On estimating predictive performance measures of risk prediction models with external validation data

Hajime Uno[*] Eisuke Inoue[†]

Abstract

Risk prediction models play an important role in selecting prevention and treatment strategies for various diseases. While it is common to observe poorer performance in a validation set compared to a development set, this difference is generally attributed to optimistic bias in measuring performance in the development set. However, this difference might be rather due to differences in the distribution of the predictors, which can strongly affect predictive performance. Conventional validation analysis does not take account of it. It could erroneously give a low rating to a useful risk prediction model even when the model is working for each subject in the validation set in the exactly same way as for those in the development set. Because results of validation studies ultimately determine which prediction models are adopted for research and clinical use, it is critical that validation methods be grounded in rigorous cross-study comparisons. We will present new inference procedures for estimating predictive performance measures in validation studies, systematically adjusting for differences in the distribution of predictors across studies.

Key Words: cross-study validation, risk prediction, prediction model, predictive performance

1. Introduction

Recent years have seen significant progress in the development and refinement of prediction models, some of which are widely used in clinical settings. Statistical methodology for risk prediction has also been extensively investigated; however, there still exist issues in validation methodology that require new methods. We have identified a potentially critical pitfall in the standard analysis approach to validating a prediction model. In current practice, performance indices are estimated by simply applying a model to a validation set; however, if the distribution of risk factors (or predictors) in the validation set is different from that on the relevant clinical population, such an approach may not provide a useful estimate of the predictive performance. For example, we sometimes observe unexpectedly poor performance in a validation set because of 1) overfitting in the development set (i.e., optimistic bias), 2) difference in distribution of the predictors between the development and validation sets, and 3) sampling variability. Unfortunately, the current analytical practice does not take the second issue into account and thus sometimes results in the inappropriate rejection of a sound and potentially useful prediction model (Steyerberg et al., 2004; Janssen et al., 2008).

2. Method

2.1 A general framework

Our proposed method is formulated in a general way, utilizing an intuitive weighting technique. Let Y be a response variable and Z be a vector variable of predictors for Y. Let $F_{\mathcal{O}}$ denote the distribution of $\mathcal{O} = (Y, Z)$ on a population. For a given prediction model,

[*] Dana-Farber Cancer Institute, Department of Biostatistics and Computational Biology, 450 Brookline Avenue, Boston, MA 02215, USA

[†] St. Marianna University School of Medicine, Division of Medical Informatics, Miyamae-ku, Kawasaki, Kanagawa, 216-8511, Japan

$\psi(Z)$, a general form of predictive performance measures is given by $D = \int Q_\psi(\mathcal{O})dF_{\mathcal{O}}$, where $Q_\psi(\mathcal{O})$ is a function to determine how to measure the performance. For example, one might use $Q_\psi(\mathcal{O}) = |Y - \psi(Z)|$, which will give the L_1-error of the prediction model $\psi(Z)$ for Y. Let $\hat{F}^{(d)}_{\mathcal{O}}$ and $\hat{F}^{(v)}_{\mathcal{O}}$ be empirical distributions $\mathcal{O}$ of a development sample and a validation sample, respectively. Estimators for D on $F_{\mathcal{O}}$, based on the development and validation samples, are then given by $\hat{D}_d = \int Q_\psi(\mathcal{O})d\hat{F}^{(d)}_{\mathcal{O}}$ and $\hat{D}_v = \int Q_\psi(\mathcal{O})d\hat{F}^{(v)}_{\mathcal{O}}$, respectively.

Now, let F^*_Z be a marginal distribution function of Z on a given target population. To estimate a performance measure on the target population, we consider the following "adjusted" estimator with the validation sample

$$\tilde{D}_v(F^*_Z) = \int Q_\psi(\mathcal{O}) \left\{ \frac{dF^*_Z(\mathcal{O})}{d\hat{F}^{(v)}_Z(\mathcal{O})} \right\} d\hat{F}^{(v)}_{\mathcal{O}}, \tag{1}$$

where $\hat{F}^{(v)}_Z(\cdot)$ is the marginal distribution function of Z of the validation sample. For example, if we consider F^*_Z to be the empirical (marginal) distribution of Z of the development sample $\hat{F}^{(d)}_Z$, we will then obtain an adjusted validation estimator for D

$$\tilde{D}_v(\hat{F}^{(d)}_Z) = \int Q_\psi(\mathcal{O}) \left\{ \frac{d\hat{F}^{(d)}_Z(\mathcal{O})}{d\hat{F}^{(v)}_Z(\mathcal{O})} \right\} d\hat{F}^{(v)}_{\mathcal{O}}, \tag{2}$$

adjusted to the observed distribution of Z in the development set. Note that, if Z includes important predictors for Y, and the two sets are similar in the sense that $\hat{F}^{(d)}_{Y|Z} \approx \hat{F}^{(v)}_{Y|Z}$, then $\frac{\hat{F}^{(d)}_Z(\cdot)}{\hat{F}^{(v)}_Z(\cdot)} \approx \frac{\hat{F}^{(d)}_{\mathcal{O}}(\cdot)}{\hat{F}^{(v)}_{\mathcal{O}}(\cdot)}$ and thus $\tilde{D}_v(\hat{F}^{(d)}_Z) \approx \int Q_\psi(\mathcal{O}) \left\{ \frac{d\hat{F}^{(d)}_{\mathcal{O}}(\mathcal{O})}{d\hat{F}^{(v)}_{\mathcal{O}}(\mathcal{O})} \right\} d\hat{F}^{(v)}_{\mathcal{O}} = \hat{D}_d$. Therefore, when the observed difference between $\tilde{D}_v(\hat{F}^{(d)}_Z)$ and $\hat{D}_d$ is large, we can determine that it is not due to the difference between $\hat{F}^{(d)}_Z(\cdot)$ and $\hat{F}^{(v)}_Z(\cdot)$, but $\hat{F}^{(d)}_{Y|Z}(\cdot)$ would be different from $\hat{F}^{(v)}_{Y|Z}(\cdot)$. Since the difference $\hat{D}_v - \hat{D}_d$ is affected by the difference between $\hat{F}^{(d)}_Z(\cdot)$ and $\hat{F}^{(v)}_Z(\cdot)$, reporting the adjusted result $\tilde{D}_v(\hat{F}^{(d)}_Z)$, in addition to $(\hat{D}_d, \hat{D}_v)$ would be useful for understanding the difference between $\hat{D}_v$ and $\hat{D}_d$ better. Furthermore, when both the development set and the validation set are considered to be samples from the same target population, one may expect that the empirical distribution of Z derived from a pooled sample $\hat{F}^{(d+v)}_Z$ is more representing F_Z on the target. In that case, $\tilde{D}_v(\hat{F}^{(d+v)}_Z)$ would be also a useful estimator for D.

2.2 Density ratio fitting approach

The key component of the general framework is estimation of the weight function

$$\frac{dF^*_Z(\cdot)}{d\hat{F}^{(v)}_Z(\cdot)}$$

in the equation (1). At first glance, this may look a difficult problem, because it consists of two density functions. It may seem that density estimation is involved. However, estimation of "ratio" of two densities can be performed without density estimation. Recently, the methodology for density ratio estimation was extensively studied in the field of machine learning Sugiyama et al. (2012).

It is interesting to note that the density ratio estimation is implicitly used in causal inference as well. In causal inference, the inverse-probability-weighting method is often used to adjust for confounding factors, where a propensity score is derived to calculate the probability of being exposed or unexposed, where the propensity score is playing a role of the density ratio estimate of the confounding factors. Often in practice, a logistic regression model is used to derive a propensity score. It is shown that the resulting logistic regression model can provide a consistent estimator for the density ratio when the model is correctly specified. There are various methods to perform density ratio estimation, other than the simple linear logistic regression modeling approach, such as kernel mean matching (Huang and Harrington, 2005), KullbackLeibler importance estimation procedure (Sugiyama et al., 2008), Least-squares importance fitting (Kanamori et al., 2009), Unconstraint least-squares importance fitting (Kanamori et al., 2012), generalized additive model (GAM)(Hastie and Tibshirani, 1995), and so on.

2.3 Estimating performance measures with the validation set for comparing results with those from the development set

Now we consider inference of the model performance D using the adjusted performance estimator with the validation set, assuming that development set well represents the target population in the sense that $\hat{F}_Z^{(d)} \approx F_Z^*$. Specifically, we derive the $\tilde{D}_v(\hat{F}_Z^{(d)})$ shown in the equation (2), using the both development and validation sets.

We apply a density ratio estimation to obtain the weight. We will use a GAM to approximate the weight $\left\{ \frac{d\hat{F}_Z^d}{d\hat{F}_Z^{(v)}} \right\}$. By applying the weight to each observations in the validation set, we will calculate $\tilde{D}_v(\hat{F}_Z^{(d)})$.

To obtain a standard error, we will apply a bootstrap method. We generate a bootstrap sample from the development and the validation set for each and calculate

$$\tilde{D}_v^{\dagger}(\hat{F}_Z^{\dagger(d)}) = \int Q_{\psi}(\mathcal{O}) \left\{ \frac{d\hat{F}^{\dagger(d)Z(\mathcal{O})}}{d\hat{F}_Z^{\dagger(v)}(\mathcal{O})} \right\} d\hat{F}_{\mathcal{O}}^{\dagger(v)},$$

where $\hat{F}_Z^{\dagger(d)}$ and $\hat{F}_Z^{\dagger(v)}$ are the distribution of the bootstrap samples from the development set and the validation set, respectively. We repeat this process M times. A confidence interval is constructed from the M of $\tilde{D}_v^{\dagger}(\hat{F}_Z^{\dagger(d)})$ by the percentile method.

3. Application to the SEER-Medicare MDS Risk Score

The Surveillance Epidemiology and End Results (SEER) program has been collecting data on MDS since 2001. These data, when linked to Medicare claims (SEER-Medicare), are an outstanding resource for comparative effectiveness research (CER) with regard to MDS. Unfortunately, existing clinical risk stratification systems (i.e., IPSS, WPSS, and so on) are not applicable for the SEER-Medicare data because many of the elements they contain do not appear in Medicare claims. Therefore, in order to make the SEER-Medicare more useful for CER, we created a prognostic risk score (i.e., SMMRS), using the 2001-2007 SEER-Medicare MDS dataset (n=9820) (Uno et al., 2014). The risk factors included in the SMMRS were cytopenias, MDS category, age, comorbidity, acute hospitalization, and red blood cell (RBC) and/or platelet (PLT) transfusion dependency. The SMMRS is given as a linear combination of these factors; the contribution of each factor was determined by fitting the SEER-Medicare data with a Cox regression model (Table 1). Three-year predicted risk score for mortality is then determined by

$$1 - \exp\{-0.025 \times \exp(\text{SMMRS})\}$$

Table 1: Components of the SMMRS with contributions to the final model (Uno et al., 2014)

Cytopenias	
No Anemia or Anemia only	0
Anemia + Neutropenia	+0.11
Anemia + Thrombocytopenia	+0.31
Anemia + Neutropenia + Thrombocytopenia	+0.39
MDS category	
RA or 5q-	0
RAEB, RAEB-T, RCMD, tMDS	+0.61
RARS	-0.08
MDS NOS	+0.23
Age at diagnosis (years)	+0.04 × age
Charlson comorbidity index*	+0.08 × index
Presence of acute hospitalization	+0.46
RBC/PLT transfusion	
No dependency	0
Transfusion dependent: either RBC or PLT	+0.57
Transfusion dependent: both RBC and PLT	+0.87

$*$ An index score greater than 6 is replaced by 6

To evaluate the performance of the SMMRS, we used a granular MDS patient database at Dana-Faber Cancer Institute (DFCI/CRIS data) as a validation set, which was independent of the development set (the SEER-Medicare MDS data; Dana-Farber is not a SEER institution).

As shown in Figure 1, the distribution of the risk score with the development dataset (solid red line) is rather different from the one with the validation dataset (blue solid line). Specifically, it is suggested that relatively more patients with higher risk are involved in the development dataset, compared to the validation dataset. The dotted blue line in Figure 1 shows the adjusted distribution of the risk score by assigning a density ratio weight to each subject in the validation dataset. The adjusted distribution with the validation set appears to be similar to the distribution of the development set.

We used C-statistics (Uno et al., 2011) as the performance index of the SMMRS. The C-statistic with the development set was 0.700, and the unadjusted C with the validation set was 0.647 (0.95CI: 0.608 to 0.688). (Table 2) The difference in C-statistic between the development and the validation sets was 0.053 was rather different between the two sets. However, it is not clear from these results if the observed difference of 0.053 is due to the difference in case-mix or something else. Interestingly, the adjusted C-statistic 0.662 (0.95CI: 0.613 to 0.711) is much closer to the C-statistic with the development set. Reporting both unadjusted and adjusted results would enhance understanding of how the model performs in different datasets.

4. Remarks

Model performance estimates depend on the distribution of the predictors. Without taking the difference in the distribution between development and validation dataset into account,

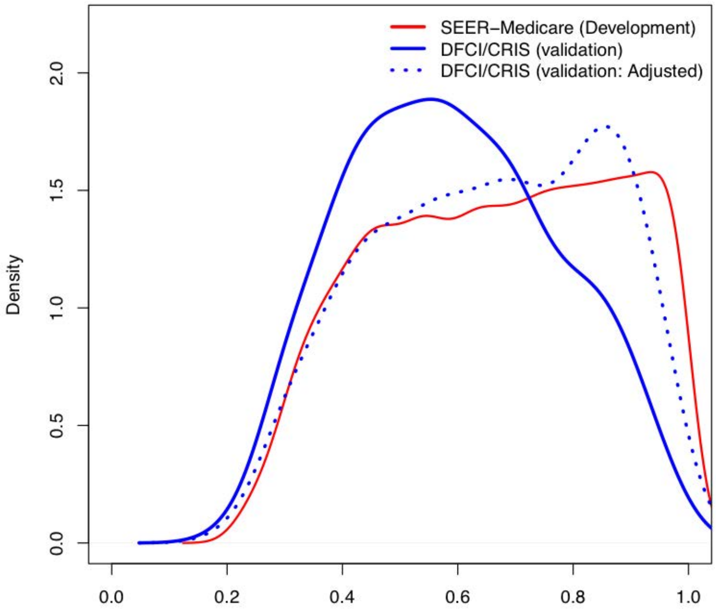

Figure 1: Distributions of the risk scores

Table 2: C-statistics and corresponding 0.95 confidence intervals of the SMMRS

Dataset	Development	Validation (unadjusted)	Validation (adjusted)
C-statistic (0.95CI)	0.700 (0.693 to 0.706)	0.647 (0.608 to 0.688)	0.662 (0.613 to 0.711)

the results sometimes would mislead us. It is recommended to calculate and report an adjusted estimate for the performance metric in validation studies, in addition to the conventional unadjusted estimate.

The proposed approach can be applied to nearly any performance measure. In fact, the equation (1) accommodates performance metrics not only for evaluating a single model but also for comparing two competing models. It can be applied in the same way, by considering $\psi(Z) = \{\psi_1(Z_1), \psi_2(Z_2)\}$ and $Z = (Z_1, Z_2)$, where $\psi_1(Z_1)$ and $\psi_2(Z_2)$ are the two prediction models to compare, and Z_1 and Z_2 are the predictors for $\psi_1(\cdot)$ and $\psi_2(\cdot)$, respectively. For example, for the integrated discrimination index, $Q_\psi(\mathcal{O})$ will be $Y\{\psi_2(Z_2) - \psi_1(Z_1)\} - (1-Y)\{\psi_2(Z_2) - \psi_1(Z_1)\}$.

References

Hastie, T. and Tibshirani, R. (1995). Generalized additive models for medical research. *Statistical methods in medical research*, 4(3):187–196.

Huang, J. and Harrington, D. (2005). Iterative partial least squares with right-censored data analysis: a comparison to other dimension reduction techniques. *Biometrics*, 61(1):17–24.

Janssen, K. J. M., Moons, K. G. M., Kalkman, C. J., Grobbee, D. E., and Vergouwe, Y. (2008). Updating methods improved the performance of a clinical prediction model in new patients. *Journal of clinical epidemiology*, 61(1):76–86.

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine learning*, 86(3):335–367.

Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C., and Habbema, J. D. F. (2004). Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine*, 23(16):2567–2586.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge University Press.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117.

Uno, H., Cronin, A. M., Wadleigh, M., Schrag, D., and Abel, G. A. (2014). Derivation and validation of the SEER-Medicare myelodysplastic syndromes risk score (SMMRS). *Leukemia research*, 38(12):1420–1424.