Open access • Posted Content • DOI:10.1101/342626

Unbiased Precision Estimation under Separate Sampling — Source link

Shuilian Xie, Shuilian Xie, Shuilian Xie, Shuilian Xie ...+3 more authors

Institutions: Texas A&M University, Texas Tech University, Lanzhou University, Texas College ...+3 more institutions

Published on: 08 Jun 2018 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Estimator, Sample size determination, Precision and recall and Population

Unbiased Precision Estimation under Separate Sampling

Shuilian Xie[1] and Ulisses M. Braga-Neto[1,2*]

1 Department of Electrical and Computer Engineering, Texas A& M University, College Station, TX, USA
2 Center for Bioinformatics and Genomic Systems Biology, Texas A& M University, College Station, TX, USA

*corresponding author (ulisses@ece.tamu.edu)

Abstract

Motivation: Precision and recall have become very popular classification accuracy metrics in the statistical learning literature. These metrics are ordinarily defined under the assumption that the data are sampled randomly from the mixture of the populations. However, observational case-control studies for biomarker discovery often collect data that are sampled separately from the case and control populations, particularly in the case of rare diseases. This discrepancy may introduce severe bias in classifier accuracy estimation.
Results: We demonstrate, using both analytical and numerical methods, that classifier precision estimates can display strong bias under separating sampling, with the bias magnitude depending on the difference between the case prevalences in the data and in the actual population. We show that this bias is systematic in the sense that it cannot be reduced by increasing sample size. If information about the true case prevalence is available from public health records, then a modified precision estimator is proposed that displays smaller bias, which can in fact be reduced to zero as sample size increases under regularity conditions on the classification algorithm. The accuracy of the theoretical analysis and the performance of the proposed precision estimator under separate sampling are investigated using synthetic and real data from observational case-control studies. The results confirmed that the proposed precision estimator indeed becomes unbiased as sample size increases, while the ordinary precision estimator may display large bias, particularly in the case of rare diseases.
Availability: Extra plots are available as Supplementary Materials.

Author summary

Biomedical data are often sampled separately from the case and control populations, particularly in the case of rare diseases. Precision is a popular classification accuracy metric in the statistical learning literature, which implicitly assumes that the data are sampled randomly from the mixture of the populations. In this paper we study the bias of precision under separate sampling using theoretical and numerical methods. We also propose a precision estimator for separate sampling in the case when the prevalence is known from public health records. The results confirmed that the proposed precision estimator becomes unbiased as sample size increases, while the ordinary precision estimator may display large bias, particularly in the case of rare diseases. In the absence of any knowledge about disease prevalence, precision estimates should be avoided under separate sampling.

1 Introduction

Biomarker discovery is typically attempted by means of observational case-control studies where classification techniques are applied to high-throughput measurement technologies, such as DNA microarrays [1], next-generation RNA sequencing (RNA-seq) [2], or "shotgun" mass spectrometry [3]. The validity and reproducibility of the results depend critically on the availability of accurate and unbiased assessment of classification accuracy [4, 5].

The vast majority of published methods in the statistical learning literature make the assumption, explicitly or implicitly, that the data for training and accuracy assessment are sampled randomly, or unrestrictedly, from the mixture of the populations. However, observational case-control studies in biomedicine typically proceed by collecting data that are sampled with restrictions. The most common restriction, and the one that is studied in this paper, is that the data are sampled separately from the case and control populations. This is always true in studies involving rare diseases, since sampling randomly from the population at large would not yield enough cases. That creates an important issue in the application of traditional statistical learning techniques to biomedical data, because there is no meaningful estimator of case prevalences under separate sampling. Therefore, any methodology that directly or indirectly uses estimates of case prevalence will be severely biased.

Precision and *Recall* have become very popular classification accuracy metrics in the statistical learning literature [6–8]. In this paper, we investigate the bias of precision and recall sample estimates when the typical separate sampling design used in case-control studies is not properly taken into account. Synthetic and real-world biomedical data are used to quantify the magnitude of the bias, which is systematic in the sense that it cannot be reduced by increasing sample size. If information about the true prevalence of cases is available, then a modified estimator is proposed that displays smaller bias, which can be decreased to zero asymptotically as sample size increases under certain regularity conditions on the classification algorithm, in a sense to be made precise.

In [9], a similar study was conducted into the accuracy of cross-validation under separate sampling. It was shown that the usual "unbiasedness" property of k-fold cross-validation does not hold under separate sampling. In fact, the bias can in fact be substantial and systematic, i.e., not reducible under increasing sample size. In [9], modified k-fold cross-validation estimators were proposed for the class-specific error rates. In the case where the true case prevalence is known, those estimators can be combined into an estimator of the overall error rate, which satisfies the usual "unbiasedness" property of cross-validation.

The present paper employs analytical and numerical methods to show that the ordinary precision estimator can display large bias under separate sampling. More specifically, while the recall estimator is asymptotically unbiased as sample size increases, under regularity conditions on the classification rule to be specified, the precision estimator may display a systematic bias, which cannot be reduced by increasing sample size if the observed prevalence of cases in the data is different from the true prevalence in the population of interest. This is a consequence of the fact that precision is a function of the prevalence, whereas recall is not. Case-control studies involving rare diseases are specially affected, since in those studies the true prevalence is small and will almost always differs substantially from the observed prevalence in the data. To address this problem, we propose a new estimator for precision, which can be applied in case the true prevalence is known. This estimator has small bias that vanishes as sample size increases under certain regularity conditions. In the absence of any knowledge about the prevalence, precision estimates should be avoided under separate sampling.

2 Materials and Methods

In this section we define and investigate the various error rates of interest in this study, including precision and recall.

2.1 Population Performance Metrics

The *feature* vector $\mathbf{X} \in R^d$ summarizes numerical characteristics of a patient (e.g, blood concentrations of given proteins). The *label* $Y \in \{0,1\}$ is defined as $Y = 0$ if the patient is from the control population, and $Y = 1$ if the patient is from the case population. The *prevalence* is defined by

$$\text{prev} \;=\; P(Y=1)\,, \tag{1}$$

i.e., the probability that a randomly selected individual is a case subject. The prevalence plays a fundamental role in the sequel.

A *classifier* $\psi : R^d \to \{0,1\}$ assigns $\mathbf{X}$ to the control or case population, according to whether $\psi(\mathbf{X}) = 0$ or $\psi(\mathbf{X}) = 1$, respectively. The classification sensitivity and specificity are defined as:

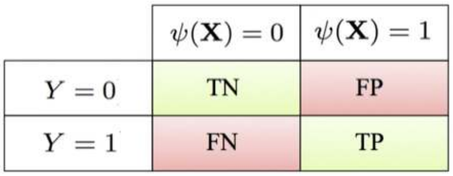
$$\begin{aligned} \text{sens} &= P(\psi(\mathbf{X}) = 1 \mid Y = 1)\,, && (2)\\ \text{spec} &= P(\psi(\mathbf{X}) = 0 \mid Y = 0)\,. && (3) \end{aligned}$$

The closer both are to 1, the more accurate the classifier is. A noteworthy property of the sensitivity and specificity is that they *do not depend on the prevalence.*

Other common performance metrics for a classifier are the *false-positive* (FP), *false-negative* (FN), *true-positive* (FP), and *true-negative* (FN) rates, given by

$$\begin{aligned} \text{FP} &= P(\psi(\mathbf{X}) = 1, Y = 0) &&= (1-\text{spec}) \times (1-\text{prev})\,, && (4)\\ \text{FN} &= P(\psi(\mathbf{X}) = 0, Y = 1) &&= (1-\text{sens}) \times \text{prev}\,, && (5)\\ \text{TP} &= P(\psi(\mathbf{X}) = 1, Y = 1) &&= \text{sens} \times \text{prev}\,, && (6)\\ \text{TN} &= P(\psi(\mathbf{X}) = 0, Y = 0) &&= \text{spec} \times \text{prev}\,. && (7) \end{aligned}$$

Unlike sensitivity and specificity, the previous performance metrics *do* depend on the prevalence. See Fig. 1 for an illustration.



	$\psi(\mathbf{X}) = 0$	$\psi(\mathbf{X}) = 1$
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Fig 1. Diagram of error (red) and accuracy (green) rates.

Notice that

$$\begin{aligned} &\text{prev} \;=\; \text{FN} + \text{TP}, \quad 1 - \text{prev} \;=\; \text{FP} + \text{TN}, && (8)\\ &\text{sens} \;=\; \frac{\text{TP}}{\text{TP}+\text{FN}}\,, \quad \text{spec} \;=\; \frac{\text{TN}}{\text{TN}+\text{FP}}\,. && (9) \end{aligned}$$

Finally, we define the precision and recall accuracy metrics. Precision measures the likelihood that one has a true case given that the classifier outputs a case:

$$\text{prec} \;=\; P(Y = 1 \mid \psi(\mathbf{X}) = 1)\,. \tag{10}$$

Applying Bayes' Theorem and using previously-derived relationships reveal that:

$$\text{prec} \;=\; \frac{\text{TP}}{\text{TP}+\text{FP}} \;=\; \frac{\text{sens}\times\text{prev}}{\text{sens}\times\text{prev}+(1-\text{spec})\times(1-\text{prev})}\,. \tag{11}$$

On the other hand, recall is simply the sensitivity:

$$\text{rec} \;=\; \text{sens} \;=\; \frac{\text{TP}}{\text{TP}+\text{FN}}\,. \tag{12}$$

It follows that precision depends on the prevalence, but recall does not.

2.2 Estimated Performance Metrics

In practice, the performance metrics defined in the previous section need to be estimated from sample data $S_n = \{(\mathbf{X}_1, Y_1), \ldots, \mathbf{X}_n, Y_n)\}$. Let $\widehat{P}$ denote the empirical probability measure defined by S_n. The estimator of prevalence is:

$$\widehat{\text{prev}} \;=\; \widehat{P}(Y=1) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_{Y_i=1}\,, \tag{13}$$

where $I_A = 1$ if A is true and $I_A = 0$ if A is false. Similarly,

$$\widehat{\text{FP}} \;=\; \widehat{P}(\psi(\mathbf{X})=1, Y=0) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=1, Y_i=0\}}\,, \tag{14}$$

$$\widehat{\text{FN}} \;=\; \widehat{P}(\psi(\mathbf{X})=0, Y=1) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=0, Y_i=1\}}\,, \tag{15}$$

$$\widehat{\text{TP}} \;=\; \widehat{P}(\psi(\mathbf{X})=1, Y=1) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=1, Y_i=1\}}\,, \tag{16}$$

$$\widehat{\text{TN}} \;=\; \widehat{P}(\psi(\mathbf{X})=0, Y=0) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=0, Y_i=0\}}\,. \tag{17}$$

The remaining performance metrics estimators are defined analogously, using (9), (11), and (12):

$$\begin{aligned}
\widehat{\text{spec}} &= \frac{\widehat{\text{TN}}}{\widehat{\text{TN}}+\widehat{\text{FP}}} = \frac{\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=0, Y_i=0\}}}{\sum_{i=1}^{n} I_{Y_i=0}}\,,\\
\widehat{\text{prec}} &= \frac{\widehat{\text{TP}}}{\widehat{\text{TP}}+\widehat{\text{FP}}} = \frac{\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=1, Y_i=1\}}}{\sum_{i=1}^{n} I_{\psi(\mathbf{X}_i)=1}}\,,\\
\widehat{\text{rec}} = \widehat{\text{sens}} &= \frac{\widehat{\text{TP}}}{\widehat{\text{TP}}+\widehat{\text{FN}}} = \frac{\sum_{i=1}^{n} I_{\{\psi(\mathbf{X}_i)=1, Y_i=1\}}}{\sum_{i=1}^{n} I_{Y_i=1}}\,.
\end{aligned} \tag{18}$$

2.3 Mixture and Separate Sampling

The usual scenario in Statistical Learning is to assume that $S_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ is an independent and identically distributed (i.i.d.) sample from the true distribution of the pair $(\mathbf{X}, Y)$.That makes S_n a sample from the *mixture* of populations, where each label Y_i is distributed as:

$$P(Y_i = 0) \;=\; 1 - \text{prev} \;\text{ and }\; P(Y_i = 1) = \text{prev}\,, \tag{19}$$

for $i = 1, \ldots, n$. Under mixture sampling, $N_0 = \sum_{i=1}^{n} I_{Y_i=0}$ and $N_1 = \sum_{i=1}^{n} I_{Y_i=1} = n - N_0$ are binomial random variables, with parameters $(n, 1-\text{prev})$ and (n, prev), respectively.

By contrast, observational case-control studies in biomedicine typically proceed by collecting data from the populations separately, where the separate sample sizes n_0 and n_1, with $n_0 + n_1 = n$, are pre-determined and nonrandom; i.e., sampling occurs with the restriction $N_0 = \sum_{i=1}^{n} I_{Y_i=0} = n_0$ (or, equivalently, $N_1 = \sum_{i=1}^{n} I_{Y_i=1} = n_1$). Therefore, all probabilities and expectations over the sample are conditional on $N_0 = n_0$. The restriction means that the labels $Y_1, \ldots, Y_n$ are no longer independent, even though the feature vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are still independent given the labels. It is not difficult to verify that

$$P(Y_i = 0 \mid N_0 = n_0) = \frac{n_0}{n} \text{ and } P(Y_i = 1 \mid N_0 = n_0) = \frac{n_1}{n}\,, \tag{20}$$

for $i = 1, \ldots, n$. Comparing (19) and (20) reveals the main difference between mixture and separate sampling.

2.4 Bias of Precision and Recall Estimators

In this subsection, we present an approximate large-sample analysis of the bias of the estimators discussed previously, focusing on the precision and recall estimators. Estimation bias is defined as the expectation over the sample data S_n of the difference between the estimated and true quantities.

The situation is clear with the estimator of the prevalence itself, given by (13). Under mixture sampling, we have

$$E[\widehat{\text{prev}}] \;=\; \frac{1}{n}\sum_{i=1}^{n} E[I_{Y_i=1}] \;=\; P(Y_1 = 1) \;=\; \text{prev}\,, \tag{21}$$

so the estimator is unbiased (in addition, as n increases, $\text{Var}(\widehat{\text{prev}}) \to 0$ and $\widehat{\text{prev}} \to \text{prev}$ in probability, by the law of large numbers). However, under separate sampling,

$$E[\widehat{\text{prev}} \mid N_0 = n_0] \;=\; \frac{1}{n}\sum_{i=1}^{n} E[I_{Y_i=1} \mid N_0 = n_0] \;=\; P(Y_1 = 1 \mid N_0 = n_0) \;=\; \frac{n_1}{n}\,, \tag{22}$$

according to (20). This also follows directly from the fact that $\widehat{\text{prev}}$ becomes a constant estimator, $\widehat{\text{prev}} \equiv n_1/n$, according to (13). Thus,

$$\text{Bias}_{\text{sep}}(\widehat{\text{prev}}) \;=\; E[\widehat{\text{prev}} - \text{prev} \mid N_0 = n_0] \;=\; \frac{n_1}{n} - \text{prev}\,. \tag{23}$$

Assuming that the ratio n_1/n is held constant as n increases (e.g., the common balanced design case, $n_0 = n_1 = n/2$), then this bias cannot be reduced with increased sample size. Furthermore, the bias is larger the further away the true prevalence is from $\frac{n_1}{n_0}$. In particular, the bias will be large when prev is small, as in case-control studies involving rare diseases.

The situation for $\widehat{\text{FP}}$, $\widehat{\text{FN}}$, $\widehat{\text{FP}}$, and $\widehat{\text{TN}}$ is more complicated. First, we are interested in a classifier ψ_n derived by a classification rule from from the sample data $S_n = \{(\mathbf{X}_1, Y_1), \ldots, \mathbf{X}_n, Y_n)\}$. Therefore, all expectations and probabilities in the previous sections are conditional on S_n. Under mixture sampling, the powerful *Vapnik-Chervonenkis Theorem* can be applied to show that all of these estimators are asymptotically unbiased, provided that classification rule has a finite *VC Dimension* [10]. This includes many useful classification algorithms such as LDA, linear SVMs, perceptrons, polynomial-kernel classifiers, certain decision trees and neural networks, but it excludes nearest-neighbor classifiers, for example. Classification rules with finite VC dimension do not cut the feature space in complex ways and are thus generally robust against overfitting.

Assuming mixture sampling and a classification algorithm with finite VC dimension $V_{\mathcal{C}}$, it can be shown that (details omitted; see [5, p. 47] for a similar argument)

$$\mathrm{Bias}_{\mathrm{mix}}(\widehat{\mathrm{FP}}) \,\leq\, 8\sqrt{\frac{V_{\mathcal{C}}\log(n+1)+4}{2n}}\,, \tag{24}$$

so that the bias vanishes as $n \to \infty$. Similar inequalities apply to $\widehat{\mathrm{FN}}$, $\widehat{\mathrm{FP}}$, and $\widehat{\mathrm{TN}}$. These are distribution-free results, hence vanishingly small bias is guaranteed if $n \gg V_{\mathcal{C}}$, regardless of the feature-label distribution. For linear classification rules, $V_{\mathcal{C}} = d+1$, where d is the dimensionality of the feature vector. In this case, the $\widehat{\mathrm{FP}}$, $\widehat{\mathrm{FN}}$, $\widehat{\mathrm{FP}}$, and $\widehat{\mathrm{TN}}$ estimators are essentially unbiased if $n \gg d$.

Next we consider the bias of the precision and recall estimators under mixture sampling (the analysis for the sensitivity and specificity estimators is similar; in fact, the latter is just the recall estimator). We will make use of the following relation for the expectation of a ratio of two random variables W and Z:

$$E\left[\frac{W}{Z}\right] \,=\, \frac{E[W]}{E[Z]} + \text{second and higher order terms.} \tag{25}$$

This equation can be proved by expanding W/Z around the point $(E[W],E[Z])$ using a bivariate Taylor series and taking expectation (details are omitted for space). The approximation obtained by dropping the higher order terms is quite accurate if W and Z are around $E[W]$ and $E[Z]$, respectively (it is asymptotically exact as $W \to E[W]$ and $Z \to E[Z]$). For the precision estimator,

$$E[\widehat{\mathrm{prec}}] = E\left[\frac{\widehat{\mathrm{TP}}}{\widehat{\mathrm{TP}}+\widehat{\mathrm{FP}}}\right] \approx \frac{E[\widehat{\mathrm{TP}}]}{E[\widehat{\mathrm{TP}}+\widehat{\mathrm{FP}}]} \approx \frac{E[\mathrm{TP}]}{E[\mathrm{TP}+\mathrm{FP}]} \approx E\left[\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\right] = E[\mathrm{prec}], \tag{26}$$

for a sufficiently large sample, where we used the previously-established asymptotic unbiasedness of $\widehat{\mathrm{TP}}$, $\widehat{\mathrm{TP}}$, and $\widehat{\mathrm{FN}}$. An entirely similar derivation shows that $E[\widehat{\mathrm{rec}}] = E[\mathrm{rec}]$. Hence, for "well-behaved" classification algorithms (those with finite VC dimension), both the precision and recall estimators are asymptotically unbiased under mixture sampling.

Unfortunately, there is not at this time a version of VC theory for separate sampling. In order to obtain approximate results for the separate sampling case, we will assume instead that, at large enough sample sizes, the classifier ψ is nearly constant, and invariant to the sample. This assumption is not unrelated to the finite VC dimension assumption made in the case of mixture sampling. Many of the same classification algorithms that have finite VC dimension, such as LDA and linear SVMs, will also become nearly constant as sample size increases. In this case, we have

$$\begin{aligned} E[\widehat{\mathrm{TP}} \mid N_0 = n_0] \,&=\, \frac{1}{n}\sum_{i=1}^{n} E[I_{\{\psi(X_i)=1,Y_i=1\}} \mid N_0 = n_0] \\ &=\, P(\psi(\mathbf{X}_1)=1, Y_1=1 \mid N_0 = n_0) \\ &=\, P(\psi(\mathbf{X}_1)=1 \mid Y_1=1)P(Y_1=1 \mid N_0=n_0) \,=\, \mathrm{sens}\times\frac{n_1}{n}\,, \end{aligned} \tag{27}$$

where we used the fact that the event $\{\psi(\mathbf{X}_1)=1\}$ is independent of N_0 given Y_1 and (20). Notice that the equality $P(\psi(\mathbf{X}_1)=1 \mid Y_1=1) = \mathrm{sens}$ depends on the fact that ψ is assumed to be constant, so that $(\mathbf{X}_1, Y_1)$ behaves as an independent test point (also because of a constant ψ, there is no expectation around sens). Hence, $\widehat{\mathrm{TP}}$ is biased under separate sampling, with

$$\mathrm{Bias}_{\mathrm{sep}}(\widehat{\mathrm{TP}}) \,=\, \mathrm{sens}\times\frac{n_1}{n} - \mathrm{TP} \,=\, \mathrm{sens}\times\left(\frac{n_1}{n} - \mathrm{prev}\right). \tag{28}$$

As in the case with the bias of $\widehat{\text{prev}}$ under separate sampling, the bias of $\widehat{\text{TP}}$ cannot be reduced with increasing sample size. The bias is in fact larger the more sensitive the classifier is. One can derive similar results for $\widehat{\text{FP}}$, $\widehat{\text{FN}}$, and $\widehat{\text{TN}}$.

Perhaps surprisingly, the recall estimator is approximately unbiased under separate sampling:

$$\begin{aligned} E[\widehat{\text{rec}} \mid N_0 = n_0] \;&=\; E\left[\frac{\widehat{\text{TN}}}{\widehat{\text{TN}}+\widehat{\text{FP}}}\middle| N_0 = n_0\right] \;=\; E\left[\frac{\widehat{\text{TP}}}{\widehat{\text{prev}}}\middle| N_0 = n_0\right] \\ &= \frac{E[\widehat{\text{TP}} \mid N_0 = n_0]}{\frac{n_1}{n}} \;=\; \frac{\text{sens} \times \frac{n_1}{n}}{\frac{n_1}{n}} \;=\; \text{sens} = \text{rec}\,. \end{aligned} \tag{29}$$

This is a consequence of recall's not being a function of the prevalence. However, for the precision estimator,

$$\begin{aligned} E[\widehat{\text{prec}} \mid N_0 = n_0] \;&=\; E\left[\frac{\widehat{\text{TP}}}{\widehat{\text{TP}}+\widehat{\text{FP}}}\middle| N_0 = n_0\right] \\ &\approx \frac{E[\widehat{\text{TP}} \mid N_0 = n_0]}{E[\widehat{\text{TP}}+\widehat{\text{FP}} \mid N_0 = n_0]} \;=\; \frac{\text{sens} \times \frac{n_1}{n}}{\text{sens} \times \frac{n_1}{n} + (1-\text{spec}) \times \frac{n_0}{n}} \\ &\neq \frac{\text{sens} \times \text{prev}}{\text{sens} \times \text{prev} + (1-\text{spec}) \times (1-\text{prev})} \;=\; \text{prec}\,. \end{aligned} \tag{30}$$

The precision estimator is thus biased under separate sampling, and the bias is larger the further away the true prevalence is from $\frac{n_1}{n_0}$. In particular, the bias will be large when prev is small, which is the case in case-control studies involving rare diseases.

2.5 Proposed Precision Estimator

In case the true prevalence is known, e.g., from public health records and government databases, then we propose the following estimator of the precision, based on (11):

$$\widehat{\text{prec}}^{\text{new}} \;=\; \frac{\widehat{\text{sens}} \times \text{prev}}{\widehat{\text{sens}} \times \text{prev} + (1-\widehat{\text{spec}}) \times (1-\text{prev})}\,. \tag{31}$$

This estimator is still asymptotically unbiased under mixture sampling (which can be seen by repeating the steps in the analysis of the ordinary precision estimator). Under separate sampling, we have

$$\begin{aligned} &E[\widehat{\text{prec}}^{\text{new}} \mid N_0 = n_0] \\ &\approx \frac{E[\widehat{\text{sens}} \mid N_0 = n_0] \times \text{prev}}{E[\widehat{\text{sens}} \mid N_0 = n_0] \times \text{prev} + (1-E[\widehat{\text{spec}} \mid N_0 = n_0]) \times (1-\text{prev})} \\ &= \frac{\text{sens} \times \text{prev}}{\text{sens} \times \text{prev} + (1-\text{spec}) \times (1-\text{prev})} \;=\; \text{prec}\,. \end{aligned} \tag{32}$$

since $E[\widehat{\text{sens}} \mid N_0 = n_0] = \text{sens}$ and $E[\widehat{\text{spec}} \mid N_0 = n_0] = \text{spec}$, as can be easily shown. Hence, $\widehat{\text{prec}}^{\text{new}}$ is an asymptotically unbiased estimator of the precision under either mixture or separate sampling. The ordinary precision estimator $\widehat{\text{prec}}$ should never be used under separate sampling, or large and irreducible bias may occur. On the other hand, in the impossibility of obtaining information on the true prevalence value, then no meaningful estimator of the precision is possible.

3 RESULTS AND DISCUSSION

In this section, we employ synthetic and real-world data to investigate the performance of the proposed precision estimator and the accuracy of the theoretical analysis in Section 2.4. We present results for the bias of the usual and proposed precision estimators under separate sampling. Corresponding results for mixture sampling and the recall estimator can be found in the Supplementary Material.

3.1 Experiments with Synthetic Data

We performed a set of experiments employing synthetic models with class-conditional 3-dimensional Gaussian distributions $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, for $i = 0, 1$, with $\boldsymbol{\mu}_0 = (0, 0, 0)$, $\boldsymbol{\mu}_1 = (0, 0, \theta)$, where $\theta > 0$ is a parameter governing the separation between the classes, and $\boldsymbol{\Sigma_0} = \boldsymbol{\Sigma_1} = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ (i.e., a matrix with $\sigma_1^2, \sigma_2^2, \sigma_3^2$ on the diagonal and zeros off diagonal). We consider two sample sizes, $n = 30$ and $n = 200$, so that we can compare the results for small and large sample sizes. All experiments with separate sampling are performed with sample prevalence $r = \frac{n_1}{n} \in [0.1, 0.9]$, where the value of n_1 is set to $n_1 = \lceil nr \rceil$. The synthetic data parameters are summarized in Table 1.

Parameter	Value
Dimensionality/ feature size	$D = 3$
Mean difference	$\theta = 2$
Covariance matrix	$\sigma_1^2 = 0.5, \sigma_2^2 = 0.5, \sigma_3^2 = 1$
Sample size	$n = 30, 200$
Sample prevalence n_1/n	$r = 0.1, 0.3, 0.5, 0.7, 0.9$
Actual prevalence	$\text{prev} = 0.1, 0.3, 0.5, 0.7, 0.9$

Table 1. Synthetic data parameters.

For each value of r and prev, we repeat the following process 1,000 times, and average the results to estimate expected error values:

1. Generate sample data S_n of size n according to r (separate sampling) or prev (mixture sampling);
2. Train a classifier using one of three classification rules [11]: Linear Discriminant Analysis (LDA), 3-Nearest Neighbors (3NN), and a nonlinear Radial-Basis Function Support Vector Machine (RBF-SVM).
3. Obtain recall and precision estimates. Compute both the usual precision estimate $\widehat{\text{prec}}$ and the proposed $\widehat{\text{prec}}^{\text{new}}$.
4. Obtain (good approximations of) the true precision values, by using a test set of size 10,000.

Fig. 2 displays the results of the experiment. Notice that there is only one curve for the traditional precision estimator $\widehat{\text{prec}}$ because it does not employ the actual value of prev. The results show that at $n = 30$, all estimators display bias, which is however much larger for the traditional precision estimator. At $n = 200$, the bias of the proposed precision estimator nearly disappears for LDA and is reduced for the other classification rules. Among these classification rules, LDA is the only one with a finite VC dimension, and so the bias in this case is predicted to shrink to zero as sample size increases, according to the theoretical analysis in Section 2.4. Notice also that the bias of the traditional precision estimator is largest when $r = n_1/n$ is far from prev, and it cannot be reduced by increasing sample size. All these observations are in agreement with

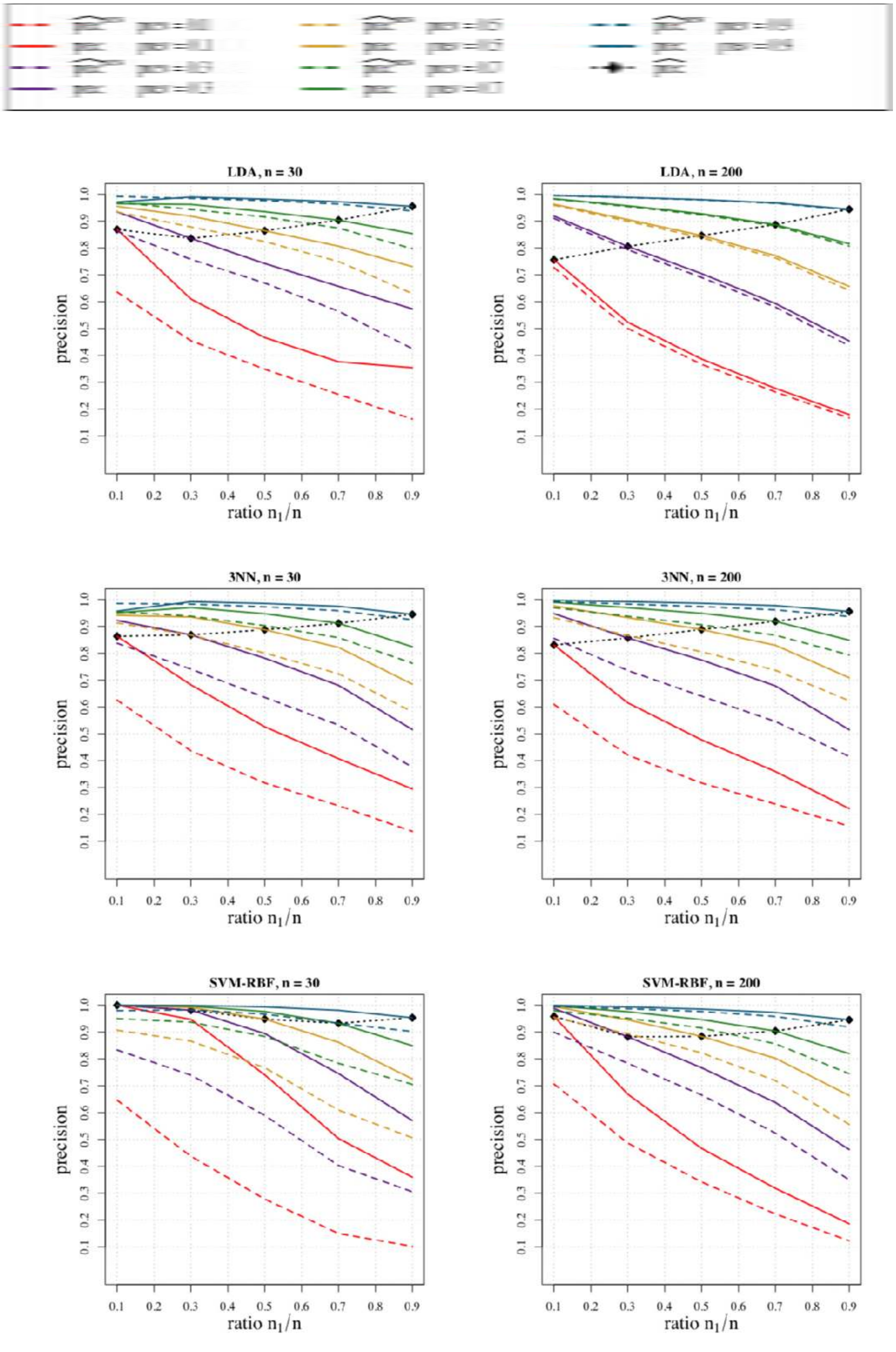


Fig 2. Average true precision values (solid curves) and average precision estimates $\widehat{\text{prec}}$ (dash-diamond curve) and $\widehat{\text{prec}}^{\text{new}}$ (dashed curves), for LDA, 3NN and RBF-SVM, sample sizes $n = 30$ (top row) and $n = 200$ (bottom row) and different prevalence values as a function of the sample prevalence $r = n_1/n$.

the theoretical analysis in Section 2.4 (the results in the Supplementary Material also confirm the theoretical analysis).

3.2 Two Case Studies with Real Data

To further investigate the bias of precision estimation under separate bias, we use real data from two published studies. The first [12] uses a tumor microarray dataset containing two types of human acute leukemia: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Gene expression measurements were taken from $15,154$ genes for 72 tissue specimens, 47 ALL type (class 0) and 25 AML type (class 1), so that the sample prevalence is $r = 0.347$. We computed the traditional the precision estimator $\widehat{\text{prec}}$, and the proposed estimator $\widehat{\text{prec}}^{\text{new}}$ by using the value $\text{prev} = 0.222$, which is the incidence rate of ALL over AML in the U.S. population [13], for four classification rules: Naive Bayes (NB) [14], C4.5 decision tree [15], 3NN and SVM. Fig. 3 displays the results. We can observe that all $\widehat{\text{prec}}$ estimates are larger than the more precise $\widehat{\text{prec}}^{\text{new}}$ estimates, pointing to an optimistic bias of the usual precision estimator.

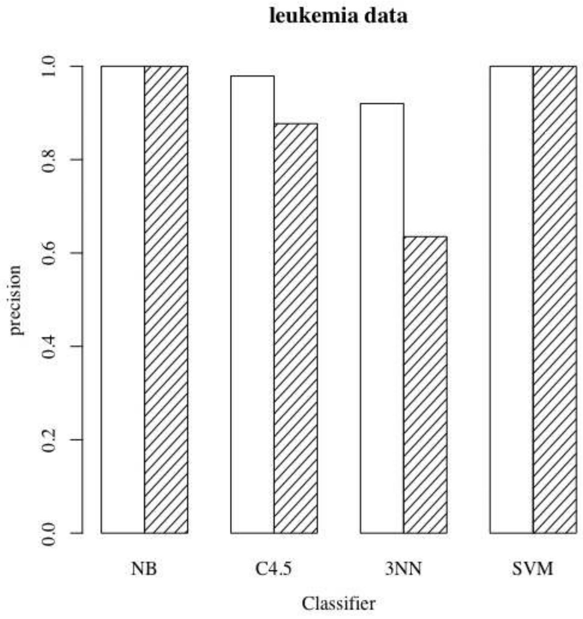


Fig 3. The white bars are the usual estimated precision on the separately-sampled ALL-AML dataset in [12] for four classification rules; the shaded bars are the precision estimates using the prevalence of ALL over AML in the U.S. population.

In the second case study, we employ the liver disease dataset in [16], taken from University of California-Irvine (UCI) Machine Learning Repository [17]. This data set contains 5 blood tests attributes and 345 records in which 145 belong to individuals with liver disease (class 0) and 200 measurements are taken from healthy individuals (class 1), so that $r = 0.42$. The authors in [16] constructed classifiers for diagnosis of liver disease based on the blood tests variables in this data set, and reported the estimated accuracy, precision, sensitivity and specificity for five classification rules: Naive Bayes (NB), C4.5, 3NN, Back-Propagated Neural Network [11] and a Linear SVM. This dataset was donated to UCI in 1990, and according to [18] , the prevalence rate for chronic liver diseases in the US was $\text{prev} = 0.1178$ between 1988 and 1994, which we use as the approximated prevalence in the computation of the $\widehat{\text{prec}}^{\text{new}}$ estimator. Fig. 4 displays the results. As in the previous study, we observe that all $\widehat{\text{prec}}$ are larger than the more precise $\widehat{\text{prec}}^{\text{new}}$ estimates, but this time the difference is much larger, indicating possible strong optimistic bias of the traditional precision estimator.

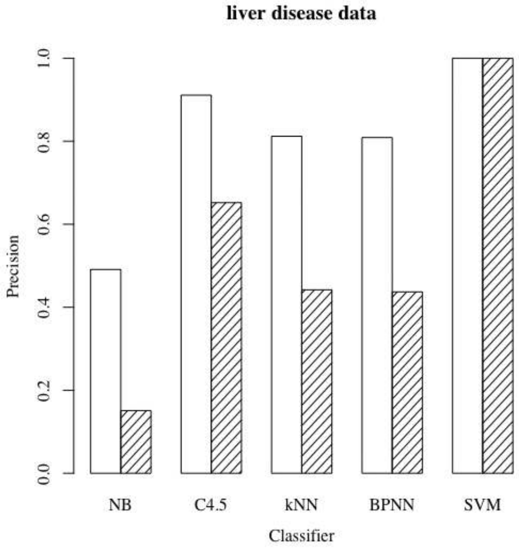


Fig 4. The white bars are the usual estimated precision on the separately-sampled liver disease dataset in [16] for four classification rules; the shaded bars are the precision estimates using the prevalence of liver disease in the U.S. population.

4 Concluding Remarks

Accuracy and reproducibility in observational studies is critical to the progress of biomedicine, in particular, in the discovery of reliable biomarkers for disease diagnosis and prognosis. In this study, we showed, using analytical and numerical methods, that the usual estimator of precision can be severely biased under the typical separate sampling scenario in observational case-control studies. This will be true especially in the case of rare diseases, when the true disease prevalence will be small and differ significantly from the apparent prevalence in the data. If knowledge of the true disease prevalence is available, or can even be approximately ascertained, then it can be used to define a new precision estimator proposed here, which is nearly unbiased at moderate sample sizes. Absence of knowledge about the true prevalence means simply that the precision cannot be reliably estimated in observational case-control studies.

Acknowledgments

The authors acknowledge the support of the National Science Foundation, through NSF award CCF-1718924.

References

1. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Journal of Biomedical optics. 1997 Oct;2(4):364-75.
2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008 Jul;5(7):621. American Journal of Roentgenology. 2017 Apr;208(4):750-3.
3. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003 Mar 13;422(6928):198.
4. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification?. Bioinformatics. 2004 Feb 12;20(3):374-80.
5. Braga-Neto UM, Dougherty ER. Error estimation for pattern recognition. John Wiley & Sons; 2015 Jul 7.
6. Ong MS, Magrabi F, Coiera E. Automated categorisation of clinical incident reports using statistical text classification. Qual Saf Health Care. 2010 Dec 1;19(6):e55-.
7. Dang HX, Lawrence CB. Allerdictor: fast allergen prediction using text classification techniques. Bioinformatics. 2014 Jan 7;30(8):1120-8.
8. Hassanpour S, Langlotz CP, Amrhein TJ, Befera NT, Lungren MP. Performance of a machine learning classifier of knee MRI reports in two large academic radiology practices: a tool to estimate diagnostic yield.
9. Braga-Neto UM, Zollanvari A, Dougherty ER. Cross-validation under separate sampling: strong bias and how to correct it. Bioinformatics. 2014 Aug 13;30(23):3349-55.
10. Devroye L, Györfi L, Lugosi G. A probabilistic theory of pattern recognition. Springer Science & Business Media; 2013 Nov 27.
11. Duda RO, Hart PE, Stork P. Pattern Classification, 2nd Ed. John Wiley & Sons, New York; 2001.
12. Hewett R, Kijsanayothin P. Tumor classification ranking from microarray data. BMC genomics. 2008 Sep;9(2):S21.
13. Howlader N and Noone AM and Krapcho M and Miller D and Bishop K and Altekruse SF and Kosary CL and Yu M and Ruhl J and Tatalovich Z and Mariotto A and Lewis DR and Chen HS and Feuer EJ and Cronin KA (eds) SEER Statistics Review, 1975-2013, National Cancer Institute Bethesda, MD, http://seer.cancer.gov/csr/1975_2013/, based on November 2015 SEER data submission, posted to the SEER web site, April 2016
14. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine learning. 1997 Nov 1;29(2-3):131-63.
15. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning. 2000 Aug 1;40(2):139-57.

16. Ramana BV, Babu MS, Venkateswarlu NB. A critical study of selected classification algorithms for liver disease diagnosis. International Journal of Database Management Systems. 2011 May 2;3(2):101-14.

17. UCI repository of machine learning databases BUPA Liver Disorders Dataset Available from ftp://ftp.ics.uci.edu/pub/machine-learningdatabasxes/liverdisorders/bupa.data, last accessed: 07 October 2010

18. Younossi ZM, Stepanova M, Afendy M, Fang Y, Younossi Y, Mir H, Srishord M. Changes in the prevalence of the most common causes of chronic liver diseases in the United States from 1988 to 2008. Clinical Gastroenterology and Hepatology. 2011 Jun 1;9(6):524-30.