University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2013 *Paper* 316

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

Benjamin A. Goldstein[*] Eric Polley[†]
Farren Briggs[‡] Mark J. van der Laan[**]

[*]Quantitative Sciences Unit, Stanford University, ben.goldstein@stanford.edu
[†]Biometric Branch, NCI, eric.polley@nih.gov
[‡]University of California, Berkeley, fbriggs@genepi.berkeley.edu
[**]Division of Biostatistics, University of California, Berkeley, laan@berkeley.edu

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

Benjamin A. Goldstein, Eric Polley, Farren Briggs, and Mark J. van der Laan

Abstract

In statistical medicine comparing the predictability or ?t of two models can help to determine whether a set of prognostic variables contains additional information about medical outcomes, or whether one of two different model ?ts (perhaps based on different algorithms, or different set of variables) should be preferred for clinical use. Clinical medicine has tended to rely on comparisons of clinical metrics like C-statistics and more recently reclassi?cation. Such metrics rely on the outcome being categorical and utilize a speci?c and often obscure loss function. In classical statistics one can use likelihood ratio tests and information based criterion if the comparisons allow for it. However, for many data adaptive models such approaches are not suitable and people have traditionally used cross-validation to choose between models in such settings. In this paper we propose a test that focuses on the "conditional" risk differences (conditional on the models being ?xed) for the improvement in prediction risk, which is valid under cross-validation. We derive Wald-type test statistics and con?dence 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 ?t, and can be used as a general test of relative ?t for any semi-parametric model alternative, using most any loss function. We apply the test to a candidate gene study to test for the association of a set of genes in a genetic pathway.

1 Introduction

An important question in statistical medicine is whether the addition of a set of predictors improves the overall prediction 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 do not always generalize well to other data settings.

Instead, we propose a general test for assessing whether a predictor or set of predictors improves model performance relative to an alternative baseline model or a null model where the comparison can be based on a wide variety of loss functions. Like many evaluation tools the proposed test involves deriving a set of predicted values using the original set of predictors and comparing those to the predicted values using the additional predictors. Inherent in this process is the derivation of a prediction model. It is well recognized that simply applying the prediction model to the same data that was used to derive the model will result in over-estimation of the predictive accuracy, referred to as *optimism* [3]. A straightforward approach would be to have an independent set of data that one can use as a validation set. However, in medical studies data is often limited, and setting aside a set of data is not an optimal use of resources. In these scenarios the typical approach is to apply cross-validation. While cross-validation (CV) produces unbiased estimates for the expectation of the loss [4], it only recently has become clear how to derive consistent estimates of the sampling distributions for types of risk estimates that CV provides [5]. It has been noted that if the parameter of interest is the risk of a fitting procedure as opposed to a fixed previously estimated model (referred heretofore as unconditional risk), there is generally an overestimation of the variances of the estimate if the variance estimate assumes the cross-validated loss over the entire data set

is independent; in fact [6] showed that the variance of the unconditional CV risk is non-identifiable.

In a pure testing context, [7] developed an algorithm using a combination of a machine learning algorithm and permutation methods to develop a general test of independence of an outcome and a vector of associated predictors. For testing, the main limitation of this method involves very long computation times, as the combination of the permutation with cross validation and machine learning methods can require impractical computation times. In addition, because it is a test of independence, it does not generalize to more general goodness-of-fit tests, such as the comparison of a semi-parametric model fit to a simpler parametric model fit. In this paper, we propose a method that relies on the asymptotic sampling distribution of a risk estimator to avoid the permutation and derive inference on the relative ability of competing models to predict the outcome, as defined by the particular loss function of interest. We present methodology that both conducts tests of the risk difference for each validation fold, we as well as estimates and inference of the average risk across folds. Though the methodology can be applied to general comparisons of the ability of estimated models to fit future data, one particularly interesting example concerns a general test of the joint association of a large vector of predictors with an outcome. Specifically, combining the general CV procedure with a particular (optimal) combination of statistical learners, one can create a powerful data-adaptive test of the association of an outcome with a vector or predictors within a semi-parametric model. As more and more studies move from traditional parametric models with clear modes of inference, to the use of supervised semi-parametric models for high dimensional problems, the implications of such a testing framework is widely applicable, in applications of high dimensional clinical data and bioinformatics.

The paper starts with a definition of the parameter of interest, the *conditional risk*, in Section 2 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 estimation generalized semi-parametric models and discuss the motivation for using the super learner [8] 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 then finish with some concluding thoughts.

2 Estimation and Inference of Conditional Risk

Before deriving the formalities of the statistical procedure, one can describe our proposal very simply. First, after dividing the sample randomly (and appropriate to the design) into V validation samples, we derive the estimated risk for competing fitted models in each validation sample, where theses models have been fit on the corresponding training sample. We then derive estimates and joint inference of these V risk differences (estimates, tests, confidence intervals). Then, we examine the average differences of the CV-risk across the V folds, and provide corresponding inference for this quantity as well. For the former, we need nothing but the standard central limit theorem for our results; for the latter, to derive a CLT we need some modest conditions [5]. We discuss both because, for the average CV-risk, these conditions can be violated in some practical situations, whereas the procedure that does not average across the folds still performs well. Though the details are important, these two approaches are very simple, and potentially very powerful for investigating a wide variety of relevant model comparisons.

Define the observed data as $O_i = (Y_i, X_i) \sim P_0, i = 1, \ldots, n$ 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 function is defined by the learning algorithm used to estimate the model. Depending on the nature of the algorithm and final estimated model, 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 . The so-called *conditional* risk (that is fixing the prediction model, $\hat{m}(\cdot)$, and looking at its performance in future random draws from the target population) is defined as [5]:

$$\theta(\hat{m}) \equiv E[L(Y, \hat{m}(x))] \tag{1}$$

for a user-chosen loss function, L, where the expectation is taken w.r.t P_0. 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:

$$\theta_n(\hat{m}) \equiv \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \hat{m}(X_i)).$$

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 via showing the estimator is a so-called asymptotically linear estimator, and thus can be written as:

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

or the standardized difference of the estimate and truth can be written as an i.i.d. sum of a random variable called the influence curve, $IC(O_i; \hat{m})$ plus a second order term. Then, the asymptotic variance of the estimated risk is:

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

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

In this case, the goal is to develop a test of two competing models for estimating $m(\cdot)$. Comparisons of model fits can cover a wide variety hypotheses. A couple examples on which we focus 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:

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

with null (typically simpler) model $\hat{m}_0(\cdot)$ and alternative $\hat{m}_1(\cdot)$.

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:

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

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

As [5] showed, Ψ_n, under the null will by asymptotically normally distributed with variance 1, given it also provides consistent estimators for the sample variance of Ψ_n. One can derive an equivalent level α confidence interval for this risk difference as:

$$\Psi_n \pm z_{1-\alpha/2} \frac{\sqrt{var_n[\hat{IC}(O;\hat{m}_0) - \hat{IC}(O;\hat{m}_1)]}}{n}. \tag{5}$$

where $z_1-\alpha/2$ is the $1-\alpha/2$ quantile of the standard normal distribution.

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 [9]. The only requirement is that the loss function is convex which excludes missclassification loss [5].

Before we turn to the specific implementation using cross-validation, we first discuss using the SL algorithm/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.

3 Optimal Estimation of Semi-parametric Models

As suggested in the outset, the utility of such a test is in the scenario where one does not know the true functions m_1 and m_0, and instead needs to estimate them adaptively. When m_1 and m_0 are known and can be estimated using a parametric model, a likelihood ratio test or information based criterion will be optimal and sufficient. However, in the more common scenario where they are unknown, a range of *machine learning* algorithms are available, many of which are semi-parametric.

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 driving 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 [8] we define the gold standard as the so-called Oracle Selector, defined as, given the data, the algorithm that chooses the estimator with the lowest true risk among all tried competitors. Let the risk of the Oracle Selector be defined 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, j, or

$$\hat{m}^* = \underset{j}{argmax}\Psi_j$$

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 [8] *stacking* procedures (particularly those that combine a wide variety of algorithms from very simple/smooth to highly data-adaptive), as implemented in the SuperLearner algorithm, meet this criterion. In stacking one uses a cross-validation procedure to combine a user-specified set of candidate prediction algorithms. The SL algorithm is available as a statistical package [10] in the R programming language. As [8] showed, the SL performs asymptotically equivalently (w.r.t. expected risk difference), up to a second order term, as the Oracle Selector. In addition, the 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. If one of the candidate learners is actually the true model, however, and thus converges at a parametric rate, the SL will converge at an close to parametric rate, implying there is not much cost for estimating under a much bigger model. Thus the SL theory encourages 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 both purposes: construction of the $\hat{m}_0, \hat{m}_1$ models, and subsequent estimation of the cross-validated risk, and the calculation of the its sampling variability. As proposed in [5], we use V-fold cross-validation, where the data are divided into $v = 1, \ldots, V$ equal-sized *testing* sets. For each independent test set, v, the prediction models, $(\hat{m}_0, \hat{m}_1)$, are fit (*trained*) on the corresponding training set and then the test statistic (4), is constructed on the test set. Thus, we fit both $\hat{m}_0^{-v}$, and $\hat{m}_1^{-v}$ within each cross-validation fold and then apply this predictions to the corresponding

testing data set to get the test-statistic of interest, or:

$$T_v = \frac{\sqrt{n_V}\Psi^v_{n_V}}{\sqrt{var^v_{n_V}[\hat{IC}(O^v;\hat{m}_0^{-v}) - \hat{IC}(O^v;\hat{m}_1^{-v})]}}, \tag{6}$$

where it is indexed to emphasize that the relevant predictor models are fit on the training data, but the difference in the risk estimates, $\Psi^v_{n_V}$ done on the testing data, which has sample size $n_V = n/V$. We first discuss a procedure, where the CLT 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 $v = 1, ..., V$ test statistics, T^v_n, we can derive a corresponding p-value of the null as $p_v = Q_0(T^v_n) = 1 - \Phi(T^v_n)$, 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 (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 CI's across folds, giving global coverage $1 - \alpha$ of

$$\begin{aligned} &\Psi^v_{n_V} \pm z_{1-\alpha/2/V} SE(\Psi^v_{n_V}) \\ \text{where } SE(\Psi^v_{n_V}) =& \sqrt{var^v_{n_V}[\hat{IC}(O^v;\hat{m}_0^{-v}) - \hat{IC}(O^v;\hat{m}_1^{-v})]/n_V} \end{aligned} \tag{7}$$

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. Of course, in practice it makes the most sense to use the cross-validated estimated of the risk, as this will coverage 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:

$$\begin{aligned}\Psi_n &= \frac{1}{V}\sum_{v=1}^{V}\Psi_{n_V}^v \\ &= \frac{1}{V}\sum_{v=1}^{V}E(L[Y,\hat{m}_0^{-v}(X)] - L[Y,\hat{m}_1^{-v}(X)])\end{aligned} \tag{8}$$

As [5] showed, under conditions in theorem 3, this estimate is asymptotically normally distributed with variance consistently estimated by:

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

where $\sigma_{v,n}^2 = var_{n_V}^v[\hat{IC}(O^v;\hat{m}_0^{-v}) - \hat{IC}(O^v;\hat{m}_1^{-v})]/n_V$, or the estimated variance of the risk difference within each validation fold. Thus, one can derive a Wald-type confidence interval and test statistic just as above, e.g.,

$$T = \frac{\Psi_n}{\sigma_n^2}.$$

Thus, 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, the competing procedures are such that, for some P_0, $\hat{m}_1^{-v}(X) \to \hat{m}_0^{-v}(X)$, as sample size gets large, then the asymptotic linearity may not hold. However, in these cases, it is because the null is "too" true, and thus we have proposed a simple solution to this degenerate case where the data-adaptive choice for deriving $\hat{m}_1$ gets very close to the null procedure $\hat{m}_0$. Thus, the situation hurts the asymptotics, but not in a substantive way that hurts the inferences from the estimate of the conditional risk difference, since the evidence so strongly points to the null.

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, it still begs the question of the finite sample performance of this procedure as a function

of 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 the true model, 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. Obviously, for this latter reason, V must be small enough to invoke asymptotic normality of T_v. In practice we have found a validation sample size of at least 30 to be necessary.

4.4 "Marginal" Risk Differences

One might want to derive a test of competing procedures, where one does not condition on the estimates of the predictors, but estimates how the competing procedures (algorithms) do in repeated samples. So, as opposed to average conditional risk, the parameter of interest is the average unconditional risk difference, but where the experiment involves not just the estimation of risk from a fixed prediction model, but where it 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 [6]. 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 $\hat{m}(X_i)$ is biased by the other two values. In later work the the authors estimated the maximum between block variance as 0.7 and suggested a t-statistic with a correction based on this value [11]. In our own previous work we suggested a Wald test that also required a correction to maintain proper error control [12]. [13] presented a method of moments estimator that while nearly unbiased, depends on the distribution errors and knowledge of the learning algorithm.

Therefore the work of [5] is important in that it avoids 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) to examine the type I error rate when the base model provides a better fit, and 3) to examine 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) To derive the "true" risk for each of these 2*V predictors, we draw a very large sample using the same distribution, representing a target population; this is used to calculate the true risks, or $\theta(\hat{m}_0^{-v}), \theta(\hat{m}_1^{-v})$ for each v, the corresponding risk differences, $\Psi\{\hat{m}_0^{-v}), \theta(\hat{m}_1^{-v})\}$, the average risks (e.g., $\bar{m}_0 = \frac{1}{V}\sum_{v=1}^{V}\theta(\hat{m}_0^{-v})$), and the difference of these average risk differences;

(6) To estimate the sample distribution and performance repeat (1) - (4) 1000 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 model contain $x1$ and $x2$ where $x2$ has an association ($b = 1$) with Y. The alternative model adds $x3$ which is 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:* The base model contain $x1$ and $x2$ where $x1$ has an association ($b = 0.5$) with Y. The alternative model adds $x3$ which *is* also associated Y, i.e. a true association. Y is categorical and both models are estimated using basic logistic regression and the risk is calculated under absolute-error loss;

- *Simulation III:* The base model was the same as in simulation I and was estimated using linear regression. The alternative model used the same X and was estimated under SL using 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 models contain only $x1$ with the true $E(Y \mid X)$ shown in Figure 1, based on a piecewise constant model. The base fit is miss-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.

1000 simulations were performed. We report bias (estimated risk difference vs "true" risk difference, type 1 error (based on a 1 tailed test), and coverage probability.

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 achieves 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 miss-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

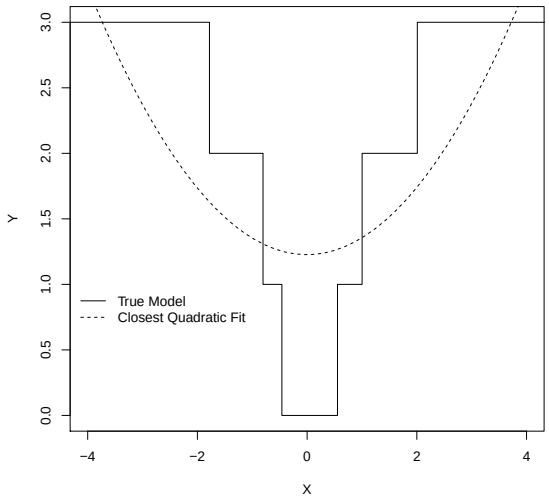

Figure 1: True model $E(Y|X)$ and closest quadratic approximation - i.e., the limit of the null estimator used in *Simulation 4*

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 noted by [6] there are additional correlation components embedded in V-fold CV. These correlation components likely add second order terms to the IC calculation (equation (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 coverage and power. This is the major advantage of performing CV as opposed to a simple sample split.

Sim	Sample Size	Fold	True Risk Difference	Estimated Risk Difference	Bias	SE	Coverage	Rejection
I	100	1	-0.003	-0.004	0.001	0.010	0.939	0.047
		2	-0.003	-0.004	0.001	0.010	0.950	0.036
		3	-0.003	-0.004	< 0.001	0.010	0.956	0.032
		4	-0.003	-0.004	< 0.001	0.010	0.948	0.029
		5	-0.003	-0.003	> -0.001	0.010	0.941	0.046
		Ave Over Folds	-0.003	-0.004	< 0.001	0.005	0.872	0.003
		Bonferonni Corrected	-0.003	-0.004	< 0.001	0.010	0.937	0.030
	1000	1	> -0.001	> -0.001	> -0.001	0.001	0.941	0.038
		2	> -0.001	> -0.001	> -0.001	0.001	0.955	0.031
		3	> -0.001	> -0.001	< 0.001	0.001	0.950	0.037
		4	> -0.001	> -0.001	> -0.001	0.001	0.928	0.050
		5	> -0.001	> -0.001	< 0.001	0.001	0.938	0.043
		Ave Over Folds	> -0.001	> -0.001	< 0.001	0.001	0.834	< 0.001
		Bonferonni Corrected	> -0.001	> -0.001	< 0.001	0.001	0.943	0.039
II	100	1	0.070	0.069	< 0.001	0.040	0.932	0.551
		2	0.070	0.069	0.001	0.040	0.933	0.539
		3	0.070	0.071	-0.001	0.040	0.942	0.542
		4	0.070	0.072	-0.002	0.040	0.938	0.561
		5	0.070	0.070	< 0.001	0.040	0.939	0.567
		Ave Over Folds	0.070	0.070	< 0.001	0.018	0.934	0.962
		Bonferonni Corrected	0.070	0.070	> -0.001	0.031	0.993	0.868
	1000	1	0.069	0.071	-0.001	0.012	0.954	1.000
		2	0.069	0.070	-0.001	0.012	0.950	1.000
		3	0.069	0.070	-0.001	0.012	0.960	1.000
		4	0.069	0.071	-0.001	0.012	0.953	1.000
		5	0.069	0.070	-0.001	0.012	0.942	1.000
		Ave Over Folds	0.069	0.070	-0.001	0.006	0.942	1.000
		Bonferonni Corrected	0.069	0.070	-0.001	0.012	1.000	1.000

Table 1: Simulation results for comparing the added benefit of a predictor. In simulation I the the additional predictor does not add information. In simulation II the added predictor is associated.

Sim	Sample Size	Fold	True Risk Difference	Estimated Risk Difference		Bias	SE	Coverage	Rejection
III	100	1	-8.249	-0.005		-8.244	0.015	0.934	0.040
		2	-0.005	-0.005		$<$ 0.001	0.014	0.935	0.035
		3	-0.005	-0.005		$>$ -0.001	0.014	0.947	0.036
		4	-0.005	-0.005		$<$ 0.001	0.015	0.949	0.032
		5	-0.005	-0.006		$<$ 0.001	0.015	0.943	0.032
		Ave Over Folds	-1.654	-0.005		-1.649	0.007	0.820	0.009
		Bonferonni Corrected	-1.654	-0.005		-1.649	0.015	0.981	0.036
	1000	1	$>$ -0.001	$>$ -0.001		$>$ -0.001	0.001	0.954	0.030
		2	$>$ -0.001	$>$ -0.001		$<$ 0.001	0.001	0.944	0.032
		3	$>$ -0.001	$>$ -0.001		$>$ -0.001	0.001	0.951	0.022
		4	$>$ -0.001	$>$ -0.001		$>$ -0.001	0.001	0.945	0.031
		5	$>$ -0.001	$>$ -0.001		$<$ 0.001	0.001	0.951	0.028
		Ave Over Folds	$>$ -0.001	$>$ -0.001		$<$ 0.001	0.001	0.828	0.007
		Bonferonni Corrected	$>$ -0.001	$>$ -0.001		$<$ 0.001	0.001	0.998	0.019
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
		Ave 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
	1000	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
		Ave Over Folds	1.153	1.148		0.006	0.047	0.941	1.000
		Bonferonni Corrected	1.153	1.148		0.006	0.105	1.000	1.000

Table 2: Simulation results for comparing across two different fitting procedures. In simulation III the base model provides the parametric fit. In simulation IV the more complex should provide a better fit.

6 Data Analysis

Gene based tests represent a unique application of this procedure. Genes are comprised of individual bases of DNA referred to as single nucleotide polymorphisms (SNPs). A typical gene may 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. Gene based tests attempt to associate the set of SNPs comprising the gene and many different procedures have been proposed [14]. Of particular relevance, the unit of interest in a gene based is not any individual SNP but instead the collection of SNPs. Therefore we are less interested in defining a specific parametric model for the relationship between the SNPs and the outcome of interest.

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) [15]. 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.

In brief the data consist of a candidate gene study on 2,722 people. In the dataset the stress response pathway consists of 409 SNPs comprising 10 genes 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 [16] method to control the False Discovery Rate ($p_{adjusted} = 0.42$).

Since our interest was only among the 10 genes (and not the 409 SNPs), we tested each gene individually. A SuperLearner was fit using a library consisting of RandomForests, LASSO, GLM and K-Nearest Neighbours 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.

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

Table 3: P-values for the overall stress response pathway as well as each individual gene. The overall pathway shows a significant association as does the CRHR1 gene.

The overall pathway had a significant association ($p < 0.012$) as did the the CRHR1 gene. After adjusting the multiple testing, the association was marginal for each test ($p_{adjusted} < 0.061$). To examine whether all of the association resided in the CRHR1 gene we compared the overall pathway to just the CRHR1. Not surprisingly there was no association (p = 1.0) confirming that CRHR1 is the only gene in the stress response pathway associated with MS. The use of the other loss functions resulted in similar conclusions. Overall, this formally confirms the conclusions in the original paper that we were only able to make by suggestion [15].

7 Conclusion

In this paper we have implemented an estimation and inferential procedure based on the theory developed in [5] 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 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 draw inference about the added predictive value from cross-validation. Previous work, both theoretical and empirical, has shown, that while cross-validation produces an unbiased estimate of the risk, the variance estimate is improper.

The test has 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 for such assessment under 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. [17] 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. While, intuitive, as our data analysis shows, often no individual predictor will be associated either because of issues of multiple testing or because a parametric model cannot be adequately specified. In this case, semi-parametric methods, like SuperLearner, become more valuable. Moreover, in this 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. We also have limited this work to hypothesis testing and have not provided a means to estimate a confidence interval for a cross-validated risk, also a potentially interesting quantity.

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

References

[1] Harrell FE, Califf RM, Pryor DB, Lee KL, RA RAR. Evaluating the yield of medical tests. *JAMA* 1982; **247**:2543–2546.

[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; **27**(2):157–172.

[3] Hastie T, Tibshirani R, Friedman J. *Elements of Statistical Learning*. 2 edn., Springer: New York, 2009.

[4] van der Laan MJ, Dudoit S, Keles S. Asymptotic optimality of likelihood-based cross-validation. *Stat Appl Genet Mol Biol* 2004; **3**:Article4.

[5] Dudoit S, van der Laan MJ. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology* 2005; **2**:131–154.

[6] Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research* 2004; **5**:1089–1105.

[7] Chaffee P, Hubbard AE, van der Laan ML. Permutation-based pathway testing using the super learner algorithm. *Technical Report 263*, U.C. Berkeley Division of Biostatistics March 2010.

[8] van der Laan MJ, Polley EC, Hubbard AE. Superlearner. *Statistical Applications in Genetics & Molecular Biology* 2007; **6**.

[9] Takenouchi T, Komori O, Eguchi S. An extension of the receiver operating characteristic curve and AUC-optimal classification. *Neural Comput* Oct 2012; **24**(10):2789–2824.

[10] Polley E, van der Laan M. *SuperLearner: Super Learner Prediction* 2012. URL `http://CRAN.R-project.org/package=SuperLearner`, r package version 2.0-6.

[11] Grandvalet Y, Bengio Y. Hypothesis testing for cross-validation. *Technical Report 1285*, Departement dInformatique et Recherche Operationnelle August 2006.

[12] Goldstein B, Hubbard A, Barcellos L. A generalized approach for testing the association of a set of predictors with an outcome: A gene based test. *Technical Report 274*, U.C. Berkeley Division of Biostatistics Working Paper Series January 2011.

[13] Markatou M, Tian H, Biswas S, Hripcsak G. Analysis of variance of cross-validation estimators of the generalization error. *J. Mach. Learn. Res.* 2005; **6**:1127–1168.

[14] Beyene J, Tritchler D, Asimit JL, Hamid J. Gene- or region-based analysis of genome-wide association studies. *Genetic Epidemiology* 2009; **33**:s105–s110.

[15] Briggs FB, Bartlett SE, Goldstein BA, Wang J, McCauley JL, Zuvich RL, De Jager PL, Rioux JD, Ivinson AJ, Compston A, *et al.*. Evidence for CRHR1 in multiple sclerosis using supervised machine learning and meta-analysis in 12,566 individuals. *Hum. Mol. Genet.* Nov 2010; **19**(21):4286–4295.

[16] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* 1995; **57**:289–300.

[17] Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. *Stat Med* Jan 2013; .