Validating prediction models for use in clinical practice: concept, steps and procedures

Mohammad Chowdhury[1] and Tanvir Turin[1]

[1]University of Calgary Cumming School of Medicine

May 5, 2020

Abstract

Prediction models are extensively used in numerous areas including clinical settings where a prediction model helps to detect or screen high-risk subjects for early interventions to prevent an adverse outcome, assist in medical decision-making to help both doctors and patients to make an informed choice regarding the treatment, and assist in healthcare services with planning and quality management. There are two main components of prediction modeling: model development and model validation. Once a model is developed using an appropriate modeling strategy, its utility is assessed through model validation. Model validation provides a true test of a model's predictive ability when the model is applied on an independent data set. A model may show outstanding predictive accuracy in a dataset that was used to develop the model, but its predictive accuracy may decline radically when applied to a different dataset. In the era of precision health where disease prevention through early detection is highly encouraged, accurate prediction of a validated model has become even more important for successful screening. Different clinical practice guidelines also recommend incorporating only those prediction models in clinical practice that has demonstrated good predictive accuracy in multiple validation studies. Our purpose is to introduce the readers with the basic concept of model validation and illustrate the fundamental steps and procedures that are necessary to implement model validation.

KEYWORDS

Prediction model, model validation, internal validation, external validation

INTRODUCTION

Prediction models also known as clinical prediction models are mathematical formula or equation that expresses the relationship between multiple variables and helps predict the future of an outcome using specific values of certain variables. Prediction models are extensively used in numerous areas including clinical settings and their application is large. In clinical application, a prediction model helps to detect or screen high-risk subjects for asymptomatic disease for early interventions, predict a future disease to facilitate patient-doctor communication based on more objective information, assist in medical decision-making to help both doctors and patients to make an informed choice regarding the treatment, and assist in healthcare services with planning and quality management.

While specific details may vary between prediction models, the goal and process of developing prediction models are mostly similar. Conventionally, a single prediction model is built from a dataset of individuals in whom the outcomes are known and then the developed model is applied to predict outcomes for future individuals. There are two main components of prediction modeling: model development and model validation. Once a model is developed using an appropriate modeling strategy, its utility is assessed through model validation. Investigators want to see through validation how the developed model works in a dataset that was not used to develop the model to ensure that the model's performance is adequate for the intended purpose.

Model validation provides a true test of a model's predictive ability when the model is applied on an independent data set. A model may show outstanding predictive accuracy in a dataset that was used to develop the model, but its predictive accuracy may decline radically when applied to a different dataset. In the era of precision health where disease prevention through early detection by monitoring health and disease based on an individual's risk is highly encouraged, accurate prediction in model validation has become even more important for successful screening.

There are numerous clinical prediction models available to serve different purposes, however, only a few found their application in clinical practice. One reason for that is lack of their validation, particularly external validation. External validity establishes generalizability of a prediction model. Generally, accuracy of a prediction model degrades from the sample in which the model was first developed to subsequent application. For a prediction model to be generalizable, the accuracy of the model need to be both reproducible and transportable. A prediction model that cannot predict outcomes accurately in a new sample is useless. Clinicians did not find confidence and trust to use prediction models in their practice that are not well validated. Despite its importance being recognized, external validation of prediction models is not common, which has largely contributed to failure to translate prediction models into clinical practice. Different clinical practice guidelines recommend incorporating only those prediction models in clinical practice that has demonstrated good predictive accuracy in multiple validation studies.

Model validation involves different aspects and our objective is to discuss those aspects in this paper to provide the readers with a basic understanding and importance of the topic. The concept of model validation is statistical. However, we tried to present a nontechnical discussion of the topic in plain language. The information provided in this paper can be helpful for anyone who wishes to be better informed, have more meaningful conversations with data analysts about their project or apply the right model validation technique given that they have advanced training in statistics. We have arranged our discussion as follows. We begin the discussion with defining model validation. Then we have outlined the major steps one needs to follow in model validation. Within the model validation steps, we discussed different ways of model validation together with their strengths and limitations which we named "model validation procedures" and how to assess the performance of a validated model which we named "model performance assessment".

METHODS

The concept of model validation

Model validation is the process of demonstration that the model can reproduce its performance with reasonable accuracy to a different population or setting that was not used to develop the model. The purpose of model validation is to demonstrate that the model is accurate for the intended population (dataset) for whom the model was developed and performs well in other populations (datasets) which were not used to develop the model.

Preferably, a model should be evaluated on samples that were not used to develop the model so that a model's effectiveness can be assessed unbiasedly. However often models are developed in one part of the sample and evaluated in the other part of the sample or the same sample is used through resampling to develop and evaluate the model. Although this kind of model evaluation belongs to model validation formally known as internal validation, this does not guarantee that the model will perform well in a different dataset from a different population. Evaluation of a model's performance in an entirely different population is formally known as external validation and is always advised to establish the generalizability of the model. Within model validation, there are different types each with its advantages and disadvantages. Once a model is validated to a different sample or population, its performance needs to be assessed. There are also different ways to assess the performance of a model. We discuss the types of model validation and how to assess the performance of a validated model within the model validation steps.

Steps of model validation

To validate a prediction model investigators need to follow a few basic steps. We broadly classify the steps

of model validation into two main categories (Figure 1):

1. Apply originally developed model into a different dataset that was not used to develop the model which we will call "model validation procedures"
2. Asses the performance of the model in the new dataset which we will call "model performance assessment"

Model validation procedures

A model can be validated either internally using the same data or data source or externally using new data from a different data source. It is important to separate these two types of validation[1].

Internal Validation

"Internal validation assesses validity for the setting where the development data originated from"[2]. Internal validity is also called "reproducibility" which means the ability to produce accurate predictions among individuals not included in the development of the model but from the same population[3]. In internal validation, generally, the dataset is divided into two categories. One category is called the "training" dataset, which is used to create the model, while the other category is called the "test" or "validation" dataset, which is used to assess the model performance. Internal validation can be performed in different ways. We discuss here some of the major internal validation procedures.

Apparent Validation.

In apparent validation, the model is validated in the same sample where the model was developed; as a result, it provides an optimistic performance of the model. This leads to a biased assessment of the model's performance as the same 100% of data are used both to build the model and to test the model[2].

Split-sample Validation.

Split-sample validation consists of dividing the sample into two parts, with model development in one part of the sample while assessing model performance in the other part of the sample. The splitting is done at random and typical splitting's are 1/2:1/2 or 2/3:1/3. For example, if 1/2:1/2 split sample is used then the model is developed in 50% of the data and the model is evaluated in the other 50% of the data.

Split-sample is an old classical approach of model validation with several limitations[2]. As splitting is done fully at random there could be an imbalance concerning the distribution of predictors and outcome in the sample[2]. Randomly splitting the data does not guarantee that the divided data is representative of the target population. This problem is serious with small samples and a predictor with rare events[2]. One way to overcome this issue is to stratify the sampling by the outcome and relevant predictors[2]. Another issue with the split-sample method is, it provides less stable results as only part of the data is used to model development. Also, small validation data provide an unreliable assessment of model performance that can be even biased because we want to know the model's performance in the full data set, but the assessment was performed only in a part[2].

Due to its several drawbacks, split-sample validation is often treated as an inefficient approach of model validation. The performance of this procedure is reasonable when the sample size is large according to some simulation studies[2]. However, it is suggested to use other efficient model validation procedures to get reliable results.

K-Fold Cross-Validation.

'K-fold cross-validation' (Figure 2) and 'bootstrapping' (Figure 3) are two popular methods that improve upon the split-sample method and produce better results in terms of bias and variability. K-fold cross-validation and bootstrapping are also better in situations where the sample size is small and when external validation is not readily available.

Cross-validation is a resampling procedure primarily used to evaluate the performance of prediction models on unseen data set, particularly, when the data set is small. The purpose is to see how the model performs in general when used to predict data that were not used to develop the model. K-fold cross-validation contains only one parameter "k" that refers to the number of groups (folds) that a given data set is to be split into. If a specific value for "k" is chosen, such as k = 10, then accordingly, the procedure is called 10-fold cross-validation.

In k-fold cross-validation, each observation in the data set is allotted to a specific subsample and remains in that subsample for the entire duration of the procedure. K-fold cross-validation starts with randomly partitioning the original sample into k roughly equal size subsamples. Then, only one subsample out of this k subsamples is kept as the validation data to test the model, and the remaining k-1 subsamples are utilized as training data to derive the model. A total of k times (the folds) this process is replicated, with each of the k subsamples used only once as the validation data. Finally, the results from the k-fold cross-validation run are summarized and a single estimate is produced by averaging (or otherwise combining) the k results from the folds.

Choosing an appropriate value for K is important to avoid misrepresentation of the performance of the model[4]. While choosing the value of k we need to be careful that each subsample (particularly, validation set) of data is large enough to reasonably represent the whole data set. More splits will reduce the size of the validation set and we will not have sufficient sample in the validation set to fairly and confidently evaluate models performance[4]. On the other hand, too few splits will not provide enough trained models to evaluate[4]. In addition, a higher k value is associated with less bias (the difference between the estimated and true values of performance) but more variability (performance of the model may change according to the data set used to fit the model) and computation. On the other hand, a lower k value is associated with more bias but less variability and computation. Though there is no formal rule, usually k is chosen between 5 or 10[4]. Often K = 5 or 10 provides a good compromise for this bias-variance tradeoff[4].

One disadvantage of k-fold cross-validation is its high variance, which makes it less attractive[4]. However, with a large training set with multiple repetitions of the whole k-fold validation-process (e.g., 50 times 10-fold cross-validation) provides true stable results that effectively increase the precision of the model estimates while still maintaining a small bias[2]. K-fold cross-validation has the big advantage that all observations are utilized for both derive and validate the model, with each observation is used only once for validation. As a result, this process has less chance to succumb to a biased division of the data.

Leave-One-Out Cross-Validation (LOOCV).

This is another version of k-fold cross validation where k = n, the number of data points. In this method, each time, only one data-point in the original dataset is held-out for model validation while the remaining data points are used to build the model. As a result, this process runs as many times as the number of data-points in the sample. This method provides negligible bias as the almost entire dataset is used for building the model, which is its advantage. However, this method has the major disadvantage that only one data point is used for validating the model every time, resulting in a high variance in the estimates of the model's performance, particularly when multiple outliers in the dataset. In addition, this method is computationally very intensive, particularly when the dataset is large[4].

Bootstrap Validation.

The bootstrap method is a resampling technique often used to estimate statistics on a population as well as validate a model by sampling a dataset with replacement. The bootstrap method allows us to use a computer to mimic the process of obtaining new data sets so that the variability of the estimates can be assessed without creating additional samples. Instead of repeatedly obtaining independent data set from the population, which is often not realistic, in bootstrapping, distinct data sets are obtained by repeatedly doing sampling from the original data set with replacement. The idea behind bootstrapping is the original observed data will take the place of the population of interest, and each bootstrap sample will represent a sample from that population.

Bootstrap samples are of the same size as the original sample and drawn randomly with replacement from the original sample. In a with replacement sampling, after a data point (observation) is selected for the subsample, it is still available for further selection. As a result, some observations represented multiple times in the bootstrap sample while others may not be selected at all. Because of such overlaps with original data, on average almost two-thirds of the original data points appear in each bootstrap sample[4]. The samples that are not included in a bootstrap sample are called "out-of-bag" samples. When performing the bootstrap, two things must be specified: the size of the sample and the number of repetitions of the procedure to perform. A common practice is to use a sample size that is equivalent to the original data set and a large number of repetitions (50-200) to get a stable performance[2, 4].

In the bootstrap method, a prediction model is developed in each bootstrap sample and measures of predictive ability such as C-statistic are estimated in each bootstrap sample. Then these models from bootstrap data are applied to the original dataset to evaluate the model and estimate the predictive measure (C-statistic) of these bootstrap models in the original data. The difference in performance in the predictive measure indicates optimism, which is estimated by averaging out all the differences in predictive measures. Finally, this estimate of optimism is subtracted from the performance of the original prediction model developed in the original data to get an optimism-adjusted measure of the predictive ability of the model[2].

Bootstrap samples have significant overlap with the original data (roughly two-third) which causes the method to underestimate the true estimate. This is considered a disadvantage of this method. However, this issue can be solved by performing prediction on only those observations that were not selected by the bootstrap and estimating model performance. Bootstrapping is more complex to analyze and interpret due to the methods used and the amount of computation required. However, this method provides stable results (less variance) than other methods with a large number of repetitions.

It is very obvious that each of the internal model validation techniques has advantages and disadvantages and no one method is uniformly better than another[4]. Researchers have a different opinion on choosing the appropriate method for internal model validation. Several factors such as sample size, finding the best indicators of a model's performance and choosing between models were asked to consider before making the choice[4].

The above-mentioned procedures for model validation pertain to internal validation, which does not examine the generalizability of the model. To ensure generalizability, it is necessary to use new data not used in the development process, collected from an appropriate (representative) patient population but using a different set of data.

External Validation

The reliability and acceptability of a prediction model largely depend on how well it performs in a validation cohort, outside of the derivation cohort where the model was developed. Internal validation of prediction models is often not sufficient for generalizability, and external validation is necessary before implementing prediction models in clinical practice. External validation of models is often considered essential to support the general applicability of a prediction model as it addresses transportability[1, 2]. Transportability requires the model to perform accurately in predicting data drawn from a different but plausibly related population or in data collected by using a little different method than those used in the development sample[3]. External validation requires data collected from a similar group of patients in a different setting and aims to address the accuracy and performance of a prediction model in a different patient population. These data (sample) are fully independent of the development data and originate from different but similar patients. The generalizability of a model becomes stronger when the model is externally validated multiple times and in a more diverse setting[2]. This is the reason perhaps why the Framingham Risk Score (FRS) for cardiovascular disease (CVD) is so widely used in the clinical setting as the model was externally validated many times with many different settings.

Most studies evaluating prediction models focus on the issue of internal validity as opposed to the important issue of external validity. Internal validation does not guarantee generalizability, and thus external validation

is necessary before implementing prediction models into clinical practice.

External validation can be assessed in different ways:

1. Temporal validation (validation in more recent individuals)
2. Geographical validation (validation in other places)
3. Fully independent validation/Strong external validation (by other investigators at other sites)

Temporal Validation.

In temporal validation, the model is typically validated in more recent individuals. The purpose of such validation is to make sure that the model maintains its accuracy when it is tested in cohorts in different time periods. A model developed way back (say 20 years ago) may not work in current patients (e.g., change in risk factor distribution and availability of large dataset on many risk factors). Temporal validation can be easily achieved just splitting the data into two parts. Develop the model in one part that contains early treated patients and validate the model to assess its performance in another part that contains patients that are more recent[2]. Temporal validation can also be achieved through the prospective application of the developed model in specifically collected cohort[2]. For example, a model can be developed in a group of patients between 2005 and 2010 and the same model can be validated in a different group of patients from the same cohort between 2012 and 2015.

Geographic Validation.

In geographic validation, the model is validated in a different location that was not used to develop the model. The purpose of geographic validation is to confirm that the model remains accurate when it is tested in data from other locations. Although it may be questioned whether a model developed in another location will work in a new location that is completely different. Geographical validation can be achieved by applying and assessing the performance of a developed model to a different site within the same region or a different cohort from a different region. For example, a model developed in the USA for predicting hypertension can be validated in a similar Canadian cohort.

Fully Independent Validation.

In fully independent validation, model validation is performed in data collected by independent investigators, usually at a different location. Generally, the validation sample is drawn from a different time. It is important to establish that the model is equally accurate when applied by independent investigators, as they are unlikely to study identically selected patients and data collecting tools[3]. In addition, the definition of predictors and outcomes and study participants selection may be slightly different compared with the development setting in fully independent validation[2].

Full independent validation often shows poor results (more unfavorable) than temporal or geographical validation. There could be several reasons for that. Some of those reasons are related to the original model's development issues such as inadequate model development strategy, small sample size, and suboptimal statistical analysis. It also happens frequently that all the variables used to build the original model may not be available at validation data, which eventually affects the model's performance in validation data. In addition, a true difference between development and validation samples may cause poor validation results. Fully independent external validation of a model is often more difficult than anticipated. However, if a model can demonstrate adequate performance in a fully independent validation in a different setting, then the results of this model's performance are more authentic, acceptable and generalizable.

Model performance assessment

Once a prediction model is developed then it needs to be validated to see or quantify how good the predictions from the models are, often referred to as model performance. There are different methods and metrics to assess the performance of a prediction model. These methods and metrics depend on the type of modeling technique used in model developing which again largely depends on the outcome of interest. We will restrict our discussion of model performance assessment for binary or survival outcomes, common in health research.

For binary and survival outcomes, the most commonly used measures include the Brier score to indicate overall model performance, the concordance statistic (also known as the C-statistic) for discriminative ability, and goodness-of-fit statistics for calibration.

Brier Score.

The model's overall performance is quantified by considering the distance between the actual outcome and the predicted outcome with better models have smaller distances[5]. The Brier score is used to calculate the model's overall performance and is measured by calculating the squared differences between actual binary outcomes and predictions calculated by the model. The range of values that the Brier score of a model can take lies between 0 and 0.25 with 0 indicating a perfect model and 0.25 indicating a non-informative model with only a 50% incidence of the outcome[2, 5]. Brier score for survival outcome is not possible to calculate directly because of censoring. However, it is possible to calculate it indirectly defining a weight function that considers the conditional probability of being uncensored during the time. One disadvantage of the Brier score is that its interpretation depends on the incidence of the outcome with lower (higher) incidence corresponds to lower (higher) Brier score[2].

Discrimination.

The discrimination is defined as the model's ability to distinguish between participants who do or do not experience the event of interest (e.g., disease outcome such as hypertension). A good prediction model can accurately discriminate between those with and without the outcome[5]. C-statistic, which is equal to the area under the receiver operating characteristic (ROC) curve for binary outcomes, is commonly employed to assess discrimination. ROC curve plots the sensitivity against (1 – specificity) for consecutive cutoffs for the probability of an outcome. The value of a C-statistic (area under ROC curve) points out to the probability that a randomly selected subject who experienced the outcome will have a higher predicted probability of having the outcome occur compared to a randomly selected subject who did not experience the event. The C-statistic can range from 0.5 to 1, with higher values indicating better predictive models. A C-statistic of 0.5 indicates the model's performance in predicting an outcome is no better than the random chance while a C-statistics of 1 indicates the model perfectly distinguishes those who will experience a certain outcome and those who will not. Generally, the C-statistic of a prediction model ranges from 0.6 to 0.85. A model with a C-statistic ranging from 0.70 to 0.80 is considered adequate, while a range of 0.80 to 0.90 is considered excellent[6].

For survival data, an extension of C-statistic called Harrell's C-statistic is suggested which indicates the proportion of all pairs of subjects who can be ordered such that the subject who survived longer will have the higher predicted survival time than the subjects who survived shorter, assuming that these subject pairs are selected at random. Although C-statistic is insensitive to outcome incidence, one disadvantage of C-statistic is, its interpretation is based on an artificial situation assumption that we have a pair of patients, one with and one without the outcome.

Calibration.

The agreement between observed outcomes and predictions made by the model is referred to as calibration[1]. Model calibration measures the validity of the predictions and determines whether the predictions based on the risk prediction model align with what is observed within the study cohort. For example, if we predict a 20% risk that a person will develop hypertension, the observed frequency of hypertension should be 20 out of 100 people with such a prediction. Calibration plot is a method that visually inspects calibration and presents plot for predicted against observed probabilities. It also uses the Hosmer-Lemeshow test to assess calibration. In a calibration plot, predictions are plotted on the x-axis and the observed outcome on the y-axis. In the y-axis, the plot contains only 0 and 1 values for binary outcomes. Different smoothing techniques (e.g., the loess algorithm) can be employed to estimate the observed probabilities of the outcome for the predicted probabilities. Perfect predictions should be on the 45° line suggesting that predicted risks are correct. An alternative assessment of calibration is to categorize predicted risk into groups (e.g., deciles) and assess whether the event rate corresponds to the average predicted risk in each risk group. The Hosmer-

Lemeshow goodness-of-fit-test makes the plot of a graphical illustration to assess whether the observed event rates match expected event rates in subgroups of the model population.

For survival data, the calibration is usually assessed at fixed time points[2]. Within each time point, survival rates are calculated by the Kaplan-Meier method for a group of patients. Then this observed survival is compared with the mean predicted survival from the prediction model[2].

Besides the above-mentioned major measures of model assessment, there are other measures occasionally used to assess a model. Although calibration and discrimination are considered the most important aspects to assess a model, they did not provide any assessment regarding the clinical usefulness of a model. Clinical usefulness assessment helps to understand the ability of a model to make better decisions compared to a situation when the model was not used. The measures associated with clinical usefulness are generally related to a cutoff, a decision threshold of the model, which classify peoples into low and high-risk groups balancing the likelihood of benefit and likelihood of harm. Net benefit (NB) is one such measure that can be used to assess the clinical usefulness of a model.

CONCLUSION

Validation of a prediction model is extremely important as it provides model applicability in different populations. The model's internal validation is quite common. However, to make a model generalizable and applicable in clinical practice, the model must need to be externally validated. Models that are externally validated multiple times with sufficient good performance are reliable and often recommended in clinical guidelines for implementation. Although, there are many prediction models only a few with external validations. We suggest investigators focus not only on just prediction model development but also on external validation of their developed model.

REFERENCES

1. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. European heart journal. 2014 Jun 4; 35(29):1925-31.
2. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer Science & Business Media; 2008 Dec 16.
3. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of internal medicine. 1999 Mar 16; 130(6):515-24.
4. Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013 May 17.
5. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.). 2010 Jan; 21(1):128.
6. Lee YH, Bang H, Kim DJ. How to establish clinical prediction models. Endocrinology and Metabolism. 2016 Mar 1; 31(1):38-44.
7. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, D'Agostino RB, Kannel WB, Vasan RS. A risk score for predicting the near-term incidence of hypertension: the Framingham Heart Study. Annals of internal medicine. 2008 Jan 15; 148(2):102-10.
8. Kivimaki M, Batty GD, Singh-Manoux A, Ferrie JE, Tabak AG, Jokela M, Marmot MG, Smith GD, Shipley MJ. Validating the Framingham hypertension risk score: results from the Whitehall II Study. Hypertension. 2009 Sep 1; 54(3):496-501.

DECLARATIONS

Acknowledgments

None.

Ethical Approval

Not applicable.

Funding

This research received no grant support from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of Interest

None.

FIGURE LEGENDS

Figure 1. Model Validation Steps

Figure 2. 5-Fold Cross-Validation

Figure 3. A graphical illustration of the bootstrap process on a hypothetical small sample containing n = 3 observations on two variables X and Y. Three bootstrap samples containing n = 3 observations drawn with replacement from the original data set. Finally, each bootstrap sample is used to obtain the prediction

Hosted file

`Figure 1.docx` available at https://authorea.com/users/294157/articles/422319-validating-prediction-models-for-use-in-clinical-practice-concept-steps-and-procedures

Hosted file

`Figure 2.docx` available at https://authorea.com/users/294157/articles/422319-validating-prediction-models-for-use-in-clinical-practice-concept-steps-and-procedures

Hosted file

`Figure 3.docx` available at https://authorea.com/users/294157/articles/422319-validating-prediction-models-for-use-in-clinical-practice-concept-steps-and-procedures