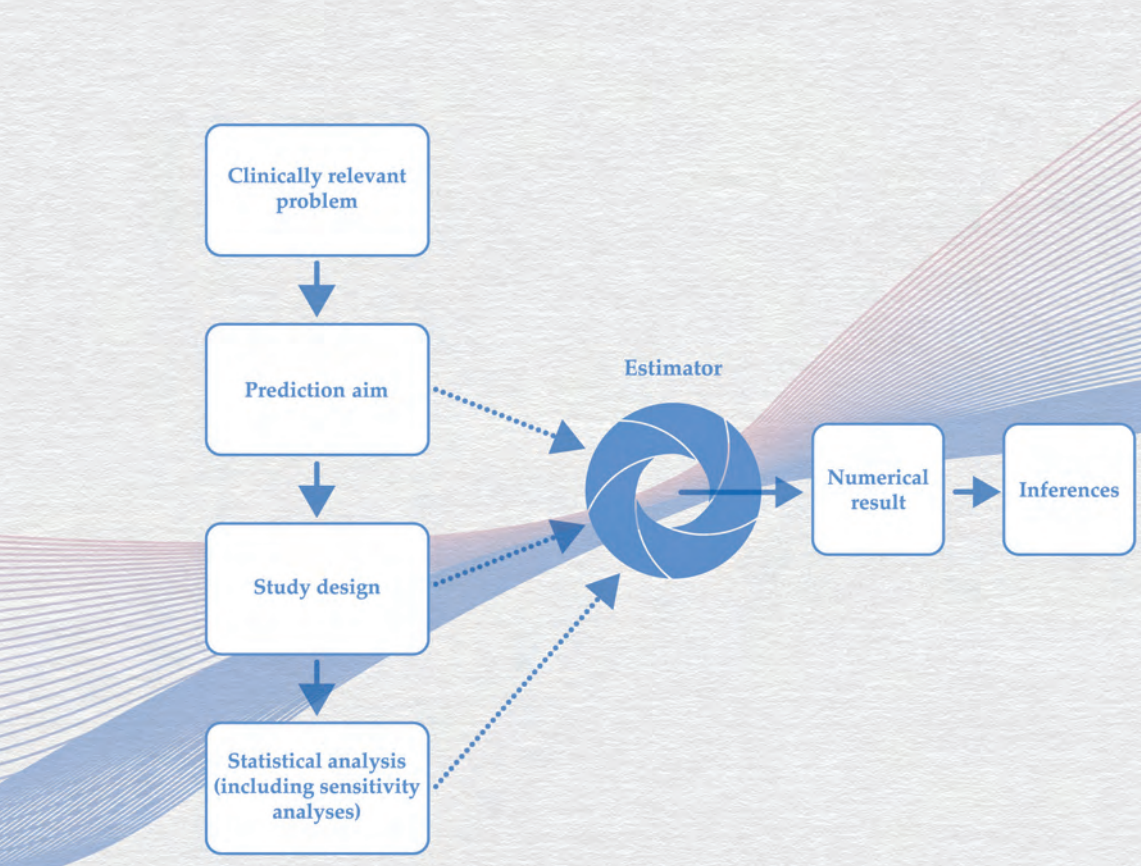
Citation

Luijken, K. (2022, May 19). *The impact of epidemiologic methods on findings in studies of causal effects and prediction modelling*. Retrieved from https://hdl.handle.net/1887/3304345

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3304345

Note: To cite this publication please use the final published version (if applicable).

8


Clinically relevant problem
Prediction aim
Study design
Statistical analysis (including sensitivity analyses)
Estimator
Numerical result
Inferences

Quantitative prediction analysis to investigate predictive performance under predictor measurement heterogeneity at model implementation

When a predictor variable is measured in similar ways at the derivation and validation setting of a prognostic prediction model, yet both differ from the intended use of the model in practice (i.e., 'predictor measurement heterogeneity'), performance of the model at implementation needs to be inferred. This study proposed an analysis to quantify the impact of anticipated predictor measurement heterogeneity. A simulation study was conducted to assess the impact of predictor measurement heterogeneity across validation and implementation setting in time-to-event outcome data. The use of the quantitative prediction analysis was illustrated using an example of predicting the risk of developing type-2 diabetes with heterogeneity in measurement of the predictor body mass index. In the simulation study, calibration-in-the-large of prediction models was poor and overall accuracy was reduced in all scenarios of predictor measurement heterogeneity. Model discrimination decreased with increasing random predictor measurement heterogeneity. Heterogeneity of predictor measurements across settings of validation and implementation reduced predictive performance at implementation of prognostic models with a time-to-event outcome. When validating a prognostic model, the targeted clinical setting needs to be considered and analyses can be conducted to quantify the anticipated impact of predictor measurement heterogeneity on model performance at implementation.

This chapter was based on: Luijken K, Song J, Groenwold RHH, Quantitative prediction error analysis to investigate predictive performance under predictor measurement heterogeneity at model implementation. *Diagnostic and Prognostic Research* (in press).

1 | Background

Clinical prognostic models aim to provide predictions of an outcome for individuals who have not been part of the modelling process[1-5]. The quantity that a clinical prediction model targets is defined by specifying the outcome, (candidate) predictors, population, setting, time of prediction, and prediction horizon as specifically as possible[6]. When the research setting does not correspond to the intended setting of application in clinical practice[7,8] or when modelling strategies are inappropriate[9,10], the predictive performance of a prognostic model may be suboptimal at implementation.

One reason for suboptimal predictive performance of a model at implementation are differences in predictor measurement procedures between model development and implementation in practice[7,11]. When discrepancies in predictor measurement procedures impact the performance of a clinical prediction model, this is referred to as *predictor measurement heterogeneity*[12]. The impact of predictor measurement heterogeneity on predictive performance at external validation has been quantified for models of binary outcome data[11-14] and illustrated in empirical data sets for logistic regression diagnostic prediction models[11,15]. However, the step towards model implementation in a target population has not been studied yet. The impact of predictor measurement heterogeneity in time-to-event data has not received adequate attention either.

In the current study, we suggest an approach to anticipate the impact of predictor measurement heterogeneity on a prognostic model when it is implemented in clinical practice. We assess the impact of predictor measurement heterogeneity in time-to-event outcome data using large-sample simulations. We propose a quantitative prediction analysis for validation studies that can be used to quantify the impact of anticipated predictor measurement heterogeneity in one of the predictors. This is illustrated using an example of a model predicting the 6-year risk of developing type-2 diabetes.

2 | Predictor measurement heterogeneity

For a prognostic model to provide correct predictions of an outcome in a clinical setting, several phases of model development should be considered, which is outlined in Figure 1[5,16-18]. Ideally, a prognostic model is derived using data that corresponds to the targeted implementation setting (derivation setting)[19,20]. Predictive performance is typically evaluated by measures of apparent performance and measures of performance after internal validation of the model, i.e., after correcting for optimism about the

performance[21,22]. When the internal predictive performance of the model is sufficient, its performance can be investigated using external data (validation setting)[23,24], which is preferably done multiple times[25-27]. When predictive performance at external validation is sufficiently well, implementation of the model in clinical practice could be considered (implementation setting), advisably after performing an impact analysis[28,29].

One aspect to consider in all phases of development of a prognostic model is predictor measurement heterogeneity, indicated in the blue box in Figure 1. Procedures to collect and measure predictor data for derivation and validation studies ideally correspond to the future implementation setting. When predictor measurement procedures at derivation and/or validation deviate from the predictor measurement procedure used in clinical practice, this can affect the predictive performance at implementation.

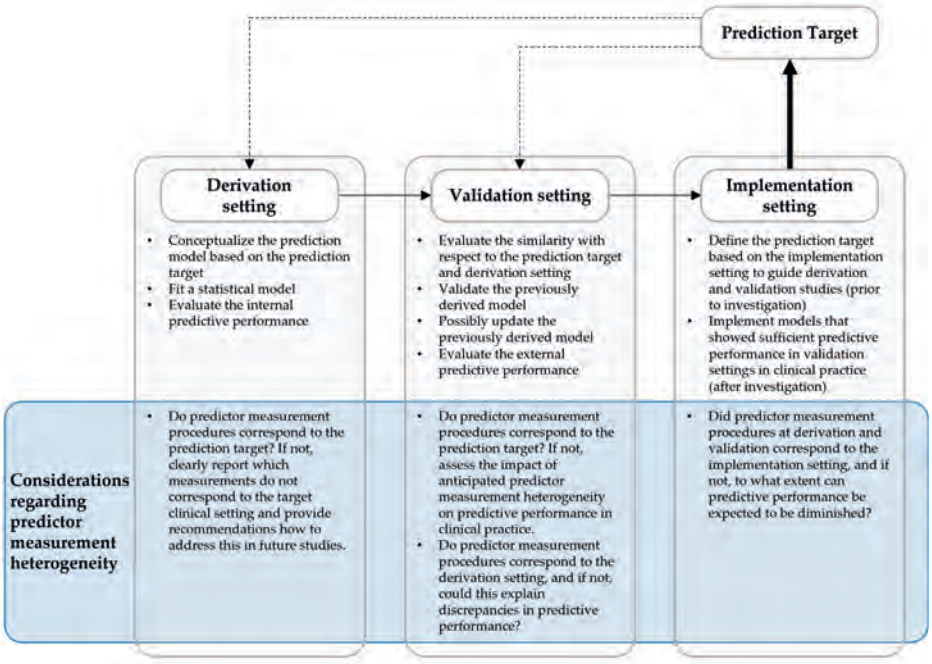


Figure 1. An overview of the derivation, validation, and implementation setting of a prognostic model, highlighting considerations regarding predictor measurement heterogeneity. Note that 'impact analysis' research is a phase between validation and implementation that is not addressed in this diagram. A prediction target is defined by specifying the target population, setting, outcome, (candidate) predictors, time of prediction, and prediction horizon as specifically as possible.

3 | Simulation study

We performed a simulation study to investigate the impact of predictor measurement heterogeneity across validation and implementation setting on out-of-sample predictive performance of a survival model derived and validated in time-to-event outcome data. We assumed that all other possible sources of discrepancy in predictive performance are not present, e.g., there are no differences in outcome prevalence and treatment assignment policy, there is no overfitting with respect to the derivation data, and the prognostic model is correctly specified in terms of functional form and included interactions. We used (very) large samples (n = 1,000,000) to minimize the role of random simulation error.

3.1 | Design of simulation study

Online Supplement 1 contains a detailed description of the simulation study. The main aspects of the design of the simulations study are described below and are reported following previous recommendations[30].

Data-generating mechanism: We simulated derivation, validation, and implementation data sets with 1,000,000 observations containing a continuous predictor variable X from a standard normal distribution. A time-to-event outcome was simulated for each subject so that outcomes followed a Cox-exponential model, using methods described by Bender and colleagues[31] (see Table 1 for simulation parameters). We generated data sets without censoring (median survival time $t = 6.6$). Additionally, data sets with administrative censoring after $t = 15$ (74% event fraction, median survival time 6.6) and with random censoring (69% event fraction, median survival time $t = 5.6$) were generated.

At implementation, a different measurement of predictor X was available, denoted W. Predictor measurement heterogeneity across validation and implementation setting was recreated using measurement error models, similar to[12]. The mean difference between X and W was denoted ψ (additive systematic measurement heterogeneity), the linear association between X and W was denoted θ (multiplicative systematic measurement heterogeneity), and the variance introduced by random deviations from X was denoted σ^2_ϵ, where non-zero values of σ^2_ϵ reflect that measurement W is less precise than X (random measurement heterogeneity).

In total, 162 scenarios were evaluated (27 scenarios of predictor measurement heterogeneity, for 2 different models under 3 different censoring mechanisms).

Prediction target: The prediction target was defined as obtaining correct predictions of the outcome risk at time point $t = 6.5$ conditional on predictor measurement W measured at moment of prediction (i.e., at $t = 0$).

Table 1 Simulation parameters.

Parameter	Value
Baseline hazard of an event	0.1
Conditional hazard ratio for association predictor *X* and survival times	2
Time point of administrative censoring	15
Baseline hazard of censoring	0.01
Conditional hazard ratio for association between random variable for censoring and censoring times	3
Mean of predictor *X* and random variable for censoring	0
Variance of predictor *X* and random variable for censoring	1
Predictor *W* at implementation*	
ψ	-0.3 to 0.3
θ	0.5 to 2
σ_ϵ	0 to √2

* At implementation, a different measurement of predictor *X* was available, denoted measurement *W*. The connection between *X* and *W* was defined using the following measurement heterogeneity model: $\mathbb{E}(W) = \psi + \theta\mathbb{E}(X) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and where ψ denotes an additive shift in *W* with respect to *X*, θ denotes a multiplicative linear association between *W* and *X*, and σ_ϵ^2 denotes random deviations from *X*.

Methods: A parametric exponential survival model and a semi-parametric Cox regression model were fitted in the derivation data set. Although a prognostic model is typically internally validated before performing external validation[1,21], we did not perform an internal validation since issues of overfitting were expected to be negligible due to the large sample sizes. The prognostic model was externally validated at time $t = 6.5$ (around median survival time) under predictor measurement homogeneity in an independent (validation) data set. Predictor measurement homogeneity refers to the situation in which predictors are measured in the same way at derivation and validation. Furthermore, the predictive performance of the prognostic model was investigated in various implementation settings under predictor measurement heterogeneity. The procedure was performed once in each scenario.

Performance metrics: Predictive performance was evaluated at $t = 6.5$, i.e., approximately at the median survival time. Calibration of the model on average,

or 'calibration in the large'[32,33], was evaluated by the ratio of the observed marginal survival at $t = 6.5$ (obtained through a Kaplan-Meier curve) versus the predicted marginal survival at $t = 6.5$ (obtained by averaging predicted survival at $t = 6.5$ of each observation), denoted the observed / expected ratio (O/E ratio). Discrimination was evaluated by the cumulative-dynamic time-dependent area under the receiver operating characteristic curve AUC(t)[34-36]. Overall accuracy was evaluated by the index of prediction accuracy at $t = 6.5$, IPA(t), which equals a Brier score[37] at $t = 6.5$ that is benchmarked to a null model ignoring all patient specific information and simply predicts the empirical prevalence to each patient[38]. A perfect model has an IPA of 1, a non-informative model has an IPA of 0 and a negative IPA indicates a harmful model.

Software: The simulation study was performed using R statistical software version 3.6.3[39]. The simulation code is available from https://github.com/KLuijken/PMH_Survival.

3.2 | Results of simulation study

Predictor measurement heterogeneity affected predictive performance at implementation. In all scenarios of predictor measurement heterogeneity, the prognostic models were miscalibrated in the large (range O/E ratio 0.89 to 1.19, compared to 1.00 under predictor measurement homogeneity), and overall accuracy was reduced (range IPA(6.5) -0.17 to 0.17, compared to 0.17 under predictor measurement homogeneity). The AUC(6.5) (range 0.58 to 0.74, compared to 0.74 under predictor measurement homogeneity) was particularly affected by random predictor measurement heterogeneity. We present results for the Cox regression model under no censoring only. The impact on the measures of predictive performance under administrative and uninformative (random) censoring and for the parametric exponential survival model were similar (data in Online Supplement 1, Section 3).

As measurement procedure W contained more random variability compared to X, i.e., a case of random measurement heterogeneity, $\sigma_\epsilon > 0$, the O/E ratio moved slightly under 1 (Figure 2, row A). The AUC(6.5) and IPA(6.5) decreased as random measurement heterogenei ty increased.

Additive systematic measurement heterogeneity, i.e., $\psi \neq 0$, affected the calibration-in-the-large coefficient at implementation, but minimally affected the AUC(6.5), and IPA(6.5) at implementation (Figure 2, row B). When measurement procedure W at implementation provided a systematically higher value of the predictor compared to

measurement procedure *X* at validation, i.e., $\psi > 0$, this resulted in overestimation of the average outcome incidence at implementation, and the O/E ratio < 1.

Multiplicative systematic measurement heterogeneity, i.e., $\theta \neq 1$, yielded an O/E ratio < 1 in case $\theta > 1$ (Figure 2, row C). Multiplicative systematic measurement heterogeneity minimally affected the AUC(6.5) in absence of additive systematic and random measurement heterogeneity. As θ was further from 1, the IPA(6.5) at implementation decreased, indicating lower overall accuracy.

Combined random, additive systematic, and/or multiplicative systematic predictor measurement heterogeneity sometimes reinforced or cancelled out effects on predictive performance (see Online Supplement 1, Section 3).

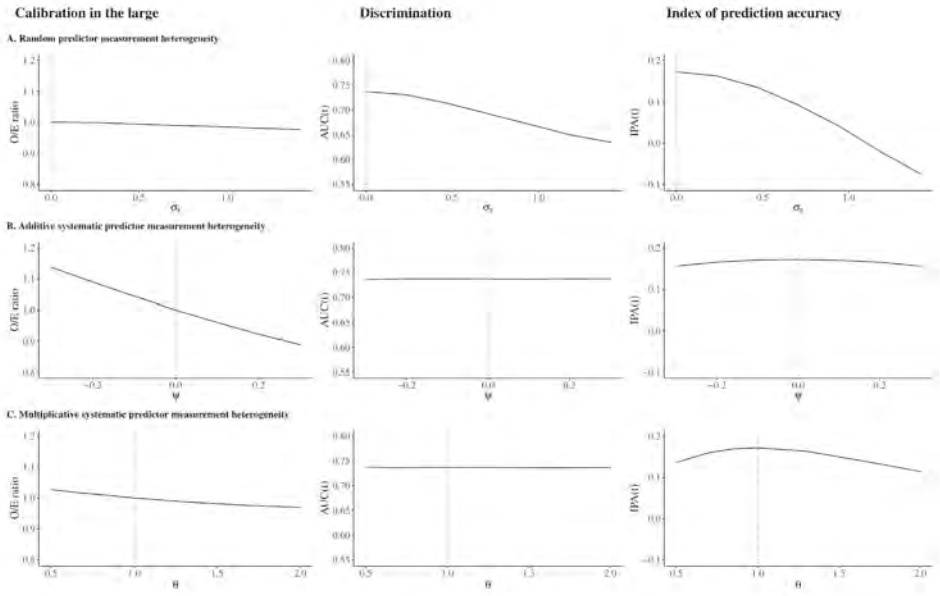


Figure 2. Measures of predictive performance under predictor measurement heterogeneity between validation and implementation setting. Results shown for random predictor measurement only (row A), additive systematic predictor measurement only (row B), and multiplicative systematic predictor measurement heterogeneity only (row C). The vertical dashed line indicates predictor measurement homogeneity between validation and implementation setting. The x-axes show measurement heterogeneity parameters describing the predictor measurement at implementation relative to the predictor measurement at validation, where σ_ϵ denotes random deviations from the measurement at validation, ψ denotes an additive shift with respect to the measurement at validation, and θ denotes a systematic multiplicative association with the measurement at validation. Note that additional simulation scenarios were run to smooth the plots.

4 | Illustration of quantitative prediction analysis

We describe an analysis that quantifies the impact of anticipated predictor measurement heterogeneity between the validation and implementation setting. This is illustrated by validating a prognostic model predicting the 6-year risk of developing type-2 diabetes in a modified example dataset. Section 4.1 describes derivation and validation of the model. The hypothetical step to implementation is described in Section 4.2 by means of the proposed analysis. A detailed description including analysis code can be found in Online Supplement 2.

4.1 | Motivating validation study

Zhang and colleagues derived a prognostic model predicting the 6-year risk of developing type-2 diabetes from the predictors age, BMI, triglyceride, and fasting plasma glucose at moment of prediction[40]. Here, we used the set of predictors identified by Zhang et al. as a starting point for derivation and validation of a prognostic model. We emphasize that this is for illustrative purposes only and is not a recommended approach in practice, where a validation study typically validates a previously derived prognostic model as-is.

The example data was obtained from a publicly available data set containing information about 15,464 individuals who participated in a medical examination program at the Murakami Memorial Hospital from 2004 to 2015, made publicly available alongside a study by Okamura and colleagues[41]. BMI was reported to be measured at medical examination; we assumed it was computed from scale and measuring-tape measurements.

We recreated a derivation and validation sample by resampling from the original dataset stratified on cumulative event fraction at 6 years (2,192 days). In the recreated derivation sample (n = 10,824), the incidence density rate was 2.83/1,000 person years (134 events in total), event times ranged from 285 to 2,191 days, and censoring times ranged from 164 to 2,192 days. In the recreated validation sample (n = 4,639), the incidence density rate was 2.88/1,000 person years (58 events in total), event times ranged from 285 to 2,191 days, and censoring times ranged from 164 to 2,192 days. We assumed censoring was non-informative.

We evaluated predictive performance at 6 years using the performance measures described in our simulation study and used a bootstrap procedure with 500 resamples to correct the AUC(6 years) and IPA(6 years) for optimism and estimate 95-percentile

confidence intervals (CIs). There was no predictor measurement heterogeneity across derivation and validation setting by construction of the samples.

At derivation, the calibration-in-the-large O/E ratio was 1.02 (95% CI, 0.76; 1.43), the optimism-corrected AUC(6 years) was 0.87 (95% CI, 0.84; 0.90), and the optimism-corrected IPA(6 years) was 0.07 (95% CI, 0.04; 0.11). At validation, the calibration-in-the-large O/E ratio was 1.01 (95% CI, 0.78 to 1.34), the AUC(6 years) was 0.89 (95% CI, 0.84 to 0.93), and the IPA(6 years) was 0.06 (95% CI, 0.01 to 0.11).

4.2 | Quantitative prediction analysis for anticipating the impact of predictor measurement heterogeneity between validation and implementation setting on predictive performance

Seven steps are described to perform a quantitative prediction analysis in a prognostic model validation study to assess the impact of anticipated measurement heterogeneity in measurement of BMI, where BMI is assumed to be measured from self-reported height and weight at implementation, instead of tape and scale measures at validation (Box 1).

First, the prediction target is stated. In this example, the prediction target would be the 6-year risk of developing type-2 diabetes in Asian adults presenting for preventive medical examination by measurements of age, BMI, triglyceride, and fasting plasma glucose at moment of prediction. Incident diabetes is defined as HbA1c $\geq$ 6.5% (48 mmol/mol) in two test results, measured using a standardized method[42]. Age is measured in years, BMI is calculated from self-reported weight and height, triglyceride is measured according to standards of the National Institute of Standards and Technology[43], and fasting plasma glucose is measured using a standardized method[44,45]. Details on procedures to measure HbA1c, triglyceride, and fasting plasma glucose are omitted here for brevity, but are ideally described in more detail in an empirical study. Treatment assignment policy was assumed to be similar in the research settings compared to the target clinical setting and interventions such as diet were not modeled explicitly (i.e., ignore-treatment strategy[46]).

Second, it is described whether predictor measurement procedures in the validation setting correspond to those that will be used at implementation. Measurements of age, triglyceride, and fasting plasma glucose roughly correspond to the target predictor measurement procedures. However, the validation study measured BMI during medical examination of a patient, which differs from self-reported measurements defined in the prediction target.

Box 1. Quantitative prediction analysis to quantify the impact of anticipated predictor measurement heterogeneity when implementing a prognostic model in clinical practice (details in Section 4.2 of the main text).

1. State the prediction target.
2. Report whether predictor measurement procedures in the validation setting correspond to those at implementation.
3. Identify one predictor that is expected to be measured using a different procedure in the implementation setting than in the validation setting.
4. Define a model for the relation between the measurement in the validation study and its equivalent in the implementation setting.
5. Perform a literature search to establish a range for the size of the possible parameters of predictor measurement heterogeneity.
6. Simulate the scenarios of anticipated measurement heterogeneity to assess the possible impact on predictive performance.
7. Report the impact of anticipated predictor measurement heterogeneity on predictive performance in clinical implementation.

Third, a predictor is identified that is expected to be measured differently (e.g., using a different procedure) in the implementation setting compared to the validation setting. Measurement heterogeneity was expected to be strongest for the predictor BMI.

Fourth, a model for the relation between the measurement of BMI in the validation study, BMI_{val}, and in the implementation setting, BMI_{imp}, is defined, e.g.:

$$BMI_{imp} = \psi + \theta BMI_{val} + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and $\psi \neq 0$ indicates that measurements of BMI in the implementation setting are systematically additively shifted with respect to BMI in the validation study, $\theta \neq 0$ indicates measurements of BMI in the implementation setting are systematically multiplicatively altered with respect to BMI in the validation study, and $\sigma_\epsilon > 0$ indicates measurements of BMI in the implementation setting contain more random variation relative to BMI in the validation study.

Fifth, the range is specified for the parameter values of the model for the anticipated predictor measurement heterogeneity, as defined in Step 4. A literature search was performed to identify studies describing measurement error in BMI. Informed by studies

comparing measured and self-reported BMI values[47-51], the range of measurement error parameters was specified as -1 to 0 for ψ, 0.9 to 1 for θ, and 0 to 1.5 for σ_ϵ. In general, we advise to use terms like 'measurement error', 'validation study', and the measurement procedures to search for relevant literature. Of note, the term 'validation study' has a different meaning in prediction literature compared to measurement error literature. In prediction literature, a validation study refers to a study that evaluates the predictive performance of an existing prediction model. In measurement error literature, a validation study refers to a study in which a perfect measurement is taken of a mismeasured covariate, usually in a subset of individuals included in the study[52]. The purpose of a measurement-error validation study is to estimate the connection between the error-prone and error-free measurement, for instance using measurement error models, to address issues introduced by measurement error in the substantive analysis. In the current study, we thus far used the term 'validation study' according to the prediction literature.

Sixth, the scenarios of anticipated measurement heterogeneity can be investigated using statistical simulations to assess the possible impact on predictive performance. Briefly, we plugged in the values found in Step 5 into the model specified in Step 4 to generate measurements of BMI that can be anticipated in the implementation setting in participants otherwise similar to the validation sample. We evaluated the O/E ratio for calibration in the large, AUC(6 years), and IPA(6 years) under the scenarios of measurement heterogeneity in BMI (see Online Supplement 2) and plotted the outcomes (Figure 3).

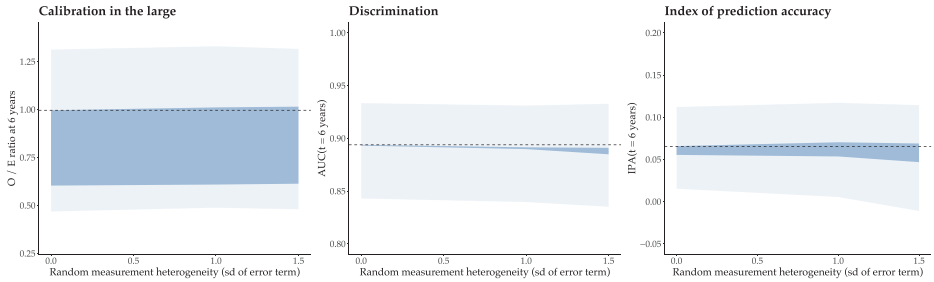


Figure 3. Impact of anticipated heterogeneity in measurement of the predictor body mass index on measures of predictive performance at implementation of a model to predict the 6-year risk of developing diabetes type 2. Dark blue indicates the impact within the range of specified predictor measurement heterogeneity and light blue indicates 95 percentile CIs from 500 bootstrap resamples. Random predictor measurement heterogeneity is presented on the x-axis, and performance measures are marginalized over scenarios of additive and multiplicative systematic predictor measurement heterogeneity.

Seventh, the impact of anticipated predictor measurement heterogeneity on predictive performance in the implementation setting can be reported in a validation study, accompanied by a description of Steps 1-6. Anticipating on the possibility that BMI may be measured differently in clinical practice compared to how data on BMI were collected in the validation study, we found that performance of the type-2 diabetes prediction model might be reduced when implemented 'as-is' in clinical practice (Figure 3). In particular, with increasing differences in BMI measurement variance between our validation study and the clinical target setting, model miscalibration increases. Possible consequences of this finding may be to either update the current prediction model using self-reported measures of BMI before implementing it in clinical practice or to collect data on BMI using scale and measuring-tape measures only when the model is used in clinical practice to predict 6-year risk of developing diabetes.

5 | Discussion

Our simulations indicated that predictor measurement heterogeneity across the validation and implementation setting of a prognostic model can substantially affect predictive performance at implementation. We illustrated how a quantitative prediction analysis can be applied in validation studies to quantify the impact of anticipated dissimilar predictor measurements in the clinical target setting on predictive performance. Based on this analysis, a validation study can inform readers about the severity of possible predictor measurement heterogeneity when the model is implemented in clinical practice.

The rationale for the quantitative prediction analysis was analogous to the quantitative bias analysis framework by Lash and colleagues, which can be applied to estimate the direction, magnitude, and uncertainty from systematic errors affecting studies of causal inference[53,54]. While Lash and colleagues encourage researchers to address multiple sources of bias[53], we focused on a single source of heterogeneity across settings that can affect performance of a clinical prediction model. In this, we focused on non-differential systematic and random measurement heterogeneity in a single predictor, where the clinical implementation setting contained more measurement variance compared to the validation setting. Future work could extend these quantitative prediction analyses to non-differential measurement heterogeneity, to settings were the clinical implementation setting contained less measurement variance compared to the validation setting – for instance through methods analogous

to the simulation-extrapolation method (SIMEX)[55,56] – and to models that take into account correlations of measurement heterogeneity structures when multiple predictors are expected to be measured heterogeneously across validation and implementation setting. Additionally, other sources of heterogeneity across settings that can affect performance of a clinical prediction model can be added to the quantitative prediction analysis, such as heterogeneity in event rate, heterogeneity in outcome measurement procedures, and heterogeneity in treatment-assignment policies during follow-up.

The example of predicting the risk of developing type-2 diabetes illustrated the impact of anticipated measurement heterogeneity in the predictor BMI. Notably, the magnitude of the impact of anticipated measurement heterogeneity strongly depends on whether the linear predictor was centered to the validation data. While many functionalities in R[39] center the linear predictor by default, centering is likely uncommon in clinical practice and obviously decreases the impact of predictor measurement heterogeneity on predictive performance. A limitation of the example is that only measurement heterogeneity in a single predictor was considered, while the predictor fasting plasma glucose can potentially be measured heterogeneous across settings as well, in particular because fasting instructions and adherence to instructions may differ across settings. Taking this into account requires consideration of the duration of fasting relative to the timing of the plasma glucose measurement[57]. Modelling the functional form of fasting plasma glucose or another (circadian) fluctuating hormone or biomarker over time to assess the impact in heterogeneity of measurement timings across time would be an interesting topic for future research.

As a limitation to our study, implementation of the quantitative prediction analysis may be hampered because literature informing the choice of measurement error parameters (Step 5) may be limited. When no information is available about predictor measurement structures in an implementation setting of interest, it might be helpful to set up a (measurement heterogeneity) validation study to estimate the predictor measurement heterogeneity parameters directly[52]. This may be an alternative approach to anticipate the performance of a prognostic model in a particular setting that is likely less cumbersome than conducting a prediction validation study in the implementation setting.

Data for derivation and validation of prognostic models are collected ideally using procedures that match the target clinical setting. When this is infeasible, the quantitative prediction analysis provides an analytical approach to quantify the anticipated impact of the discrepancies between available research data and clinical practice.

Online Supplementary Files

The supplementary files referred to in this Chapter are available online at https://github.com/KLuijken/Dissertation_Online_Supplements/tree/main/Chapter_8

References

1. Steyerberg EW. *Clinical Prediction models*. Springer; 2019.
2. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine*. 1999;130(6):515-524.
3. Shmueli G, Koppius OR. Predictive analytics in information systems research. *MIS Quarterly*. 2011;35(3):553-572.
4. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *British Medical Journal*. 2013;346.
5. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381.
6. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. 2014;11(10):e1001744.
7. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine*. 2015;162(1):W1-W73.
8. Debray TP, 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. *Statistics in Medicine*. 2013;32(18):3158-3180.
9. Steyerberg EW, Uno H, Ioannidis JP, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. *Journal of Clinical Epidemiology*. 2018;98:133-143.
10. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in Medicine*. 2016;35(2):214-226.
11. Pajouheshnia R, Van Smeden M, Peelen L, Groenwold R. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *Journal of Clinical Epidemiology*. 2019;105:136-141.
12. Luijken K, Groenwold RH, 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(18):3444-3459.
13. Khudyakov P, Gorfine M, Zucker D, Spiegelman D. The impact of covariate measurement error on risk prediction. *Statistics in Medicine*. 2015;34(15):2353-2367.
14. 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. *Population Health Metrics*. 2012;10(1):1-11.
15. Luijken K, Wynants L, van Smeden M, et al. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *Journal of Clinical Epidemiology*. 2020;119:7-18.
16. Wynants L, Collins GS, Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG: An International Journal of Obstetrics & Gynaecology*. 2017;124(3):423-432.
17. Cowley LE, Farewell DM, Maguire S, Kemp AM. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagnostic and Prognostic Research*. 2019;3(1):1-23.
18. Toll D, Janssen K, Vergouwe Y, Moons K. Validation, updating and impact of clinical prediction rules: a review. *Journal of Clinical Epidemiology*. 2008;61(11):1085-1094.
19. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *British Medical Journal*. 2009;338.
20. Riley RD, Ensor J, Snell KI, et al. Calculating the sample size required for developing a clinical prediction model. *British Medical Journal*. 2020;368.

21. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine.* 1996;15(4):361-387.

22. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans M, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology.* 2001;54(8):774-781.

23. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *British Medical Journal.* 2009;338.

24. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine.* 2000;19(4):453-473.

25. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *British Medical Journal.* 2009;338.

26. Vergouwe Y, Nieboer D, Oostenbrink R, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Statistics in Medicine.* 2017;36(28):4529-4539.

27. Ensor J, Snell KI, Debray TP, et al. Individual participant data meta-analysis for external validation, recalibration, and updating of a flexible parametric prognostic model. *Statistics in Medicine.* 2021;40(13):3066-3084.

28. Adams ST, Leveson SH. Clinical prediction rules. *British Medical Journal.* 2012;344.

29. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of Internal Medicine.* 2006;144(3):201-209.

30. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine.* 2019;38(11):2074-2102.

31. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine.* 2005;24(11):1713-1723.

32. 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. *Journal of Clinical Epidemiology.* 2016;74:167-176.

33. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Medicine.* 2019;17(1):1-7.

34. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics.* 2005;61(1):92-105.

35. Uno H, Cai T, Tian L, Wei L-J. Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association.* 2007;102(478):527-537.

36. Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of-year predicted risks. *Biostatistics.* 2019;20(2):347-357.

37. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Review.* 1950;78(1):1-3.

38. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic and Prognostic Research.* 2018;2(1):1-7.

39. R Core Team. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2020.

40. Zhang M, Zhang H, Wang C, et al. Development and validation of a risk-score model for type 2 diabetes: a cohort study of a rural adult Chinese population. *PloS one.* 2016;11(4):e0152054.

41. Okamura T, Hashimoto Y, Hamaguchi M, Obora A, Kojima T, Fukui M. Ectopic fat obesity presents the greatest risk for incident type 2 diabetes: a population-based longitudinal study. *International Journal of Obesity.* 2019;43(1):139-148.

42. American Diabetes Association. 2. Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care.* 2021;44(Supplement 1):S15-S33.

43. Warnick GR, Kimberly MM, Waymack PP, Leary ET, Myers GL. Standardization of measurements for cholesterol, triglycerides, and major lipoproteins. *Laboratory Medicine.* 2008;39(8):481-490.

44. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation. 2006.

45. D'Orazio P, Burnett RW, Fogh-Andersen N, et al. Approved IFCC recommendation on reporting results for blood glucose: International Federation of Clinical Chemistry and Laboratory Medicine Scientific Division, Working group on selective electrodes and point-of-care testing (IFCC-SD-WG-SEPOCT). *Clinical Chemistry and Laboratory Medicine (CCLM).* 2006;44(12):1486-1490.

46. van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *European Journal of Epidemiology.* 2020;35:619-630.

47. Nawaz H, Chan W, Abdulrahman M, Larson D, Katz DL. Self-reported weight and height: implications for obesity research. *American Journal of Preventive Medicine.* 2001;20(4):294-298.

48. Allison C, Colby S, Opoku-Acheampong A, et al. Accuracy of self-reported BMI using objective measurement in high school students. *Journal of Nutritional Science.* 2020;9:e35.

49. Dekkers JC, van Wier MF, Hendriksen IJ, Twisk JW, van Mechelen W. Accuracy of self-reported body weight, height and waist circumference in a Dutch overweight working population. *BMC Medical Research Methodology.* 2008;8(1):1-13.

50. Villarini M, Acito M, Gianfredi V, et al. Validation of self-reported anthropometric measures and body mass index in a subcohort of the dianaweb population study. *Clinical Breast Cancer.* 2019;19(4):e511-e518.

51. Ortiz-Panozo E, Yunes-Díaz E, Lajous M, Romieu I, Monge A, López-Ridaura R. Validity of self-reported anthropometry in adult Mexican women. *Salud publica de Mexico.* 2017;59:266-275.

52. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement error in nonlinear models: a modern perspective.* Chapman and Hall/CRC; 2006.

53. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *International Journal of Epidemiology.* 2014;43(6):1969-1985.

54. Lash TL, Fox MP, Fink AK. *Applying quantitative bias analysis to epidemiologic data.* Springer Science & Business Media; 2011.

55. Cook JR, Stefanski LA. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association.* 1994;89(428):1314-1328.

56. Stefanski LA, Cook JR. Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association.* 1995;90(432):1247-1256.

57. Whittle R, Royle K-L, Jordan KP, Riley RD, Mallen CD, Peat G. Prognosis research ideally should measure time-varying predictors at their intended moment of use. *Diagnostic and Prognostic Research.* 2017;1(1):1-9.