

## MIT Open Access Articles

*Evaluating discrete choice prediction models when the evaluation data is corrupted: analytic results and bias corrections for the area under the ROC*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Stein, Roger M. "Evaluating Discrete Choice Prediction Models When the Evaluation Data Is Corrupted: Analytic Results and Bias Corrections for the Area under the ROC." Data Mining and Knowledge Discovery 30.4 (2016): 763–796.

**As Published:** <http://dx.doi.org/10.1007/s10618-015-0437-7>

**Publisher:** Springer US

**Persistent URL:** <http://hdl.handle.net/1721.1/106979>

**Version:** Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

**Terms of Use:** Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



# Evaluating discrete choice prediction models when the evaluation data is corrupted: Analytic results and bias corrections for the area under the ROC

Roger M. Stein

Current draft: August 27, 2015

**Abstract** There has been a growing recognition that issues of data quality, which are routine in practice, can materially affect the assessment of learned model performance. In this paper, we develop some analytic results that are useful in sizing the biases associated with tests of discriminatory model power when these are performed using corrupt (“noisy”) data. As it is sometimes unavoidable to test models with data that are known to be corrupt, we also provide some guidance on interpreting results of such tests. In some cases, with appropriate knowledge of the corruption mechanism, the true values of the performance statistics such as the area under the ROC curve may be recovered (in expectation), even when the underlying data have been corrupted. We also provide estimators of the standard errors of such recovered performance statistics. An analysis of the estimators reveals interesting behavior including the observation that “noisy” data does not “cancel out” across models even when the same corrupt data set is used to test multiple candidate models. Because our results are analytic, they may be applied in a broad range of settings and this can be done without the need for simulation.

**Mathematics Subject Classification (2010)** 62-07 · 62G10

**Keywords** ROC · AUC · model validation · model evaluation · prediction · data corruption · bias correction · missing data · misclassification · discrete choice models · credit models · machine learning

---

I am grateful to Sanjiv Das, David Fagnan, Lisa Goldberg and Mitchell Petersen for detailed comments on earlier drafts of this paper. I am particularly grateful to Foster Provost who provided extensive and detailed suggestions on improving the exposition and extending the results – including suggesting the idea of a recovered ROC. This article was greatly improved by the observations and suggestions of three anonymous reviewers. All errors are, of course, my own. The views expressed in this article are those of the author and do not represent the views of former employers or any of their affiliates.

---

MIT Laboratory for Financial Engineering, 100 Main Street, Cambridge, MA 02142. E-mail: [steinr@mit.edu](mailto:steinr@mit.edu)

## 1 Introduction

In the past two decades, a number of standard approaches have been developed and come into common use to test the accuracy of discrete choice prediction models. At the same time, there has been a growing recognition that issues of data quality can materially affect the assessment of model performance. There now exist many texts and articles, in both the academic and professional literatures, that deal with how to specify and estimate discrete choice models, even in the presence of noisy data. However, there are relatively fewer that deal in any detail with model *evaluation*, and fewer still that deal with issues of data corruption or “noise” in model testing.<sup>1</sup> Data corruption is prevalent in practice. For example it may occur when analyzing combined transaction data from firms that have been party to acquisitions or mergers.

Recently, Russell et al. [2012] reported the results of a series of simulation experiments that sought to demonstrate the impact of data corruption on tests of the power and calibration of credit default prediction models. (In the context of this paper, *power* refers to the degree to which a discrete choice model differentiates between one or more classes, e.g., defaulting firms and non-defaulting firms, and includes measures such as the area under the ROC curve; while *calibration*, which we do not consider in detail in this article, refers to the degree to which the probabilities produced by a model match observed probabilities, e.g., whether the *ex ante* conditional probability of default produced by a model matches the observed *ex post* default frequency of the loans in a bank’s portfolio). The experiments reported in Russell et al. [2012] examined differences in performance statistics calculated using artificially generated data that were pristine versus using that same data set after it had been corrupted in some way. The main results involved descriptive guidance on the direction of the bias for certain performance statistics resulting from certain corruption schemes.

While such experiments are informative, the range of cases that the authors test in their experiments is not easily generalizable to the broader range of situations encountered in practice. The results also offer little in the way of normative guidance (beyond the admonition to use uncorrupted data for evaluation). In this paper, we attempt to provide some more general guidance by developing analytic estimators that may be used even when data are corrupted. We will illustrate key concepts through a series of examples relating to the evaluation of bankruptcy prediction models.

This paper concerns the impact of *data corruption* – in this case either the omission from the evaluation database of some records or the mislabeling

---

<sup>1</sup> This is not universally the case. For example, Chapter 7 of Bohn and Stein [2009] deals extensively with model evaluation and data issues are also discussed in Chapters 3, 4, 6 and 9.

of the outcome (e.g., default or no-default) for some records. In particular, we focus on the impact of such corruption on tests of model *power*. Power measures the degree to which a discrete choice model (e.g., a bankruptcy model) differentiates between one or more classes based on relative score. As an example, a bankruptcy model with high power would be one that produced scores for bankrupt firms that were much worse than those it produced for non-bankrupt firms. A lower power model would not differentiate as much between these two classes, making it less useful for screening firms.

We derive analytic estimators for power statistics when data are corrupted. These estimators can be used to allow researchers to make use of corrupted data to assess model quality. We also provide some analysis of how data corruption affects estimates of model power in a number of realistic settings such as those in which data are corrupted due to the integration of incompatible database management systems, conversion of legacy files from paper or obsolete data formats, poor record keeping, and so forth.

We find that data corruption impacts measures of model power in both obvious and sometimes less obvious ways. For example, if minority class data are *missing* from the evaluation sample, this will have little impact on the expected value of statistics that measure power. In contrast, *misabeled* data do affect such measures, though to different degrees. An important result of our analysis is that mislabeling errors *do not* “cancel out” across models, contrary to the beliefs of some practitioners.

It is not uncommon for practitioners to ignore data corruption due to a (mistaken) belief that testing a set of candidate models on the same data exposes all models to the same data problems, thereby neither favoring nor disadvantaging any one model more than any other. Our results show that this is a misapprehension: mislabeling errors impact higher power models more severely than weaker ones, moving estimates of the power of better models closer to those of poorer ones. As we noted earlier, and despite the widespread assumption by many industry participants, this implies that it is not the case the same (noisy) data set handicaps all competing models similarly.

## 1.1 Data corruption and inference

Throughout this paper, we will focus on the impact on model evaluation of two types of data corruption: those that arise due to missing records and those that arise due to mislabeled records.<sup>2</sup> In both cases, we assume that the data

---

<sup>2</sup> Because our focus is on evaluation rather than estimation (learning) we assume that a model to be evaluated has already been estimated and that the task at hand is to evaluate this model using the available data.

are either missing at random (MAR) in the usual sense [cf., Heitjan and Rubin, 1991] or that the mislabeling mechanism is equivalent to MAR conceptually. *Missing at random* implies that likelihood of an observation being missing is not related to the value of the observation, though it may be related to other factors.

Because we will primarily be discussing *corrupted* rather than missing data, we use the term *corrupted at random* (CAR) to connote a similar notion, albeit with a slightly different mechanism. If the class indicator (e.g., a flag indicating whether a firm defaulted or did not) is changed, we assume that this is not related to the model prediction or score for that record. Thus, *that a record is corrupted provides no additional information about whether the model would have correctly predicted the true class or not.*<sup>3</sup>

More formally, if  $s_i^m$  is the score produced by model  $m$  for record  $i$ ,  $f_s^m(x)$  is the distribution function for the scores produced by that model, and  $\kappa_i \in \{0, 1\}$  is an indicator of whether the  $i^{th}$  record has been corrupted, the CAR assumption implies

$$f_s^m(s_i | \kappa_i = 1) = f_s^m(s_i), \forall s_i.$$

We also assume that the positive class is the minority class and that the distribution of outcomes is heavily skewed so that there are many more negative instances in the data set than positive instances, though this is not required for our results to hold. (For example, a bank might wish to test a loan default model using a dataset in which on the order of 1% of the borrowers default. In this case, the minority class would be the positive class, and the class distribution would heavily skewed.) For convenience, we often use the term “bad” to refer to the minority class, as these rare instances are typically associated with abnormal outcomes that require some form of intervention or attention (e.g., bankruptcy, diseased status, quality failure, fraud, etc.), and we use the term “good” to refer to the majority class of normal outcomes.

**Example 1 (Credit records missing from evaluation database)** An example of a *missing* record in credit analysis would be the record of a defaulting firm to which a bank had made a loan that has been dropped from the database of financial statements. Records may become missing during the default resolution process since it is not uncommon for a defaulted firm’s information to be moved from the lending or monitoring department in a bank, which collects financial statement data and other information on the borrower, to a specialized “workout” group that deals only with troubled loans and which is charged with recovering the loan principal through liquidation and other means.

■

---

<sup>3</sup> This assumption represents a key area for future research. See the discussion in Section 4.3.

*Example 2 (Credit records mislabeled in evaluation database)* An example of a *mislabeled* record, again from the credit analysis domain, is one that occurs commonly in matching a firm's financial statements to the firm's delinquency status. It is typical to join financial statement records in one database to performance data in another database that contains information on the loan status (current, delinquent, etc.). For example, in the case of commercial lending, it is not uncommon for the details of loan performance to be stored in a "servicing" database that is separate from the database containing details of the financial status of the borrower. Because it can be difficult to merge such databases, it is not unusual for some records to be unmatched or mismatched, particularly if a bank has merged with or acquired another bank. Thus, the default indicator associated with the financial records of a defaulted firm may erroneously indicate "no default." This phenomenon is sometimes called the "hidden default" problem in industry.

An example of a case in which a non-defaulted firm's records may have an indicator inadvertently labeled as a default would be cases in which two databases are merged but in which the default indicators were defined differently. For example, one definition may consider a single missed interest payment to be an event of default, while another may only set the flag if a borrower is more than 120 or 180 days past due.<sup>4</sup>

■

It is important to note that if the mechanism that causes the mislabeling is known with a high level of specificity (e.g., record #35 was mislabeled), then it is trivial to address. However, if the mechanism is even a little bit less transparent (e.g., it is known that 3 in 100 records in the database are mislabeled, but it is not known which records are erroneous), it is no longer trivial to unwind the corruption when calculating statistics that depend on comparing the model outputs to the corrupted fields. It is this latter case with which this paper deals.

Because real-world data sets are often imperfect, it is useful to have a robust set of tools to estimate the impact of corrupted data on evaluation statistics. A natural alternative to simulation-based trial and error is to derive analytic results for the bias caused to performance statistics in the presence of data noise. For some measures, this is challenging. However, for many of the most commonly used measures of model performance (e.g., measures of power such as the area under the ROC curve), such analytic results can be derived analytically through reference to the underlying probability models on which the statistics are based.

In this paper, we develop some analytic results that are useful in sizing the biases associated with tests of discriminatory power done using corrupt ("noisy") data. Because our results are analytic rather than computational, we are also able to invert them to provide recommendations on how to adjust observed performance statistics if we suspect that the data used to calculate

---

<sup>4</sup> Note that depending on the circumstances, this may violate the CAR assumption. See Section 4.3.

the statistics are corrupt. In this way, *even corrupted data may be used to obtain performance estimates* in many cases. Alternatively, when we have less knowledge of the corruption mechanism, the analytic results can be used to perform sensitivity analysis on the observed statistics to determine a range of likely values, given the data are known to be corrupted.

Our analytic results may be applied in a broad range of settings and we demonstrate this for a variety of parameter values. We also show, in an appendix, how a corrected ROC curve may be generated using the results of our bias corrections for the AUC.

The remainder of this paper is organized as follows. In Section 2 we examine the impact of imperfect data on measures of model power. We decompose the area under the ROC curve (AUC) into individual components and then demonstrate how the AUC changes as we introduce noise in one form or another to the various components. We also demonstrate how to recover an estimate of the true AUC even when the observed data are corrupted. In addition, we provide variance estimators for AUCs calculated on corrupted data. Section 3 demonstrates, through simulation, that the analytic results align well with their simulated counterparts. Though simulation is not required to apply our analytic results, the experiments serve to confirm our conclusions, and provide insight into the coverage of confidence intervals derived from the analytic estimates of variance. Section 4 discusses a number of implications and also reviews some related work from the biostatistics literature. We also provide suggestions on how the levels of data noise may be estimated if they are not known *a priori*. Finally, the paper contains two appendices: The first appendix discusses some of the literature on data corruption in the development sample, rather than the evaluation sample. Though these results are largely outside of the scope of this paper, they are of interest nonetheless and provide some context. A second appendix demonstrates how our results may be used to recover ROC curves in the presence of noisy data.

## 2 Impact of missing and mislabeled data on estimates of the area under the ROC

We begin by reviewing a basic measure of model power which describes a model’s ability to discriminate between positive and negative instances: the area under the *receiver operating characteristic* (ROC) curve [Peterson et al., 1954], often denoted  $A$  and sometimes referred to as the *AUC*. A related measure that is often used in the finance industry is the *accuracy ratio* or *AR* [Sobehart et al., 2000]. Both statistics measure the degree to which a predictive model assigns high scores to positive instances (“bads” that require attention) and low scores to negative ones (“goods”).

The accuracy ratio,  $AR$ , is equivalent to the Gini coefficient, and both are related to the area under the curve,  $A$ , through the identity [Engelmann et al., 2003, Hand and Till, 2001]:

$$A = \frac{AR}{2} + 0.5,$$

or, equivalently

$$AR = 2A - 1,$$

which implies that analytic results for  $A$  can be related to results for  $AR$  trivially. (For the remainder of this paper, we focus our analysis on  $A$ .)

Recall that the ROC curve generalizes the notion of a contingency table, where the generalization is done over all possible cut-off scores. The ROC plots, for every cut-off  $k$ , the false positive rate against the true positive rate for a particular classification model when instances are classified as belonging to one class for scores above  $k$  and the other for scores below  $k$ .  $A$  is simply the area under the curve, with a larger area denoting better discrimination between two classes based on the specific scoring scheme. Figure 6 in Appendix B provides an example of an ROC curve.

It can be shown [Bamber, 1975, Hanley and McNeil, 1982] that when  $A$  is calculated empirically using the trapezoid rule (one standard approach),  $A$  is equivalent to the Wilcoxon-Mann-Whitney statistic. This observation allows us to draw on a rich set of results from the literature on  $U$ -statistics [Hoeffding, 1948] in our analysis of the AUC.<sup>5</sup>

A formulation of  $A$  that is useful for our purposes is a generalization of Bamber [1975] due to DeLong et al. [1988]:

$$\hat{A} = \frac{1}{mn} \sum_j^n \sum_i^m I(G_j < B_i) \quad (1)$$

where  $G_j$ ,  $j = 1 \dots n$ , is the model output for the  $j^{th}$  majority class instance in a dataset (“good”),  $B_i$ ,  $i = 1 \dots m$  is the model output for the  $i^{th}$  minority class instance (“bad”) and  $I(c)$  is an indicator function that takes the value of 1 if condition  $c$  is true and 0 if condition  $c$  is false. (For simplicity, we assume no tied values, though relaxing this assumption is not difficult).

---

<sup>5</sup> Note that the AUC is a summary statistic. As such, there are many cases in which important information is lost in summarization. In particular, if the ROC curves for two models cross, the AUC may not provide information on the best model for a specific application. In fact, depending on the specific application, it may be possible to arrive at a higher AUC using the combined models than either can achieve on its own [Provost and Fawcett, 2001].



We now derive results on the impact of various types of data corruption on tests of model power.

### 2.1 Impact on estimates of model power when positives are missing in the evaluation sample

In the case in which some number of positive instances are *omitted* from the evaluation sample (e.g., defaulted firms were omitted from the database because upon default their files were transferred to a “workout” group, but some records were lost), we can determine the impact on  $A$  directly from (1). The new corrupted AUC,  $A_c$ , is given as:

$$\hat{A}_c = \frac{1}{n(m-k)} \sum_j^n \sum_i^{m-k} I(G_j < B_i) \quad (2)$$

From the relationship between the AUC and Mann-Whitney statistic we have  $E[I(G_j < B_i)] = \hat{P}(G_j < B_i) = \hat{A}$ . Substituting and collapsing the summations yields:

$$E[\hat{A}_c] = \frac{1}{n(m-k)} \left[ n(m-k) E[\hat{A}] \right] = E[\hat{A}],$$

implying that random omission of minority class records should not affect the expected value of  $A$ , though model calibration would clearly be affected. (A similar result can be obtained for missing majority class instances.) This result is anticipated since the expected value of the AUC is well known to be insensitive to changes in class proportion, which is how deletion of either class impacts the data set.

Furthermore, there is no explicit adjustment required for the variance ( $\text{Var}(\hat{A}_c)$ ) since the variance of the AUC already explicitly scales in the number of minority and majority instances [Bamber, 1975]. Although no additional adjustment is required, the variance of  $\hat{A}$  is not the same as the variance of  $\hat{A}_c$ . When there are missing positives,  $\text{Var}(\hat{A}_c)$  will be higher than that of  $\text{Var}(\hat{A})$  (the “true” variance of  $A$ ), particularly due to the sensitivity of the variance estimator of  $A$  to the number of observations in the minority class in the evaluation data [cf., Stein, 2007].

## 2.2 Impact on estimates of model power when records are mislabeled in the evaluation sample

In the case in which some portion of positive instances are *mislabeled* (e.g., because of inaccurate record keeping or faulty merging), we can again determine the impact on  $A$  directly from (1), due to the relationship between the AUC and the Mann-Whitney statistic.

Without loss of generality, in the remainder of what follows, we assume that if  $l$  goods are corrupted, then the records labeled “good” are ordered such that the first  $n - l$  are those that are not corrupted and the remaining  $l$  are those that are corrupted. Similarly, we assume that if  $k$  bads are corrupted, then the first  $m - k$  of the records labeled “bad” are uncorrupted with the remaining  $k$  being corrupted. Finally, if  $l$  goods are corrupted and  $k$  bads are corrupted, then for the records in the database labeled “good” the first  $n - l$  are true goods while records  $n - l + 1$  through  $n + k$  are actually mislabeled true bads. Similarly, for records labeled “bad” in the database, the first  $m - k$  are true bads, while records  $m - k + 1$  through  $m - k + l$  are actually mislabeled true goods.

Assume first that  $k$  of the positive records in the evaluation data set have had their status changed from “bad” to “good.” The new estimate for AUC using the corrupted data ( $\hat{A}_c$ ) is:

$$\begin{aligned}\hat{A}_c &= \frac{1}{(m-k)(n+k)} \sum_j^{n+k} \sum_i^{m-k} I(G_j < B_i) \\ &= \frac{1}{(m-k)(n+k)} \left[ \sum_j^n \sum_i^{m-k} I(G_j < B_i) + \sum_{j=n+1}^{n+k} \sum_i^{m-k} I(G_j < B_i) \right]\end{aligned}$$

Note that for  $j > n$ , each  $G_j$  is actually drawn from the set of true “bad” records so the estimate of  $\hat{A}_c$  can be written:

$$\hat{A}_c = \frac{1}{(m-k)(n+k)} \left[ n(m-k) \hat{A} + k(m-k) \hat{A}_0 \right],$$

or, after canceling terms:

$$\hat{A}_c = \frac{n\hat{A} + k\hat{A}_0}{(n+k)}, \quad (3)$$

where  $\hat{A}_0$  is the estimated value of the AUC for a model with no discriminatory power.

By inspection,  $E[\hat{A}_c] < E[\hat{A}]$  as long as  $A > 0.5$  (i.e., the model has positive predictive power) and  $k > 0$  (i.e., there is at least one mislabeled positive instance). Intuitively, the effect of mislabeling to “mix” the true value of  $\hat{A}$  with the value associated with a random model.

If instead,  $l$  of the “goods” are mislabeled, then a similar analysis gives:

$$\hat{A}_c = \frac{m\hat{A} + l\hat{A}_0}{(m+l)}. \quad (4)$$

(Note that in the case that  $k$  or  $l$ , respectively, equal zero, (3) or (4) reduce to  $\hat{A}$ .)

Finally, if  $k$  of the “bads” are mislabeled and  $l$  of the “goods” are mislabeled we have:

$$\hat{A}_c = \frac{1}{(m-k+l)(n+k-l)} \sum_j^{n+k-l} \sum_i^{m-k+l} I(G_j > B_i)$$

which can be rewritten as :

$$\hat{A}_c = \frac{Q_1 + Q_2 + Q_3 + Q_4}{(m-k+l)(n+k-l)} \quad (5)$$

where

$$\begin{aligned} Q_1 &= \sum_j^{n-l} \sum_i^{m-k} I(G_j > B_i) \\ Q_2 &= \sum_{j=1}^{n-l} \sum_{i=m-k+1}^{m-k+l} I(G_j > B_i) \\ Q_3 &= \sum_{j=n-l+1}^{n+k-l} \sum_{i=1}^{m-k} I(G_j > B_i) \\ Q_4 &= \sum_{j=n-l+1}^{n+k-l} \sum_{i=m-k+1}^{m-k+l} I(G_j > B_i). \end{aligned}$$

Following the reasoning used in (3) and (4), note that the first term ( $Q_1$ ) in the numerator of (5) can be rewritten as a function  $\hat{A}$  of while the middle two terms ( $Q_2$  and  $Q_3$ ) are analogous to the  $\hat{A}_0$  terms in (3) and (4). The last term in the numerator ( $Q_4$ ) is interesting in that it compares true “bads” against true “goods”, so the model’s power,  $\hat{A}$ , is expressed, but in this case,

because the labels have been switched, the relationships are inverted, and the term becomes a function of the complement of  $\hat{A}$ ,  $(1 - \hat{A})$ . Rewriting this way yields:

$$\hat{A}_c = \frac{[(n-l)(m-k)\hat{A} + l(n-l)\hat{A}_0^{nl} + k(m-k)\hat{A}_0^{mk} + kl(1-\hat{A})]}{(m-k+l)(n+k-l)}, \quad (6)$$

where  $\hat{A}_0^{nl}$  and  $\hat{A}_0^{mk}$  take the place of  $\hat{A}_0$  in (4) and (3), respectively. This explicit designation will become useful in Section 2.6 (see Footnote 9). More formally

$$\begin{aligned} \hat{A}_0^{mk} &= \frac{1}{(m-k)k} \sum_j^{m-k} \sum_{i=m-k+1}^m I(B_j > B_i), \\ \hat{A}_0^{nl} &= \frac{1}{(n-l)l} \sum_j^{n-l} \sum_{i=n-l+1}^n I(G_j > G_i), \end{aligned}$$

and

$$E[\hat{A}_0^{mk}] = E[\hat{A}_0^{nl}] = 0.5, \quad (7)$$

with Equation (7) following from the CAR assumption.

Figure 1 gives some examples of how  $\hat{A}_c$  changes for different values of  $k$  and  $l$ , plugging in constant values for  $\hat{A}$  and setting  $\hat{A}_0^{nl} = E[\hat{A}_0^{nl}] = 0.5$  and  $\hat{A}_0^{mk} = E[\hat{A}_0^{mk}] = 0.5$  for the specific values of  $\hat{A}_0$ , respectively.<sup>6</sup>

From the plots in Figure 1, it is evident that *mislabelings of the minority class impact the calculation of  $\hat{A}$  less severely than do mislabelings of the majority class*. In a sense, this is not surprising as in absolute terms, a percentage change to the minority class affects fewer records than does the same percentage change to the majority class.<sup>7</sup>

<sup>6</sup> For a random model,  $E[A] = 0.5$ , but  $\hat{A}_0$  will almost surely be different than 0.5 due to sampling error. While it is sometimes convenient to drop the expectation notation for  $\hat{A}_0$ ,  $\hat{A}_0$  cannot be calculated practically without knowledge of the specific data corruption process, so we can *only* work with expectations. The same is true of estimates of  $\hat{A}$ , the true, unobserved, value of  $A$ . This introduces practical issues in calculating higher moments such as covariance. (See Section 2.6.)

<sup>7</sup> Note that we show these results in percentage terms (e.g., a percentage of positive observations mislabeled). We do this for convenience, however, the results are similar directionally regardless of whether we measure in percentage or absolute terms. For example, assume that  $A=0.8$ ,  $n=10,000$ , and  $m=2000$ . First examine the case of constant mislabelings if we keep the number of mislabelings fixed at 200. In this case, by (5) we get that mislabelings of

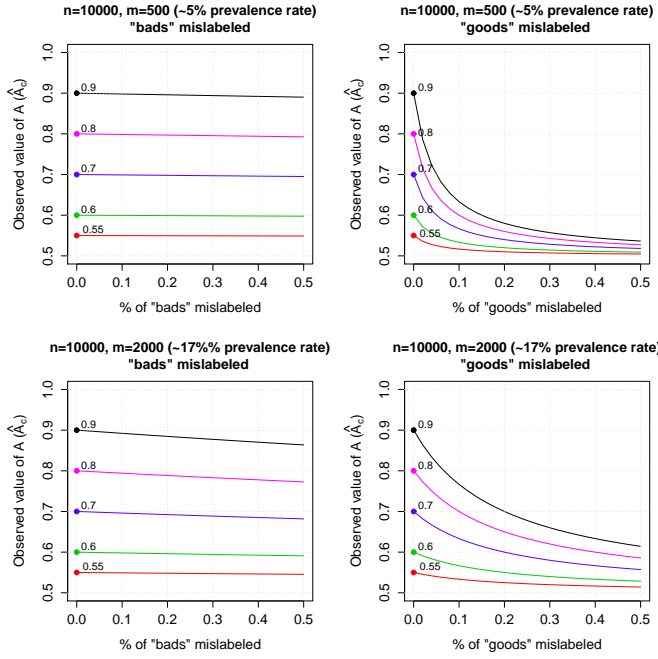


Fig. 1: Changes to  $\hat{A}_c$  for different values of  $k$  and  $l$

The figures show the impact of mislabeling  $k$  (left) or  $l$  (right) “bads” or “goods”, respectively using the analytic estimator of  $\hat{A}_c$  given in (5). The plots in the top row show the impact on a lower prevalence rate sample, while the plots in the bottom row show the impact on a higher prevalence rate sample. The true value of  $A$  is given near the  $y$ -axis in each case. In the plots, each line represents a hypothetical model with the power indicated.

### 2.3 Mislabeling errors do not cancel across models

The plots in Figure 1 also suggest that *corruption affects higher power models more severely than lower power models*. This too is intuitive. Since the lower power models extract less information from the each data record, they lose less information when some records are corrupted.

These observations can be seen analytically as well:

$$\frac{\partial^2 A_c}{\partial k \partial A} = \frac{l}{(m-k+l)^2} + \frac{n-l}{(n+k-l)^2} > 0 \text{ and } \frac{\partial^2 A_c}{\partial l \partial A} = \frac{k}{(n+k-l)^2} + \frac{m-k}{(m-k+l)^2} > 0$$

“bads” ( $k=200, l=0$ ) results in  $A_c = 0.794$ ; in contrast when we mislabel the same number of “goods” ( $k=0, l=200$ ) we get a greater degradation and  $A_c = 0.773$  results. Now we repeat but keep the percentage of mislabelings constant at 10%. In this case, we get that mislabelings of “bads” ( $k=200, l=0$ ) results in  $A_c = 0.794$  as before; mislabeling the same percentage of “goods” ( $k=0, l=1000$ ) again yields a greater degradation as  $A_c = 0.7$ .

Because  $\partial^2 A_c / \partial k \partial A$  and  $\partial^2 A_c / \partial l \partial A$  are both positive everywhere (since by construction  $l \leq n$  and  $k \leq m$ ), as  $A$  increases, the impact of data corruption is more pronounced.

This implies a profound (yet obvious in hindsight) result: mislabeling affects assessments of power differentially and this happens in the least helpful manner. Mislabeling moves the observed estimates of  $A$ ,  $\hat{A}_c$ , for powerful models closer to those of weaker models, blurring the distinction between the two. Said differently, contrary to the beliefs (and aspirations) of some practitioners, it is not the case the same (noisy) data set will handicap competing models similarly. Errors do not “cancel out” across models.

To gain further intuition for this result, consider the two panels in Figure 2. These plots show, schematically, the distribution of model scores for two hypothetical models, one weaker and one more powerful, respectively. In each plot, the distribution of model scores for the majority class is shown as a set of green “+” symbols, while the distribution of the minority class is shown as red “o” symbols. Because the more powerful model (bottom) is able to differentiate better between the two classes, the two distributions are separated to a higher degree than is the case with the weaker model.

The solid black point represents a hypothetical mislabeled record in the data set. In the case of the weaker model, this corrupted record is in the region in which both distributions overlap substantially. Thus, before the record was corrupted, the model could not determine well to which class the record belonged. For similar records in this region, the conditional AUC would therefore be close to 0.5: neither “bad” nor “good” instances are well predicted, so changing the class has little impact. In the case of the more powerful model, this record is better discriminated, so its mislabeling impacts the overall assessment of discriminatory power to a higher degree since information that was useful in discrimination is being corrupted. By definition, weaker models have a higher proportion of records that the model cannot well discriminate. (Note that the location of the black record in each set of distributions would differ between models as, by definition, models with differing power would not rank records identically.)

Importantly, although *in expectation* the ranks of two models will not be changed as a result of corruption, the differences between the more and less powerful models will shrink, resulting in a smaller perceived difference in performance. This may be *practically* important in settings in which a new or alternative model is being considered to replace an existing model or in which a more costly better quality model is being compared to a cheaper lower quality one. Furthermore, as the noise level increases, it becomes more likely that the *observed* rankings of candidate models may become inverted, even when substantial differences exist, due to increased variance and decreased differences in AUC.

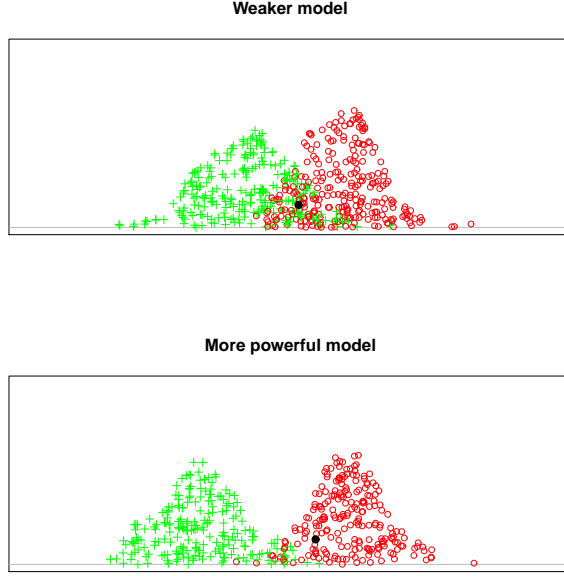


Fig. 2: Impact of mislabeling on weaker and more powerful models

The figures show the impact of mislabeling a record on weaker (top) or more powerful (bottom) models respectively. The plots show the distribution of model scores for the majority class (green “+”) and the minority class (red “o”). The solid black point shows a mislabeled record in the data set. In the case of the weaker model, the corruption affects a record which the model cannot well discriminate, so the impact is small. In the case of the more powerful model, a well discriminated record is mislabeled, impacting the overall assessment of discriminatory power. By definition, weaker models have a higher proportion of records that the model cannot well discriminate.

#### 2.4 Best AUC possible given mislabeling in the evaluation sample

We may also ask what the maximum attainable observed AUC,  $[A_c]$ , might be, given some level of mislabeling. We can calculate this upper bound by assuming that  $A = 1$  (the largest possible value for  $A$ ) and again setting  $A_0 = E[A_0^{nl}] = E[A_0^{mk}] = 0.5$  and using (6) to calculate  $A_c$ , given a specified level of mislabeling. (Note that because  $[A_c]$  is a theoretical upper bound, it may be calculated without reference to a specific data set.)

**Example 3 (Calculating an upper-bound on  $A$ )** Consider an evaluation dataset of anonymized financial statement data and corresponding default flags. The dataset contains 35,021 non-default records and 2,023 default records. Before beginning his evaluation, an analyst is told that 541 of the default records are actually “technical defaults” (i.e., related to non-credit issues) and that they should not be considered defaults but mislabelings. Because the data are anonymized,

however, the identities of the technical defaults cannot be determined. The analyst can determine an upper-bound on the value of  $\hat{A}_c$  that he will be able to observe using this data:

$$\begin{aligned}
 n &= 35,562 = 35,021 + 541 \\
 m &= 1,482 = 2,023 - 541 \\
 k &= 0 \\
 l &= 541 \\
 \hat{A}_c &= \frac{[(n-l)(m-k)\hat{A} + l(n-l)\hat{A}_0^l + k(m-k)\hat{A}_0^{mk} + kl(1-\hat{A})]}{(m-k+l)(n+k-l)} \\
 \lceil A_c \rceil &= \frac{[(35,562 - 541)(1,482 - 0) \times 1 + 541(35,562 - 541) \times 0.5 + 0 \times 0.5 + 0 \times 0.5]}{(1,482 - 0 + 541)(35,562 + 0 - 541)} \\
 &\approx 0.866
 \end{aligned}$$

■

Figure 3, below, shows some examples of  $\lceil A_c \rceil$ , the upper bound on  $A_c$  in the case that either the “bads” or “goods” are mislabeled. The figure shows the value of  $\lceil A_c \rceil$  for various sample prevalence rates by holding one of either  $n$  or  $m$  constant and adjusting the other of  $m$  or  $n$  to produce the appropriate prevalence rate. Separate curves are shown for each prevalence rate, with the rate noted to the right of each curve.

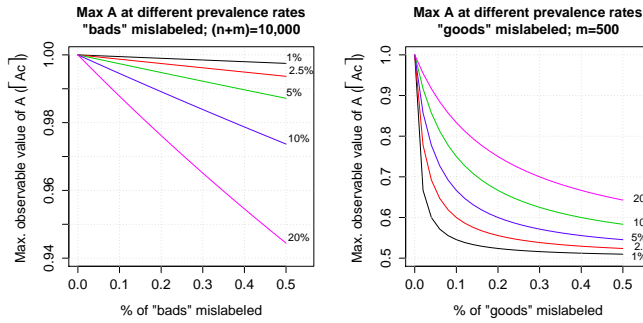


Fig. 3: Upper bounds on  $A_c$ , denoted  $\lceil A_c \rceil$ , for different levels of  $k$  and  $l$

The figures show the maximum possible value of the observed AUC,  $\lceil A_c \rceil$ , given a known level of mislabeling of  $k$  “bad” (left) or  $l$  “good” (right) records, respectively, using the analytic estimate of  $A_c$  given in (5), and assuming that  $A = 1.0$ . The observed value of  $A_c$  is shown on the  $y$ -axis and the corruption rate is shown on the  $x$ -axis. Each line represents a different prevalence rate sample.



## 2.5 Interpretation and bias correction for observed $\hat{A}_c$ given known mislabeling in test data (Estimating “true” values of $\hat{A}$ from corrupted testing data)

It is interesting to ask how we might arrive at an estimate of  $\hat{A}$ , given that we only had noisy (partially mislabeled) data on which to test a model. We can answer this question directly by solving for the estimate  $\hat{A}$  (the “accurate” estimate of  $A$ ) as a function of  $\hat{A}_c$ ,  $n$ ,  $m$ ,  $k$  and  $l$ . Solving and collecting terms yields a (somewhat messy) estimate of the recovered value of  $\hat{A}$ , denoted  $\hat{A}_r$ :

$$\hat{A}_r = \frac{(k-l-m)(k-l+n)\hat{A}_c + [l(n-l) + k(m-k)]\hat{A}_0 + kl}{nk - nm + lm}.$$

More conveniently, we can write<sup>8</sup>:

$$\hat{A}_r = \gamma_c \hat{A}_c + \gamma_0 \hat{A}_0 + \gamma_1 \quad (8)$$

where

$$\gamma_c = \frac{(k-l-m)(k-l+n)}{nk - nm + lm},$$

$$\gamma_0 = \frac{[l(n-l) + k(m-k)]}{nk - nm + lm},$$

and

$$\gamma_1 = \frac{kl}{nk - nm + lm}.$$

Note that because  $\hat{A}_r$  is an estimate, it is not naturally bounded between zero and one. It is therefore convenient to enforce these boundaries in practice.

Figure 4 shows the relationship between misclassification rates and  $\hat{A}_r$  for different constant values of  $\hat{A}$ , where we have again substituted  $E[\hat{A}_0] = 0.5$  for  $\hat{A}_0$ .

---

<sup>8</sup> Note that the variances of the two estimates of  $\hat{A}_0$ , corresponding to  $l$  and  $k$  mislabelings of  $n$  and  $m$  respectively, will be different. See footnote 9.

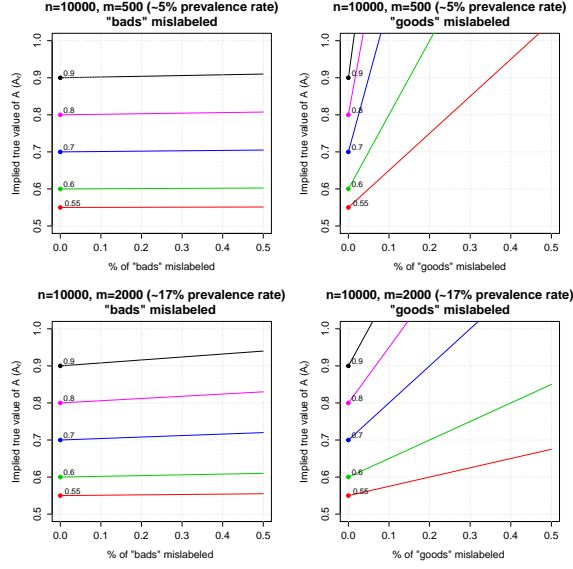


Fig. 4: Implied values of  $A$  ( $\equiv A_r$ ) given known mislabeling rate and  $A_c$

The figures show the expected implied value of  $A$ ,  $A_r$ , for mislabeling of  $k$  (left) or  $l$  (right) “bad” or “good” records, respectively (Eq.8). The plots in the top row show the impact for a lower prevalence rate sample, while the plots on the bottom row the impact on a higher prevalence rate sample. The value of  $A_c$  is given near the  $y$ -axis in each case.

**Example 4 (Recovering an estimate of the true value of  $A$  from a corrupted estimate ( $\hat{A}_c$ ))** Consider an analyst evaluating a bankruptcy model who is given only a dataset containing anonymized financial statement data and corresponding default flags. On this data set, the estimated value of the AUC is 0.73. As before, dataset contains 35,021 non-default records and 2,023 default records with 541 of the default records actually being “technical defaults” that should not be considered defaults but that cannot be identified due to anonymization. Assuming that the analyst believes the technical defaults to be non-informative, he can use the data given to estimate the true power of the the model:

$$\begin{aligned}\hat{A}_c &= 0.73 \\ n &= 35,562 = 35,021 + 541 \\ m &= 1,482 = 2,023 - 541 \\ k &= 0, \quad \text{and} \\ l &= 541.\end{aligned}$$

Now, using (8), he calculates that the expected true value of  $\hat{A}$  is:

$$\hat{A}_r = 0.814,$$

or about 12% higher than originally estimated.

■

**Example 5 (A new type of error-bound for  $\hat{A}_r$ )** Imagine instead the analyst is only told that between 1-2% of the “goods” (non-defaults) are mislabeled. He can caveat his results by noting that while the estimated AUC was  $\hat{A}_c = 0.73$  the true AUC *estimate* is likely in the range

$$\hat{A}_r \in [0.785, 0.840]_r$$

(where the notation  $[\cdot]_r$  indicates a *recovered range*).

Now suppose that a second model was tested on the same dataset. This model’s AUC was  $\hat{A}_c = 0.82$  giving

$$\hat{A}_r \in [0.897, 0.974]_r.$$

Here again, we see that corruption affects more powerful models (those with higher  $A$ ) more severely than weaker ones: the recovered range for the more powerful model is about 0.077 points of AUC, while the recovered range for the weaker model is only about 0.055.

■

Note that in Example 5 the “error bound” is not the same as a traditional confidence bound. Rather the variability in this estimator derives from the uncertainty around the level of data corruption rather than sampling variability. As such it is distinct from the notion of a confidence bound in the traditional statistical sense. In order to calculate more standard confidence bounds, we next derive the variance of  $\hat{A}_r$ .

## 2.6 The variance of $\hat{A}_r$

The variance of  $\hat{A}_r$  will be different from that of  $\hat{A}$ , both due to the natural difference in variance between estimates when  $A = A_1$  versus when  $A = A_2$  ( $A_1 \neq A_2$ ), and due to the additional “mixing” coefficients on  $\hat{A}_c$  and  $\hat{A}_0$  in (8).

If we collect terms in (8) and drop those that do not depend on  $\hat{A}_c$  or  $\hat{A}_0$ , then by the fundamental properties of variance, it can be shown that:

$$\text{Var}(\hat{A}_r) = \gamma_c^2 \times \text{Var}(\hat{A}_c) + \gamma_0^2 \times \text{Var}(\hat{A}_0) + 2\gamma_c\gamma_0\text{Cov}(\hat{A}_c, \hat{A}_0). \quad (9)$$

It is instructive to note that when  $l = k = 0$  (no corruption)  $\gamma_c^2 = 1$  and  $\gamma_0^2 = 0$ , so (9) gives  $\text{Var}(\hat{A}_r) = \text{Var}(\hat{A}_c) = \text{Var}(\hat{A})$ . Similarly, when  $l = n$  and  $k = m$ , we again get  $\gamma_c^2 = 1$  and  $\gamma_0^2 = 0$  (in this case  $\hat{A}_c = 1 - \hat{A}$ ).

$\text{Var}(\hat{A}_c)$  can be calculated in the standard way, non-parametrically [cf., Bamber, 1975, DeLong et al., 1988], via bootstrap [Stein, 2007] or via a parametric approximation as in [Hanley and McNeil, 1982].

However, note that in the case of higher moments (such as covariance), we cannot substitute expected values for the estimands (e.g., by substituting  $\hat{A}$  for  $E[\hat{A}]$ ), making analytic approaches impractical without specific side information about which records are corrupted. For example, to apply the approach of Bamber [1975] to calculate  $\text{Var}(\hat{A})$  one must determine, among other things, the probability of a random true “bad” being scored worse than two randomly selected true “goods”. (Of course to do this would require knowledge of the corrupted records, so that they may be excluded from this probability estimate, which we do not have.)

The situation is somewhat better with respect to  $\text{Var}(\hat{A}_0)$ . We cannot estimate  $\text{Var}(\hat{A}_0)$  directly without side information. However, since we know  $E[\hat{A}_0]$ , under mild assumptions we can approximate  $\text{Var}(\hat{A}_0)$  analytically, assuming we also know  $n, m, k$ , and  $l$ .<sup>9</sup> (In the event that  $k$  and  $l$  are not known *a priori*, we discuss some possible approaches to estimating them in Section 4.1.) Returning now to the covariance term, note that we cannot assume that  $\text{Cov}(\hat{A}_c, \hat{A}_0) = 0$  since  $A_c$  is a mixture of  $A_0$  and  $A$  (Eq. 6). When  $k$  and  $l$  are not too large, as  $l+k$  increases,  $\hat{A}_0$  and  $\hat{A}_c$  may become increasingly positively correlated by virtue of the inclusion of  $\hat{A}_0$  in  $\hat{A}_c$ .

The signs of  $\gamma_0$ ,  $\gamma_c$  and  $\gamma_1$  depend on the values of  $m, n, k$  and  $l$ , but we expect the correlation between  $A_0$  and  $A_c$  to generally be positive, given the mixing of  $A_0$  into  $A_c$ ; the larger the values of  $k$  and  $l$ , the more mixing will occur and thus, the higher should be the correlation.  $\hat{A}_c$  and  $\hat{A}_0$  depend (in the same direction) on the data sample which implies that both depend on the same distribution of the underlying model scores for “bad” and “good” records

---

<sup>9</sup> When  $E[\hat{A}_0] = 0.5$ , it can be shown that a reasonable analytic approximation is given by  $\text{Var}(\hat{A}_0) = \text{Var}(A_0^{mk}) + \text{Var}(A_0^{nl}) \approx \frac{m+1}{12(m-k)k} + \frac{n+1}{12(n-l)l}$ . See Bamber [1975], Eq. (4) for a general form.

and on the true value of  $A$ .<sup>10</sup>

We can refine  $\text{Cov}(\hat{A}_c, \hat{A}_0)$  a bit more. To do so, it is helpful to rewrite  $\hat{A}_c$  (Eq.6) more compactly:

$$\hat{A}_c = \eta_0 \hat{A}_0 + \eta_A \hat{A} + \eta_1 \quad (10)$$

where

$$\eta_0 = \frac{l(n-l) + k(m-k)}{(m-k+l)(n+k-l)},$$

$$\eta_A = \frac{(n-l)(m-k) - kl}{(m-k+l)(n+k-l)},$$

and

$$\eta_1 = \frac{kl}{(m-k+l)(n+k-l)}.$$

Then

$$\begin{aligned} \text{Cov}(\hat{A}_c, \hat{A}_0) &= \text{Cov}(\hat{A}_0, \eta_0 \hat{A}_0 + \eta_A \hat{A} + \eta_1) \\ &= \eta_0 \text{Cov}(\hat{A}_0, \hat{A}_0) + \eta_A \text{Cov}(\hat{A}_0, \hat{A}) \\ &= \eta_0 \text{Var}(\hat{A}_0) + \eta_A \text{Cov}(\hat{A}_0, \hat{A}) \end{aligned} \quad (11)$$

---

<sup>10</sup> For example, assume a credit model produces a high score if it predicts that a firm is likely to default and a low score if it predicts the firm will not default. Intuitively, if the set of  $k$  corrupted default records *happens to* include a large proportion of records with higher model scores (relative to the mean of the defaulted records),  $\hat{A}_c$  will be decreased (because the mean score of the “default” records is decreased and the mean score of the “non-default” records is increased, so on average the separation between the two classes is reduced). In this case,  $\hat{A}_0 = \hat{A}_0^{mk}$  will also decrease (there will now be more separation on average between the set of (high score) false “non-default” records and the set of (low score) true “default” records, but the relationship will be inverted since the false “non-default” records have higher scores on average than the true “default” records to which they are compared. See the definition of  $\hat{A}_0^{nk}$ ). Similarly, if the set of  $k$  corrupted default records *happens to* include a high proportion of very low model scores,  $\hat{A}_c$  should increase (since the mean score of the true “default” records is increased more than is the mean score of the set of all records labeled “non-default”) while  $\hat{A}_0$  will also increase (since there will now be more separation on average between the false “non-default” records and the true “default” records, but this time in the correct direction.). The degree to which this positive correlation is expressed will depend on the true value of  $A$ , which measures the model’s ability to discriminating between the two classes, the distributions of scores in each class, and the values of  $m$ ,  $n$ ,  $k$ , and  $l$ .

(Note that for convenience we have collapsed  $\hat{A}_0^{nl}$  and  $\hat{A}_0^{mk}$  into  $\hat{A}_0$ . For our purposes, this does not affect our results since in the examples and experiments only one of  $k, l$  is non-zero.)

Similarly, we can estimate  $\text{Cor}(\hat{A}_c, \hat{A}_0)$ . Again denoting by  $\hat{\sigma}_c$  and  $\hat{\sigma}_0$ , the standard errors of  $\hat{A}_c$  and  $\hat{A}_0$ , respectively:

$$\text{Cor}(\hat{A}_c, \hat{A}_0) = \frac{\text{Cov}(\hat{A}_c, \hat{A}_0)}{\hat{\sigma}_c \hat{\sigma}_0} \quad (12)$$

$$\begin{aligned} &= \frac{\eta_0 \text{Var}(\hat{A}_0) + \eta_A \text{Cov}(\hat{A}_0, \hat{A})}{\hat{\sigma}_c \hat{\sigma}_0} \\ &= \frac{\eta_0 \sigma_0}{\hat{\sigma}_c} + \frac{\eta_A \text{Cov}(\hat{A}_0, \hat{A})}{\hat{\sigma}_c \hat{\sigma}_0} \\ &= \frac{\eta_0 \hat{\sigma}_0}{\hat{\sigma}_c} + \frac{\eta_A \text{Cor}(\hat{A}_0, \hat{A}) \hat{\sigma}_A}{\hat{\sigma}_c}. \end{aligned} \quad (13)$$

Unfortunately, because in practice we would not know the specific  $k$  “bads” and  $l$  “goods” that were corrupted, we cannot directly calculate  $\text{Cor}(\hat{A}_0, \hat{A})$  from the data. Indeed, in this expression,  $\hat{A}$  is better written  $\hat{A}_{\bar{c}}$ , as it is the estimate of  $A$  from the uncorrupted sample using only those records that are not subsequently corrupted, so although  $\mathbb{E}[\hat{A}_{\bar{c}}] = A$ , the underlying sample is shared by  $\hat{A}_{\bar{c}}$  and  $\hat{A}_0$ .

While we cannot estimate (11) directly, in our simulation experiments, we have found that a value of  $\text{Cor}(\hat{A}_c, \hat{A}_0) \approx 0.5$  produces  $\hat{\sigma}_r$  with reasonable coverage for the parameter values we study here.

## 2.7 Applications of $\text{Var}(\hat{A}_r)$

The estimate of  $\text{Var}(\hat{A}_r)$  may be used in place of the standard (uncorrupted) variance estimator in many applications. For example, it may be used for assessing confidence bounds on  $A_r$  or for hypothesis tests (though in some cases, further adjustments may be required, e.g., in cases in which two models are correlated). Importantly, the approach described here is most suited to parametric applications such as parametric hypothesis tests of model power.

For example, Appendix B shows how  $\text{Var}(\hat{A}_r)$  may be used to construct proxy ROC confidence bands using the parametric approach described in Macskassy et al. [2005] (by simply replacing  $A_r$  with the upper and lower confidence

levels for  $\widehat{A}_r$  at the desired confidence level in Equations (15)-(19)).

On the other hand, for example,  $\text{Var}(\widehat{A}_r)$  is relatively less useful in *directly* constructing the semi-parametric and non-parametric confidence bands described in Macskassy et al. [2005].

Note however, that even here, although  $\text{Var}(\widehat{A}_r)$  cannot be used directly to implement such confidence bands,  $\widehat{A}_r$  itself can!  $\widehat{A}_r$  can be utilized to estimate the parameters of (15) (as described in Appendix B) and the resulting equation can then be used to generate the corresponding ROC. At this point some of the methods described in Macskassy et al. [2005] may be directly applied to this  $\widehat{A}_r$ -based ROC curve (under the associated assumptions discussed).

### 3 Simulation experiments

In this section, we present some simulation results in which we compare performance statistics calculated based on empirical data, both before and after corruption, with those calculated using only corrupted data, but then adjusted based on the analytic results from Section 2.

#### 3.1 Simulations of the agreement between analytic and true values of $A$

We first generated two large data sets by simulating a synthetic single factor discrete-choice process. In the simulations, we then added varying levels of factor noise to the process output to simulate a model with either high power (for our experiment,  $A \approx 0.9$  or, equivalently  $AR \approx 0.8$ ) or low power ( $A \approx 0.8$  or, equivalently,  $AR \approx 0.6$ ). Prevalence rates for the samples are given as  $m/(n+m)$  or approximately 4.8% for the “High” case and approximately 1% for the “Low” case. Simulation is done using the algorithm shown in Figure 5, with appropriate parameter values for  $\mu$  and  $\gamma$ .

We then generated random samples from each “model” repeatedly. For example, if we were examining the behavior of high power models, we used the model with  $A \approx 0.9$  and if we were examining lower power models, we use the model set with  $A \approx 0.8$ . For each simulation we generate a synthetic data set and then use this data set to test the estimators of model power.

From each synthetic data set, we also create a second corrupted data set. To do this, we corrupt the data by flipping the labels on some number of records so that “bads” become “goods” or “goods” become “bads”. We then

---

**1. Simulate class membership probabilistically**


---

$x \sim \Phi(0, 1)$	<i>{random noisy true value for record}</i>
$truescore = \Phi^{-1}(\mu + x)$	<i>{probability of <math>x</math> (true score)}</i>
$u \sim U(0, 1)$	<i>{class given <math>x</math> (probabilistic)}</i>
if ( $u < truescore$ ) $class=1$	
else $class=0$	

**2. Then simulate model accuracy**

$\varepsilon \sim \Phi(0, \gamma)$	<i>{noise level ( simulated accuracy)}</i>
$score = \Phi^{-1}(\mu + x + \varepsilon)$	
return ( $score, class$ )	<i>{realized score and actual data class}</i>

---

Fig. 5: Data simulation algorithm used for generating synthetic data for experiments.  $U(\cdot)$  and  $\Phi(\cdot)$  are cumulative uniform and Gaussian distribution functions, respectively,  $\Phi^{-1}(\cdot)$  is the inverse Gaussian distribution function and  $\mu$  is the true data distribution mean.  $\gamma$  is the level of model noise used to determine the model's estimate of the class membership.

calculate performance statistics on each data set and compare the results.

In the case of the corrupted data set, we then apply Eq. (8) to recover  $\hat{A}_r$  from  $\hat{A}_c$ . Our primary objective is to demonstrate that the analytic results are borne out by simulation experiments.

In each experiment we calculate:

- “true”  $\hat{A}$ : the estimate of the AUC from the uncorrupted sample using Eq. 1 and the uncorrupted data;
- observed  $\hat{A}_c$ : the AUC calculated from the corrupted sample using Eq.1 and corrupted data;
- analytic  $\hat{A}_c$ : the estimate of the corrupted AUC calculated based on  $m, n, k$  and  $l$ , the “true”  $\hat{A}$ , using Eq. 5; and
- $\hat{A}_r$ : the analytic estimate of the true AUC calculated based on  $m, n, k$  and  $l$  and the observed  $\hat{A}_c$  (Eq. 8).

$\hat{A}_r$  is the most potentially useful quantity as it is the estimate of the true value of  $\hat{A}$  derived from statistics calculated on corrupted data. Said differently, if one had only a corrupted sample, but had some sense of the degree of corruption, the recovered  $\hat{A}_r$  could be used as a proxy for what the true value of  $\hat{A}$  would be, if the data were not corrupted.



---

The results of the first of these experiments are shown in Tables 1 and 2. As the tables show, the analytic and simulated results match quite closely as we would expect. In general, the mean values for both the simulated and analytic results are identical to the third decimal (typically the limit of precision for many practical model evaluation exercises). In cases in which the means do differ, they differ by less than 0.5% in relative terms.

Table 1: Simulation results for tests of power in case of mislabeled records:  $l$  “goods” are mislabeled as “bads”

Model	Baseline Prevalence Rate	Mean “true” $\hat{A}$	Mean analytic $\hat{A}_r$	Mean observed $\hat{A}_c$	Mean analytic $\hat{A}_c$	$l$	$m$	$n$
Powerful	Low	0.900	0.900	0.700	0.700		100	100
		0.899	0.899	0.633	0.633		200	100
		0.900	0.900	0.567	0.567		500	100
		0.900	0.899	0.536	0.536		1000	100
		0.900	0.897	0.515	0.515		2500	100
	High	0.900	0.900	0.833	0.833		100	500
		0.900	0.900	0.786	0.785		200	500
		0.900	0.899	0.700	0.700		500	500
		0.900	0.900	0.633	0.633		1000	500
		0.900	0.900	0.567	0.567		2500	500
Weak	Low	0.800	0.800	0.650	0.650		100	100
		0.799	0.798	0.599	0.600		200	100
		0.800	0.799	0.550	0.550		500	100
		0.800	0.800	0.527	0.527		1000	100
		0.799	0.799	0.511	0.512		2500	100
	High	0.800	0.799	0.750	0.750		100	500
		0.800	0.799	0.714	0.714		200	500
		0.800	0.800	0.650	0.650		500	500
		0.800	0.799	0.600	0.600		1000	500
		0.800	0.800	0.550	0.550		2500	500

Table 2: Simulation results for tests of power in case of mislabeled data:  $k$  “bads” are mislabeled as “goods”

Model	Baseline Prevalence Rate	Mean “true” $\hat{A}$	Mean analytic $\hat{A}_r$	Mean observed $\hat{A}_c$	Mean analytic $\hat{A}_c$	$k$	$m$	$n$
Powerful	Low	0.899	0.899	0.899	0.899	1	100	10,000
		0.899	0.899	0.899	0.899	2	100	10,000
		0.900	0.900	0.899	0.899	5	100	10,000
		0.900	0.900	0.899	0.899	10	100	10,000
		0.899	0.899	0.898	0.898	25	100	10,000
Powerful	High	0.899	0.899	0.899	0.899	5	500	10,000
		0.900	0.900	0.899	0.899	10	500	10,000
		0.899	0.899	0.898	0.898	25	500	10,000
		0.900	0.899	0.897	0.898	50	500	10,000
		0.900	0.900	0.895	0.895	125	500	10,000
Weak	Low	0.800	0.800	0.800	0.800	1	100	10,000
		0.799	0.800	0.799	0.800	2	100	10,000
		0.800	0.800	0.799	0.800	5	100	10,000
		0.800	0.800	0.800	0.799	10	100	10,000
		0.799	0.801	0.799	0.800	25	100	10,000
Weak	High	0.800	0.800	0.800	0.799	5	500	10,000
		0.800	0.800	0.799	0.799	10	500	10,000
		0.800	0.800	0.799	0.799	25	500	10,000
		0.800	0.800	0.798	0.798	50	500	10,000
		0.800	0.800	0.796	0.796	125	500	10,000

Results of simulated estimates of the area under the ROC curve (AUC). The “Powerful” model has  $A \approx 0.9$  ( $AR \approx 0.8$ ) and the “Weak” model has  $A \approx 0.8$  ( $AR \approx 0.6$ ). Prevalence rates for uncorrupted samples are given as  $m/(n+m)$  or approximately 4.8% for the “High” case and 1% for the “Low” case. Means for the variables are defined as follows: “true”  $\hat{A}$  is the estimate of the AUC from the uncorrupted sample; observed  $\hat{A}_c$  is the AUC calculated from the corrupted sample; analytic  $\hat{A}_c$  is the value of the AUC calculated based on  $m$ ,  $n$ ,  $k$  and  $l$ , using the “true”  $\hat{A}$ ; and recovered  $\hat{A}_r$  is the estimated value of the true AUC calculated based on  $m$ ,  $n$ ,  $k$  and  $l$ , using the observed  $\hat{A}_c$ . ( $\hat{A}_r$  is the estimate of the true power of the model based on statistics calculated on the corrupted sample.)

### 3.2 Simulations of the coverage of $\hat{\sigma}_r$

To explore the variance estimators, we calculate the six different estimates of the standard error of  $\hat{A}_r$ . First ( $HM_1$ ) we use the [Hanley and McNeil, 1982] estimator, using  $\hat{A}_r$  as the value of  $A$  and taking the observed counts in place of  $m$ , and  $n$  (i.e, we use the values calculated from the corrupted data. In the observed data these are  $(m - k + l)$  and  $(n - l + k)$ , respectively). Second, ( $HM_2$ ), we use a modified version of the [Hanley and McNeil, 1982] estimator

in which we again take  $\hat{A}_r$  as the value of  $A$ , but now correct for the corruption in calculating  $m$  and  $n$ , by adding back the corrupted records so that  $m$  and  $n$  take on their original values, before the corruption. Third,  $(HM_3)$ , we use the [Hanley and McNeil, 1982] estimator, but take  $\hat{A}_c$  as  $A$  and use the observed (corrupted) values of  $(m - k + l)$  and  $(n - l + k)$ , respectively. Fourth  $(HM_4)$ , we use the [Hanley and McNeil, 1982] estimator, but take  $\hat{A}_c$  as  $A$  and use the corrected (true) values of  $m$  and  $n$ . Fifth,  $(BS)$ , we calculate bootstrap estimates of the s.e. of  $\hat{A}_c$  via a balanced bootstrap. (Thus, in each iteration of the simulation, we perform a bootstrap on the simulated corrupted but ensure that each bootstrap replication contains the same number of records marked as “bad”  $(m - k)$  and “good”  $(n - l)$ , respectively.) Finally we estimate (9), the bias-corrected analytic standard error ( $\hat{\sigma}_r$ ) via a two-step method. In step 1, we estimate  $(BS)$  and in step 2, we use this estimate of  $\text{Var}\hat{A}_c$  to then estimate  $\text{Var}[\hat{A}_r]$  in (9).

To give a sense of how appropriate these estimates are, we can calculate the frequencies with which the actual values of  $A$  exceed confidence bounds implied by the various standard error estimates of the recovered value  $\hat{A}_r$ . In principle, these *exceedance rates* should be approximately the same as (one minus) the nominal confidence levels implied by the standard errors.<sup>11</sup> However, calculating bootstrap estimates of the variance of  $A_c$ , requires a large number of bootstrap replications (we use 20,000) per simulation path and we are thus constrained in the number of paths we can generate per simulation and in our corresponding choice for  $\alpha$ .<sup>12</sup> Because we are constrained in the number of simulation we perform, we simulate  $N_S = 500$  paths per simulation run and set  $\alpha = 0.05$ .

For a binomial process with probability  $p = 0.05$ , the nominal coverage for the CIs we will be examining, we expect the observed proportion of cases that fall outside the CI to be in range of  $[0.034, 0.073]$  for  $N_S = 500$  trials. In Table 3, below, we show some examples of how the coverage of different estimators of  $\text{Var}[\hat{A}_r]$  changes as  $\hat{A}_c$ ,  $k$  and  $l$  change.

Note that in all cases but one, the realized exceedance rate are within the tolerances of the nominal levels for  $\hat{\sigma}_r$ . In contrast the exceedance rates of the other estimators appear to be generally inconsistent with the nominal coverage levels. Note also that  $\hat{\sigma}_r$  is sometimes smaller and sometimes larger

<sup>11</sup> We estimate the “true” value of  $A$  as the mean value of  $\hat{A}$  across all realizations in each simulation.

<sup>12</sup> We bootstrap each simulated corrupted data set  $B$  times. Thus, for  $N_S$  simulations paths, we calculate a total of  $N \times B$  calculations of the AUC. In Table 3 we use  $B = 20,000$  and  $N_S = 500$  so the total number of AUC calculations is 10 million per simulation (row). We parallelize the bootstrap to permit multi-threaded evaluation.

Table 3: Exceedance rates for various estimates of the standard error of ( $\hat{A}_r$ )

Model	Mean $\hat{A}$	Mean $\hat{A}_c$	Exceedence rates for $\hat{A}_r$							$k$	$l$
			$HM_1$	$HM_2$	$HM_3$	$HM_4$	$BS$	$\sigma_r$			
Powerful	0.899	0.899	0.000	0.000	0.000	0.000	0.060*	0.060*	5	0	
	0.899	0.898	0.004	0.006	0.004	0.006	0.064*	0.066*	25	0	
	0.899	0.897	0.004	0.006	0.004	0.006	0.064*	0.066*	50	0	
	0.899	0.894	0.006	0.010	0.006	0.010	0.060*	0.066*	125	0	
	0.899	0.833	0.060*	0.042*	0.020	0.012	0.038*	0.034*	0	100	
	0.899	0.700	0.380	0.208	0.206	0.064*	0.200	0.018	0	500	
	0.899	0.634	0.622	0.380	0.446	0.186	0.434	0.038*	0	1000	
	0.899	0.566	0.826	0.554	0.698	0.414	0.696	0.044*	0	2500	
Weaker	0.799	0.798	0.010	0.010	0.010	0.010	0.060*	0.060*	5	0	
	0.798	0.798	0.008	0.010	0.008	0.010	0.060*	0.064*	25	0	
	0.799	0.798	0.010	0.018	0.010	0.018	0.046*	0.050*	50	0	
	0.799	0.795	0.010	0.038*	0.010	0.036*	0.058*	0.058*	125	0	
	0.798	0.749	0.046*	0.036*	0.036*	0.026*	0.062*	0.052*	0	100	
	0.798	0.649	0.322	0.176	0.262	0.144	0.276	0.062*	0	500	
	0.799	0.599	0.512	0.282	0.462	0.230	0.460	0.058*	0	1000	
	0.798	0.550	0.792	0.534	0.754	0.486	0.756	0.048*	0	2500	

Exceedance rates ( $\alpha = 0.05$ ) of  $\hat{A}_r$  using various s.e. calculations. High prevalence rate ( $m=500$ ,  $n=10,000$ ). \*=within nominal coverage at,  $\alpha = 0.05$  level,  $N_S = 500$  paths per simulation;  $\text{Cor}(A_c, A_0) = 0.5$ .

than that of, say, the bootstrap estimator that it modifies. These experiments suggest that the bounds implied by the standard errors calculated using (9) have reasonable coverage when records are mislabeled. We can also observe that when  $k$  and  $l$  are small (i.e., data set is relatively “clean”), the bootstrap estimator also has reasonable coverage. However, as error rates increase, the bootstrap estimator understates variance (exceedances are higher than nominal). It is also obvious that mislabelings of the minority class (e.g., defaults) have less impact on the variance than mislabelings of the majority class. This may be a result of the relatively small impact that mislabeling “bads” has on the estimates of  $A$  for very skewed distributions. This can be seen in the relatively slight differences between  $\hat{A}$  and  $\hat{A}_c$  in the case of mislabeled “bads”, a behavior we noted in Figure 1 of Section 2.

## 4 Discussion

In this section, we provide suggestions on how the levels of data noise may be estimated if they are not known *a priori*. We also discuss some related work from the biostatistics literature and consider briefly the CAR assumption.

### 4.1 Some approaches to estimating $k$ and $l$

Much of the preceding has assumed that we knew, or could find out, the values of  $k$  and  $l$ . We have left until now the question of the best method for determining whether the data have been corrupted, and, if so, how and to what degree. In the case in which one has concrete prior knowledge (e.g., of the data collection process or the data generating process itself) this is not an issue. However, in more general cases, inferential approaches are required. While these approaches can be limited in some settings, they still provide a starting point for estimating the degree and forms of data corruption. (Note that the use of estimates of corruption rates, rather than known quantities, will increase  $\text{Var}(\hat{A}_c)$ . This may also impact the derivation of Eq. (8).)

#### 4.1.1 Inferring the number of missing records of one class

In section 2.1, we briefly discussed the impact of missing “bads.” Although these do not affect the expected value of  $\hat{A}$ , they may affect calibration. It is feasible to estimate the number of missing “bads” (or “goods”) in a database where some records have been dropped. The crudest approach is to estimate the number of missing “bad” records by simply calculating the difference in baseline prevalence rates in the development and testing samples (which we denote  $\hat{\pi}_S$  and  $\hat{\pi}_T$ , respectively), if these are known. We can then solve for  $k$ , subject to knowing  $\hat{\pi}_S - \hat{\pi}_T$ . This method is a coarse one, particularly when multiple forms of data corruption are present.

We can potentially do better than this if we have access to multiple databases. In this setting, it may be possible to infer the number of missing “bads” (and/or “goods”) using capture/recapture techniques [cf., Dwyer and Stein, 2006].

#### **Example 6 (Estimating missing default records in a credit database)**

Assume an analyst had access to two default databases (e.g., from two banks that merged within a common market) and that the first database contained 2500 defaulted firms, the second contained 1200 defaults, and there were 1000 defaults common to both databases, then (assuming independence) she could use these counts [see, Dwyer and Stein, 2006] to infer the true number of defaults in the population:

$$\hat{N} = 1000 + 1500 + 200 + \frac{(1500 \times 200)}{1000} = 3000,$$

or that there were a total of

$$3000 - (1500 + 1200) = 300$$

missing defaults in the combined dataset. If the data collection methods were correlated between the two data sets, more involved calculations would be required [Dwyer and Stein, 2006].

■

Of course, both of these methods break down if the evaluation database used to calculate  $\hat{\pi}_T$  (in the first case) or the two databases used to calculate the missing “bads” (in the second case) are themselves subject to multiple forms of corruption.

#### 4.1.2 Inferring the number of mislabeled “bads” and “goods”

If we suspect that the labels of some of the “bad” records in the evaluation sample have been mislabeled, we can make use of a number of results from the econometrics literature on mislabeled dependent variables. For example, the procedure proposed in Hausman et al. [1998] provides estimates of the probability that a “bad” (“good”) instance is mislabeled as “good” (“bad”) instance and may be implemented as part of a straightforward discrete choice model.

It can be shown that under not-unreasonable conditions, unbiased parameter estimates can be obtained from models of the form,

$$y_i = g(\alpha + \alpha_0 + \alpha_1 + \beta'x_i + \varepsilon_i), \quad (14)$$

where  $g(\cdot)$  is a link function,  $y_i$  is the status flag for record  $i$ ,  $y_i \in \{0, 1\}$  (e.g., 0 = “good”, 1 = “bad”),  $\alpha_0$  and  $\alpha_1$  are the mislabeling rates for “bads” and “goods”, respectively, and the other terms have their conventional meanings.

By applying a learned discrete choice model to the evaluation sample (e.g., if  $p_i$  is the output of a model for record  $i$ , and  $y_i$  is the true status for the record, then we let  $x_i = p_i$  and apply the procedure described in Hausman et al. [1998]). Using this approach, we can obtain estimates of  $\hat{\alpha}_0$  and  $\hat{\alpha}_1$ , the estimated rates of mislabeling inferred for each class.

**Example 7 (Estimating mislabeling rate in a credit database)** To give a sense of the degree to which we are able to recover reasonable estimates of mislabeling rates, we applied the approach of Hausman et al. [1998] to a test data set in which there were 10,000 non-defaults, 500 defaults and in which 1,000 of the non-defaults were mislabeled as defaults, the procedure yielded a misclassification estimate for non-defaults (i.e., non-defaults that were misclassified as defaults) of 0.101 which accords well with the true 10% misclassification rate.

■

While the procedure works well in detecting misclassifications of the majority class, the estimates of the mislabeling rates for the minority class tend to be poor when the prevalence rate is very low, or even just low. However, recall that mislabeling of the minority class tends to have a much smaller impact on estimates of  $\hat{A}$  than does mislabeling of the majority class (see, Section 3). Some experimentation with starting values may be required to achieve stability/convergence.

#### 4.2 Relationship to gold standard and verification bias literature

Our results with respect to mislabeling of the  $y_i$  are related to a stream of research in the biostatistics literature that investigates the power of diagnostic tests in the absence of a “gold standard”. This literature concerns itself with situations in which the methods required to verify the true status (e.g., healthy or sick) of an individual are too expensive (e.g., MRI imaging) or invasive (e.g., autopsy) to permit its use for some or all of the individuals in the evaluation set. Thus a second, imperfect test is used to define the pseudo-status of the individuals, and then the test under study is evaluated against this pseudo-status. A related stream of literature concerns itself with “verification bias” in which the true status of some individuals is known with certainty, but is unknown for others in the test set.

While these literature streams are related to the problem we study, we note some important differences in our approach relating both to our assumptions and to the computational complexity of the methods.

Our assumptions about the data differ from those of the typical gold standard set-up which require that each record in the data set be labeled as either verified (i.e., a gold-standard record) or un-verified (i.e., a non-gold standard record), depending on whether the  $y_i$  has been confirmed or not, respectively. Other variants of gold-standard adjustments require contingency tables at each point in the ROC (and for each rating scale class) showing the correct classifications of the model at that point on the ROC, broken out for the verified and non-verified records.

In principle, the gold standard and verification bias approaches and the approach we propose in this paper may be brought closer together by assuming that all records are unverified, however, this greatly reduces the efficiency of the estimator as in most cases, as the large majority of the records in a dataset are typically not corrupted. In the absence of this (strong) assumption, our approach provides a method for making inference on corrupted data, with lower information requirements.



Secondly, applying the bias corrections and variance estimators we propose here is trivial computationally, in contrast to many of the methods in these literatures, which require somewhat more involved statistical machinery such as Markov Chain Monte Carlo or the application of the EM algorithm. While such approaches are appropriate in settings that closely match the underlying assumptions, our approach may be used in a variety of alternative settings and may be applied without extensive computation. Indeed, the estimators described in this paper do not necessarily require even the underlying data from which the AUC was calculated.

### 4.3 The CAR assumption

### 4.4 Limitations of current approach

One of the main limitations of our approach (and thus an area for future work) is the CAR assumption. While there are many settings in which this is a reasonable approximation, there are others where it may not be. For example, some default prediction problems are characterized by the presence of non-trivial selection bias (see the example, below).

This limitation is not, however, unique to the approaches described here and is rather a characteristic of non-MAR (and non-CAR) problems. While there is often little that can be done to correct for dependence in the corruption mechanism (unless the missingness mechanism is known exogenously), in some cases, experimental design can improve statistics.

**Example 8 (Violation of CAR in a corrupted credit database)** Historically, it is common for banks to track the details of defaulted loans in separate files and different business units to administer defaulted loans and defaulted loans that were still performing. (Different skill sets are required for resolving foreclosures than for servicing loans). Analysis of these historical data sometimes suggest possible under-reporting of defaults on smaller loans. One explanation for this is that larger loans tend to be more easily recalled by loan officers (and are more likely to have detailed records) when banks perform archival data reconstruction. Furthermore, smaller loans are often written-off or consolidated with other losses. This mechanism may result in a higher prevalence of underreporting of defaults (mislabeling) among smaller loans than among larger loans. This is at odds with the CAR assumption.

One response to this is for analysts to calculate separate performance statistics for different cohorts within the data set, e.g., data may be stratified by loan size and the AUC calculated for each of strata. This type of analysis is not uncommon, and may be performed over many different dimensions (e.g. size, industry, etc.). While it does not completely address the non-CAR nature of the corruption, it does provide a form of baseline sensitivity analysis [cf., Stein et al., 2003].

■

#### 4.5 Impact of relaxing the CAR assumption

While beyond the scope of this work, it is intriguing to consider ways in which this approach could be extended to in the absence the CAR assumption.

Consider again Figure 2. In both the top and bottom plot, the distributions of model scores for the “goods” and “bads,” respectively overlap in some regions and not in others.

Imagine that now, rather than being corrupted at random, the corruption occurs only in the *overlapping* region of the two distribution of model scores. In this case, as we discussed in Section 2.3, corruption will have little impact on the observed value of  $\hat{A}$ ,  $\hat{A}_c$ , since records with model scores in the overlapping region are not well discriminated by the model and thus *conditional on being in that region* records from this region produce AUCs near 0.5.

On the other hand, if the corruption occurs in one of the non-overlapping regions, it will impact the observed value of  $\hat{A}_c$  more severely since records in these regions of the distributions are well discriminated by the model.

**Example 9 (Violation of CAR in a corrupted credit database – continued)** Continuing Example 8, assume that the corruption rate is higher for smaller firms for the reasons described above. It is empirically the case that very large firms are less likely to default than other firms. However, this tends to be true for only the largest firms. For most of the distribution of firm size, there is little difference in the conditional default rate between various sized firms and thus conditioning on size does not materially affect model discriminatory power in this region. Were there a systematic corruption of the smaller firms in the database (and assuming this were the *only* source of systematic corruption) it is likely that this corruption would look as if it were “at random” with respect to  $\hat{A}_c$ . In contrast, were the corruption to occur in the very large sized firms, the impact could be much more pronounced.

■

## 5 Conclusion

*“In theory, theory and practice are the same. In practice they are different.”*

-Yogi Berra

Many of the statistics in routine use for assessing model performance assume theoretically pure data even as they are calculated on inconveniently noisy samples. Given the noise levels endemic in real-world data, it is useful to develop approaches for correcting the biases that are introduced when performance statistics are calculated on corrupted data. The results of this article suggest that in a number of practical settings, we can derive useful analytic results to do this.

Our results provide a direct means to use corrupted data for model evaluation in a variety of situations. This makes it easier to draw reasonable conclusions from experiments based on noisy data. Our analysis also provides some intuition for the behavior of test statistics in noisy-data environments. In addition, we have provided analytic results that adjust *observed* performance statistics when data corruption is suspected. In doing so, our goal has been both to offer insight into the impact that noisy data can have on performance assessments of discrete choice models, and to provide computationally simple methods correcting for the effects of the noisy data.

This permits researchers to make estimates of model performance, even when using corrupted data in many cases, as long as some information about the corruption mechanism is known. Alternatively, when researchers have less knowledge of the corruption mechanism, these analytic results can be used to perform sensitivity analysis on the observed performance statistics to determine a range of likely values.

Our key results are first, that data corruption impacts measures of model discriminatory power in predictable ways. Second, that different types of data corruption affect measures of model performance in different ways. For example, if “bad” (minority class) instances are *missing* from the evaluation sample, this will have little impact on statistics that measure power such as the area under the AUC (though, the efficiency of the estimators declines in such cases). In contrast, *misabeled* data do affect such measures, though to different degrees. Third, different data corruption schemes affect power statistics in different ways. Moderate levels of *minority class* mislabeling appear to have only minor impact on estimates of the AUC, while the same percentage of mislabeling of *majority class* instances can affect estimates of model power statistics more materially.

Importantly, (fourth), our analytic results in Section 2.3 demonstrate that it is not enough to simply “test all models on the same data,” when the data are corrupted. Mislabeling errors do not “cancel out” across models. To the contrary, mislabeling errors impact higher power models more severely than weaker ones, moving estimates of the power of better models closer to those of poorer ones. As we noted earlier, this implies that it is not the case the same (noisy) data set handicaps all competing models similarly. To the contrary, data corruption implies that it becomes more likely in practice that a powerful model will look statistically similar to (or even worse than in some cases), a weaker one.

Fifth, we developed bias corrections that can be used in a number of realistic settings to allow researchers to make use of corrupted data to assess model quality, albeit in some cases at the cost of increased variance. Individual measures of performance are often less useful than cross-model comparisons. For this reason, we also derived analytic results for the variance of the adjusted

statistics in order to facilitate more natural hypothesis tests. We expect that this will be an area of ongoing future research. An attractive feature of these bias corrections is that they may be calculated quite simply and applied even without access to the underlying data on which the AUC and its variance were calculated.

We believe our results will permit new types of analyses and allow for a richer understanding of the true performance of candidate models. For example, if a researcher tests a model and observes an AUC of  $\hat{A}_c = 0.81$  on a data set containing 15,000 “goods” and 5,000 “bads”, but is told that on the order of 3-5% of the “goods” are mislabeled, she can caveat her result by noting that the true AUC *estimate* is in likely in the range  $[83.8, 85.7]_r$ . Such caveats serve both to highlight the (sometimes large) impact of data corruption on model evaluation and to precipitate discussions of data quality control.

While our results give insights into evaluating models on corrupt testing data, they are useful only in so far as information about the corruption mechanism, or at least the range of possible parameter values relating to the corruption, is understood. We have tried to provide some guidance on how these parameters may be estimated empirically, but much work remains to be done in this area as well as to more fully explore the impact of relaxing the “at random” assumptions on the corruption.

That said, in certain common situations this background information is known or can be estimated with some confidence. In these situations, various of our results may be useful in reducing bias in tests of model power and thereby lead to better informed decisions on model performance.

## References

- Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387 – 415, 1975.
- Jeffrey R. Bohn and Roger M. Stein. *Active Credit Portfolio Management in Practice*. Wiley, 2009.
- E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, September 1988. ISSN 0006-341X.
- Douglas W. Dwyer and Roger M. Stein. Inferring the default rate in a population by comparing two incomplete default databases. *Journal of Banking and Finance*, 30(3):797 – 810, 2006.
- Charles Elkan. The foundations of cost-sensitive learning. *Proceedings of the Joint Conference on Artificial Intelligence (IJCAI’01)*, pages 973–978, 2001.
- B. Engelmann, E. Hayden, and D. Tasche. Testing rating accuracy. *RISK*, 16: 82–862, 2003.

- David J. Hand and Robert J. Till. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2):171–186, November 2001. ISSN 08856125. doi: 10.1023/A:1010920819831.
- J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, April 1982. ISSN 0033-8419.
- J. A. Hausman, Jason Abrevaya, and F. M. Scott-Morton. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87(2):239–269, 1998.
- Daniel F. Heitjan and Donald B. Rubin. Ignorability and Coarse Data. *Annals of Statistics*, 19:2244–2253, 1991. doi: 10.1214/aos/1176348396.
- Wassily Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.2307/2235637.
- Sofus A. Macskassy, Foster Provost, and Saharon Rosset. ROC confidence bands: An empirical evaluation. In *Proceedings of the 22st International Conference on Machine Learning*, pages 537–544, Bohn, Germany, August 2005. ICML.
- W. W. Peterson, T. G. Birdsall, and W. C. Fox. The theory of signal detectability. *Transactions of the IRE Professional Group in Information Theory*, 2-4: 171–212, 1954.
- Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(2), 2001.
- H. Russell, Q. K. Tanng, and D. W. Dwyer. The Effect of Imperfect Data on Default Prediction Validation Tests. *Journal of Risk Model Validation*, 6(1):1–20, 2012.
- J. Sobehart, S. Keenan, and R. Stein. Validation methodologies for default risk models. *Credit*, 16:51–56, 2000.
- R. M. Stein, A. E. Kocagil, J. Bohn, and J. Akhavain. Systematic and Idiosyncratic Risk in Middle-Market Default Prediction: A Study of the Performance of the RiskCalc and PFM Models. *MKMV Special Comment*, 2003.
- Roger M. Stein. Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation. *Journal of Risk Model Validation*, 1(1):77–113, 2007.
- D. York. Least Squares Fitting of a Straight Line with Correlated Errors. *Earth Science and Planetary Science Letters*, 5:320–324, 1969.

## 6 Appendix A: Some comments on data corruption in the development sample

Although beyond the scope of this paper, we note that the case of data corruption of the development sample represents a more involved problem as it can potentially affect the model parameters themselves. Fortunately, a number of aspects of this problem have been studied extensively in the statistics and econometrics literature, so in many cases, the effects of such mislabeling on model estimation are well understood and fix-ups have been developed for many common problems.

For example, it is well known that in general, random noise to the *independent* variables in regression problems will serve to bias downward the estimates of the model coefficients since the sum of squares of the independent variables ( $\mathbf{X}'\mathbf{X}$ ) will increase faster than the sum of squares of the dependent and independent variables ( $\mathbf{X}'\mathbf{y}$ ) implying that the ratio of the two will decline. This type of bias is often termed regression dilution.

For standard regression, corruption of the *dependent* variable will not bias the estimates of model coefficients, however it will reduce the efficiency of the estimation by adding additional error to the right-hand side. Unbiasedness does not hold though when the dependent variable is limited in certain ways, as it is in the case of a logit or probit specification where  $y_i \in \{0, 1\}$ . In such cases, the coefficients estimated via maximum likelihood will be biased.

As discussed in Section 4.1, there are known methods for correcting for this bias by simultaneously estimating both the model coefficients and the rate of mislabeling in the sample [cf., Hausman et al., 1998].

Finally, if there are “bad” records missing from the development sample, the situation is that of a sample base rate that differs from the true base rate (and is lower). In this case, estimation is still feasible, but the predicted probabilities will be biased downward due to the “lower” prevalence in the estimation data. Elkan [2001] provides a proof that when the base rates of the development sample and the evaluation sample differ, the observed probability  $p_i^*$ , for the  $i^{th}$  observation in the evaluation sample is calculated as:

$$p_i^* = \pi_T \frac{p_i - p_i \pi_S}{\pi_S - p_i \pi_S + p_i \pi_T - \pi_S \pi_T},$$

where  $p_i$  is the raw probability produced by the model,  $\pi_S$  is the baseline prevalence rate in the development sample, and  $\pi_T$  is the baseline prevalence in the evaluation sample. (Note that we adopt the notation used in Bohn and Stein [2009] in preference to that of Elkan [2001].)

Bohn and Stein [2009] provides a discussion of this approach as well as examples.

Note that the observations in most of this Appendix is premised on the data corruption being CAR or MCAR. If it is not the case that one of these assumption holds, alternative approaches are available in some cases, but the estimators typically become much more involved. This is a problem area that has been studied extensively (see, e.g., [York, 1969] for an early reference).

## 7 Appendix B: Recovering an approximate ROC curve (under parametric assumptions)

Some readers may find it useful to construct ROC curves in addition to calculating AUC statistics. It is feasible to generate representative ROC curves that are, under certain assumptions, consistent with the recovered values of the AUC,  $\widehat{A}_r$ . Because, in general, we do not know which records are corrupted (if we did, we would correct them), we cannot easily determine the recovered values of  $(\widehat{FN}_r, \widehat{TP}_r)$ , at individual points on the ROC. However, subject to assumptions about the distribution of the model scores, we can generate a parametric ROC.

If we make the assumption that  $G$  and  $B$  are distributed normally with means  $\mu_G, \mu_B$  and variances  $\sigma_G^2, \sigma_B^2$ , respectively, it can then be shown that

$$ROC(u) = \Phi(a + b\Phi^{-1}(u)), \quad 0 \leq u \leq 1, \quad (15)$$

where  $a = (\mu_B - \mu_G)/\sigma_B^2$ ,  $b = \sigma_G^2/\sigma_B^2$  and  $\Phi(\cdot)$  and  $\Phi^{-1}(\cdot)$  are the cumulative and inverse cumulative normal distribution functions, respectively. It then follows that

$$A = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right). \quad (16)$$

Finally, we can solve for each of either  $a$  or  $b$  in terms of the other (and  $A$ ):

$$a = \Phi^{-1}(A) \times \sqrt{1+b^2} \quad (17)$$

and

$$b = \sqrt{\left(\frac{a}{\Phi^{-1}(A)}\right)^2 - 1}. \quad (18)$$

We *do not* know which records are corrupted so there is little guidance on how to adjust the estimates of the  $\mu$  and  $\sigma^2$  parameters. Though there is no theoretical motivation for them, we present two heuristic approaches for plotting the recovered ROC curve. Both focus on estimating  $a$  based on assumptions about  $b$ .

The first is the simpler of the two and assumes that the variances of  $G$  and  $B$  do not change as a result of the corruption. Thus, we would plug in the estimates of  $\sigma_G^2$  and  $\sigma_B^2$ , from the corrupted data.

The second, and more restrictive, assumes that the variances are equal, and thus that  $b = \sigma_G^2/\sigma_B^2 = 1$ . Then

$$a = \Phi^{-1}(A) \times \sqrt{2}. \quad (19)$$

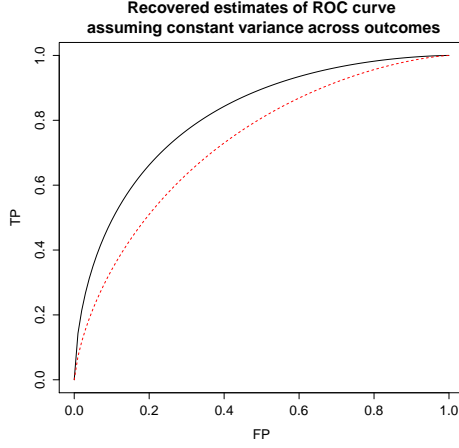


Fig. 6: Recovered estimate of ROC curve based on analysis shown in Example 10. Solid line shows ROC based on recovered value of  $A$  ( $\hat{A}_r$ ), dashed line shows ROC based on original estimate ( $\hat{A}_c$ ). Assumes constant variance ( $\sigma_G^2 = \sigma_B^2$ ).

In either case, the ROC curve may be generated by assuming a value for  $b$ , solving for  $a$ , and then calculating Eq. 15 for values of  $u$  in  $[0, 1]$ .

**Example 10 (Recovering an estimated ROC from corrupted data using an estimate of  $\hat{A}_r$ )** Consider again the analyst from Example 4 who is evaluating a bankruptcy model using a corrupted dataset of anonymized financial statement data and corresponding default flags. In Example 4 the analyst calculated that  $\hat{A}_r = 0.814$  (based on the original estimate of  $\hat{A}_c = 0.73$ ).

If the analyst wished to produce an estimated ROC curve, and were comfortable making the assumption that the variances of the model scores for the defaulting and non-defaulting firms were equal, he could use Eq. 19 to estimate  $a$ , and then use Eq. 15 with  $b = 1$  to generate the ROC curve.

$$\begin{aligned} a &= \Phi^{-1}(A) \times \sqrt{2} \\ &= \Phi^{-1}(0.814) \times \sqrt{2} \\ &\approx 1.26 \\ ROC(u) &= \Phi(1.26 + \Phi^{-1}(u)), \quad 0 \leq u \leq 1. \end{aligned}$$

The resulting curve is shown in Figure 6. For comparison, the same curve, plotted for the original estimate of  $\hat{A}_c = 0.73$  is shown as a dashed line.

■

We note that because of the various assumptions, ROC curves generated in this fashion will likely differ from those that would have been recovered, e.g., nonparametrically, were the corruption mechanism known. However, for some applications, and with caveats about the strong assumptions above, this approach may still provide helpful visual guidance.