Title

Testing the Relative Performance of Data Adaptive Prediction Algorithms: A Generalized Test of Conditional Risk Differences

Permalink

https://escholarship.org/uc/item/7cn4j6m5

Journal

The International Journal of Biostatistics, 12(1)

ISSN

2194-573X

Authors

Goldstein, Benjamin A
Polley, Eric C
Briggs, Farren BS
et al.

Publication Date

2016-05-01

DOI

10.1515/ijb-2015-0014

Peer reviewed

Published in final edited form as:
Int J Biostat. 2016 May 1; 12(1): 117–129. doi:10.1515/ijb-2015-0014.

Testing the Relative Performance of Data Adaptive Prediction Algorithms: A Generalized Test of Conditional Risk Differences

Benjamin A. Goldstein,
Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA

Eric C. Polley,
National Institute of Health, National Cancer Institute – Biometric Research Branch, 6130 Executive Blvd EPN RM 8146, Rockville, MD 20892, USA

Farren B. S. Briggs,
Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

Mark J. van der Laan, and
Division of Biostatistics, UC Berkeley, School of Public Health, Berkeley, CA 94720, USA

Alan Hubbard
Division of Biostatistics, UC Berkeley, School of Public Health, Berkeley, CA 94720, USA

Benjamin A. Goldstein: ben.goldstein@duke.edu

Abstract

Comparing the relative fit of competing models can be used to address many different scientific questions. In classical statistics one can, if appropriate, use likelihood ratio tests and information based criterion, whereas clinical medicine has tended to rely on comparisons of fit metrics like C-statistics. However, for many data adaptive modelling procedures such approaches are not suitable. In these cases, statisticians have used cross-validation, which can make inference challenging. In this paper we propose a general approach that focuses on the “conditional” risk difference (conditional on the model fits being fixed) for the improvement in prediction risk. Specifically, we derive a Wald-type test statistic and associated confidence intervals for cross-validated test sets utilizing the independent validation within cross-validation in conjunction with a test for multiple comparisons. We show that this test maintains proper Type I Error under the null fit, and can be used as a general test of relative fit for any semi-parametric model alternative. We apply the test to a candidate gene study to test for the association of a set of genes in a genetic pathway.

Keywords

risk prediction; cross-validation; semi-parametric models; machine learning

Correspondence to: Benjamin A. Goldstein, ben.goldstein@duke.edu.

1 Introduction

An important question in statistical medicine is whether the addition of a set of predictors improves the predictive accuracy of an outcome. Often in these scenarios the unit of interest is not a single predictor but instead a set of predictors (e.g. SNPs in a gene, a set of laboratory values). Numerous metrics and procedures have been developed to assess prediction. Discrimination and reclassification statistics are often used to assess an added predictor to a model. Model building tools such as likelihood ratio tests and Akaike Information Criterion/Bayesian Information Criterion (AIC/BIC) can be used to choose between a model with and without an added predictor. Furthermore, many fields have developed domain specific tests (e.g. gene based tests). While numerous, these methods are all somewhat specialized. Evaluation metrics like C-statistics [1] and Net Reclassification Improvement (NRI [2]); often require the outcome to be categorical. Model building tools while more general require one to be able to specify the number of degrees of freedom, not always feasible when more complicated learning algorithms are used. While domain specific tests are able to leverage specific data structures, they may not always generalize well to other data settings.

In these scenarios a common alternative is to use cross-validation (CV) to compare the fits of two proposed models. Here, one computes the estimated cross-validated risk (average loss), on the two (or more) proposed models. As in information based approaches, one prioritizes the model with smallest estimated risk. One drawback of this approach is that there is typically no accounting for the sampling distribution of these risk estimates. In many medical scenarios there is clearly a simpler (null) model and a more complex (alternative) one. Moreover, there is often a cost associated with the more complex model either via needing to collect additional data (e.g. biomarkers, genes) or computation (e.g. a simple score vs. a more complex one). In these cases, in addition to asking whether one model fits better than another, it is also worthwhile to estimate the probability of an observed difference if the two fits were equally good, i.e. estimating a *p*-value.

With this in mind, we propose a general test for comparing two model fits. The test can be used to answer one of two related questions: (1) does a set of predictors improve model performance?; (2) is one predictor "better" than another? In addition, the proposed test allows the user to specify their own loss-function of interested instead of relying on specific definitions of loss. As more and more studies move from traditional parametric models with clear modes of inference, to the use of supervised semi-parametric (machine learning) approaches for high dimensional problems, demand for such testing procedures will grow.

The paper starts with a definition of the parameter of interest, the *conditional risk*, and we present this first in the context where one has an independent validation set. In Section 3 we then place this test within the context of estimating generalized semi-parametric models. We discuss the motivation for using the super learner [3] procedures for constructing predictors. In Section 4 we introduce our procedure for drawing inferences on the risk differences via cross-validation. We report the results of a simulation study in Section 5 and illustrate a potential use of the proposed procedure via an application to genetic data in Section 6; we end with a conclusion section.

2 Estimation and inference of conditional risk

Before deriving the formalities of the statistical procedure, we intuitively describe the analytic problem. We start by considering a set of predictors and candidate fitting procedure. We divide the sample randomly (and appropriate to the design) into V training and corresponding validation samples. Next, we derive the estimated risk for the fitted model in each validation sample, where the model has been fit on the corresponding training sample. This results in a set of cross-validated predictions, through which we can derive cross-validated losses. The mean of these cross-validated losses, the risk, is an unbiased estimate for the risk on an independent validation set [4]. If we wanted to get a standard error along with the estimate (for example to derive a confidence interval), we may consider applying asymptotic normality and dividing the sample standard deviation by $\sqrt{n}$, the square root of the sample size. However, as others have shown [5], this does not provide an unbiased estimate of the standard deviation on an independent validation set. The goal of this work, then, is to resolve this problem. However to do so, we begin with the *easier* scenario where one has an independent validation set and generalize to the case where one only has a cross-validated dataset.

Define the observed data as $O_i = (Y_i, X_i) \sim P$, $i = 1, \ldots, n$ where Y_i is the outcome of interest and can be a real number or class variable and X_i are a p-dimensional vector of predictors. The unknown model representing $E(Y|X)$ is denoted by $m(X)$. The model's functional form is defined by the data adaptive learning algorithm. Depending on the nature of the algorithm, all, some or none of the input vectors may be used to estimate $E(Y|X)$.

Once the learning algorithm is fit to the data, define $\hat{m}(X)$ to be a prediction based on this model for a randomly drawn (new) X. We call this the *conditional* risk, where one fixes the prediction model, $m(\cdot)$, and looks at its performance in future random draws from the target population. We define it as:

$$\theta(\hat{m}) \equiv E_P[L(Y, \hat{m}(X))] \quad (1)$$

for a user-chosen loss function, L, where the expectation is taken w.r.t P_0, the observed data, O. Possible loss functions include standard ones like squared-error (ℓ_2), absolute error (ℓ_1), or less common ones such as negative log-likelihood, and AUC based loss [6]. The only requirement is that the loss function is convex which excludes missclassification loss [4].

Let the plug-in estimate of the risk be, for an i.i.d. sample of O_i of size n (independent of that used to derive $\hat{m}$) be:

$$\theta_n(\hat{m}) \equiv E_{P_n}[L(Y, \hat{m}(X))] = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \hat{m}(X_i)),$$

where P_n is defined as the empirical distribution. In this case, because the estimator is just a simple average of i.i.d. random variables the asymptotic normality of this estimator is trivially established. More generally (and relevant to the average risk), the asymptotic normality can be established by showing the estimator is an asymptotically linear estimator, and therefore can be written as:

$$\sqrt{n}(\theta_n(\hat{m}(\cdot))-\theta(\hat{m}(\cdot))=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}IC(O_i;\hat{m})+o_p(1/\sqrt{n}). \quad (2)$$

This is the standardized difference between the estimated and true risk of predictor $\hat{m}(\cdot)$. If we can write this as a sum of i.i.d. random variable called the influence curve (IC), $IC(O_i;\hat{m})$ plus a second order term, then, the asymptotic variance of the estimated risk is:

$$var_P[\theta_n(\hat{m})]=\frac{var[IC(O;\hat{m})]}{n}.$$

In the case of average risk, or the case of risk within each validation fold, the IC is simply $IC(O;\hat{m}) = L(Y, \hat{m}(X)) - \theta(\hat{m})$.

For our purposes, the goal is to develop a test of two competing models for estimating $m(\cdot)$. There are a wide variety of scenarios where one would be interested in comparing models fits. Two general cases that we focus on are: the added improvement of a set of predictors of one model fit relative to another, and the relative fit between two competing fitting procedures (e.g., parametric versus data adaptive procedure).

Regarding a statistical test, it is clear that a null of equality will never be true unless one uses the same set of predictors and the same fitting procedure; otherwise one model will always fit better. Thus, for the implied nulls, of whether a data adaptive procedure provides a significantly improved fit relative to a pre-specified parametric model, one can never hope for perfect type I error rate under the null. However we can consider a base model or fitting procedure as the referent (i.e. “null”) and construct a one-sided test of the form:

$$H_0:\theta(\hat{m}_1) \geq \theta(\hat{m}_0), \quad (3)$$

with null (typically simpler) model $\hat{m}_0(\cdot)$ and alternative $\hat{m}_1(\cdot)$. Here, the null hypothesis implies that $\hat{m}_1(\cdot)$ has a larger risk (i.e. worse fit) than $\hat{m}_0(\cdot)$.

The parameter of interest which motivates the test is $\Psi\{\theta(\hat{m}_0), \theta(\hat{m}_1)\} = \theta(\hat{m}_0) - \theta(\hat{m}_1)$, estimated by plug-in estimator $\Psi\{\theta_n(\hat{m}_0), \theta_n(\hat{m}_1)\}$, or Ψ_n for short, which leads naturally to a Wald-type statistic:

$$T_n=\frac{\sqrt{n}\Psi_n}{\sqrt{var_n[\hat{I}C(O;\hat{m}_0)-\hat{I}C(O;\hat{m}_1)]}}, \quad (4)$$

where $\hat{I}C(O;\hat{m}_1) = L(Y, \hat{m}_1(X)) - \theta_n(\hat{m}_1)$.

As Dudoit and van der Laan ([4]) showed, under the null, Ψ_n will by asymptotically normally distributed with variance 1. An equivalent level confidence interval for this risk difference is:

$$\Psi_n \pm z_{1-\alpha/2} = \frac{\sqrt{var_n[\hat{I}C(O;\hat{m}_0) - \hat{I}C(O;\hat{m}_1)]}}{n} \qquad (5)$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

Before we turn to the specific implementation using cross-validation, we first discuss using the super learner (SL) theory for estimating when the goal is either a test of association of a vector of covariates and outcome in a semi-parametric model or to have a standard by which to compare a smaller sub-model via a goodness-of-fit test using the inferential procedure implied by the previous discussion.

3 Optimal estimation of semi-parametric models

As suggested in the outset, the proposed test is most useful in the scenario where one does not have a pre-specified model for m_1 and m_0, and instead needs to estimate them adaptively. When the functional form of m_1 and m_0 are known and the parameters can be estimated using a parametric model, a likelihood ratio test or information based criterion can be optimal and sufficient. However, in the more common scenario where they are unknown, a range of *machine learning* algorithms are available.

Ideally, we want to choose a procedure for deriving $\hat{m_1}$ that is not arbitrary, but has some optimality properties with regards to doing "as good as job as possible" at fitting the true model of $E(Y|X)$. Therefore, when compared to a fit from a null model (or some smaller model), we want to have maximal power for detecting departures from this null. Though no theory exists for deriving an optimal estimator of the predictor among all possible estimators in a semi-parametric model, we can at least define a gold standard among a finite set of competitors. As others have done [3] we define the gold standard as: given the data, the algorithm that chooses the estimator with the lowest true risk among all tried competitors (indexed by j). This is referred to as the "Oracle Selector":

$$\hat{m}^* = \arg\max_j \Psi_j.$$

We define the risk of the Oracle Selector as $\theta(\hat{m^*})$ based on an independent training sample of size n_{Tr}. An estimator that converges in risk to the Oracle Selector will result from maximizing the true $\Psi_j \equiv \theta(\hat{m_0}) - \theta(\hat{m_j})$ over different competing estimators. Thus, such a procedure should also result in a relatively powerful test compared to other procedures used to derive $\hat{m_1}$.

As shown in van der Laan et al. [3] *stacking* procedures [7, 8] (particularly those that combine a wide variety of algorithms from very simple/smooth to highly data-adaptive), as implemented in the SL algorithm, meet this criterion. In stacking one typically uses a V-fold cross-validation procedure to combine a user-specified set of candidate prediction algorithms. The SL algorithm is available as a statistical package [9] in the R programming language. The SL performs asymptotically equivalently (w.r.t. expected risk difference), up to a second order term, to the oracle selector. In addition, an Oracle Inequality suggests that this relative optimality occurs if the number of learning algorithms included as candidates in the SL is polynomial in sample size, and none of its candidate learners (and oracle estimator) converges at a parametric rate [3]. If one of the candidate learners is actually the true model, however, and thus converges at a parametric rate, the SL will converge at a close to parametric rate, implying there is not much cost for building the predictor within a much bigger model (assuming the loss function is bounded). Thus theory on stacking procedures suggests the use of a very large number of possible learning algorithms. If stacking is used to derive $\hat{m_1}$ then the Oracle Inequality can be invoked as a heuristic argument as to why the resulting test is relatively powerful compared to other procedures that would use a different procedure for estimating $\hat{m_1}$.

4 Cross-validated multiple testing procedure

Given a procedure described (for simplicity) as designed for an independent test set, we now turn to the more relevant situation where a single data set must serve multiple purposes: construction of the $\hat{m_0}$, $\hat{m_1}$ models, subsequent estimation of the cross-validated risk, and the calculation of the its sampling variability. As is typical, we use V-fold cross-validation, where the data are divided into V equal-sized *testing* sets. For each independent test set indexed by v, the prediction models, ($\hat{m_0}$, $\hat{m_1}$), are fitted (*trained*) on the corresponding training set and then the test statistic (4), is constructed on the test set. Thus, both $\hat{m}_0^{-v}$, and $\hat{m}_1^{-v}$ represent fits on the *vth* training sample where those observations labelled (randomly) v have been removed. We can then define:

$$T_v = \frac{\Psi^v_{n_V}}{SE(\Psi^v_{n_V})}$$

$$\text{where } SE(\Psi^v_{n_V}) = \sqrt{var^v_{P_{n,v}}\left[\hat{I}C(O;\hat{m}_0^{-v}) - \hat{I}C(O;\hat{m}_1^{-v})\right]} / \sqrt{n_V} \qquad (6)$$

where we have indexed (6) to emphasize that the relevant predictor models are fit on the training data, but the difference in the risk estimates, $\Psi^v_{n_V}$ is performed on the testing data, which has sample size $n_V = n/V$. We first discuss a procedure, where the Central Limit Theory follows without any special conditions – that is, we treat each validation sample as the basis of a separate test and estimate, and combine them only via standard multiple testing procedures. We then discuss estimating an average risk across the folds which does require some mild conditions for the asymptotic distributional results to hold.

4.1 Estimates by validation sample

For each of the $v = 1, \ldots, V$ test statistics, T^v_n, we can derive a corresponding p-value of the null as $p_v = 1 - \Phi(T^v_n)$, where $\Phi(\cdot)$ is the standard normal distribution function. To draw inference we can perform a multiple testing correction across the V tests, rejecting the null hypothesis if the minimum corrected p-value is less than the prescribed α. In practice, due to the relatively few number of tests, we find that the standard Bonferonni correction is not overly conservative – obviously it could be generalized to other procedures. Therefore, in the same way that we derive inference from the fold with the greatest association (i.e. risk difference) in eq. (6), we can similarly choose that same fold as our estimate of the risk difference. To get appropriate coverage we calculate Bonferonni corrected confidence intervals. This leads to a global confidence interval for the set of Confidence Intervals (CIs) across folds, giving global coverage 1 − α of

$$\Psi^{v}_{n_V} \pm z_{1-\alpha/(2V)} SE(\Psi^{v}_{n_V}) \quad (7)$$

Thus, by defining the parameter of interest as the conditional risk, one gets V-different estimates of an experiment where two different competing procedures are used to generate predictors for which the risks are estimated. However, in practice it makes the most sense to use the cross-validated estimated of the risk, as this will converge to the true risk faster.

4.2 Averaging across validation samples

Instead of joint inference across the V-folds keeping the validation estimates separate, we can also combine them into an average conditional risk:

$$\begin{aligned}\Psi_n &= \frac{1}{V}\sum_{v=1}^{V}\Psi^{v}_{n_V} \\ &= \frac{1}{V}\sum_{v=1}^{V} E_{P_{n,v}}\left(L[Y,\hat{m}_o^{-v}(X)] - L[Y,\hat{m}_1^{-v}(X)]\right)\end{aligned} \quad (8)$$

Dudoit and van der Laan ([4]) showed that the normalized estimator $\sqrt{n}(\psi_n - \psi)$ converges to a normal distribution with mean 0 and variance, $\sigma^2 = \frac{1}{V}\sum_{v=1}^{V}\sigma_v^2$. We define $SE(\psi_n) = \frac{\sigma_n}{\sqrt{n}}$ where

$$\sigma_n^2 = \frac{1}{V}\sum_{v=1}^{V}\sigma_{v,n}^2$$

and $\sigma_{v,n}^2$ is the estimated variance of [$\hat{I}C(O;\hat{m}_0^{-v}) - \hat{I}C(O;\hat{m}_1^{-v})$] within each validation fold. Thus, one can derive a Wald-type confidence interval and test statistic just as above:

$$T = \frac{\sqrt{n\Psi_n}}{\sqrt{\sigma_n}}.$$

This provides a single overall test and/or confidence interval that might be a more efficient summary of the evidence related to the guiding hypothesis of interest. However, it does require more assumptions, and these are not trivial. For instance, it is possible that under H_0, as the sample size gets large $\hat{m}_1^{-v}(X) \to \hat{m}_0^{-v}(X)$, and the asymptotic linearity will no longer hold. However, in these cases, it is because the null is “too” true, resulting in a point mass at 0. In this degenerative case where $m_1\hat{}$gets very close $m_0\hat{}$, one can simply choose the null model. Therefore, the situation hurts the asymptotics, but in practice it does not impact inference on the conditional risk differences.

4.3 Finite sample considerations

While the asymptotic statistical inference is straightforward when one views the parameter of interest as the difference in conditional risks, there is still the question of finite sample performance. For such considerations, the primary tuning parameter is the number of splits, V, one should choose. The performance of the test will be a function of two competing goals: 1) having the training samples as large as possible in order to get closer convergence of the fit, $m_{1,n}^{-v}$ to its limit, and 2) having the validation samples as large as possible in order to get estimates of $\Psi_{n_V}^{v}$ with variances as small as possible as well as being able to invoke the asymptotic sampling distribution under the null. Therefore, V must be *small enough* to invoke asymptotic normality of T_V. Through our simulations we noted that it is typically necessary to have a validation sample size of at least 30. While we did not explicitly seek to derive such limits, this is worthy of further research.

4.4 “Marginal” risk differences

In this work, we focus on comparing risks conditional on the fitting procedure. One might want to instead derive a test of competing procedures, where one does not condition on the estimates of the predictors, but measure the performance of competing procedures (algorithms) in repeated experiments. So, as opposed to average conditional risk, the parameter of interest is the average unconditional risk difference. In this case, in addition to estimating the risk from a fixed prediction model, the experiment also involves re-fitting of the prediction model as well, something we will refer to as average marginal risk differences. The most complete discussion of this problem was performed by Bengio and Grandvalet [5]. The authors showed that the variance of the loss can be broken down into three components:

1. The variability of the prediction within each validation block
2. The covariance between predictions within each block
3. The covariance between predictions in different blocks

The first value is the parameter of interest, however the empirical variance of $m(X_i)$ is confounded by the other two values. In later work the authors estimated the maximum between block variance as 0.7 and suggested a t-statistic with a correction based on this value [10]. In our own previous work we suggested a Wald test that also required a correction to maintain proper error control [11]. Markatou et al. [12] presented a method of moments estimator that while nearly unbiased, depends on the distribution errors and knowledge of the learning algorithm.

Therefore this work is important in that it circumvents these issues by proposing a type of random parameter, the conditional risk, which both has some appeal as the quantity of interest (typically, for practical performance of an estimated predictor, one is interested in how a fixed model fit will do in the future) and also avoids these intractable problems.

5 Simulation

Simulations were performed to examine 1) the asymptotic sampling distribution of the cross-validated risk estimates, 2) the type I error rate when the base model provides a better fit, and 3) the coverage of the confidence interval. The simulations all had the same structure, which complies with the experiment under which the theory is developed for the asymptotic distribution of the risk estimator.

5.1 Methods

1. Generate a random sample of size n from the data generating distribution and break into V equal validation samples of size $n_V = n/V$ with corresponding training samples of size $n - n/V$;

2. For each training sample, estimate the models using both the base model and the alternative model, resulting in V pairs of model fits. These are the sets ($\hat{m}_0^{-v}, \hat{m}_1^{-v}$, $v = 1, .., V$) discussed above, and are considered the set of fixed predictors of which we evaluate the sampling distribution of the risk estimates in future draws from the target population;

3. For each of the V leave-out sets, calculate the risk difference, Ψ_{nv}^{v} and associated standard error;

4. Calculate the Wald test from (6) and the associated p-value (using a Bonferonni correction). Using the most significant Wald test and the average of the Ψ_{nv}^{v} calculate the corresponding CI from (7);

5. As opposed to deriving the "true" risk parameter analytically, we did so empirically. For each of these 2 × V predictors (V prediction for m_0 and m_1), we draw a very large sample (100,000 observations) using the same distribution. This represents the target population and is used to calculate the true risk and the parameters of interest: $\theta(\hat{m}_0^{-v}), \theta(\hat{m}_1^{-v})$ for each v, the corresponding risk differences, $\Psi\{(\hat{m}_0^{-v}), \theta(\hat{m}_1^{-v})\}$, the average risk and the difference of these average risk differences;

6. To estimate the sample distribution and performance repeat (1)–(4) 1,000 times, drawing new samples and generating new fits. Compare the mean of the risk differences, coverage probabilities for confidence intervals, and rejection (at 0.05 level) probability, to the true risk difference from (5).

We present results for two sample sizes (n = 100, 1000) and V equal to 5 (though other n and V were explored) within four simulations.

- *Simulation I:* The base dataset contains variables, $X = (x_1, x_2)$. Conditional on X, $Y \sim N(x_1\beta, 1)$. The alternative dataset contains 8 additional variables such that: $X = (x_1 \ldots x_{10})$, where the additional variables are not associated with Y, i.e. a null association. Both models are estimated using basic linear regression and the risk is calculated under squared-error loss;

- *Simulation II:* Like Simulation I, the base dataset contains $X = (x_1, x_2)$ and the alternative dataset contains $X = (x_1 \dots x_{10})$. Y is binary such that $logit(P(Y = 1|X)) \sim x_1\beta_1 + x_3\beta_3$. Therefore the alternative model, with both x_1 and x_3, should provide a better fit for estimating Y. Both models are estimated using basic logistic regression and the risk is calculated under absolute-error loss;
- *Simulation III:* Both the base and alternative datasets contain only x_1 and x_2, with linearly x_1 associated with a continuous outcome Y: $E(Y) = \beta X_1$. The base model was estimated using linear regression. The alternative model was estimated using a SL with linear regression, general additive model, decision tree and an intercept model, i.e. an overly complex model. Absolute-error loss was used to calculate the risk;
- *Simulation IV:* Both datasets contain only x_1 with the true $E(Y|X)$ shown in Figure 1, based on a piecewise constant model. The base fit is mis-specified using only linear regression where the alternative is fit under a SL using linear regression, general additive model, decision tree and an intercept function. The risk is calculated under squared-error loss.

1,000 simulations were performed. We report bias (estimated risk difference vs. “true” risk difference, type 1 error (based on a 1 sided test), and coverage probability. Since simulations I & II involve comparison of parametric fits we also report the results of a likelihood ratio test and AIC.

5.2 Results

Tables 1 and 2 show the results, and are as predicted based on the theory: 1) the coverage probabilities of the risk differences with independent validation samples achieve close to the specified coverage rate, 2) when the base model is correctly specified (and is a simple parametric model) then the procedure has a very high probability of failing to reject the null hypothesis (that is, it suggests the simpler model is a sufficient/superior fit), 3) if the null model is mis-specified, the procedure has very high power.

When looking across the validation folds, the Bonferonni corrected statistic provides slightly conservative coverage and type 1 error rate. However the need for such a statistic is seen when one simply takes the empirical variance across the validation folds. In this scenario the coverage is not always appropriate particularly when the base model is correctly specified (simulations I & III), as discussed in Section 4.3. As noted by Bengio and Grandvalet [5] there are additional correlation components embedded in V-fold CV. These correlation components likely add second order terms to the IC calculation (eq. (2)).

We note, that when the sample size is relatively small, so the validation sample is very small (20 in this case), then borrowing across the validation samples to get an average risk gives better performance, both with respect to type I error (Simulations 1 & III) and power (Simulations II & IV). This is the major advantage of performing CV as opposed to a simple sample split.

Since simulation II consisted in comparison of two parametric fits we were able to explicitly compare the power of the proposed method to both a likelihood ratio test (LRT) as well as a model choice via AIC. For the LRT we rejected the null at an α level of 0.05 and for the AIC we chose the model that minimized the AIC. Results are shown in Figure 2. With the smaller sample, the parametric methods show more power (0.859 and 0.845, respectively), while all methods maintain full power with the larger sample size.

6 Data application

Gene based tests represent a particular application of this procedure. Genes are comprised of individual bases of DNA referred to as single nucleotide polymorphisms (SNPs). Most genes consist of 10s or 100s of SNPs. Typical methodology involves testing individual SNPs in a gene to determine whether variation in the gene as a whole may be associated with the outcome (typically disease) of interest. The goal of gene based tests is to associate the set of SNPs comprising the gene with some outcome, with many different proposed procedures [13]. Of particular relevance to the present discussion is that the unit of interest in a gene based test is the collection of SNPs instead of any individual SNP. Therefore we are not interested in defining a model that estimates a parameter for each individual SNPs. Instead we want to globally assess the association of a set of SNPs.

To illustrate the flexibility of the proposed method we will show how it can be used as a gene based test for association. We return to a data analysis we previously performed where we explored whether genes from the stress response pathway, a set of predefined genes, are associated with Multiple Sclerosis (MS) [14]. Using a combination of machine learning procedures and logistic regression we identified one gene, CRHR1, to be associated with disease. However, at the time we were unable to formally test this association. Using the proposed method, we revisted this analysis.

The data consist of a candidate gene study on 2,722 people. The stress response pathway consists of 10 genes, which are comprised of 409 SNPs across the genome. In univariate testing rs171442 in CRHR1 has the smallest p-value ($p < 0.003$). However the association is no longer significant after controlling for multiple testing using the Benjamini-Hochberg [15] method to control the False Discovery Rate ($p_{adjusted} = 0.42$).

Since our unit of interest were the 10 genes (and not the 409 SNPs), we tested each gene individually. The outcome Y is MS disease state and the X are the set of SNPs in each of the genes. In this case $\hat{m}_0$ is the baseline fit with no genes (SNPs) in the model. We tested each of the 10 genes: $\hat{m}_1^1, \ldots, \hat{m}_1^{10}$, where $\hat{m}_1^i$ is the fit for the i^{th} gene. In addition we assessed the full pathway, fitting a global model $\hat{m}_1$ for all the genes. We used squared-error loss as the primary loss function (L), though other loss functions resulted in similar conclusions. A SuperLearner was fit using a library consisting of RandomForests [16], LASSO [17], GLM and K-Nearest Neighbours [18] along with an intercept. Twenty-fold cross-validation was performed and the p-value based on (6) was calculated in each fold. The results for the entire pathway and each individual gene is shown in Table 3.

The overall pathway had a significant association ($p < 0.012$) as did the CRHR1 gene. To examine whether all of the association resided in the CRHR1 gene we compared the overall pathway to just the CRHR1. In this case, $\hat{m}_0$ is the fit based on just CRHR1 and $\hat{m}_1$ is the fit based on all SNPs. Not surprisingly there was no added benefit of any individual gene above and beyond that of CRHR1 ($p = 1.0$), suggesting that CRHR1 is the only gene in the stress response pathway associated with MS. Overall, this formally confirms the conclusions in the original paper that we were only able to make by suggestion [14].

7 Conclusion

In this paper we have implemented an estimation and inferential procedure based on the theory developed in Dudoit and van der Laan [4] for testing the risk difference in two competing fitted prediction models. The proposed test can be interpreted as a comparison of the fit of two models or test of association for a set of predictors. It can also be used as a goodness of fit test for a semi-parametric or data adaptive model. This test of risk difference can be based upon almost any bounded loss function. In constructing the test we utilized the independent validation sets that exist within V-fold cross-validation. This work then also addresses an open question in statistical learning: how to derive estimators and inference for cross-validated risk, and use those to make decisions about competing models.

The procedures proposed have important application to statistical medicine. Many studies are interested in whether a set of values (e.g. biomarkers, clinical measurements etc.) improve the assessment of an outcome. Typical methods for such assessment (e.g. ROC, recalibration statistics), rely on the outcome being binary (generally disease state). However, this is not always the case as one may want to predict a continuous outcome such as a laboratory value or a survival outcome. Moreover these methods each use a specific loss function that the user has little control over. The proposed method allows one to use any convex loss function. The test can be interpreted as a test of association for the set of predictors, similar to likelihood ratio tests or information based methods, but not restricted to nested models or models where one knows the degree of freedom. This allows one to use any semi-parametric model to estimate the functional form and derive inference on the overall fit. If that model has certain optimality properties, as does the SL stacking based algorithm, then the test represents an asymptomatically most powerful goodness of fit test for semi-parametric models.

In our example we applied the test to a previous analysis of a candidate gene study. The scientific question was the association of particular genes with disease state, as opposed to simply individual SNPs. A previous analysis concluded that only the CRHR1 gene in the stress response pathway was associated with MS, but this could not be quantified. Our formal test, confirmed this conclusion.

This procedure also presents a means to test a group of variables for association with an outcome. Pepe et al. [19] showed that if any of the variables in a prediction model are associated with the outcome then a test for prediction will always be significant. However, as our data analysis shows, often no individual predictor will be associated. This may be due to issues of multiple testing or because a parametric model cannot be adequately specified. In this case, semi-parametric methods, like SuperLearner, become more valuable. In our presented data analysis, the unit of interest for inference was less focused on the individual SNPs, but more focused on the gene and pathway level, for which no general test would exist.

While this test has great application and is fairly intuitive, it is limited by the ability to derive a strong predictive model. Depending on the learning algorithm used one may reach different conclusions. It is for this reason that we framed the test within the context of semi-parametric models with optimality properties. While the test is robust to the size of the validation set, too small of a set size will result in an anti-conservative test.

In all, this test fills a gap both in the machine learning literature as well as the statistical medicine literature.

Acknowledgments

Funding BAG was funded in part by National Institute of Diabetes and Digestive and Kidney Diseases K25 DK097279, a National Institutes of Health NRSA Trainee appointment on grant T32 HG 00047 and the Russell M. Grossman Endowment. FBSB is a National Multiple Sclerosis Society Post-Doctoral Fellow (FG 1847A1/1).

References

1. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982; 247:2543–6. [PubMed: 7069920]
2. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. Jan.2008 270:157–72. [PubMed: 17569110]
3. van der Laan MJ, Polley EC, Hubbard AE. Superlearner. Stat Appl Genet Mol Biol. 2007; 6
4. Dudoit S, van der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Stat Methodol. 2005; 2:131–54.
5. Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation. J Mach Learn Res. 2004; 5:1089–105.
6. Takenouchi T, Komori O, Eguchi S. An extension of the receiver operating characteristic curve and AUC-optimal classification. Neural Comput. Oct.2012 240:2789–824. [PubMed: 22734493]
7. Breiman L. Stacked regressions. Mach Learn. 1996; 24:49–64.
8. Wolpert DH. Stacked generalization. Neural Networks. 1992; 5:241–59.
9. Polley, E.; van der Laan, M. SuperLearner: Super Learner Prediction. 2012. Available at: http://CRAN.R-project.org/package=SuperLearner. R package version 2.0-6
10. Grandvalet, Y.; Bengio, Y. Technical Report. Vol. 1285. Departement dInformatique et Recherche Operationnelle; Aug. 2006 Hypothesis testing for cross-validation.
11. Goldstein BA, Hubbard AE, Barcellos LF. A generalized approach for testing the association of a set of predictors with an outcome: A gene based test. Technical Report. Jan.2011 274 U.C. Berkeley Division of Biostatistics Working Paper Series.
12. Markatou M, Tian H, Biswas S, Hripcsak G. Analysis of variance of cross-validation estimators of the generalization error. J Mach LearnRes. 2005; 6:1127–1168.
13. Beyene J, Tritchler D, Asimit JL, Hamid JS. Gene- or region-based analysis of genome-wide association studies. Genet Epidemiol. 2009; 33:s105–s110. [PubMed: 19924708]
14. Briggs FB, Bartlett SE, Goldstein BA, Wang J, McCauley JL, Zuvich RL, et al. Evidence for CRHR1 in multiple sclerosis using supervised machine learning and meta-analysis in 12,566 individuals. Hum Mol Genet. Nov.2010 190:4286–95. [PubMed: 20699326]
15. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc. 1995; 57:289–300.

16. Liaw A, Wiener M. Classification and regression by randomforest. R News. 2002; 20:18–22. Available at: http://CRAN.R-project.org/doc/Rnews/.

17. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Software. 2010; 330:1–22. Available at: http://www.jstatsoft.org/v33/i01/.

18. Venables, WN.; Ripley, BD. Modern Applied Statistics with S. 4. New York: Springer; 2002. Available at: http://www.stats.ox.ac.uk/pub/MASS4

19. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. Stat Med. Jan; 2013 32(9):1467–1482. [PubMed: 23296397]

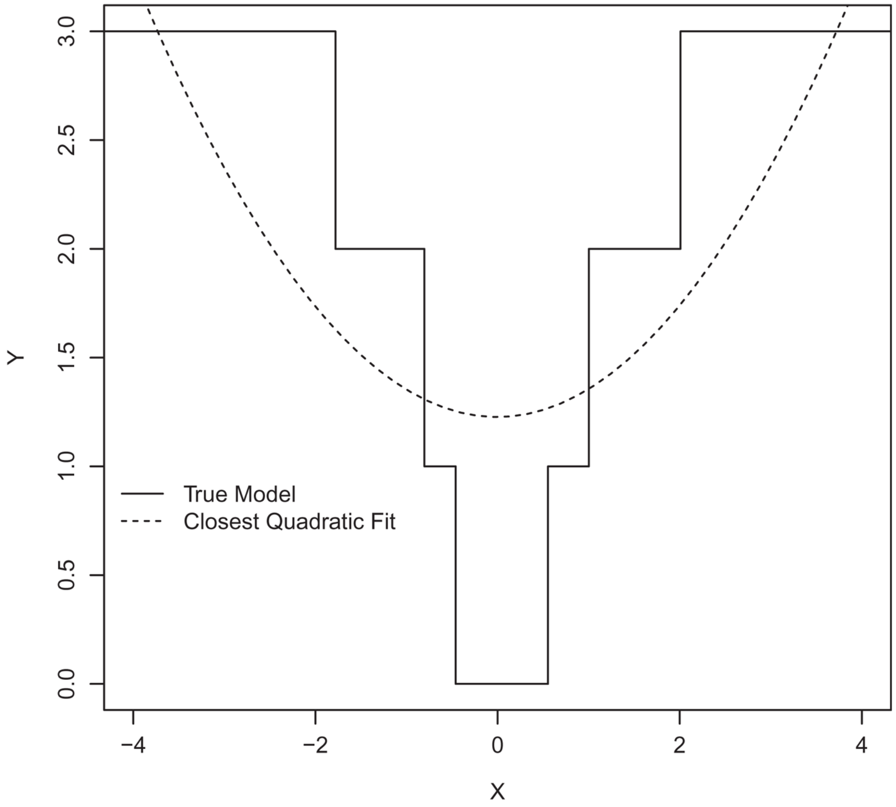

Figure 1.

The true model is Simulation IV: $E(Y|X)$ and closest quadratic approximation.

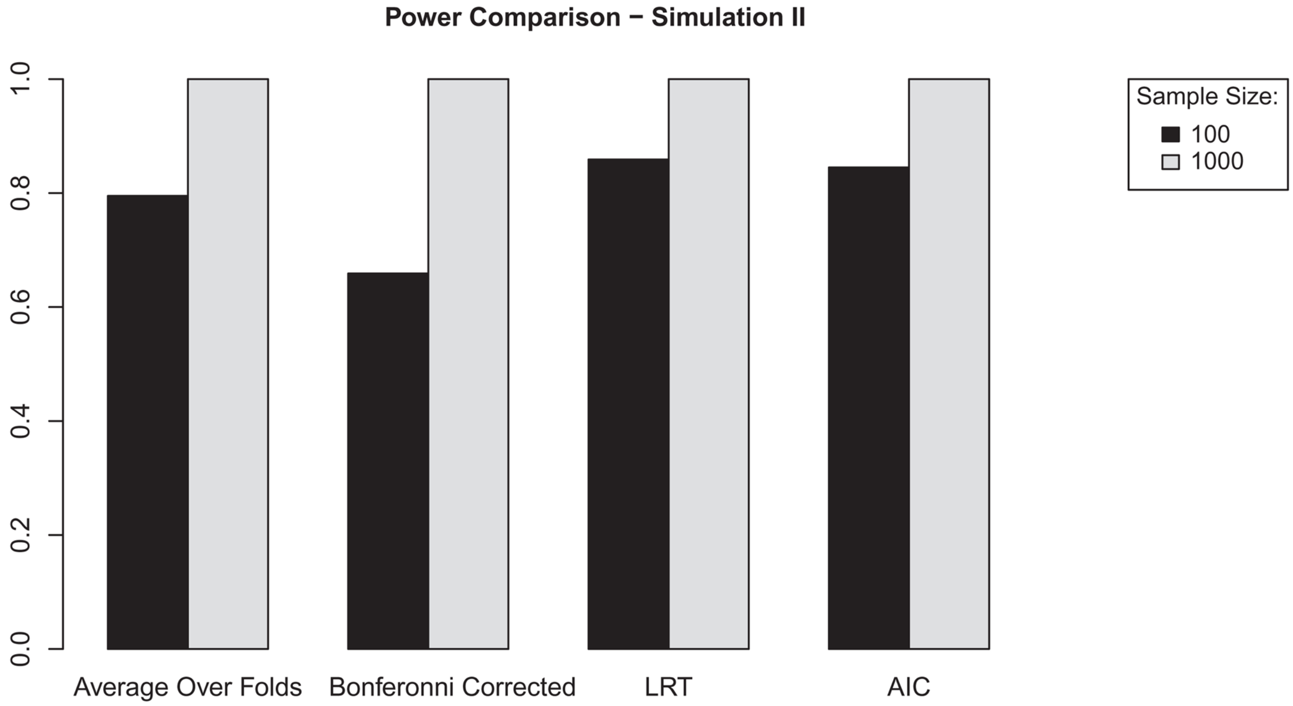

Figure 2.

Comparison of power (rejection rate) of proposed method to parametric method in *Simulation II*.

Table 1

Simulations I & II: The question of interest is the added benefit of a set of predictor. In simulation I the additional predictors does not add information. In simulation II the added predictors are associated.

Sim	Sample size	Fold	True risk diff	Est risk diff	Bias	SE	Coverage	Rejection
I	100	1	−0.032	−0.032	−0.000	0.039	0.940	0.016
		2	−0.031	−0.029	0.002	0.038	0.943	0.011
		3	−0.031	−0.029	0.002	0.038	0.925	0.018
		4	−0.031	−0.029	0.002	0.038	0.930	0.012
		5	−0.032	−0.031	0.001	0.038	0.923	0.020
		Average over folds	−0.031	−0.030	0.001	0.018	0.835	0.001
		Bonferonni corrected	−0.031	−0.030	0.001	0.042	0.994	0.011
I	1,000	1	−0.003	−0.003	−0.000	0.003	0.941	0.010
		2	−0.003	−0.002	0.000	0.003	0.946	0.016
		3	−0.003	−0.003	−0.000	0.004	0.947	0.006
		4	−0.003	−0.003	0.000	0.003	0.953	0.009
		5	−0.003	−0.003	0.000	0.003	0.947	0.016
		Average over folds	−0.003	−0.003	0.000	0.002	0.818	0.000
		Bonferonni corrected	−0.003	−0.003	0.000	0.004	1.000	0.010
II	100	1	0.069	0.067	−0.002	0.052	0.936	0.373
		2	0.069	0.068	−0.001	0.052	0.926	0.383
		3	0.069	0.067	−0.002	0.052	0.943	0.364
		4	0.069	0.070	0.002	0.052	0.939	0.395
		5	0.069	0.065	−0.004	0.052	0.927	0.368
		Average over folds	0.069	0.068	−0.001	0.024	0.915	0.795
		Bonferonni corrected	0.069	0.068	−0.001	0.043	0.999	0.659
II	1,000	1	0.070	0.070	0.000	0.013	0.947	1.000
		2	0.070	0.071	0.000	0.013	0.939	0.999
		3	0.071	0.070	−0.001	0.013	0.962	1.000
		4	0.070	0.071	0.000	0.013	0.942	1.000
		5	0.070	0.070	−0.001	0.013	0.960	1.000
		Average over folds	0.070	0.070	−0.000	0.006	0.940	1.000

Sim	Sample size	Fold	True risk diff	Est risk diff	Bias	SE	Coverage	Rejection
		Bonferroni corrected	0.070	0.070	−0.000	0.012	1.000	1.000

Table 2

Simulation II & IV: The question of interest is to compare across two different fitting procedures. In simulation III the alternative model is overly complex. In simulation IV the more complex should provide a better fit.

Sim	Sample size	Fold	True risk diff	Est risk diff	Bias	SE	Coverage	Rejection
III	100	1	−0.008	−0.007	−0.001	0.016	0.959	0.028
		2	−0.008	−0.008	0.000	0.017	0.958	0.033
		3	−0.009	−0.008	−0.001	0.018	0.957	0.040
		4	−0.008	−0.007	−0.001	0.017	0.963	0.033
		5	−0.009	−0.011	0.002	0.018	0.956	0.029
		Average over folds	−0.008	−0.008	0.000	0.010	0.950	0.011
		Bonferonni corrected	−0.008	−0.008	0.000	0.025	0.929	0.013
III	1,000	1	−0.001	−0.001	0.000	0.001	0.963	0.020
		2	−0.000	−0.001	0.000	0.001	0.960	0.024
		3	−0.001	−0.001	0.000	0.001	0.953	0.037
		4	−0.001	−0.001	0.000	0.001	0.957	0.022
		5	−0.001	−0.001	0.000	0.001	0.973	0.026
		Average over folds	−0.001	−0.001	0.000	0.001	0.934	0.012
		Bonferonni corrected	−0.001	−0.001	0.000	0.002	0.959	0.014
IV	100	1	1.143	1.122	0.021	0.335	0.932	0.980
		2	1.140	1.141	−0.001	0.341	0.925	0.983
		3	1.144	1.147	−0.003	0.340	0.933	0.981
		4	1.143	1.129	0.013	0.337	0.925	0.973
		5	1.142	1.133	0.009	0.338	0.931	0.987
		Average over folds	1.142	1.134	0.008	0.154	0.942	1.000
		Bonferonni corrected	1.142	1.134	0.008	0.324	0.998	1.000
IV	1,000	1	1.154	1.149	0.004	0.106	0.939	1.000
		2	1.154	1.148	0.006	0.106	0.939	1.000
		3	1.153	1.143	0.011	0.106	0.952	1.000
		4	1.153	1.152	0.001	0.106	0.957	1.000
		5	1.153	1.146	0.008	0.106	0.944	1.000
		Average over folds	1.153	1.148	0.006	0.047	0.941	1.000

Sim	Sample size	Fold	True risk diff	Est risk diff	Bias	SE	Coverage	Rejection
		Bonferonni corrected	1.153	1.148	0.006	0.105	1.000	1.000

Table 3

P-values for the overall stress response pathway as well as each individual gene. The overall pathway ($\hat{m}_1$) shows a significant association as does the CRHR1 gene ($\hat{m}_1^4$). However the remaining genes show no evidence of association with MS.

All genes	BDNF	BDNFOS	CRHBP	CRHR1	CRHR2
0.0111	0.6528	1.0000	1.0000	0.0061	0.6058
	GDNF	**HCRTR1**	**HCRTR2**	**OPRD1**	**OPRK1**
	1.0000	0.1185	1.0000	0.7442	0.2028