# Evaluating Reproducibility and Generalizability of a Medical Predictive Model in Clinical Research

## Fengnan Li* and Shein Chung Chow

*Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, USA*

## Abstract

In clinical research, a medical predictive model is intended to provide insight into the impact of risk factors (predictors) such as demographics and patient characteristics on clinical outcomes. A validated medical predictive model informs disease status and treatment effects under study. More importantly, it can be used for disease management. However, a gap in the development process of these models is often observed. That is, most studies only focus on the internal validation for the model's reproducibility but overlook the external validation needed for evaluating generalizability. To solve this issue, this article proposes several methods for assessing both the reproducibility and generalizability of a developed/ validated medical predictive model. The generalizability estimation approaches allow for sensitivity analysis in situations where data on new populations is not available, which provides valuable insights into the model's applicability to patients from a different population.

**Keywords:** Internal validation • Reproducibility • External validation • Generalizability

## Introduction

Medical predictive models have long played a crucial role in clinical research and practice, especially in recent times when personalized medicine has become increasingly popular. Usually, a medical predictive model is designed to express a clinical outcome (response) as a function of a multivariate set of predictors (risk factors). It offers valuable insight into the association between clinical outcomes and predictors and more importantly, facilitates disease management. Such models can be either diagnostic or prognostic. A diagnostic model reflects the patient's current health conditions, while a prognostic model estimates the probability or risk of developing an outcome over a specific period. For example, the Framingham Risk Score is a prognostic medical predictive model that takes a multivariate set of risk factors as input and estimates the 10-year risk of developing cardiovascular disease [1].

The procedure for developing a medical predictive model involves the following steps:

(i) Predictor Identification, where potential predictors are identified and a subset to be included in the model is selected *via* regression methods.

(ii) Collinearity testing, which involves detecting and addressing collinearity among predictors to enhance the model's parsimony and interpretability.

(iii) Model Building, which estimates the coefficients by fitting a regression model based on the identified predictors and assesses the goodness-of-fit.

(iv) Model Validation, comprising interval validation for reproducibility and external validation for generalizability.

Among the four critical steps discussed above, the current issue often

arises from the phase of model validation. Internal validation for reproducibility refers to the model's ability to replicate the results when applied to new data from the same population (e.g., data collected from the same medical center). External validation for generalizability refers to the extent to which the predictive model is applicable across different populations (e.g., different age groups or ethnic groups). Ideally, a comprehensive validation of a medical predictive model should include both internal and external validations. However, in practice, we find that most of the model's development only performs the internal validation. The process of external validation to assess the model's generalizability is often hindered by the unavailability or the prohibitive cost of acquiring the new population's data. In response to the issue in the model's validation, we propose several methods for estimating both the reproducibility and generalizability probability of a medical predictive model. The approaches for generalizability assessment are developed by adapting the reproducibility probability estimation methods, which effectively estimate the model's generalizability to new populations.

The remaining parts of the paper are organized as follows. In the first section, we summarize the procedure for developing a medical predictive model and discuss the critical issue regarding the model's validation. Then we address this issue by proposing methods for reproducibility and generalizability assessment. Concluding remarks are given in the last section of this article.

## Predictive model building and validation

### Procedure for model development

Usually, the procedure for developing a medical predictive model in practice involves the following four steps: predictor identification, collinearity testing, model building, and model validation, which are discussed in detail as follows.

**Predictor identification:** The foundational step in building a medical predictive model is to identify the predictors (risk factors) to be included in the model. Potential predictors are obtained from demographics and patient characteristics information at baseline. To ascertain the relationship between a predictor and the clinical outcome, the Pearson correlation coefficient can be applied as a measure of correlation. A commonly used rule of thumb is that a correlation coefficient greater than 0.3 indicates a moderate correlation and we can consider including the predictor in the model. After collecting the set of correlated predictors, we need to further assess whether the inclusion of a predictor can improve the model's predictive ability. A subset of all the correlated predictors is obtained by sequential selection methods like forward, backward, or stepwise selection, which forms the final set of predictors in the model.

***Address for Correspondence***: *Fengnan Li, Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina, USA, E-mail: fengnan.li@duke.edu, fengnan.li.77@gmail.com*

**Collinearity testing:** Particular attention should be paid to the collinearity among predictors. Even though the collinearity issue does not affect the overall predictive power of the model, it seriously impedes the model's interpretability by inflating the coefficient and confounding the corresponding t-statistics [2,3]. Also, the correlated predictors add unnecessary complexity to the model. By removing or combining these predictors, we can reduce the number of predictors in the model and thus improve the model's performance in fitting new, unseen data.

In addition to the correlation coefficient that measures the collinearity between two predictors, the Variance Inflation Factor (VIF) is applied to quantify the correlation between one predictor and the collective of other predictors in the model. Furthermore, the Condition Number (CN) provides a measure of the overall multicollinearity level within the model [4]. Once a high level of collinearity within the predictors is detected, various techniques can be employed to address the collinearity. We suggest the use of a composite index approach proposed by Chow SC, et al. [5]. This method develops a composite index *via* ordinary least squares regression to combine multiple correlated predictors into a single predictor. This approach stands out among other collinearity resolution techniques like PCA because it accounts for the outcome variable and preserves the predictors' correlation with the outcome.

**Model building:** Previous steps yield the final set of predictors, with the collinearity issue being addressed. Statistical models utilized to build a medical predictive model are usually generalized linear models. A linear regression model is chosen for continuous clinical outcomes. Logistic regression is often utilized for binary outcomes. For time-to-event outcomes, Cox regression is applied. After fitting the data upon a specified model, the goodness-of-fit needs to be measured. For continuous outcomes, $R^2$ or adjusted $R^2$ is commonly applied to evaluate the model's performance in fitting the data. For categorical clinical outcomes, the evaluation of the model's performance involves two aspects: discrimination and calibration. Discrimination measures how well the model distinguishes between different outcome categories and calibration assesses how well the predicted probabilities match the observed frequencies.

**Model validation:** The final step is model validation which involves internal validation for reproducibility and external validation for generalizability. Reproducibility denotes the ability to replicate the results when the model is applied to new data from the same population (e.g., data collected from the same medical center). In contrast, generalizability refers to the extent to which the model's predictions are applicable across various populations [6].

## Critical issue in model development

A critical issue in the predictive model development process is the absence of external validation. After a model is validated in one population, it is often of interest to evaluate whether the model is applicable across different populations (e.g., different age groups or different ethnic groups). Although researchers always aspire to broaden the model's applicability and theoretically an external validation should be performed to evaluate the model's generalizability, in practice, we find that most of the predictive model development efforts halt at the phase of internal validation. This premature termination in model development is often due to the unavailability or the prohibitive cost of acquiring the new population's data. To solve this issue, we first propose three methods for estimating the reproducibility probability of a predictive model. Subsequently, we adapt these reproducibility methods to enable the assessment of generalizability, which serves as a substitute sensitivity analysis when the data of other populations are not present. More details about the proposed methods are discussed in below section.

## Statistical methods for assessing reproducibility and generalizability

Internal validation for reproducibility assesses whether results from one study are reproducible in another study where data are collected from the same population, whereas external validation for generalizability evaluates whether the model is applicable across various populations. We

find most of the validation process for medical predictive models in practice focuses solely on the internal validation within the same medical centre and the external validation for generalizability is missing. This section aims to resolve this issue by proposing three approaches to estimating the model's generalizability probability. The organization of this section is as follows: In first, section estimation of reproducibility probability proposes three approaches to estimating the model's reproducibility probability. Building upon these reproducibility estimation methods, In second, section estimation of generalizability probability adjusts the approaches in section estimation of reproducibility probability to allow for generalizability probability estimation. The reproducibility or generalizability probability can be defined by either the relative difference or the absolute difference between the true clinical outcome and the predicted clinical outcome. That is,

$$ P\{|\ \hat{y} - y\ | < \delta\} \text{ or } P\{|\ \frac{\hat{y} - y}{y}\ | < \delta\} $$

Where $\delta$ is pre-determined, clinically negligible value. The following sections use $P\{|\ \hat{y} - y\ | < \delta\}$ as the reproducibility probability as $\hat{y} - y$ follows a normal distribution, which yields more analytical results. Moreover, the methods proposed are in the context of a linear regression model, where a closed-form solution for coefficients is utilized to estimate the probability.

## Estimation of reproducibility probability

This section provides three approaches to estimating reproducibility probability, which aims to establish a foundation for the generalizability estimation methods proposed in this section Shao J and Chow SC, et al. [7] proposed three methods for estimating reproducibility probability in clinical trials, the estimated power approach, the confidence bound approach, and the Bayesian approach, to assess the probability of replicating results of the first clinical trial in a subsequent trial with the same study protocol. The key ideas are as follows: the estimated power approach directly replaces the model's parameter in the second trial with its estimate obtained from the first trial; the confidence bound approach accounts for the variation in the estimated parameter and thus uses a lower confidence bound for more conservative estimation and the Bayesian approach calculates the reproducibility probability as a posterior mean. Although the methodologies in medical predictive models and clinical trials are completely different, these ideas can be referred to when estimating the reproducibility probability for predictive models by analogizing the model-building process to the first clinical trial and the validation phase to the second clinical trial.

**Approach 1:** Similar to the estimated power approach proposed in Shao, et al.'s paper, approach 1 substitutes the estimated regression coefficients $\hat{\beta}$ for the true regression coefficients $\beta$. Given an observation $(X^*, y^*)$ in the validation dataset, we have

$$ y^* = X^{*T}\beta + \epsilon^* \qquad (1) $$

Where $\epsilon^* \sim N(0, \sigma^2), X^*, \beta \in \mathbb{R}^p$ and $p$ is the number of parameters in the model

The estimated response $\hat{y}^*$ equals $X^{*T}\hat{\beta}$. If we replace $\beta$ by the estimated coefficient $\hat{\beta}$ in Eq. (1), the difference between the estimated response value $\hat{y}^*$ and observed response value $y^*$ would solely result from the error term $\epsilon^*$. Then the reproducibility probability $p_1$ is:

$$ p_1 = P\{|\widehat{y^*} - y^*| < \delta|\hat{\beta}\} = P\{|\epsilon^*| < \delta\} = 1 - 2\Phi\left(-\frac{\delta}{\sigma}\right) \qquad (2) $$

$\sigma^2$ can be estimated by $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \left(X_i^T\hat{\beta}\right)\right)^2$, where $(X_i, y_i)$ is the $i$th data point in the training set that has $n$ observations in total.

$p_1$ estimates the reproducibility probability for a given observation. Utilizing $p_1$, the overall reproducibility of the validation dataset can be evaluated. Assuming the validation dataset has $n$ observations, the count of reproducible observations follows a Binomial distribution $M \sim Bin(n, p_1)$. For large sample sizes, this approximates a normal distribution $N(np_1, np_1(1-p_1))$. For instance, to estimate the probability that at least 95% of the data points are reproducible, we use

$$P\left(M \geq 0.95n\right) = 1 - \Phi\left(\frac{0.95n - np_1}{\sqrt{np_1\left(1 - p_1\right)}}\right),$$

where $\Phi$ denotes the cumulative distribution function of the standard normal distribution.

**Approach 2:** The reproducibility probability estimated by approach 1 is rather optimistic, especially when the variance of $\hat{\beta}$ is large. In contrast, approach 2 accounts for the variation of the estimated coefficient $\hat{\beta}$ rather than directly substitute it for the true parameter.

The notations are as follows: in the dataset used for model building, we have $Y = X\beta + \epsilon$, where $Y \in \mathbb{R}^n$ represents the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix for predictors, $\beta \in \mathbb{R}^p$ is the coefficient vector, and $\epsilon \in \mathbb{R}^n$ is the error term with each element $\epsilon_i$ following a normal distribution $N(0, \sigma^2)$. In the validation dataset, for a new observation $\left(X^*, y^*\right)$, The model is

$$y^* = X^{*T}\beta + \epsilon^*$$

Where $\epsilon^* \sim N\left(0, \sigma^2\right)$ and $X^*, \beta \in \mathbb{R}^p$. The estimated coefficient vector is $\hat{\beta} = (X^TX)^{-1}X^TY$. The same notations will be used throughout this section. $\hat{\beta}$ is an unbiased estimator, the expected value of $\widehat{y^*} - y^*$ is 0. The variance of the residual is given by

$$var\left(\widehat{y^*} - y^*\right) = var\left(X^{*T}\hat{\beta}\right) + var\left(\epsilon\right) = X^{*T}var\left(\hat{\beta}\right)X^* + \sigma^2 = \sigma^2 X^{*T}\left(X^TX\right)^{-1}X^* + \sigma^2$$

Since both $\hat{y}^*$ and $y*$ are normally distributed, we have

$$\hat{y}^* - y^* \sim N\left(0, \sigma^2 X^{*T}\left(X^TX\right)^{-1}X^* + \sigma^2\right).$$

Therefore, for approach 2, the reproducibility probability for an observation $\left(X^*, y^*\right)$ is

$$p_1 = P\left\{\left|\widehat{y^*} - y^*\right| < \delta\right\} = 1 - 2\Phi\left(-\frac{\delta}{\zeta_1}\right),$$

$$\text{where } \zeta_1^2 = \sigma^2 X^{*T}\left(X^TX\right)^{-1}X^* + \sigma^2 \ (\zeta_1 > 0) \tag{3}$$

**Approach 3 (Bayesian approach):** Approach 3 utilizes Bayesian inference and prediction for reproducibility probability allows us to estimate calculation, which the uncertainty in the regression coefficients $\beta$.

Denote data points in the training set as $D$ and assume $\sigma^2$, the variance of the error term $\epsilon$, is known already. From the Bayesian perspective, the posterior distribution $\pi\left(\beta|D, \sigma\right)$ is expressed as

$$\pi(\beta|D, \sigma) = \pi(\beta|X, Y, \sigma) = \frac{p(Y|X, \beta, \sigma), \pi(\beta|\sigma)}{\int p(Y|X, \beta, \sigma)P(\beta|\sigma)d\beta} \tag{4}$$

The likelihood $P(Y|X, \beta, \sigma)$ is normal and $Y|X, \beta, \sigma \sim N\left(X\beta, \sigma^{-2}I\right)$. The posterior distribution will be normal if a conjugate normal prior is specified. Suppose our choice for the prior is $\pi(\beta|\sigma) = N(\mu_0, \Sigma_0)$. The posterior distribution $\pi\left(\beta|D, \sigma\right)$ can be derived subsequently, which is given in the form of

$$\beta|D, \sigma \sim N\left(\mu_\beta, \Sigma_\beta\right)$$
$$\text{where } \Sigma_\beta = \left(\frac{1}{\sigma^2}X^TX + \Sigma_0^{-1}\right)^{-1}$$
$$\mu_\beta = \Sigma_\beta\left(\frac{1}{\sigma^2}X^TY + \Sigma_0^{-1}\mu_0\right) \tag{5}$$

Similarly, if we choose a non-informative prior of $P\left(\beta|\sigma\right) \equiv 1$, the posterior distribution will be $N\left(\mu_\beta, \Sigma_\beta\right)$ with $\Sigma_\beta = \left(\frac{1}{\sigma^2}X^TX\right)^{-1}$ and $\mu_\beta = \frac{1}{\sigma^2}\Sigma_\beta X^TY$.

Based on the posterior distribution $N\left(\mu_\beta, \Sigma_\beta\right)$, the posterior predictive distribution for a new observation $(X^*, y^*)$ in the validation dataset can be derived. Given $y^* = X^{*T}\beta + \epsilon^*$, with $\beta|\sigma, D \sim N\left(\mu_\beta, \Sigma_\beta\right)$ and $\epsilon^* \sim N(0, \sigma^2)$, the posterior predictive distribution $P\left(y^*|X^*, D, \sigma\right)$ is then expressed as

$$y^*|X^*, D, \sigma \sim N\left(X^{*T}\mu_\beta, X^{*T}\Sigma_\beta X^* + \sigma^2\right) \tag{6}$$

Our estimation $\widehat{y^*}$ for the response value $\hat{y}$ is the mean of posterior predictive distribution, which is $\widehat{y^*} = X^{*T}\mu_\beta$ Consequently, the distribution of $\widehat{y^*} - y^*$ should be $N\left(0, X^{*T}\Sigma_\beta X^* + \sigma^2\right)$. And the reproducibility probability $p_1$ is given by

$$p_1 = P\left\{\left|\widehat{y^*} - y^*\right| < \delta\right\} = 1 - 2\Phi\left(-\frac{\delta}{\zeta_2}\right),$$

$$\text{where } \zeta_2^2 = X^*\Sigma_\beta X^{*T} + \sigma^2 \ (\zeta_2 > 0) \tag{7}$$

For a more realistic case in practice where $\sigma^2$ is unknown, we need to specify a prior for the error term variance. Now the prior distribution is

$$\pi\left(\beta, \sigma^2\right) = \pi\left(\beta|\sigma^2\right) \times \pi\left(\sigma^2\right) \tag{8}$$

A conjugate prior that leads to an analytical posterior distribution would be the Normal-Inverse-Gamma distribution, where

$$\beta|\sigma^2 \sim N\left(\mu_0, \sigma^2\Sigma_0\right)$$
$$\sigma^2 \sim \text{Inv-Gamma}\left(a_0, b_0\right) \tag{9}$$

Given the conjugate prior, the posterior distribution $\pi\left(\beta, \sigma^2|X, Y\right) = \pi\left(\beta|X, Y, \sigma^2\right) \times \pi\left(\sigma^2|X, Y\right)$ can be derived. The result is expressed as

$$\beta|X, Y, \sigma^2 \sim N\left(\mu_\beta, \sigma^2\Sigma_\beta\right)$$
$$\sigma^2|X, Y \sim \text{Inv-Gamma}\left(a_\sigma, b_\sigma\right) \tag{10}$$

Where,

$$\Sigma_\beta = \left(\Sigma_0^{-1} + X^TX\right)^{-1}$$
$$\mu_\beta = \Sigma_\beta\left(\Sigma_0^{-1}\mu_0 + X^TY\right)$$
$$a_\sigma = a_0 + \frac{n}{2}, \tag{11}$$
$$b_\sigma = b_0 + \frac{1}{2}\left(Y^TY + \mu_0^T\Sigma_0^{-1}\mu_0 - \mu_\beta^T\Sigma_\beta^{-1}\mu_\beta\right)$$

Subsequently, for a new observation $y^* = X^{*T}\beta + \epsilon^*$, the posterior predictive distribution $p\left(y^*|X^*, D\right)$ is

$$p\left(y^*|X^*, D\right) = \int p\left(\sigma^2|X, Y\right)\left[\int p\left(y^*|X^*, \beta, \sigma^2\right)p\left(\beta|X, Y, \sigma^2\right)d\beta\right]d\sigma^2 \tag{12}$$

Integrating both the uncertainty in $\beta$ and $\sigma^2$, the posterior predictive distribution would follow a Student's t distribution, which is given by

$$y^*|X^*, D \sim t_{2a_\sigma}\left(X^{*T}\mu_\beta, \frac{b_\sigma + b_\sigma X^{*T}\Sigma_\beta X^*}{a_\sigma}\right) \tag{13}$$

where $X^{*T}\mu_\beta$ is the location parameter and $\frac{b_\sigma + b_\sigma X^{*T}\Sigma_\beta X^*}{a_\sigma}$ is the scale parameter. We take $X^{*T}\mu_\beta$, the mean of the posterior distribution as our prediction for $y^*$. The difference $\hat{y}^* - y^*$ would follow a t-distribution of $t_{2a_\sigma}\left(0, \frac{b_\sigma + b_\sigma X^{*T}\Sigma_\beta X^*}{a_\sigma}\right)$. The reproducibility probability can be subsequently estimated by transforming the random variable into a standard t-distribution, which is expressed as

$$p_1 = P\left\{\left|\widehat{y^*} - y^*\right| < \delta\right\} = 1 - 2 * T_{2a_\sigma}\left(-\frac{\delta}{\zeta_3}\right),$$

$$\text{where } \zeta_3^2 = \frac{b_\sigma + b_\sigma X^{*T}\Sigma_\beta X^*}{a_\sigma} \ (\zeta_3 > 0) \tag{14}$$

$T_{2a_\sigma}$ denotes the cumulative distribution function of a standard t-distribution with $2a_\sigma$ degrees of freedom.

Approach 3 presents the estimation of reproducibility probability by employing conjugate priors for scenarios with known and unknown error term variance. In addition, in cases where a non-conjugate prior is selected, we can apply computational methods like Markov Chain Monte Carlo (MCMC)

to sample from the posterior distributions and approximate the posterior predictive distribution for new observations.

## Estimation of generalizability probability

An application of the reproducibility estimating approach is to assess the model's generalizability. In practice, following the evaluation of a medical predictive model's reproducibility through internal validation, there is commonly an interest in assessing the model's generalizability to a similar but distinct population (for example, varying age groups or ethnic groups). For these populations, variations in regression coefficients are anticipated. Based on these variations, we adapt the reproducibility estimating approaches to assess generalizability.

**Adjusted approach 1 for generalizability:** Approach 1 estimates the reproducibility probability from the frequentist perspective, where regression coefficients $\beta$ is viewed as a set of fixed, unknown values. We use $\lambda$ to denote the deviation of coefficients when the medical predictive model is applied to a new population. The coefficient vector for the new population will be $\beta' = \beta + \lambda$, where the $i^{\text{th}}$ element of $\lambda$, $\lambda_i$, is the difference between the $i^{\text{th}}$ regression coefficient between the two populations. For the first approach, the reproducibility probability $p_1$ is given by

$$p_1 = P\left\{ \left| \hat{y}^* - y^* \right| < \delta \middle| \hat{\beta} \right\} = P\left\{ \left| X^{*T}\hat{\beta} - \left(X^{*T}\hat{\beta} + \epsilon\right) \right| < \delta \right\}$$

Considering the deviation of coefficients, the generalizability probability $p_2$ can be estimated by

$$\begin{aligned} P_2 &= P\left\{ \left| \widehat{y^*} - y^* \right| < \delta \middle| \hat{\beta} \right\} \\ &= P\left\{ \left| X^{*T}\hat{\beta} - \left( X^{*T}\left(\hat{\beta} + \lambda\right) + \epsilon \right) \right| < \delta \right\} \\ &= P\left\{ \left| X^{*T}\lambda + \epsilon \right| < \delta \right\} \end{aligned}$$

$X^{*T}\lambda + \epsilon \sim N\left(X^{*T}\lambda, \sigma^2\right)$, the generalizability probability is

$$p_2 = \Phi\left(\frac{\delta - X^{*T}\lambda}{\sigma}\right) - \Phi\left(\frac{-\delta - X^{*T}\lambda}{\sigma}\right) \qquad (15)$$

Where $\sigma^2$ can be estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \left(X_i^T\hat{\beta}\right)\right)^2$$

**Adjusted approach 2 for generalizability:** Similar to approach 1, for approach 2, we replace $y^* = X^{*T}\beta + \epsilon$ in Eq. (1) by $y^* = X^{*T}(\beta + \lambda) + \epsilon$. Then the expectation $E[\hat{y}^* - y^*]$ equates to $E[X^{*T}\hat{\beta} - (X^{*T}(\beta + \lambda) + \epsilon)] = X^{*T}\lambda$ The variance $var\left(\widehat{y^*} - y^*\right)$ remains the same as it is in Eq. (3), given that the added term $X^{*T}\lambda$ is a constant. Therefore, the distribution of $\widehat{y^*} - y^*$ becomes $N\left(X^{*T}\lambda, \sigma^2 X^{*T}\left(X^T X\right)^{-1}X^* + \sigma^2\right)$ and the generalizability probability $p_2$ is given by

$$p_2 = P\left\{ \left| \widehat{y^*} - y^* \right| < \delta \right\} = \Phi\left(\frac{\delta - X^{*T}\lambda}{\zeta_1}\right) - \Phi\left(\frac{-\delta - X^{*T}\lambda}{\zeta_1}\right),$$
$$\text{where } \zeta_1^2 = \sigma^2 X^{*T}\left(X^T X\right)^{-1}X^* + \sigma^2 \; (\zeta_1 > 0) \qquad (16)$$

$\sigma^2$ can be estimated by $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \left(X_i^T\hat{\beta}\right)\right)^2$.

**Adjusted approach 3 for generalizability:** For approach 3, the Bayesian method, the coefficients $\beta$ are not fixed values but random variables assigned to a probability distribution. Suppose the transition of regression coefficients from the original population to a new population is characterized by a shift of mean, denoted by $\lambda$ and a recalibration of variance, denoted by $c^2 (c > 0)$. In other words, $N\left(\mu_{\beta_i}, \sigma_{\beta_i}^2\right)$ the distribution of the $i^{\text{th}}$ coefficient $\beta_i$ is replaced by $N\left(\mu_{\beta_i} + \lambda_i, c_i^2 \sigma_{\beta_i}^2\right)$. Consequently, the covariance between $\beta_i$ and $\beta_j$ will be scaled to $c_i c_j \text{cov}\left(\beta_i, \beta_j\right)$. Now we can construct a diagonal matrix $C$, where the diagonal elements $C_{ii} = c_i$ and non-diagonal elements are 0. The

covariance matrix $\Sigma_\beta$ is then transformed into $C\Sigma_\beta C$ for a new, generalized population.

Therefore, under the premise that $\sigma^2$ is known, the posterior distribution, initially $N\left(\mu_\beta, \Sigma_\beta\right)$ undergoes a transition to

$$\beta \mid D, \sigma^2 \sim N\left(\mu_\beta + \lambda, C\Sigma_\beta C\right)$$

In the scenario that $\sigma^2$ is unknown, the posterior distribution for $\beta$, initially $N(\mu_\beta, \sigma^2 \Sigma_\beta)$, is modified to

$$\beta \mid D, \sigma^2 \sim N\left(\mu_\beta + \lambda, \sigma^2 C\Sigma_\beta C\right)$$

Where

$$\lambda = \left(\lambda_1, \lambda_2, \lambda_3 \ldots, \lambda_p\right)^T \text{ and } C = \begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_p \end{bmatrix}$$

Consequently, when $\sigma^2$ is known, the posterior predictive distribution for a similar but distinct population is given by substituting $\mu_\beta + \lambda$ and $C\Sigma_\beta C$ respectively for $\mu_\beta$ and $\Sigma_\beta$ in Eq. (6), which results in $N\left(X^{*T}\left(\mu_\beta + \lambda\right), X^{*T}C\Sigma_\beta CX^* + \sigma^2\right)$. Thus, the distribution for $y^* - \hat{y}^*$ changes to $N\left(X^{*T}\lambda, X^{*T}C\Sigma_\beta CX^* + \sigma^2\right)$

Then the generalizability probability $p^2$ is estimated by

$$p_2 = P\left\{ \left| \widehat{y^*} - y^* \right| < \delta \right\} = \Phi\left(\frac{\delta - X^{*T}\lambda}{\zeta_2'}\right) - \Phi\left(\frac{-\delta - X^{*T}\lambda}{\zeta_2'}\right),$$
$$\text{where } \zeta_2'^2 = X^{*T}C\Sigma_\beta CX^* + \sigma^2 \; \left(\zeta_2' > 0\right) \qquad (17)$$

Similarly, for the cases when $\sigma^2$ is unknown, replace $\mu_\beta$ and $\Sigma_\beta$ with $\mu_\beta + \lambda$ and $C\Sigma_\beta C$ respectively and the posterior predictive distribution in Eq. (13) becomes

$$y^* \mid X^*, D \sim t_{2a_\sigma}\left( X^{*T}\left(\mu_\beta + \lambda\right), \frac{b_\sigma + b_\sigma X^{*T}C\Sigma_\beta CX^*}{a_\sigma} \right)$$

The generalizability probability $p^2$ is estimated by

$$p_2 = P\left\{ \left| \widehat{y^*} - y^* \right| < \delta \right\} = T_{2a_\sigma}\left(\frac{\delta - X^{*T}\lambda}{\zeta_3'}\right) - T_{2a_\sigma}\left(\frac{-\delta - X^{*T}\lambda}{\zeta_3'}\right),$$
$$\text{where } \zeta_3'^2 = \frac{b_\sigma + b_\sigma X^{*T}C\Sigma_\beta CX^*}{a_\sigma} \; \left(\zeta_3' > 0\right) \qquad (18)$$

In practice, when $\lambda$ and $C$ (or $\lambda$ in approaches 1 and 2) are unknown, the generalizability of the medical predictive model can be estimated by performing a sensitivity analysis that calculates $p_2$ based on a range of potential $\lambda$ and $C$ (or $\lambda$) values. For example, if we want to assess the model's generalizability when each regression coefficient may deviate by up to 10 percent from its original value, we can take values of $\lambda_i$ from the range $\left(-0.1 * \widehat{\beta_i}, 0.1 * \widehat{\beta_i}\right)$. These selected $\lambda_i$ values are then used to construct a set of $\lambda$ vectors. Using these $\lambda$ vectors, we compute a series of $p_2$ values to evaluate the model's capability in generalizing to data from new populations. The results of the sensitivity analysis provide valuable insights into the model's applicability to new patients, who come from populations distinct from the one upon which the model was developed.

## Conclusion

In this paper, we have summarized the procedure for developing a medical predictive model. Within the outline development process, we identify one critical issue that most of the medical predictive development in practice halts at the internal validation phase. Due to the lack of available data from new populations, there is a notable gap in external validation and assessment of the model's generalizability. To address this challenge, we propose methods for

estimating the reproducibility and generalizability of medical predictive models. We first design three approaches to estimating the model's reproducibility. Subsequently, methods for generalizability estimation are established by adjusting the three proposed methods. When the data of the new population is not accessible, our approaches allow for a sensitivity analysis that provides valuable insight into whether the model can be applied to patients from different populations. Under a validated medical predictive model, we are able to determine an optimal dose based on the risk factors included in the model for achieving the best clinical outcome for disease management. Such a model, one rigorously established and validated, can assist disease management from various aspects, including disease prevention, diagnosis, treatment, and clinical decision-making. Regarding future work, our approaches are proposed in the context of a linear regression model, where a closed-form solution for regression coefficients is utilized. Future opportunities to extend this research could aim to propose generalizability probability estimation methods for more complex modelling scenarios like logistic regression where a closed-form solution is not present.

## Acknowledgement

None.

## Conflict of Interest

None.

## References

1. Sullivan, Lisa M., Joseph M. Massaro and Ralph B. D'Agostino Sr. "Presentation of multivariate data for clinical use: The Framingham study risk score functions." *Stat Med* 23 (2004): 1631-1660.

2. Dormann, Carsten F., Jane Elith, Sven Bacher and Carsten Buchmann, et al. "Collinearity: A review of methods to deal with it and a simulation study evaluating their performance." *Ecography* 36 (2013): 27-46.

3. Mason, Charlotte H. and William D. Perreault Jr. "Collinearity, power and interpretation of multiple regression analysis." *J Mark Res* 28 (1991): 268-280.

4. Belsley, David A., Edwin Kuh and Roy E. Welsch. "Regression diagnostics: Identifying influential data and sources of collinearity." John Wiley & Sons (2005).

5. Chow, S. C., P. J. Lee and Gao, et al. "Statistical method for development of composite index in clinical research." *Am J Biomed Res* 4 (2020): 10.

6. Bazett, H. C. "An analysis of the time-relations of electrocardiograms." *Ann Noninvasive Electrocardiol* 2 (1997): 177-194.

7. Shao, Jun and Shein-Chung Chow. "Reproducibility probability in clinical trials." *Stat Med* 21 (2002): 1727-1742.