Imperfect gold standard gene sets yield inaccurate evaluation of causal gene identification methods

Lijia Wang, Xiaoquan Wen, Jean Morrison

Department of Biostatistics, University of Michigan

Abstract. Causal gene discovery methods are often evaluated using gold-standard (GS) sets of causal genes. However, GS gene sets are always incomplete, leading to mis-estimation of sensitivity, specificity, AUC. Labeling biases in GS gene sets can also lead to inaccurate ordering of discovery methods. We argue that evaluation of these methods should rely on statistical techniques like those used for variant discovery, rather than on comparison with GS gene sets.

Identifying causal genes for complex diseases can highlight disease-specific dysregulated pathways, improve disease classification, and identify drug targets [3]. In genetic association analysis, it has become a common practice to implicate putative causal genes (PCG) computationally by linking variant-level genetic association evidence and the existing biological knowledge base [7]. Some of those computational methods approach PCG implication as a supervised learning problem, aiming to predict unknown binary causal/non-causal gene labels for a given trait. Others rank the candidate genes by their likelihood of being PCGs, returning a continuous probability estimate or ranking for each gene. In this article, we focus on the common practice of evaluating PCG implication methods in reference to known (gold standard, GS) sets of causal genes. A critical challenge for this assessment strategy is that all GS gene sets are incomplete and known causal genes may differ meaningfully from as-yet unidentified causal genes. We show that when the GS gene set is incomplete, estimates of power, specificity, receiver operating characteristic (ROC) will be inaccurate and may even misorder the relative quality of two different classifiers. This phenomenon can occur even if the GS gene set contains no erroneously labeled non-causal genes. We argue that no true GS sets of labeled genes are currently available and therefore urge caution in interpreting comparisons of causal gene classifiers made based on existing sets of labels.

Table 1 summarizes major GS gene set curation methods. We note that the labeled causal genes by these methods are typically reliable, as the standards for causality claim are usually stringent. However, some causal genes will inevitably be missed (thus motivating continued PCG discovery research). When the curated GS gene sets are used as reference sets for evaluation, genes not labeled as positive are implicitly treated as non-causal. However, the more confident we are in the positive labels of a GS gene set, the more mislabeled non-causal genes we can expect [7]. GS gene sets also tend to favor genes with particular features determined by the method of constructing the set. For example, GS gene sets derived from the set of causal coding variants favor genes that act through protein-coding changes rather than expression regulatory mechanisms [8]. Most classification methods also have gene-feature-related biases due to the type of data they use as input. A PCG implication method will appear more accurate if it is evaluated using a GS gene set with similar feature-related biases to its own and less accurate if the GS gene set has different biases. Authors naturally select a GS gene set constructed using features they feel are important and may, therefore, unintentionally tilt the scales toward their own proposed method.

A more accurate view of genes outside the constructed GS set is not as negatives (N) but as unlabeled (U). Combined with accurately positively labeled genes in the GS set, the overall GS gene set should be regarded as positive-unlabeled (PU) data, a term used in semi-supervised machine learning. Using PU data to evaluate performance as though they were positive-negative (PN) labeled data results in inaccurate evaluations [4]. A PU-labeled gene set with perfect positive labeling consists of three subsets of genes: true causal genes that are correctly identified (labeled positives, LP), true causal genes that are not labeled and therefore assumed to be non-causal (unlabeled positives, UP), and non-causal genes that are unlabeled and therefore correctly assumed to be non-causal (unlabeled negatives, UN) (Fig. 1a). Evaluation treating PU labels (with perfect positive labeling) as PN labels will always underestimate the precision (or positive predictive value) and overestimate the negative predictive value (NPV) of a classifier (Fig. 1b). If we further assume that inaccuracies in the PU data are random with respect to features that influence classification accuracy,

evaluation based on PU-labeled data will correctly estimate sensitivity and overestimate specificity (Fig. A.1a). However, if LP and UP genes differ on important features, sensitivity and specificity may be either over- or under-estimated (Appendix A). In genetics research, we expect labeling biases because there are multiple molecular mechanisms by which a causal gene can affect complex diseases, and different classification and GS identification methods will favor different mechanisms. Error in estimating sensitivity and specificity results in error in the ROC curve and therefore error in the area under the ROC curve (AUC). This means that this error applies to evaluation of ranking methods as well as methods that return only hard classifications.

To illustrate this issue, we consider a hypothetical example in which causal genes fall into two functional categories, expression-acting (gene-to-trait effect mediated by RNA expression level) and protein-acting (gene-to-trait effect mediated by variation in protein products). Only protein-acting genes are labeled positive in the GS gene set, mimicking that most LP genes are Mendelian genes in current practice. While LP genes are all protein-acting, UP genes are a mixture of protein-acting and expression-acting genes (Fig. 1a). We then consider two hypothetical PCG classifiers. The first, based on expression-related features (E-Classifier), has higher sensitivity to detect expression-acting genes; the second, based on protein-related features (P-Classifier), has a higher sensitivity to detect protein-acting causal genes. Figure 1c shows that the ROC curve for the E-Classifier is below the true ROC curve for this classifier, resulting in a downward bias in estimated AUC. In comparison, the evaluation for the P-Classifier, is overly optimistic. In this particular setting, the E-Classifier has a better overall ability to identify PCGs than the P-Classifier but is evaluated being much worse.

Due to the myriad biological mechanisms leading to complex phenotypes, it is currently impossible to confidently determine a comprehensive GS gene set that includes all causal genes for any trait. Several studies have acknowledged that supervised ML methods designed to classify PCGs should not be trained on incomprehensive GS gene sets [2], [5]. Here, we want to draw attention to the fact that sensitivity and specificity estimated using incomplete labels are also inaccurate, making it inappropriate to compare and evaluate methods using these measures or AUC estimated using GS gene sets. To address a similar issue in other fields, researchers have proposed incorporating negative controls (i.e., known negatives) or weights estimating each unit's probability of being detected based on its features ([6], [1]). However, whether either approach is feasible for the PCG implication problem is unclear. An alternative approach that circumvents the issue is to use a statistical model-based approach for causal gene identification. Methods based on estimating parameters in probabilistic models provide model-based measures of uncertainty, such as posterior inclusion probabilities or confidence intervals. These methods can also be evaluated in simulations to test their sensitivity to violations of modeling assumptions. Such a paradigm has been established as a long-standing practice in genetics research to identify causal variants. Finally, we re-emphasize that researchers should avoid concluding the relative performance of different classifiers based on performance measures estimated using PU-labeled gene sets and should clearly acknowledge the limitations of any labeled gene set used in evaluation or training.

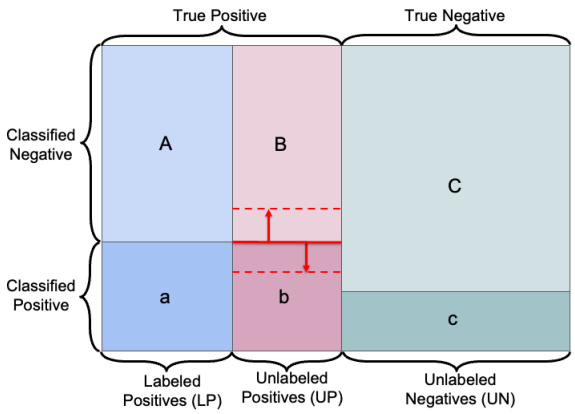
Simulation procedure and results can be found at https://lijiaw.gitlab.io/GS-gene-sets/comparison.html.

Construction scheme	**Publication**
Genes harboring large-effect-size coding variants for selected phenotypes derived from OMIM and additional literature.	Connally et. al. (2021) [S1]
Genes that are targets of any known drug against the disease of interest (reported in Open Targets).	Picart-Armada et. al. (2019) [S7]
Genes annotated to phenotypes of interest in databases (OMIM, GO, DrugBank, literature derived drug targets)	Greene et. al. (2015) [S3] Tranchevent et. al. (2016) [S8]
Phenotype-associated genes compiled from GWAS catalog mining, publication-based, and expert curation.	Kolosov et. al (2021) [S5] [1]
Genes derived from expert domain knowledge, drug target-disease pairs, experimental alteration, and functional data.	Mountjoy et. al. (2021) [S6]
Genes with fine-mapped protein coding variants within loci non-coding credible sets.	Weeks et. al. (2020) [S9]
Top 10% genes ranked by PoPS for validation.	Gazal et. al. (2022) [S2]
Phenotype associated genes in recent reports from the Broad, TOPMed consortium, and IntOGen project.	Kessler et. al. (2021) [S4] [2]

Table 1: **Constructions of Gold Standard gene sets** All of the listed methods used their constructed gold standard gene sets to compare the precision and recall of classification methods, or evaluate the AUC of ranking methods.

[1] This article adopts a PU scheme when using the GS gene set in benchmarking classification methods.

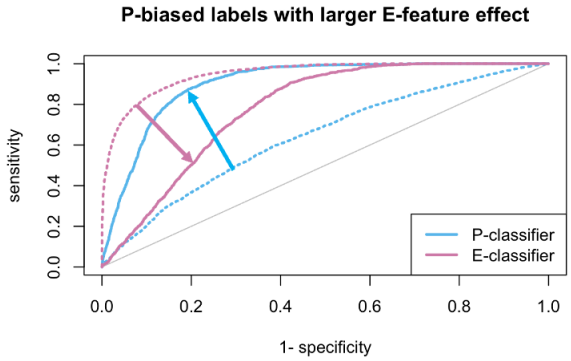
[2] Causal genes selected here are not used to evaluate classification methods but to narrow down putative causal variants.



(a) Illustration of mislabeled data

	True Value	Measured Value
Sensitivity	$\frac{a+b}{a+b+A+B}$	$\frac{a}{A+a}$
Specificity	$\frac{C}{C+c}$	$\frac{B+C}{b+c+B+C}$
PPV	$\frac{a+b}{a+b+c}$	$\frac{a}{a+b+c}$
NPV	$\frac{C}{A+B+C}$	$\frac{B+C}{A+B+C}$

(b) Benchmark metric calculations



(c) Classifier ROC comparison

Fig. 1: **Comparison of Classifiers on True vs. Biased Labels** (a) Illustration of classifier performance relative to a typical GS gene set. The classifier may not have the same performance on UP genes as it has on LP genes (red lines and arrows), which affects estimated sensitivity and specificity. (b) Calculation for true vs. measured accuracy in mislabeled data. PPV stands for positive predictive value and NPV stands for negative predictive value. (c) ROC curves for simulated classifiers. Dotted lines show true classifier performance, solid lines show performance estimated using an incomplete GS gene set. Arrows indicate the difference between true and estimated curves.

References

1. Marthinus C Du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems*, 27, 2014.
2. Marlena Duda, Hongjiu Zhang, Hong-Dong Li, Dennis P. Wall, Margit Burmeister, and Yuanfang Guan. Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. *Translational Psychiatry*, 8(1):1–9, March 2018. Number: 1 Publisher: Nature Publishing Group.
3. Farhad Hormozdiari, Gleb Kichaev, Wen-Yun Yang, Bogdan Pasaniuc, and Eleazar Eskin. Identification of causal genes for complex traits. *Bioinformatics*, 31(12):i206–i213, June 2015.
4. Nikita Kolosov, Mark J. Daly, and Mykyta Artomov. Prioritization of disease genes from GWAS using ensemble-based positive-unlabeled learning. *European Journal of Human Genetics*, 29(10):1527–1535, October 2021. Number: 10 Publisher: Nature Publishing Group.
5. Arjun Krishnan, Ran Zhang, Victoria Yao, Chandra L. Theesfeld, Aaron K. Wong, Alicja Tadych, Natalia Volfovsky, Alan Packer, Alex Lash, and Olga G. Troyanskaya. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature Neuroscience*, 19(11):1454–1462, November 2016. Number: 11 Publisher: Nature Publishing Group.
6. Lu Liu and Tao Peng. Clustering-based method for positive and unlabeled text categorization enhanced by improved tfidf. *J. Inf. Sci. Eng.*, 30(5):1463–1481, 2014.
7. Sergio Picart-Armada, Steven J. Barrett, David R. Willé, Alexandre Perera-Lluna, Alex Gutteridge, and Benoit H. Dessailly. Benchmarking network propagation methods for disease gene identification. *PLOS Computational Biology*, 15(9):e1007276, September 2019. Publisher: Public Library of Science.
8. Elle M. Weeks, Jacob C. Ulirsch, Nathan Y. Cheng, Brian L. Trippe, Rebecca S. Fine, Jenkai Miao, Tejal A. Patwardhan, Masahiro Kanai, Joseph Nasser, Charles P. Fulco, Katherine C. Tashman, Francois Aguet, Taibo Li, Jose Ordovas-Montanes, Christopher S. Smillie, Moshe Biton, Alex K. Shalek, Ashwin N. Ananthakrishnan, Ramnik J. Xavier, Aviv Regev, Rajat M. Gupta, Kasper Lage, Kristin G. Ardlie, Joel N. Hirschhorn, Eric S. Lander, Jesse M. Engreitz, and Hilary K. Finucane. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases, September 2020. Pages: 2020.09.08.20190561.

1 Appendix

A Illustrating Effect of Mislabeling on Classification Accuracy

Receiver operating characteristics (ROC) analysis is one of the most common diagnostic method for binary classifications, and is constructed using a combination of sensitivity and specificity estimations. In this section we will demonstrate how mislabeled positive-negative (PN) data affects the observed sensitivity and specificity compared to perfectly labeled data.

A set of binary, labeled data is illustrated in figure A.1, where a portion of the true positives are labeled as negative. By definition, key accuracy measures can be calculated in figure 1b. We can easily establish that:

$$\begin{aligned} \mathrm{PPV}_T &\geq \mathrm{PPV}_M, \\ \mathrm{NPV}_T &\leq \mathrm{NPV}_M \end{aligned} \tag{1}$$

always hold. Here PPV_T denotes positive predictive value of classifier on true data labels, where PPV_M denotes PPV of classifier using incorrect labels. Additionally, if the mislabeling among true positives occurs randomly, then we have:

$$\frac{a}{A+a} = \frac{b}{B+b}. \tag{2}$$

Setting $\gamma = \frac{B+b}{A+a}$ as the raio of mislabeled positives, we have:

$$\begin{aligned} B+b &= \gamma(A+a), \\ \text{sensitivity} = \frac{a+b}{A+a+B+b} &= \frac{(1+\gamma)a}{(1+\gamma)(A+a)} = \frac{a}{A+a}. \end{aligned} \tag{3}$$

This implies that sensitivity estimate is accurate when error is complete random. However, if the classifier has some discrimination among the mislabeled genes, then:

$$\frac{b}{B+b} > \frac{c}{C+c}. \tag{4}$$

Setting $\alpha = \frac{B+b}{C+c}$ as the ratio of mislabeled negatives, we have:

$$\begin{aligned} \frac{B}{B+b} &< \frac{C}{C+c}, \\ B &< C\alpha, \\ \frac{B+C}{B+b+C+c} = \frac{B+C}{(1+\alpha)(C+c)} &< \frac{(1+\alpha)C}{(1+\alpha)(C+c)} = \frac{c}{C+c}. \end{aligned} \tag{5}$$

This indicates that specificity will be underestimated as long as the classifier has any discriminating ability among the mislabeled genes.

We further illustrate a few different other mislabeling scenarios:

When classifier among mislabeled genes performs worse than correctly-labeled true positives, but better than on the correctly-labeled true negatives (Fig. A.1b), we would have:

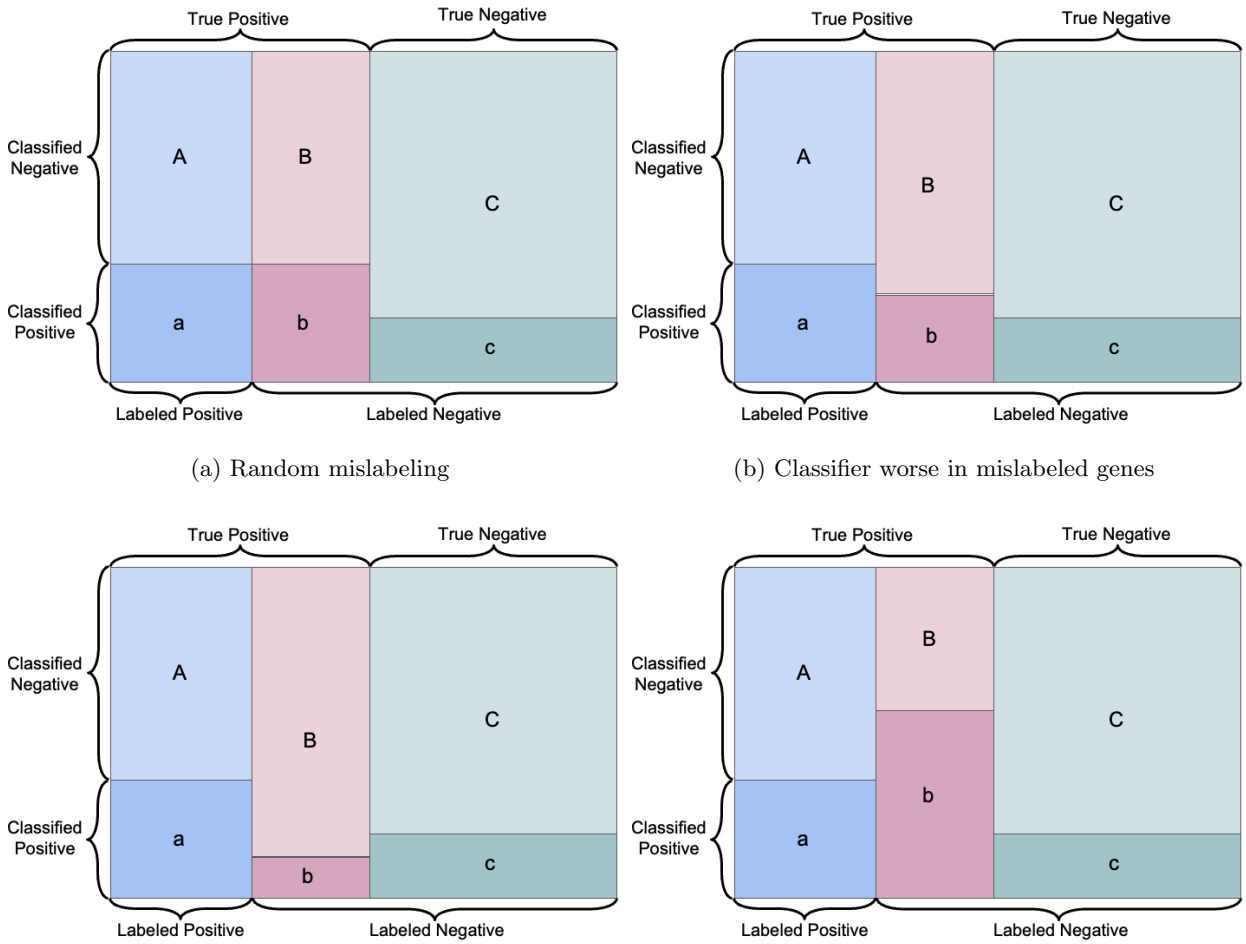
$$\begin{aligned} \mathrm{Sensitivity}_T &< \mathrm{Sensitivity}_M, \\ \mathrm{Specificity}_T &> \mathrm{Specificity}_M. \end{aligned} \tag{6}$$

If classifier has the worst performance in mislabeled genes ($\frac{b}{B+b} < \frac{c}{C+c}$, Fig. A.1c), we would expect:

$$\begin{aligned} \mathrm{Sensitivity}_T &< \mathrm{Sensitivity}_M, \\ \mathrm{Specificity}_T &\leq \mathrm{Specificity}_M. \end{aligned} \tag{7}$$

If in turn the classifier has the best performance in mislabeled genes ($\frac{b}{B+b} > \frac{a}{A+a}$, Fig. A.1d), we would expect:

$$\begin{aligned} \text{Sensitivity}_T &> \text{Sensitivity}_M, \\ \text{Specificity}_T &> \text{Specificity}_M. \end{aligned} \tag{8}$$



(a) Random mislabeling

(b) Classifier worse in mislabeled genes

(c) Classifier is worst in mislabeled genes

(d) Classifier is best in mislabeled genes

Fig. A.1: **Illustration of mislabeled data** Binary data is labeled such that a portion of the true positives are labeled as negatives, and classifier has varied performance on true positive and true negative data.

B Mis-ordered Classifiers in ROC Analysis

We illustrate the biased evaluation results for data sets where a subset of genes are preferentially labeled. We simulate a gene set consisting of only protein-acting genes (P) and expression-acting (E) genes. We assume genetic variants within P genes and E genes to have different effect sizes due to varying degree of penetrance. Our goal here is to demonstrate that if our gold-standard (GS) set contains mostly one category of genes (e.g. mostly protein-acting genes), it could result in a classifier that also preferentially selects for protein-acting genes demonstrating superior performance when evaluated through ROC analysis.

We simulate the truth to be under:

$$logit(Y_i) = \beta_0 + \beta_P P_i + \beta_E E_i + \epsilon_i,$$

where each gene $i \in [1, n]$ has binary "causal" classification $Y_i \in (0, 1)$, a set of "protein-acting features" (e.g. SNPs within a protein coding gene), and a set of "expression-acting features" (e.g. eQTL), and their corresponding effect sizes β. In particular, we will simulate the following two cases:

Model 1: Protein-acting (P) features have greater effect ($\beta_P > \beta_E$)

$$logit(Y_{ti}) = -3 + 6P_i + 2E_i,$$

Model 2: Expression-acting (E) features have greater effect ($\beta_P < \beta_E$)

$$logit(Y_{ti}) = -3 + 2P_i + 6E_i.$$

Two types of classifiers are trained using this simulated data:
P-classifier: a classifier using only P-features:

$$Y \sim P,$$

and **E-classifier**: a classifier using only E-features:

$$Y \sim E.$$

Subsequently, we simulate biased labels to be biased towards either of the two classes of features: protein-acting or expression-acting features. **P-feature biased labels** are generated under the following model:

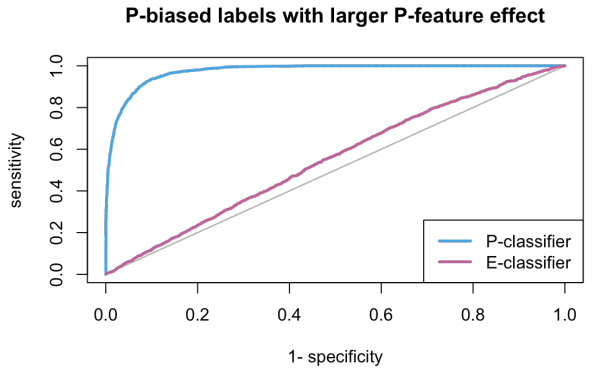
$$logit(Y_{bi}) = -3 + 4P_i,$$

and **E-feature biased labels** are generated under:

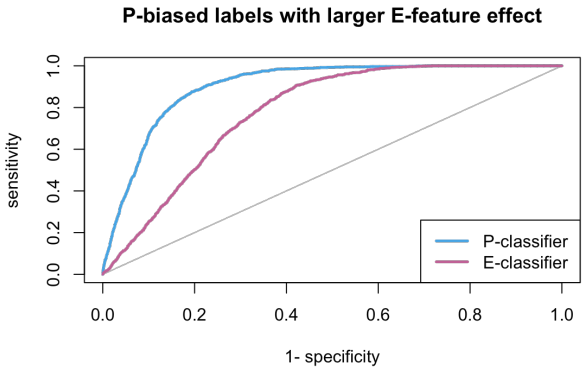
$$logit(Y_{bi}) = -3 + 4E_i.$$

Lastly, we multiply the simulated true labels Y_{ti} and simulated biased labels Y_{bi} so that our final observed truths among $Y_i = Y_{ti} \cdot Y_{bi}$ are a subset of the real truths. The classifiers are subsequently evaluated separately using the P-feature and E-feature biased labels (Fig. B.2).

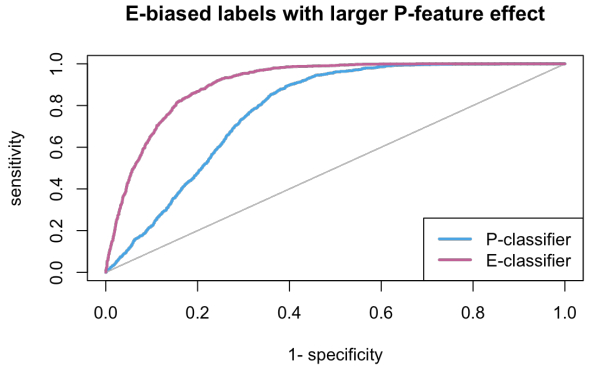
We compare the classifiers evaluated against biased labels and classifiers evaluated against true labels. In particular, we want to demonstrate that depending on the effect size of the different features, we could have an opposite conclusion of the classifier performance. In figure B.3, when labels are biased towards the feature with larger effect size in the underlying data, we notice that ROC curve measured according to biased labels are not far from its evaluation according to the true labels, and would lead us to the same conclusion regarding relative performance of the two classifiers (Fig. B.3a, B.3d). However, when the labels are biased towards feature with smaller effect size, the observed classifier performance is the opposite from the underlying truth (fig. B.3b, B.3c)



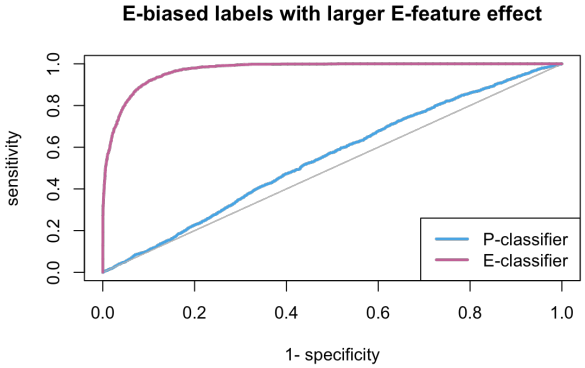
(a) P-feature biased labels



(b) P-feature biased labels

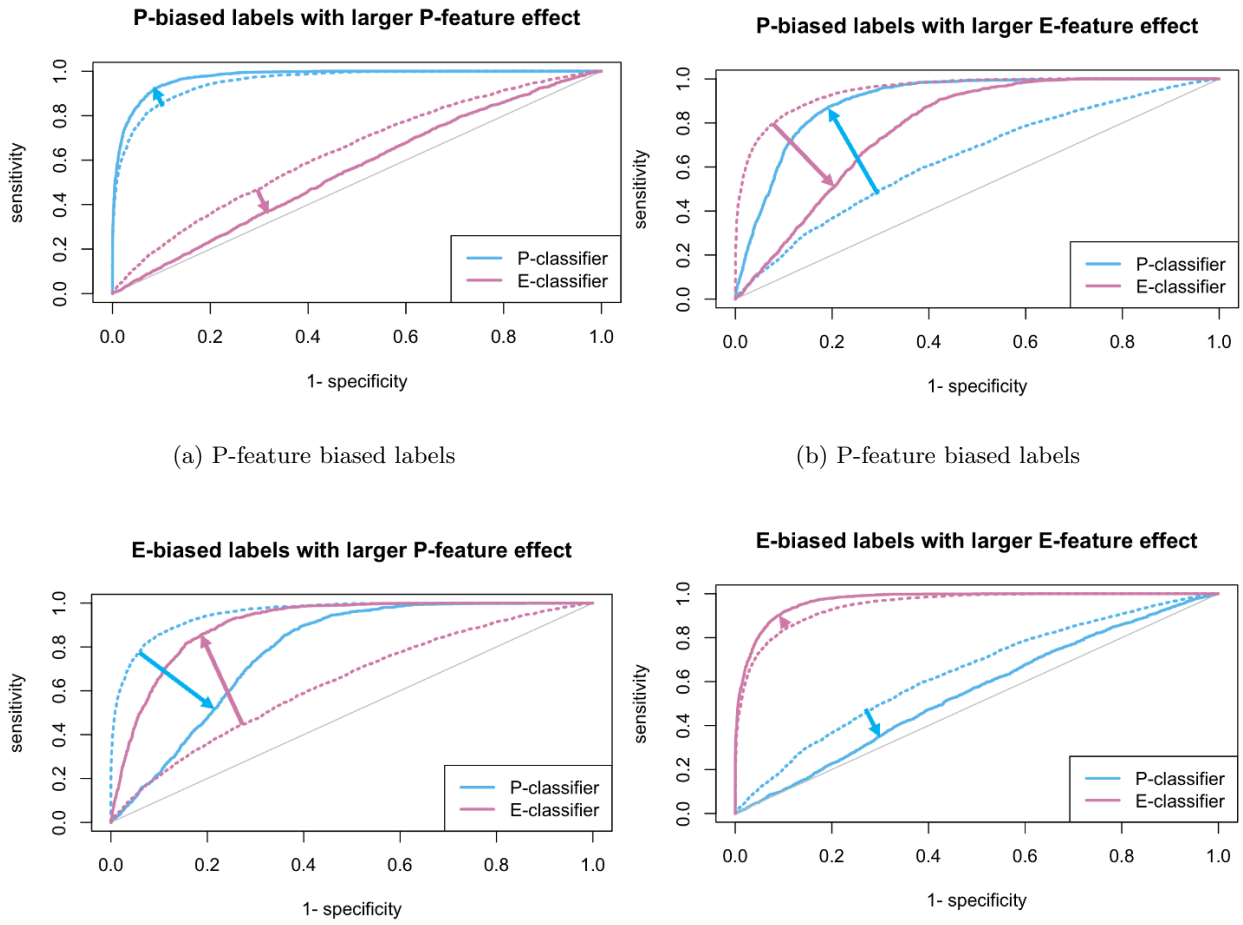


(c) E-feature biased labels



(d) E-feature biased labels

Fig. B.2: **Comparison of Classifiers on Biased Labels** Classifiers are trained on true data labels and evaluated against P-feature and E-feature biased labels. (a), (c) true data are simulated under model 1, where P-feature effect sizes are greater; (b), (d) true data are simulated under model 2, where E-feature effect sizes are greater.



(a) P-feature biased labels

(b) P-feature biased labels

(c) R-feature biased labels

(d) R-feature biased labels

Fig. B.3: **Comparison of Classifiers on True vs. Biased Labels** Classifiers evaluated against true labels (dotted lines) and biased labels (solid lines). Arrows highlight the change in ROC curve when some true positive labels are missed.

Table 1 References

1. Noah James Connally, Sumaiya Nazeen, Daniel Lee, Huwenbo Shi, John Stamatoyannopoulos, Sung Chun, Chris Cotsapas, Christopher A Cassa, and Shamil R Sunyaev. The missing link between genetic association and regulatory function. *Elife*, 11:e74970, 2022.
2. Steven Gazal, Omer Weissbrod, Farhad Hormozdiari, Kushal K Dey, Joseph Nasser, Karthik A Jagadeesh, Daniel J Weiner, Huwenbo Shi, Charles P Fulco, Luke J O'Connor, et al. Combining snp-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nature Genetics*, 54(6):827–836, 2022.
3. Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, 47(6):569–576, 2015.
4. Michael D Kessler, Amy Damask, Sean O'Keeffe, Michael Van Meter, Nilanjana Banerjee, Stefan Semrau, Dadong Li, Kyoko Watanabe, Julie Horowitz, Yariv Houvras, et al. Exome sequencing of 628,388 individuals identifies common and rare variant associations with clonal hematopoiesis phenotypes. *MedRxiv*, pages 2021–12, 2022.
5. Nikita Kolosov, Mark J Daly, and Mykyta Artomov. Prioritization of disease genes from gwas using ensemble-based positive-unlabeled learning. *European Journal of Human Genetics*, 29(10):1527–1535, 2021.
6. Edward Mountjoy, Ellen M Schmidt, Miguel Carmona, Jeremy Schwartzentruber, Gareth Peat, Alfredo Miranda, Luca Fumis, James Hayhurst, Annalisa Buniello, Mohd Anisul Karim, et al. An open approach to systematically prioritize causal variants and genes at all published human gwas trait-associated loci. *Nature genetics*, 53(11):1527–1533, 2021.
7. Sergio Picart-Armada, Steven J Barrett, David R Willé, Alexandre Perera-Lluna, Alex Gutteridge, and Benoit H Dessailly. Benchmarking network propagation methods for disease gene identification. *PLoS computational biology*, 15(9):e1007276, 2019.
8. Léon-Charles Tranchevent, Amin Ardeshirdavani, Sarah ElShal, Daniel Alcaide, Jan Aerts, Didier Auboeuf, and Yves Moreau. Candidate gene prioritization with endeavour. *Nucleic acids research*, 44(W1):W117–W121, 2016.
9. Elle M Weeks, Jacob C Ulirsch, Nathan Y Cheng, Brian L Trippe, Rebecca S Fine, Jenkai Miao, Tejal A Patwardhan, Masahiro Kanai, Joseph Nasser, Charles P Fulco, et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *MedRxiv*, pages 2020–09, 2020.