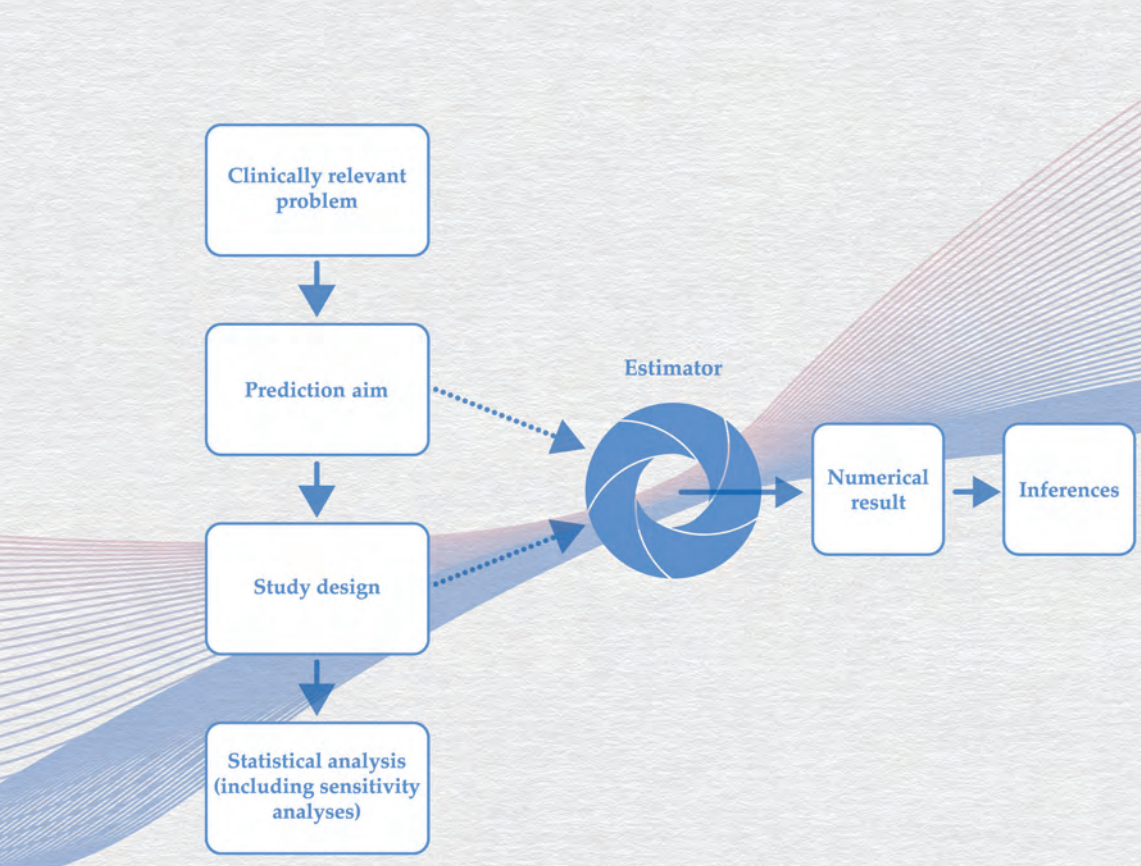
The impact of epidemiologic methods on findings in studies of causal effects and prediction modelling

Luijken, K.

7


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

Changing predictor measurement procedures affected the performance of prediction models in clinical examples

The objective of this study was to quantify the impact of predictor measurement heterogeneity on prediction model performance. Predictor measurement heterogeneity refers to variation in the measurement of predictor(s) between the derivation of a prediction model and its validation or application. It arises, for instance, when predictors are measured using different measurement instruments or protocols. We examined effects of various scenarios of predictor measurement heterogeneity in real-world clinical examples using previously developed prediction models for diagnosis of ovarian cancer, mutation carriers for Lynch syndrome, and intrauterine pregnancy. Changing the measurement procedure of a predictor influenced the performance at validation of the prediction models in nine clinical examples. Notably, it induced model miscalibration. The calibration intercept at validation ranged from -0.70 to 1.43 (0 for good calibration), while the calibration slope ranged from 0.50 to 1.67 (1 for good calibration). The difference in c-statistic and scaled Brier score between derivation and validation ranged from -0.08 to +0.08 and from -0.40 to +0.16, respectively. This study illustrates that predictor measurement heterogeneity can influence the performance of a prediction model substantially, underlining that predictor measurements used in research settings should resemble clinical practice. Specification of measurement heterogeneity can help researchers explaining discrepancies in predictive performance between derivation and validation setting.

This chapter was based on: Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *Journal of Clinical Epidemiology*. 2020;119:7-18.

1 | Background

Clinical prediction models are commonly applied in clinical practice to assist healthcare professionals in determining a patient's diagnosis or prognosis[1]. Clinical prediction models are applied to patients that were not part of the data used to derive the model, often with the aim to estimate a probability for the presence of a disease or future health state[2]. When applied on new patients, the performance in estimating these probabilities is often different from the performance in the derivation data. This is commonly explained by model overfitting with respect to the derivation data[3-6] and differences in patient characteristics (case-mix) between derivation and validation settings[7-9].

Previous studies have identified imprecise predictor measurement procedures as another reason for a suboptimal performance of prediction models at derivation[10,11] and highlighted that differences in predictor measurement procedures between derivation and validation setting substantially affected performance at validation[12-14]. Predictor variables may be measured by different procedures in external validation data than those applied in derivation data, that is, according to different measurement protocols, measurement instruments or by applying different predictor definitions. We refer to these differences in measurement across settings as predictor measurement heterogeneity. Simulation studies have shown that predictor measurement heterogeneity can induce miscalibration of prediction models and affect discrimination and accuracy at external validation[12]. While predictor measurement heterogeneity across derivation and validation samples appears to be common in clinical (research) settings (see for example [4,15,16]), its impact on the performance of prediction models at validation is not well-studied using empirical data.

In this study, we quantify the impact of predictor measurement heterogeneity on predictive performance in a series of real-world clinical examples.

2 | Illustrating and defining predictor measurement heterogeneity

We briefly illustrate predictor measurement heterogeneity here using measurements of the predictor body mass index (BMI). We fitted a logistic regression model to predict the presence of pre-stage diabetes containing only two parameters for a linear and a quadratic term of BMI besides the intercept (this example was adapted from Rosella and colleagues[11]). Data were available on 1,264 participants from the NHANES Study 2013-

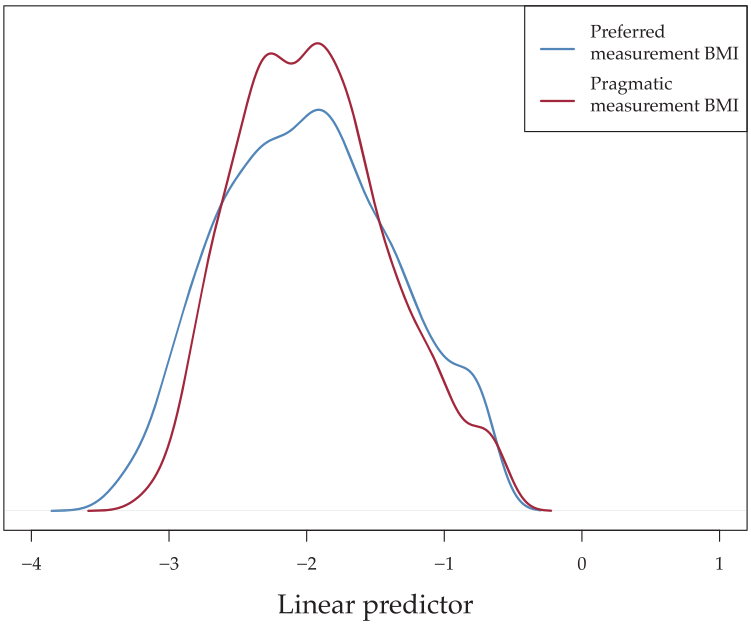


Figure 1. Impact of predictor measurement heterogeneity on distributions of linear predictors. Density of the logit transformation of the predicted risks (linear predictor) from a logistic regression model predicting the probability of a pre-stage of diabetes using the predictor BMI. BMI was obtained as an instrumental (preferred) and self-reported (pragmatic) measure. Distributions of the linear predictors for both procedures are presented. The prediction model was: $\text{logit}\left(P(Y_i = 1|BMI_i)\right) = \beta_0 + \beta_1 BMI_i + \beta_2 BMI_i{}^2$

2014[17]. BMI data were computed from participants' height and weight measurements, obtained by a trained examiner who followed a standardized protocol[18]. Since this measurement is close to what we would consider the ideal method of measurement, we will refer to it as the *preferred measurement*. A second measurement of BMI was computed via self-reported weight and height by the participants, which we will refer to as the *pragmatic measurement*. The concept *predictor measurement heterogeneity* refers to the phenomenon where the predictor measurement strategy at derivation differs from the measurement strategy at validation or application of the prediction model.

A second regression model was fitted with a linear and quadratic term for BMI using the pragmatic measurement of BMI. Comparing the output of the two regression models, it becomes clear that substituting the preferred measurement of BMI with the pragmatic measurement changed the distribution of the linear predictor (Figure 1). To better understand how substitution of pragmatic by preferred measurements (and vice versa) can affect predictive performance, we present empirical case studies in the next sections.

3 | Methods

We examined effects of predictor measurement heterogeneity in previously established prediction models, using three empirical datasets on the diagnosis of ovarian cancer, hereditary non-polyposis colorectal cancer (Lynch syndrome), and intrauterine pregnancy, respectively. Scenarios from various clinical domains were investigated in order to provide a general assessment of the potential impact of predictor measurement heterogeneity.

3.1 | Example dataset 1: diagnosis of ovarian cancer

The International Ovarian Tumor Analysis (IOTA) dataset includes clinical and ultrasound information on 5,914 non-pregnant women with at least one persistent adnexal mass[19]. We used data from IOTA phases I-III (1999-2012) in which we studied two prediction models, here referred to as Model 1 and Model 2. Model 1 is a logistic regression model that estimates the probability of presence of ovarian mass malignancy from pre-operatively measured predictors: age (years); maximal diameter of the tumor (mm); personal history of ovarian cancer (yes/no); current use of hormonal therapy (yes/no); experience of pain during examination (yes/no); presence of ascites (yes/no); presence of blood flow within a solid papillary projection (yes/no); maximal diameter of the largest solid component (mm); presence of irregular cyst walls (yes/no); presence of acoustic shadows (yes/no); color score of intratumoral blood flow (ordinal, ranging 1-4); and presence of entirely solid tumors (yes/no). Model 1 is based on the LR1 model, which was developed and internally validated in IOTA phase-I data[20] and has been externally validated several times[21-23]. Model 2 is a logistic regression model to preoperatively diagnose ovarian mass malignancy by age (years), the proportion of solid tissue, the presence of more than ten locules (yes/no), the number of papillary structures, the presence of acoustic shadows (yes/no), and the presence of ascites (yes/no). It is a previously described reduction of Model 1, developed for methodological illustrations[24].

3.2 | Example dataset 2: prediction of mutation carrier status (Lynch syndrome)

We analysed data from 19,866 patients with colorectal cancer (CRC), who were tested for mutations in Lynch syndrome–related mismatch repair genes. We studied a simplification of the $PREMM_{1,2}$-model[25] and MMRpredict model[5,26] in the Lynch syndrome dataset, which we refer to as Model 3. Model 3 is a logistic regression model that predicts the prevalence of MLH1/MSH2 mutations from the following predictors measured at baseline: sex; age at CRC diagnosis (years); and family history of CRC

and endometrial cancer. Family history was defined as a weighted sum of positive first- and second-degree relatives, where second-degree relatives were weighted half times the first-degree relatives. The sum ranged from 0-3, with family history coded as 0, 1, or 2+ affected relatives.

3.3 | Example dataset 3: prediction of intrauterine pregnancy

We analysed data from 75 consecutive patients at the Early Pregnancy and Acute Gynaecology Unit (EPAGU) at Queen Charlottes' and Chelsea Hospital from November 2013 to May 2014. We studied a logistic regression model in the pregnancy data, here referred to as Model 4, that predicts the probability of an ongoing intrauterine pregnancy based on measurements of hCG level at presentation (pmol/L) and an hCG ratio of hCG at 48h after presentation to hCG at presentation. hCG Levels could be measured using two different measurement instruments, named the 'ria kit' and the 'imm kit'. Model 4 is adapted from an existing multinomial logistic regression model (named M4)[27], by grouping the outcome categories 'ectopic pregnancy' and 'pregnancy of unknown location'.

3.4 | Models and assessment of predictive performance

To separate the impact of predictor measurement heterogeneity from other possible external validation effects on predictive performance, such as changes in case-mix and outcome incidence, we focus on derivation and validation within the same study population and evaluate predictive performance[28]. In each example, we defined scenarios of measurement heterogeneity by identifying two measurement procedures of a single predictor: a *preferred* measurement and a *pragmatic* measurement (Table 1). The terms '*preferred*' and '*pragmatic*' are only meant in a relative sense: a preferred measurement may still be far from the ideal measurement of a particular phenomenon, but as a predictor of a particular outcome it could be preferable over the pragmatic measurement in terms of a lower measurement error or anticipated better predictive potential for the particular outcome.

For each scenario, we assessed the optimism-corrected predictive performance of a regular maximum likelihood logistic regression model under both predictor measurement *homogeneity* and *heterogeneity*. The optimism correction was performed since measures of predictive performance based on the derivation data may give an over-optimistic assessment of model performance, as maximum likelihood models are generated to provide the best fit for the derivation data[28]. Measures of predictive performance were obtained by deriving and validating a prediction model in 500 bootstrap samples and averaging optimism-corrected measures of performance over the bootstrap samples

Table 1 Scenarios of measurement heterogeneity in four clinical prediction models

Scenario	Dataset	Model	Measurement heterogeneity		Explanation
			Preferred procedure (scale)	Pragmatic procedure (scale)	
1	IOTA	1	Maximal diameter tumor (continuous)	Mean diameter tumor (continuous)	In the original model, the diameter of the tumor was measured as the maximum of three measurements of the tumor lesion in different dimensions. Alternatively, the mean of these three measurements could be used as model input.
2	IOTA	1	Maximal diameter solid component tumor, non-truncated (continuous)	Mean diameter solid component tumor, non-truncated (continuous)	In the original model, the diameter of the largest solid component of the tumor was measured as the maximum of three measurements of the solid component of the tumor lesion in different dimensions. Alternatively, the mean of these three measurements could be used as model input.
3	IOTA	1	Diameter solid component truncated at 50mm (continuous)	Original diameter solid component (continuous)	The diameter of the largest solid component of the tumor was truncated at 50 mm in the original model. In application of this model, the truncation can be ignored or forgotten.
4	IOTA	1	Color-score (ordinal, 1-4)	Color-score at extremes (dichotomous, 1 or 4)	The intratumoral blood flow was scored by a colorscore ranging 1-4 in the original model. Alternatively, the extremes of this score (1 or 4) could be used as model input, because: - A colorscore is a subjective measurement; at model application, physicians could score the colors at the extremes (either no or high blood flow). - Researchers could use a (public) dataset for model validation in which only a binary version of the score is available, rather than a categorical score, and recode this variable into scores 1 or 4.
5	IOTA	2	≥10 locules (binary)	≥5 locules (binary)	The original model included a dichotomized version of the number of locules, where the cut-off was at 10 locules. At model validation or application, the cut-off value for dichotomization could be different.

Table 1 *(continued)*

Scenario	Dataset	Model	Measurement heterogeneity		Explanation
			Preferred procedure (scale)	Pragmatic procedure (scale)	
6	Lynch syndrome	3	Family history CRC summarized by counting 0,1,2+ FDRs and 0,1,2+ SDR, weighted by a half (categorical, 0-3)	Family history CRC summarized by counting only FDRs (categorical, 0-3)	Family history of CRC is computed as a weighted count of diagnoses of CRC in first- and second-degree relatives. Possibly, the history of CRC is recorded for first-degree relatives only and used as model input.
7	Lynch syndrome	3	Family history EC summarized by counting 0,1,2+ FDRs and 0,1,2+ SDR, weighted by a half (categorical, 0-3)	Family history EC summarized by counting only FDRs (categorical, 0-3)	Family history of EC is computed as a weighted count of diagnoses of EC in first- and second-degree relatives. Possibly, the history of EC is recorded for first-degree relatives only and used as model input.
8	Pregnancy	4	hCG level measured in serum, using the ria kit (continuous)	hCG level measured in urine, using the ria kit (continuous)	A hCG measurement is preferably obtained from serum samples but could alternatively be obtained from urine samples.
9	Pregnancy	4	hCG level measured in serum, using the ria kit (continuous)	hCG level measured in serum, using the imm kit (continuous)	A hCG measurement can be obtained using different measurement kits, e.g., the ria kit or imm kit.

Abbreviations: CRC = colorectal cancer, EC = endometrial cancer, FDR = first-degree relative, SDR = second-degree relative.

(see Online Supplementary File 1 for detailed explanation)[28]. To assess predictor measurement homogeneity, the prediction model was derived and validated based on the same predictor definitions. To assess predictor measurement heterogeneity, a derivation and validation setting were recreated by deriving the model using the *preferred* measurement and validating the model using the *pragmatic* measurement, denoted scenario 1a-9a, or by deriving the model using the *pragmatic* measurement and validating the model using the *preferred* measurement, denoted scenario 1b-9b. Note that we isolate the impact of measurement heterogeneity here by keeping all other factors besides measurement heterogeneity constant (i.e., the modelling strategy, included predictors, and patient characteristics are equal at derivation and validation).

Measures of predictive performance were the calibration-in-the-large coefficient and calibration slope from a logistic recalibration model, the c-statistic (area under the ROC curve) and the Brier score. Model calibration refers to the agreement between observed outcomes and risk estimates[1,29]. The calibration-in-the-large coefficient evaluates whether there is a difference between the observed event fraction and the average predicted risk (0 for perfect calibration) and is estimated as the intercept of the recalibration model while the calibration slope is fixed at a value of 1. The calibration slope (<1 indicating overfitting, i.e., predicted risks that are too extreme, and >1 indicating underfitting) was computed by regressing the observed outcome on the logit transformation of the predicted risks and evaluated graphically by plotting lowess calibration curves. We considered the scaled Brier score, in which the Brier score is scaled by its maximum score under a non-informative model, $Brier_{scaled} = 1 - Brier / Brier_{max}$, so that it ranges from 0 for perfect predictions to 1 for non-informative predictions[1,29].

To quantify the resemblance between the predictor measurement procedures, the partial correlation between the *preferred* and *pragmatic* predictors was estimated by correlating residuals of two linear regression models regressing each of the predictor measurements on the outcome and other covariates in the model. Shrunken regression coefficients from a Ridge logistic regression model were estimated, for which the tuning parameter (necessary for shrinkage) was determined by the value minimizing the deviance in 10-fold cross-validation[30]. All analyses were performed in R 3.5.1.[31] and R code is available at https://github.com/KLuijken/Prediction_Measurement_Heterogeneity_Examples. Measures of predictive performance were obtained using the `rms` package[32].

4 | Results

Measures of predictive performance in all scenarios are presented in Table 2 (under measurement homogeneity) and Table 3 (under measurement heterogeneity). The latter results are presented graphically in Figure 2-4, for the IOTA dataset, Lynch syndrome dataset and pregnancy dataset, respectively. Each scenario is discussed in detail in the Online Supplementary File 2. Results after regression shrinkage (Ridge regression, Online Supplementary File 3) did not differ from results without; we only discuss the latter here.

4.1 | Predictive performance under measurement homogeneity

Measures of predictive performance varied between models. However, within scenarios a switch in measurement strategy for a single predictor did not materially impact the predictive performance (Table 2), with the exception of scenario 8, where the c-statistic and scaled Brier score decreased when the *pragmatic* measurement was used (pregnancy dataset, $N = 75$).

4.2 | Predictive performance under measurement heterogeneity

Table 3 shows estimates of predictive performance under predictor measurement heterogeneity across the different models. The calibration-in-the-large coefficient at validation ranged from -0.70 (95% CI, -1.26; -0.21) to 1.43 (95% CI, 0.31; 2.54), suggesting systematic over- or underestimation of the predicted risks. The calibration slope at validation ranged from 0.50 (95% CI, 0.22; 0.91) to 1.67 (95% CI, 0.83; 3.47), consistent with overfitting (too extreme predictions) and underfitting (too narrow range of predictions), respectively. The differences in c-statistic between derivation and validation were small to moderate, ranging from -0.08 (95% CI, -0.11; -0.07) to +0.08 (95% CI, 0.07; 0.11). The change in scaled Brier score between derivation and validation ranged from -0.40 (95% CI, -0.80; -0.15) to +0.16 (95% CI, 0.15; 0.17). In what follows, we provide a detailed discussion of predictive performance, where we group the scenarios by type of predictor measurement heterogeneity.

In settings where a different measure of aggregation for defining the predictor was used (scenario 1ab-2ab), the direction of miscalibration was related to the shift in aggregational measure (Figure 2). When the maximum tumor diameter was used at derivation and the mean at validation, the calibration-in-the-large coefficient was larger than zero, indicating systematic underestimation of the predicted risks at validation

(scenario 1a, 2a). The reverse occurred in scenario 1b and 2b. Calibration-in-the-large was more strongly affected in scenario 2ab, where the predictor-outcome association was higher than in scenario 1ab.

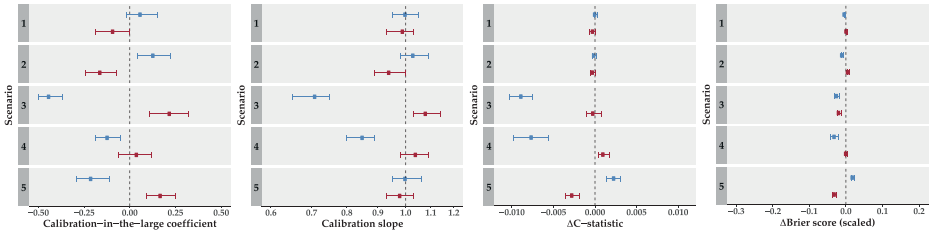


Figure 2. Measures of predictive performance under predictor measurement heterogeneity of a model predicting the probability of having ovarian mass malignancy. The model is applied to the International Ovarian Tumor Analysis (IOTA) dataset, containing information on 5,914 non-pregnant women (1999-2012). Error bars represent the 95-percentile interval over 500 bootstrap samples. Blue error bars indicate scenario 1a-5a, meaning the model was derived using the preferred measurement and validated using the pragmatic measurement, red error bars indicate scenario 1b-5b, meaning the model was derived using the pragmatic measurement and validated using the preferred measurement.

Truncation of a continuous predictor measurement showed the following effects on calibration (scenario 3ab; Figure 2). When the truncated value was used for model derivation and the non-truncated value at validation, the calibration-in-the-large coefficient indicated systematic overestimation of the predicted risks at validation, and the calibration slope was smaller than one, indicating overfitting with respect to the derivation data; predicted risks were too extreme compared to the observed proportions (and vice versa in scenario 3b).

When the categories of an ordinal predictor were collapsed into a binary variable by using only the extremes of the scale (scenario 4a; Figure 2), the calibration-in-the-large coefficient indicated systematic overestimation of the predicted risks, the calibration slope indicated overfitting with respect to the derivation data, and the c-statistic decreased (and vice versa in scenario 4b).

When a more stringent dichotomization was used at validation by shifting the cut-off of a count upward (scenario 5b; Figure 2) or including only first-degree relatives in a summary score on family history, rather than both first- and second-degree relatives (scenario 6a,7a; Figure 3), risks were systematically underestimated, as indicated by the calibration-in-the-large coefficient (and vice versa in 5a, 6b, 7b). In scenario 6a, the calibration slope indicated model underfitting, the c-statistic decreased, and the scaled Brier score decreased (and vice versa in scenario 6b).

Switching from serum to urine hCG measurements (scenario 8ab) showed the following effects on predictive performance (Figure 4). When the predictor measurement had a smaller variance at derivation compared to validation (scenario 8a), the calibration-in-the-large coefficient indicated systematic overestimation of the predicted risks and the calibration slope indicated model overfitting. The c-statistic and scaled Brier score decreased. The reverse occurred when the predictor measurement had lower variance at validation compared to derivation (scenario 8b), except for the scaled Brier score, which decreased again.

A switch in measurement instrument, i.e., using the ria kit versus using the imm kit for hCG measurement in serum (scenario 9ab; Figure 4), minimally affected predictive performance. The large uncertainty around measures of predictive performance in scenario 8ab and 9ab can largely be explained by the limited sample size.

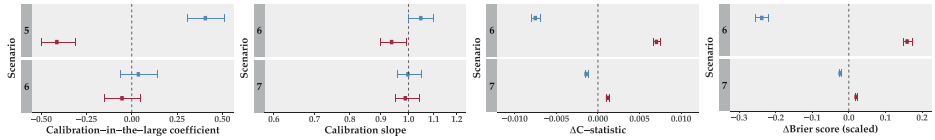


Figure 3. Measures of predictive performance under predictor measurement heterogeneity of a model predicting the probability of having Lynch syndrome–related mismatch repair genes. The model is applied to the Lynch syndrome dataset, containing information on 19,866 patients with colorectal cancer who were tested for mutations. Error bars represent the 95-percentile interval over 500 bootstrap samples. Blue error bars indicate scenario 6a and 7a, meaning the model was derived using the preferred measurement and validated using the pragmatic measurement, red error bars indicate scenario 6b and 7b, meaning the model was derived using the pragmatic measurement and validated using the preferred measurement.

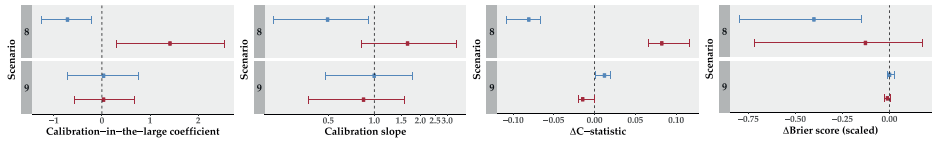


Figure 4. Measures of predictive performance under predictor measurement heterogeneity of a model predicting the probability of intrauterine pregnancy. The model is applied to the pregnancy dataset, containing information on 75 patients at the Early Pregnancy and Acute Gynaecology Unit (EPAGU) at Queen Charlottes' and Chelsea Hospital (2013 – 2014). Error bars represent the 95-percentile interval over 500 bootstrap samples. Blue error bars indicate scenario 8a and 9a, meaning the model was derived using the preferred measurement and validated using the pragmatic measurement, red error bars indicate scenario 8b and 9b, meaning the model was derived using the pragmatic measurement and validated using the preferred measurement.

Table 2 Measures of optimism-corrected predictive performance under predictor measurement homogeneity

Scenario	Event fraction	Measurement strategy		Mean value measurement	Standard deviation measurement	C-statistic	Scaled Brier score
1	0.33	Preferred	Maximal diameter tumor	82.06	52.07	0.94 (0.94; 0.95)	0.62 (0.60; 0.64)
	0.33	Pragmatic	Mean diameter tumor	68.98	42.68	0.94 (0.94; 0.95)	0.62 (0.60; 0.65)
2	0.33	Preferred	Maximal diameter solid component tumor, non-truncated	27.55	39.34	0.94 (0.93; 0.95)	0.60 (0.58; 0.62)
	0.33	Pragmatic	Mean diameter solid component tumor, non-truncated	22.73	32.71	0.94 (0.93; 0.95)	0.60 (0.58; 0.62)
3	0.33	Preferred	Diameter solid component truncated at 50mm	18.91	21.31	0.94 (0.94; 0.95)	0.62 (0.60; 0.64)
	0.33	Pragmatic	Original diameter solid component	27.55	39.34	0.94 (0.93; 0.95)	0.60 (0.58; 0.62)
4	0.33	Preferred	Color-score 4 categories	2.25	0.99	0.94 (0.94; 0.95)	0.62 (0.60; 0.64)
	0.33	Pragmatic	Color-score dichotomous	2.20	1.47	0.94 (0.94; 0.95)	0.61 (0.59; 0.64)
5	0.33	Preferred	≥10 locules	0.08	0.27	0.89 (0.89; 0.90)	0.46 (0.44; 0.49)
	0.33	Pragmatic	≥5 locules	0.19	0.40	0.90 (0.89; 0.91)	0.47 (0.44; 0.49)
6	0.10	Preferred	Family history CRC both FDR and SDR	0.64	0.76	0.78 (0.77; 0.79)	0.16 (0.14; 0.17)
	0.10	Pragmatic	Family history CRC FDR only	0.45	0.66	0.77 (0.76; 0.78)	0.14 (0.13; 0.16)
7	0.10	Preferred	Family history EC both FDR and SDR	0.10	0.31	0.78 (0.77; 0.79)	0.16 (0.14; 0.17)
	0.10	Pragmatic	Family history EC FDR only	0.07	0.28	0.78 (0.77; 0.79)	0.16 (0.14; 0.17)

Table 2 *(continued)*

Scenario	Event fraction	Measurement strategy		Mean value measurement	Standard deviation measurement	C-statistic	Scaled Brier score
8*	0.40	Preferred	hCG level measured in serum, ria kit	2.74 and -0.23	1.44 and 0.86	0.90 (0.81; 0.97)	0.54 (0.32; 0.78)
	0.40	Pragmatic	hCG level measured in urine, ria kit	4.32 and -0.16	1.74 and 1.12	0.81 (0.70; 0.91)	0.27 (0.05; 0.52)
9*	0.40	Preferred	hCG level measured in serum, ria kit	2.74 and -0.23	1.44 and 0.86	0.90 (0.81; 0.97)	0.54 (0.31; 0.78)
	0.40	Pragmatic	hCG level measured in serum, imm kit	2.90 and -0.27	1.41 and 0.84	0.91 (0.83; 0.98)	0.56 (0.32; 0.79)

Measures of predictive performance were averaged over 500 bootstrap samples and corrected for optimism. Confidence intervals for the c-statistic and scaled Brier score were obtained by subtracting the optimism from the 95-percentile interval over the 500 bootstrap estimates of the performance measure under predictor measurement homogeneity. Scaled Brier score is computed as: $1 - Brier/Brier_{max}$.

Abbreviations: CRC = colorectal cancer, EC = endometrial cancer, FDR = first-degree relative, SDR = second-degree relative.

* The hCG measurements are included in the model as a log-transformed hCG measurement at presentation plus a log-transformed ratio of hCG at 48h to hCG-at-presentation measurement.

Table 3 Measures of predictive performance under predictor measurement heterogeneity

Scenario	Measurement strategy at derivation	Measurement strategy at validation	ρ_{part}	Calibration-in-the-large	Calibration slope	ΔC-statistic * 100	Δ scaled Brier score
1a	Maximal diameter tumor	Mean diameter tumor	0.98	0.06 (-0.02; 0.15)	1.00 (0.95; 1.05)	0.00 (-0.00; 0.02)	-0.00 (-0.00; -0.00)
1b	Mean diameter tumor	Maximal diameter tumor		-0.09 (-0.19; -0.00)	0.99 (0.93; 1.03)	-0.02 (-0.06; 0.00)	0.00 (0.00; 0.00)
2a	Maximal diameter solid component tumor, non-truncated	Mean diameter solid component tumor, non-truncated	0.98	0.13 (0.04; 0.22)	1.03 (0.98; 1.09)	-0.00 (-0.00; 0.00)	-0.01 (-0.01; -0.01)
2b	Mean diameter solid component tumor, non-truncated	Maximal diameter solid component tumor, non-truncated		-0.16 (-0.24; -0.07)	0.94 (0.89; 1.00)	-0.00 (-0.00; 0.00)	0.01 (0.00; 0.01)
3a	Diameter solid component truncated at 50mm	Original diameter solid component	0.65	-0.44 (-0.50; -0.37)	0.71 (0.65; 0.75)	-0.88 (-1.02; -0.75)	-0.02 (-0.03; -0.02)
3b	Original diameter solid component	Diameter solid component truncated at 50mm		0.22 (0.11; 0.32)	1.08 (1.03; 1.14)	-0.02 (-0.10; 0.07)	-0.02 (-0.02; -0.01)
4a	Color-score 4 categories	Color-score dichotomous	0.81	-0.12 (-0.19; -0.05)	0.85 (0.80; 0.89)	-0.76 (-0.98; -0.56)	-0.03 (-0.04; -0.02)
4b	Color-score dichotomous	Color-score 4 categories		0.04 (-0.06; 0.12)	1.04 (0.98; 1.09)	0.10 (0.04; 0.17)	0.00 (-0.00; 0.00)
5a	≥10 locules	≥5 locules	0.56	-0.21 (-0.29; -0.11)	1.00 (0.95; 1.06)	0.23 (0.14; 0.31)	0.02 (0.01; 0.02)
5b	≥5 locules	≥10 locules		0.17 (0.09; 0.25)	0.98 (0.93; 1.03)	-0.27 (-0.35; - 0.18)	-0.03 (-0.04; -0.03)
6a	Family history CRC both FDR and SDR	Family history CRC FDR only	0.90	0.41 (0.31; 0.51)	1.05 (1.00; 1.10)	-0.74 (-0.79; -0.70)	-0.24 (-0.26; -0.22)
6b	Family history CRC FDR only	Family history CRC both FDR and SDR		-0.41 (-0.50; -0.31)	0.94 (0.90; 0.99)	0.71 (0.67; 0.76)	0.16 (0.15; 0.17)
7a	Family history EC both FDR and SDR	Family history EC FDR only	0.93	0.04 (-0.06; 0.14)	1.00 (0.96; 1.05)	-0.13 (-0.15; -0.12)	-0.02 (-0.02; -0.02)

Table 3 *(continued)*

Scenario	Measurement strategy at derivation	Measurement strategy at validation	ρ_{part}	Calibration-in-the-large	Calibration slope	ΔC-statistic * 100	Δ scaled Brier score
7b	Family history EC FDR only	Family history EC both FDR and SDR		-0.05 (-0.15;0.05)	0.99 (0.95; 1.04)	0.12 (0.11; 0.14)	0.02 (0.02; 0.03)
8a*	hCG level measured in serum, using the ria kit	hCG level measured in urine, using the ria kit	0.61 and 0.92	-0.70 (-1.26; -0.21)	0.50 (0.22; 0.91)	-7.97 (-10.80; -6.59)	-0.40 (-0.80; -0.15)
8b*	hCG level measured in urine, using the ria kit	hCG level measured in serum, using the ria kit		1.43 (0.31; 2.54)	1.67 (0.83; 3.47)	8.34 (6.67; 11.63)	-0.12 (-0.72; 0.17)
9a*	hCG level measured in serum, using the ria kit	hCG level measured in serum, using the imm kit	0.98 and 0.997	0.05 (-0.71; 0.75)	1.01 (0.48; 1.78)	1.31 (0.07; 2.00)	0.00 (-0.01; 0.03)
9b*	hCG level measured in serum, using the imm kit	hCG level measured in serum, using the ria kit		0.05 (-0.56; 0.68)	0.86 (0.37; 1.58)	-1.36 (-2.02; 0.00)	-0.01(-0.03; 0.01)

Performance measures under predictor measurement heterogeneity: median calibration coefficients and mean difference scores of c-statistic and scaled Brier score over 500 bootstrap samples with 95-percentile intervals. Δ indicates that the measure of predictive performance under predictor measurement homogeneity is subtracted from the performance measure under predictor measurement heterogeneity. Scaled Brier score is computed as: $1 - Brier/Brier_{max}$.
Abbreviations: CRC = colorectal cancer, EC = endometrial cancer, FDR = first-degree relative, SDR = second-degree relative.
* The hCG measurements are included in the model as a log-transformed hCG measurement at presentation plus a log-transformed ratio of hCG at 48h to hCG-at-presentation measurement.

5 | Discussion

In this study, we evaluated the impact of predictor measurement heterogeneity in nine different scenarios in three clinical datasets. A change in measurement strategy of a predictor within the derivation set, from *preferred* measurement to *pragmatic* measurement or vice versa, minimally affected measures of predictive performance in our example studies. We found that heterogeneity of measurements across settings of derivation and validation can have a substantial impact on the performance of a prediction model, most notably on overall accuracy and calibration of risk predictions, resulting in systematic over- or under estimation of predicted risks and risk models that are consistent with overfitting (systematically too extreme predictions) or underfitting (systematically a too narrow range of predictions).

In the examples, the impact on calibration was larger when predictors were stronger associated with the outcome or when the partial correlation between predictor measurement strategies was lower. Using Ridge regression as a shrinkage method or correcting for optimism did not compensate for effects of measurement heterogeneity in our study. The variety of effects on predictive performance in the examples illustrated the difficulty of anticipating the exact impact of predictor measurement heterogeneity, emphasising the need to be generally mindful of (dis)similarities of predictor measurement strategies between derivation and validation studies.

We observed small effects of predictor measurement heterogeneity on the discriminatory power of the model at validation in our examples. Previous simulation studies found larger effects on the c-statistic[10-12]. Our finding may be explained by the fact that we focused on within-sample predictive performance under measurement heterogeneity in a single predictor. With a larger number of predictors subject to measurement heterogeneity, we anticipate the combined effect on the discrimination performance can be larger. Also, given that the c-statistic is a rank order statistic it is possible that this metric is less affected by measurement heterogeneity[33].

Our findings showed that internal predictive performance may not be affected by changes in predictor measurement strategy within the same dataset, in line with previous studies[10,11,34]. Previous research showed that variations in measurement error did not affect risk calibration[10], but these findings were restricted to within-sample effects on predictive performance only. Within the derivation dataset, models derived using logistic regression achieve, by definition, a calibration-in-the-large coefficient of zero and calibration slope of one regardless of the measurement error structure of

predictors[29]. Our study highlights that this does not apply when the degree or structure of measurement error varies across settings of derivation and validation, the case of measurement heterogeneity.

It is common practice in validation studies to quantify the relatedness of derivation and validation samples by inspecting the distribution of the linear predictors, also referred to as comparison of *case-mix* distributions[8,9,35]. Dissimilarities in the distributions of the linear predictor between derivation and validation may rise from both actual differences in patient characteristics and differences in the procedures used to measure patient characteristics. By identifying predictor measurement heterogeneity as a separate explanation of discrepancies in linear-predictor distributions across settings, our findings can facilitate implementation of the influential TRIPOD statement in clinical prediction research[36].

Our study has several limitations. Firstly, it was limited to three empirical datasets with a diagnostic outcome modelled using logistic regression. One dataset, from the IOTA study, was a multicenter study in which homogeneous measurement strategies across centers was among its hallmark characteristics, see[19]. Measurement heterogeneity within development and validation studies, e.g., due to variability in measurement precision between clinicians or centres[37], is an important topic for future research. Given the potential impact and limited attention to date[38], research is needed on the effect of measurement heterogeneity for other statistical models and outcomes (e.g., survival models for time-to-event outcomes) and the impact on more flexible prediction modelling strategies. Finally, the similarity between the preferred and pragmatic measurement of a predictor was quantified using a partial correlation coefficient. This measure quantifies the conditional association between predictor measurements rather than agreement[39]. Since the current paper aimed to examine whether variation in predictor measurement strategies across settings can have an effect on predictive performance of any degree or direction, we presented a single measurement of similarity of predictor measurements and left out further quantification. One way to visualize agreement between measurement could be Bland-Altman plots[40].

The following recommendations follow from our work. When a prediction model is derived, predictor measurements should be clearly defined and ideally resemble procedures in the intended setting of application as closely as possible. For prediction model validation studies, we encourage researchers to investigate to which extent predictor measurement procedures are homogeneous and may have contributed to differences in predictive performance between the validation and derivation setting.

Accurate reporting of predictor measurement heterogeneity in both derivation and validation studies is therefore essential. Furthermore, we take the position that addressing measurement heterogeneity at the data collection stage is preferred over statistical correction for measurement error in predictors. Corrections – typically aiming to alleviate measurement-error bias in regression coefficients – may increase rather than reduce the measurement heterogeneity[12].

We emphasize that consideration of predictor measurement heterogeneity is crucial also in the implementation stage of a prediction model in clinical practice. Deployment of a prediction model might alter predictor measurement heterogeneity. For example, after implementation of a prediction model, physicians may be recommended to use a more precise or standardized measurement (or even routinely measure predictors that were not measured in all patients up to that point). For implementation of prediction models in clinical practice, our findings indicate that measurement procedures should follow the measurements in derivation and validation datasets as closely as possible.

In summary, our findings highlight that predictor measurement heterogeneity can have a substantial influence on the performance of a prediction model, most notably on risk calibration. Explicit reporting of the procedures and timing involved in measurement of predictors in derivation and validation studies is vital to improve the performance and applicability of prediction models in clinical practice.

Online Supplementary Files

The supplementary files referred to in this Chapter are available online at https://doi.org/10.1016/j.jclinepi.2019.11.001

References

1. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Springer Science & Business Media; 2008.
2. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine.* 2000;19(4):453-473.
3. 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.
4. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *British Medical Journal.* 2015;350:g7594.
5. 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.
6. 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.
7. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine.* 1999;130(6):515-524.
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. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology.* 2010;172(8):971-980.
10. 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.
11. 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):20.
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. 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.
14. Pajouheshnia R, Groenwold RH, Peelen LM, Reitsma JB, Moons KG. When and how to use data from randomised trials to develop or validate prognostic models. *British Medical Journal.* 2019;365:l2154.
15. Te Velde E, Nieboer D, Lintsen A, et al. Comparison of two models predicting IVF success; the effect of time trends on model performance. *Human Reproduction.* 2013;29(1):57-64.
16. Smith T, Muller DC, Moons KG, et al. Comparison of prognostic models to predict the occurrence of colorectal cancer in asymptomatic individuals: a systematic literature review and external validation in the EPIC and UK Biobank prospective cohort studies. *Gut.* 2019;68(4):672-683.
17. Control CfD, Prevention. National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Questionnaire (or Examination Protocol, or Laboratory Protocol). *http://wwwcdcgov/nchs/nhaneshtm.* 2006.
18. Control CfD, Prevention. National health and nutrition examination survey (NHANES): Anthropometry procedures manual. *National Center for Health Statistics, editor Atlanta, GA: Centers for Disease Control.* 2007.

19. Timmerman D, Valentin L, Bourne T, Collins W, Verrelst H, Vergote I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 2000;16(5):500-505.
20. Timmerman D, Testa AC, Bourne T, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. *Journal of Clinical Oncology*. 2005;23(34):8794-8801.
21. Van Holsbeke C, Van Calster B, Testa AC, et al. Prospective internal validation of mathematical models to predict malignancy in adnexal masses: results from the international ovarian tumor analysis study. *Clinical Cancer Research*. 2009;15(2):684-691.
22. Timmerman D, Van Calster B, Testa AC, et al. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. *Ultrasound in Obstetrics and Gynecology*. 2010;36(2):226-234.
23. Nunes N, Ambler G, Hoo W-L, et al. A prospective validation of the IOTA logistic regression models (LR1 and LR2) in comparison to subjective pattern recognition for the diagnosis of ovarian cancer. *International Journal of Gynecological Cancer*. 2013;23(9):1583-1589.
24. Wynants L, Vergouwe Y, Van Huffel S, Timmerman D, Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. *Statistical Methods in Medical Research*. 2018;27(6):1723-1736.
25. Balmaña J, Stockwell DH, Steyerberg EW, et al. Prediction of MLH1 and MSH2 mutations in Lynch syndrome. *Journal of the American Medical Association*. 2006;296(12):1469-1478.
26. Barnetson RA, Tenesa A, Farrington SM, et al. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *New England Journal of Medicine*. 2006;354(26):2751-2763.
27. Van Calster B, Abdallah Y, Guha S, et al. Rationalizing the management of pregnancies of unknown location: temporal and external validation of a risk prediction model on 1962 pregnancies. *Human Reproduction*. 2013;28(3):609-616.
28. 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.
29. 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.
30. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 1992;41(1):191-201.
31. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020.
32. Harrell Jr FE. *rms: regression modeling strategies*. R package version 5.1-4. 2016.
33. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clinical Chemistry*. 2008;54(1):17-23.
34. Carroll RJ, Ruppert D, Crainiceanu CM, Stefanski LA. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC; 2006.
35. Kundu S, Mazumdar M, Ferket B. Impact of correlation of predictors on discrimination of risk models in development and external populations. *BMC Medical Research Methodology*. 2017;17(1):63.
36. 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.
37. Wynants L, Timmerman D, Bourne T, Van Huffel S, Van Calster B. Screening for data clustering in multicenter studies: the residual intraclass correlation. *BMC Medical Research Methodology*. 2013;13(1):128.

38. 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. *Journal of Clinical Epidemiology*. 2018;102:38-49.

39. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. 1996;1(1):30.

40. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*. 1986;327(8476):307-310.