A Permutation Method to Assess Heterogeneity in External Validation for Risk Prediction Models

Ling-Yi Wang, Wen-Chung Lee

1 Research Center for Genes, Environment and Human Health and Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan
2 Department of Medical Research, Tzu Chi General Hospital, Hualien, Taiwan

*wenchung@ntu.edu.tw

Abstract

The value of a developed prediction model depends on its performance outside the development sample. The key is therefore to externally validate the model on a different but related independent data. In this study, we propose a permutation method to assess heterogeneity in external validation for risk prediction models. The permutation p value measures the extent of homology between development and validation datasets. If \( p < 0.05 \), the model may not be directly transported to the external validation population without further revision or updating. Monte-Carlo simulations are conducted to evaluate the statistical properties of the proposed method, and two microarray breast cancer datasets are analyzed for demonstration. The permutation method is easy to implement and is recommended for routine use in external validation for risk prediction models.

Introduction

A risk prediction model estimates the probability that a certain outcome is present (diagnosis) or will occur (prognosis) in a new subject [1–3]. Once a prediction model has been constructed in a development population, the next step is to evaluate its prediction performance. This can be done by internal validation (e.g. bootstrapping [4] or cross-validation [5]), that is, constructing the model in one part (training dataset) and then evaluating its performance on another non-overlapping part (testing dataset) of the model development dataset.

Although internal validation can assess the reproducibility of a model, the value of a developed (diagnostic and prognostic) prediction model depends on its performance outside the development sample (transportability). The key is therefore to externally validate the model on a different but related independent data. Debray et al. [6] recently proposed a three-step framework to enhance the interpretation of external validation studies of prediction models. This should help researchers judge whether a prediction model is clinically practicable or merely statistically reproducible.

Following Debray et al.‘s framework [6], we propose a permutation method to assess heterogeneity in external validation for risk prediction models. Monte Carlo simulation is
implemented for the evaluation of our method. We demonstrate the application of the proposed method using two microarray breast cancer datasets.

**Methods**

Suppose that a model development dataset (Data D) which consists of cases (subjects with the outcome) and controls (subjects without the outcome) is used to develop a prediction model (Model M). For external validation, Model M is tested on another independent validation dataset (Data V) to obtain a performance estimate: the externally validated AUC (area under the receiver operating characteristic curve), denoted as AUC\text{ext}.

To assess heterogeneity between Data D and V, we permute the subjects between these two datasets, separately for cases and controls. At the jth permutation, let D\text{j} and V\text{j} denote the permuted development and validation datasets, respectively. Data D\text{j} is used to develop a prediction model: M\text{j}. Data V\text{j} is then used to evaluate the performance of this Model M\text{j} to give a validated AUC, denoted as AUC\text{j}. The permutation process is repeated for a total of k times.

The permutation p value is calculated as the proportion of the \{AUC\text{1}, AUC\text{2},..., AUC\text{k}\} that are smaller than the previously calculated AUC\text{ext}.

The permutation p value measures the extent of homology between Data D and V. If the permutation p value is less than 0.05, we conclude that there is significant heterogeneity (at a significance level of \(\alpha = 5\%\)) between the two datasets. If the value is larger, we may transport the prediction model developed in Data D to Data V.

**Simulation Studies**

**Simulation Setup**

Suppose that there are three model development datasets (Data D\text{A}, Data D\text{B}, and Data D\text{C}), each with a different data structure. The variables in Data D\text{A} and D\text{B} are generated using the multivariate normal distributions for cases and controls, respectively (the means: detailed in S1 Exhibit; the variances: 1 for all variables; the correlation coefficients: 0 between any two variables in D\text{A} and 0.2 between any two variables in D\text{B}). The variables in Data D\text{C} are generated using a two component mixture of multivariate normal distributions for both cases and controls. (The variances for all variables are set to 1, and the correlation coefficients between any two variables, 0, for each component. Each component contributes 50\% of the whole data. The means of these two multivariate normal distributions are detailed in S1 Exhibit.)

In each dataset, we use support vector machines (SVM) to construct a prediction model. SVM is an efficient learning algorithm for high-dimensional data in classification, regression and pattern recognition. The basis of SVM is to implicitly map data to a higher dimensional space via a kernel function to identify an optimal hyperplane that maximizes the margin between the two groups [7]. In this study, we use the e1071-package of R with a default radial basis function kernel to obtain the prediction scores [8].

We consider three validation datasets (Data V\text{A}, Data V\text{B}, and Data V\text{C}). The data generating process and parameter setting for V\text{A}, V\text{B}, and V\text{C} are the same as the aforementioned D\text{A}, D\text{B}, and D\text{C}, respectively. For homogeneity scenarios, we let the prediction models developed in D\text{A}, D\text{B}, and D\text{C} be tested on V\text{A}, V\text{B}, and V\text{C}, respectively. For heterogeneity scenarios, we let the prediction model developed in one type of data be tested on a different type of data.

In the simulation, we consider prediction models with 3 and 10 predictors, respectively. We also consider three different sample sizes (small, medium, and large): \(N = 30\) (cases)+30 (controls), 50+50, 100+100, respectively, for the model development datasets. We assume equal sample sizes for the development and the validation datasets. The number of permutations is set at \(k = 500\), and the significance level is set at \(\alpha = 0.05\). We conduct a total of 5000
simulations for each scenario. In the simulation, we additionally create a very large validation dataset (1000 cases and 1000 controls) for each data type. These are used to determine a true AUC value for a prediction model as applied to the same model development population. We refer to these as the reproducibility AUCs.

Simulation Results

Table 1 presents the results of homogeneity scenarios. We see that the externally validated AUCs and the corresponding reproducibility AUCs are approximately equal. We also see that the proposed permutation test has permutation p values that are around 0.5 and type I error rates close to the nominal α level of 0.05.

Table 2 presents the results of heterogeneity scenarios. We see that now the externally validated AUCs are smaller than the corresponding reproducibility AUCs. We also see that the permutation p value decreases when sample size increases and that the power (for detecting heterogeneity) of the permutation test increases when sample size increases.

Real Data Application

Two independent microarray breast cancer datasets, W (Wang et al. [9]) and S (Sotiriou et al. [10]), were used to demonstrate the proposed method. The gene expression data and patient profile are available at the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo) with accession code GSE2034 (Data W) and GSE2990 (Data S). Both datasets were generated from the same Affymetrix-HG-U133A microarray platform. In the study of Wang et al. [9], Data W (consisting of 107 breast cancer patients with distant relapse and 197 without distant relapse) was divided into training (115 patients) and testing (171 patients) by concentration of the estrogen receptor, and a 76-gene signature was identified with an internally validated AUC of 0.694. Here we use Data S (consisting of 120 breast cancer patients without
relapse and 67 with relapse; 2 patients with unknown relapse status are omitted from our analysis) to validate the prediction performance of the 76-gene signature developed in Data W, and the externally validated AUC is 0.534.

Next, we conduct the permutation test. We performed a total of 100,000 permutations and found that all the permuted AUCs are larger than 0.534 (permutation p value < $10^{-5}$). Hence we conclude a significant heterogeneity between Data W and S. For this example, we know that the 76-gene signature developed in Data W cannot be directly transported to Data S, unless further model updating or revision was done.

### Discussion

Debray et al. [6] suggested using the following three steps to interpret the results of external validation of a prediction model: 1) to assess the extent of relatedness between development and validation datasets, 2) to assess the performance of the model on the external validation dataset, and 3) to interpret the model’s predictive accuracy given the results from 1) and 2).
Our permutation method integrates the above steps 1 and 2. The permutation \( p \) value measures the extent of homology between development and validation datasets (step 1), while at the same time the homology/heterogeneity judgment is based directly on model performance comparison between development and validation datasets (step 2). This should greatly facilitate the interpretation of external validation studies of prediction models.

If the purpose of the model is purely to make predictions for new individuals in the same population or future patients in the same clinical setting (the temporal validation [11]), then we need a model that has good reproducibility. To estimate the reproducibility AUC, one can use an internal validation method, or better still, to sample more subjects in the same population for an 'external' validation; external here to be taken relative to the model development data at hand but not to the study population at large. Our permutation method can be applied in this situation to help check whether there is significant temporal variation in case-mix in the population that will curtail the utility of the prediction model.

But more often, the purpose of the model is for making predictions for subjects outside the model development population. We encourage the model developers to pursue as many external datasets as possible to validate the model, if transportability of the model is intended. Here the permutation \( p \) value from our proposed permutation test is a measure of homology between a chosen external dataset and the model development dataset. If the permutation \( p \) value of an external dataset from a certain population is less than 0.05, there is significant heterogeneity between the two datasets and the model may not be directly transported to that external population without further revision or updating [12,13].

In summary, the value of a developed prediction model depends on its performance outside the development sample. The permutation method proposed in this paper assesses heterogeneity in external validation for risk prediction models by integrating the step 1 and step 2 of Debray et al.'s three-step framework [6]. This should greatly facilitate the interpretation of external validation studies of prediction models. The method is easy to implement and is recommended for routine use in external validation for risk prediction models.

Supporting Information

S1 Exhibit. The means of the three model development datasets Data \( D_A \), Data \( D_B \), and Data \( D_C \) in simulation studies.

(DOCX)

Acknowledgments

The authors wish to thank Dr. Yung-Hsiang Huang for technical supports.

Author Contributions

Conceived and designed the experiments: WCL. Performed the experiments: LYW. Analyzed the data: LYW. Contributed reagents/materials/analysis tools: WCL. Wrote the paper: LYW WCL.

References

1. Eagle KA, Lim MJ, Dabbous OH, Pieper KS, Goldberg RJ, et al. (2004) A validated prediction model for all forms of acute coronary syndrome: estimating the risk of 6-month postdischarge death in an international registry. JAMA 291: 2727–2733. doi: 10.1001/jama.291.22.2727 PMID: 15187054

2. Barlow WE, White E, Balard-Barbash R, Vacek PM, Titus-Ernstoff L, et al. (2006) Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst 98: 1204–1214. doi: 10.1093/jnci/djj331 PMID: 16854473
3. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, et al. (2012) Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 98: 683–690. doi: 10.1136/heartjnl-2011-301246 PMID: 22397945

4. Efron B, Tibshirani RJ (1994) An introduction to the bootstrap: CRC press.

5. Refaelzadeh P, Tang L, Liu H (2009) Cross-validation. Encyclopedia of database systems: Springer. pp. 532–538.

6. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, et al. (2014) A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol (in press).

7. Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10: 988–999. doi: 10.1109/72.788640 PMID: 18252602

8. Karatzoglou A, Meyer D, Hornik K (2005) Support vector machines in R.

9. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671–679. doi: 10.1016/S0140-6736(05)17947-1 PMID: 15721472

10. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98: 262–272. doi: 10.1093/jnci/djl052 PMID: 16478745

11. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, et al. (2012) Risk prediction models: II. External validation, model updating, and impact assessment. Heart 98: 691–698. doi: 10.1136/heartjnl-2011-301247 PMID: 22397946

12. Toll DB, Janssen KJ, Vergouwe Y, Moons KG (2008) Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol 61: 1085–1094. doi: 10.1016/j.jclinepi.2008.04.008 PMID: 19208371

13. Vergouwe Y, Moons KG, Steyerberg EW (2010) External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 172: 971–980. doi: 10.1093/aje/kwq223 PMID: 20807737