Journal of the American Medical Informatics Association, 2024, **31(3)**, 640–650
https://doi.org/10.1093/jamia/ocad226
Advance access publication 21 December 2023
Research and Applications

Research and Applications

Semi-supervised ROC analysis for reliable and streamlined evaluation of phenotyping algorithms

Jianhui Gao, MSc[1], Clara-Lea Bonzel, MSc[2], Chuan Hong, PhD[3], Paul Varghese, MD, MMSc[4], Karim Zakir, MSc[1], Jessica Gronsbell, PhD[1,5,6,*]

[1]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada, [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States, [3]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, United States, [4]Health Informatics, Verily Life Sciences, Cambridge, MA, United States, [5]Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada, [6]Department of Computer Science, University of Toronto, Toronto, ON, Canada

*Corresponding author: Jessica Gronsbell, PhD, Department of Statistics, University of Toronto, 700 University Ave, Toronto, ON M5G 1Z5, Canada (j.gronsbell@utoronto.ca)

Author Contributions: J. Gao and C.- L. Bonzel contributed equally to this work.

Abstract

Objective: High-throughput phenotyping will accelerate the use of electronic health records (EHRs) for translational research. A critical roadblock is the extensive medical supervision required for phenotyping algorithm (PA) estimation and evaluation. To address this challenge, numerous weakly-supervised learning methods have been proposed. However, there is a paucity of methods for reliably evaluating the predictive performance of PAs when a very small proportion of the data is labeled. To fill this gap, we introduce a semi-supervised approach (ssROC) for estimation of the receiver operating characteristic (ROC) parameters of PAs (eg, sensitivity, specificity).

Materials and Methods: ssROC uses a small labeled dataset to nonparametrically impute missing labels. The imputations are then used for ROC parameter estimation to yield more precise estimates of PA performance relative to classical supervised ROC analysis (supROC) using only labeled data. We evaluated ssROC with synthetic, semi-synthetic, and EHR data from Mass General Brigham (MGB).

Results: ssROC produced ROC parameter estimates with minimal bias and significantly lower variance than supROC in the simulated and semi-synthetic data. For the 5 PAs from MGB, the estimates from ssROC are 30% to 60% less variable than supROC on average.

Discussion: ssROC enables precise evaluation of PA performance without demanding large volumes of labeled data. ssROC is also easily implementable in open-source R software.

Conclusion: When used in conjunction with weakly-supervised PAs, ssROC facilitates the reliable and streamlined phenotyping necessary for EHR-based research.

Key words: electronic health records; phenotyping; semi-supervised; ROC analysis.

Background and significance

Electronic health records (EHRs) are a vital source of data for clinical and translational research.[1] Vast amounts of EHR data have been tapped for real-time studies of infectious diseases, development of clinical decision support tools, and genetic studies at unprecedented scale.[2–10] This myriad of opportunities rests on the ability to accurately and rapidly extract a wide variety of patient phenotypes (eg, diseases) to identify and characterize populations of interest. However, precise and readily available phenotype information is rarely available in patient records and presents a major barrier to EHR-based research.[11,12]

In practice, phenotypes are extracted from patient records with either rule-based or machine learning (ML)-based phenotyping algorithms (PAs) derived from codified and natural language processing (NLP)-derived features.[13,14] While PAs can characterize clinical conditions with high accuracy, they traditionally require a substantial amount of medical supervision that limits the automated power of EHR-based studies.[15] Several research networks have spent considerable effort developing PAs, including i2b2 (Informatics for Integrating Biology & the Bedside), the eMERGE (Electronic Medical Records and Genomics) Network, and the OHDSI (Observational Health Data Sciences and Informatics) program that released APHRODITE (Automated PHenotype Routine for Observational Definition, Identification, Training, and Evaluation).[16]

Typically, PA development consists of 2 key steps: (i) algorithm estimation and (ii) algorithm evaluation. Algorithm estimation determines the appropriate aggregation of features extracted from patient records to determine phenotype status. For a rule-based approach, domain experts assemble a comprehensive set of features and corresponding logic to assign patient phenotypes.[12,13] As this manual assembly is highly laborious, significant effort has been made to automate algorithm estimation with ML. Numerous studies have demonstrated success with PAs derived from standard supervised learning methods such as penalized regression, random forest, and deep neural networks.[17–24] The scalability of a

Received: May 3, 2023; **Revised:** September 22, 2023; **Editorial Decision:** October 30, 2023; **Accepted:** November 20, 2023

supervised approach, however, is limited by the substantial number of gold-standard labels required for model training. Gold-standard labels, which require time-consuming manual medical chart review, are infeasible to obtain for a large volume of records.[25,26]

In response, semi-supervised (SS) and weakly-supervised methods for PA estimation that substantially decrease or eliminate the need for gold-standard labeled data have been proposed. Among SS methods, self-training and surrogate-assisted SS learning are common.[27–29] For example, Zhang et al[15] introduced PheCAP, a common pipeline for SS learning that utilizes silver-standard labels for feature selection prior to supervised model training to decrease the labeling demand. Unlike gold-standard labels, silver-standard labels can be automatically extracted from patient records (eg, ICD codes or free-text mentions of the phenotype) and serve as proxies for the gold-standard label.[30,31] PheCAP was based on the pioneering work of Agarwal et al[32] and Banda et al,[33] which introduced weakly-supervised PAs trained entirely on silver-standard labels. These methods completely eliminate the need for chart review for algorithm estimation and are the basis of the APHRODITE framework. Moreover, this work prompted numerous developments in weakly-supervised PAs, including methods based on non-negative matrix/tensor factorization, parametric mixture modeling, and deep learning, which are quickly becoming the new standard in the PA literature.[14,29]

In contrast to the success in automating PA estimation, there has been little focus on the algorithm evaluation step. Algorithm evaluation assesses the predictive performance of a PA, typically through the estimation of the receiver operating characteristic (ROC) parameters such as sensitivity and specificity. At a high-level, the ROC parameters measure how well a PA discriminates between phenotype cases and controls relative to the gold-standard. As phenotypes are the foundation of EHR-based studies, it is critical to reliably evaluate the ROC parameters to provide researchers with a sense of trust in using a PA.[34–36] However, complete PA evaluation is performed far too infrequently due to the burden of chart review.[14,25,37]

To address this challenge, Gronsbell and Cai[38] proposed the first semi-supervised method for ROC parameter estimation. This method assumes that the predictive model is derived from a penalized logistic regression model and was only validated on 2 PAs with relatively large labeled data sets (455 and 500 labels). Swerdel et al[25] later introduced PheValuator, and its recent successor PheValuator 2.0, to efficiently evaluate rule-based algorithms using "probabilistic gold-standard" labels generated from diagnostic predictive models rather than chart review.[37] Although the authors provided a comprehensive evaluation for numerous rule-based PAs, PheValuator can lead to biased ROC analysis, and hence a distorted understanding of the performance of a PA, when the diagnostic predictive model is not correctly specified.[39–41] PheValuator can also only be applied to rule-based PAs.

To fill this gap in the PA literature, we introduce a SS approach to precisely estimate the ROC parameters of PAs, which we call "ssROC". The key difference between ssROC and classical ROC analysis (supROC) using only labeled data is that ssROC imputes missing gold-standard labels in order to leverage large volumes of unlabeled data (ie, records without gold-standard labels). By doing so, ssROC yields less variable estimates than supROC to enable reliable PA evaluation with fewer gold-standard labels. Moreover, ssROC imputes the missing labels with a nonparametric calibration of the predictions from the PA to ensure that the resulting estimates of the ROC parameters are unbiased regardless of the adequacy of PA.

Objective

The primary objectives of this work are to:

1) Extend the proposal of Gronsbell and Cai[38] to a wider class of weakly-supervised PAs that are common in the PA literature, including a theoretical analysis and development of a statistical inference procedure that performs well in finite-samples.
2) Provide an in-depth real data analysis of PAs for 5 phenotypes from Mass General Brigham (MGB) and extensive studies of synthetic and semi-synthetic data to illustrate the practical utility of ssROC.
3) Release an implementation of ssROC in open-source R software to encourage the use of our method by the informatics community.

Through our analyses of simulated, semi-synthetic, and real data, we observe substantial gains in estimation precision from ssROC relative to supROC. In the analysis of the 5 PAs from MGB, the estimates from ssROC are approximately 30% to 60% less variable than supROC on average. Our results suggest that, when used together with weakly-supervised PAs, ssROC can facilitate the reliable and streamlined PA development that is necessary for EHR-based research.

Materials and methods

Overview of ssROC

We focus on evaluating a classification rule derived from a PA with ROC analysis. ROC analysis assesses the agreement between the gold-standard label for a binary phenotype (eg, disease case/control), Y, and a PA score, S, indicating a patient's likelihood of having the underlying phenotype (eg, the predicted probability of being a case). Y is typically obtained from chart review and S can be derived from various phenotyping methods. We focus on scores derived from parametric models fit with a weakly-supervised approach due to their ability to automate PA estimation and increasing popularity in the informatics literature.[14,32,42–44] For ease of notation, we suppress the dependence of S on the estimated model parameter and provide more details on the PA in Section S3.

In classical supROC analysis, the data are assumed to contain information on both Y and S for all observations. However, in the phenotyping setting, Y is typically only available for a very small subset of patients due to the laborious nature of chart review. This gives rise to the *semi-supervised setting* in which a small labeled dataset is accompanied by a much larger unlabeled dataset. To leverage all of the available data and facilitate more reliable (ie, lower variance) evaluation of PAs, ssROC imputes the missing Y with a nonparametric recalibration of S, denoted as $\widehat{m}(S)$, to make use of the unlabeled data. An overview of ssROC is provided in Figure 1.

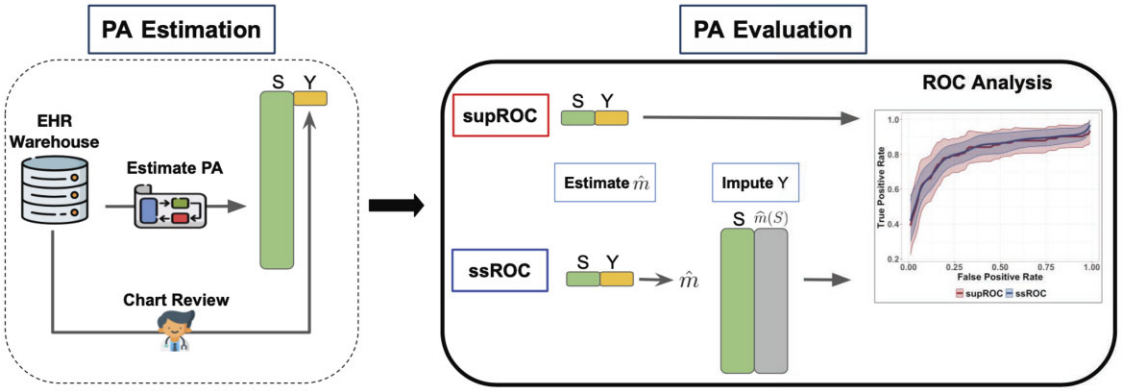

Figure 1. Overview of PA estimation and evaluation. The phenotyping algorithm (PA) is first estimated to obtain the scores (S). Patient charts from the electronic health record (EHR) warehouse are reviewed to obtain the gold-standard label (Y) for PA evaluation. In classical supervised ROC analysis (supROC), only the labeled data from chart review is used to evaluate the PA's performance. Semi-supervised ROC analysis (ssROC) uses the labeled data to impute the missing Y as $\widehat{m}(S)$ so that the unlabeled data can be utilized for estimation to yield more precise estimates of the ROC parameters.

Data structure and notation

More concretely, the available data in the SS setting consists of a small labeled dataset

$$\mathcal{L} = \{(Y_i, S_i) | i = 1, \ldots, n\}$$

and an unlabeled dataset

$$\mathcal{U} = \{S_i | i = n+1, \ldots, n+N\}.$$

In the classical setting, it is assumed that (i) $\mathcal{U}$ is a much larger than $\mathcal{L}$ so that $n \ll N$ and (ii) the observations in $\mathcal{L}$ are randomly selected from the underlying pool of data. Throughout our discussion, we suppose that a higher value of S is more indicative of the phenotype. An observation is deemed to have the phenotype if $S > c$, where c is the threshold for classification.

ROC analysis

More formally, ROC analysis evaluates a PA with the true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), and negative predictive value (NPV). In diagnostic testing, the TPR is referred to as sensitivity, while the FPR is 1 minus the specificity.[45] For a given classification threshold, one may evaluate the ROC parameters by enumerating the correct and incorrect classifications. This information can be summarized in a confusion matrix as shown in Figure 2.

In practice, it is the task of the researcher to estimate an appropriate threshold for classification. This is commonly done by summarizing the trade-off between the TPR and FPR, defined respectively as

$$\text{TPR}(c) = P(S > c \,|\, Y = 1) \text{ and } \text{FPR}(c) = P(S > c \,|\, Y = 0).$$

The ROC curve, $\text{ROC}(u) = \text{TPR}[\text{FPR}^{-1}(u)]$, summarizes the TPR and FPR across all possible choices of the threshold. In the context of PAs, c is often chosen to achieve a low FPR.[22] An overall summary measure of the discriminative power of S in classifying Y is captured by the area under the ROC curve (AUC),

$$\text{AUC} = \int_0^1 \text{ROC}(u) du.$$

The AUC is equivalent to the probability that a phenotype case has a higher value of S than a phenotype control.[46] For a given threshold, the predictive performance of the classification rule derived from the PA is assessed with the PPV and NPV, defined respectively as

$$\text{PPV}(c) = P(Y = 1 \,|\, S > c), \quad \text{and}$$
$$\text{NPV}(c) = P(Y = 0 \,|\, S < c).$$

Supervised ROC analysis

With only labeled data, one may obtain supervised estimators of the ROC parameters (supROC) with their empirical counterparts. For example, the TPR and FPR can be estimated as

$$\widehat{\text{TPR}}_{sup}(c) = \frac{\sum_{i=1}^n Y_i I(S_i > c)}{\sum_{i=1}^n Y_i} \text{ and}$$
$$\widehat{\text{FPR}}_{sup}(c) = \frac{\sum_{i=1}^n (1-Y_i) I(S_i > c)}{\sum_{i=1}^n (1-Y_i)}.$$

The remaining parameters are estimated in a similar fashion. Variance estimates can be obtained from a resampling procedure, such as bootstrap or perturbation resampling.[47]

ssROC: semi-supervised ROC analysis

Unlike its supervised counterpart that relies on only the labeled data, ssROC contains 2 steps of estimation to make use of the unlabeled data and provide a more reliable understanding of PA performance. In the first step, the missing labels are imputed by recalibrating the PA scores using a model trained with the labeled data. In the second step, the imputations are used in lieu of the gold-standard labels to evaluate the ROC parameters based on the PA scores in the unlabeled data in an analogous manner to supROC. Below we provide an overview of these 2 steps using the TPR as an example.

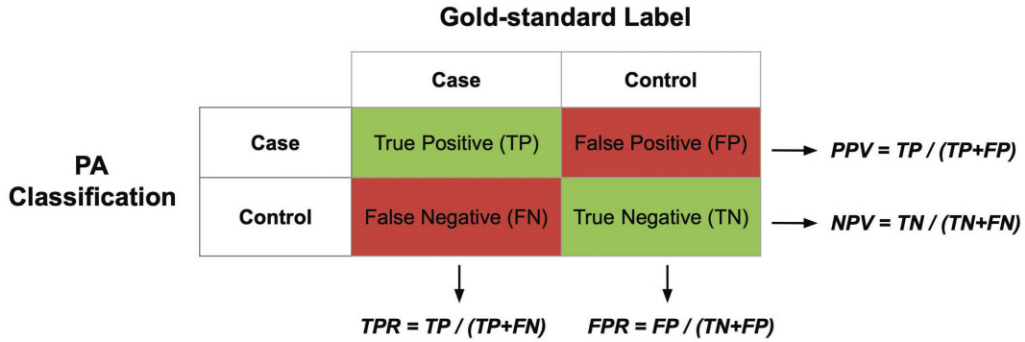

Figure 2. Confusion matrix. The algorithm score from the phenotyping algorithm (PA) is used to determine phenotype case/control status based on the classification threshold. The ROC parameters are evaluated by enumerating the number of correct and incorrect classifications relative to the gold-standard label.(A) Percent bias of supROC. (B) Percent bias of ssROC. (C) Relative efficiency (supROC: ssROC).

Step 1. Recalibrate the PA scores by fitting the model $m(S) = P(Y = 1|S)$ with the labeled data. Obtain the imputations, $\{\widehat{m}(S_i)|i = n+1, \ldots, n+N\}$, for the unlabeled data based on the fitted model.

Step 2. Use the imputations to estimate the TPR with the unlabeled data as

$$\widetilde{\text{TPR}}_{ssROC}(c) = \frac{\sum_{i=n+1}^{n+N} \widehat{m}(S_i) I(S_i > c)}{\sum_{i=n+1}^{n+N} \widehat{m}(S_i)}.$$

The purpose of the first step is to ensure that the imputations do not introduce bias into the ROC parameter estimates. For example, utilizing the PA scores directly for imputation can distort the ROC parameter estimates due to potential inaccuracies of S in predicting Y. We propose to use a kernel regression model to nonparametrically impute the missing labels to prevent biasing the ssROC estimates.[40] Technical detail related to fitting the kernel regression model is provided in Section S1. In contrast, the purpose of the second step is to harness the large unlabeled dataset to produce estimates with lower variance than supROC. Similar to supROC, we propose a perturbation resampling procedure for variance estimation and detail 2 commonly used confidence intervals (CIs) based on the procedure in Section S2. In Section S3, we also provide a theoretical justification for the improved precision of ssROC relative to supROC for a wide range of weakly-supervised PAs.

Data and metrics for evaluation

We assessed the performance of ssROC using simulated, semi-synthetic, and real-world EHR data from MGB. All analyses used the R software package, `ssROC`, available at https://github.com/jlgrons/ssROC.

Simulation study

Our simulations cover PAs with high and low accuracy and varying degrees of calibration. For each accuracy setting, we simulated PA scores that (i) were perfectly calibrated, (ii) overestimated the probability of Y, and (iii) underestimated the probability of Y.[40] In all settings, Y was generated from a Bernoulli distribution with a prevalence of 0.3. To generate S, we first generated a random variable Z from a normal mixture model with $Z|Y = y \sim N(\alpha_y, \sigma^2)$ and an independent noise variable from a Bernoulli mixture model with $\epsilon|Y = y \sim \text{Bern}(p_y)$ for $y = 0, 1$. The PA score was obtained as

Table 1. Parameter configurations for the 6 simulation studies.

	High PA accuracy	Low PA accuracy
Perfectly calibrated PA	(−0.5, 0.5, 0.5)	(−0.25, 0.25, 0.5)
Overestimated PA	(1, 2.3, 0.5, 0.3, 0.3)	(0.5, 1.2, 0.5, 0.5, 0.5)
Underestimated PA	(−2.6, −1.5, 0.5, 0.1, 0.1)	(−2.5, −1.5, 1, 0.3, 0.3)

The simulation settings were derived by varying the parameters $(\alpha_0, \alpha_1, \sigma, p_0, p_1)$ in Model 1.
Abbreviation: PA, phenotyping algorithm.

$$S = \begin{cases} \text{expit}(\gamma_0 + \gamma_1 Z) & \text{for perfect calibration} \\ \text{expit}(Z + \epsilon) & \text{otherwise} \end{cases}. \quad (1)$$

where $\gamma_0 = (\alpha_1^2 - \alpha_0^2)/2\sigma^2 + \log[(1-\mu)/\mu)]$, $\gamma_1 = (\alpha_1 - \alpha_0)/\sigma^2$, $\mu = P(Y = 1)$, and $\text{expit}(x) = \frac{1}{1+e^{-x}}$. The values of γ_0 and γ_1 ensure that $S = P(Y = 1|Z)$ for perfect calibration. Six simulation settings were obtained by varying $(\alpha_0, \alpha_1, \sigma, p_0, p_1)$, shown in Table 1. We also considered the extreme setting when S is independent of Y by permuting S generated from the model with high accuracy and perfect calibration. The calibration curves for each setting are presented in Figure S2. Across all settings, $N = 10\,000$, $n = 75$, 150, 250 and 500, and results are summarized across 5000 simulated datasets.

Semi-synthetic data analysis

To better reflect the complexity of PAs in real data, we generated semi-synthetic data for phenotyping depression with the MIMIC-III clinical database. MIMIC-III contains structured and unstructured EHR data from patients in the critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.[21,48] As depression status is unavailable in patient records, it was simulated for all observations using a logistic regression model. That is, $Y \sim \text{Bern}[\text{expit}(\boldsymbol{\beta}^T \mathbf{X})]$ where

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3),$$
$$\mathbf{X} = (1, \log(X_{NLP} + X_{ICD} + 1), X_{age}, \log(X_{HU} + 1))^T$$

X_{NLP} is the number of depression related clinical concepts, X_{ICD} is the number of depression related ICD-9 codes, X_{age} is age at admission, and X_{HU} is a measure of healthcare utilization based on the total number of evaluation and management Current Procedural Terminology (CPT) codes and the

length of stay. The list of depression-related ICD-9 codes and clinical concepts are presented in Section S6.

Given the PA score is obtained from complex EHR data, we focus on simulating the phenotype to achieve high and low PA accuracy and present the calibration in Figure S4. We set $\boldsymbol{\beta} = (1, 4, 0.05, -3)$ and $\boldsymbol{\beta} = (1, 1, 0.01, -1)$ and to mimic a PA with high and low accuracy (AUC = 90.1 and 72.6, respectively). The prevalence of Y in both settings was approximately 0.3. For both settings, the unlabeled set consisted of one visit from $N = 32\,172$ unique patients and $n =$ 75, 150, 250, and 500 visits were randomly sampled 5000 times to generate labeled datasets of various sizes. We obtained the PA for depression by fitting PheNorm without the random corruption denoising step. PheNorm is a weakly-supervised method based on normalizing silver-standard labels with respect to patient healthcare utilization using a normal mixture model.[43] PheNorm is also used in our real-data analysis and described in detail in Section S4. X_{NLP} and X_{ICD} were used as silver-standard labels to fit PheNorm with X_{HU} as the measure of healthcare utilization.

Real-world EHR data application

We further validated ssROC using EHR data from MGB, a Boston-based healthcare system anchored by 2 tertiary care centers, Brigham and Women's Hospital and Massachusetts General Hospital. We evaluated PAs for 5 phenotypes, including cerebral aneurysm (CA), congestive heart failure (CHF), Parkinson's disease (PD), systemic sclerosis (SS), and type 1 diabetes (T1DM). The data are from the Research Patient Data Registry which stores data on over 1 billion visits containing diagnoses, medications, procedures, laboratory information, and clinical notes from 1991 to 2017.

The full data for each phenotype consisted of patient records with at least one phenotype-related PheCode in their record.[49] A subset of patients was randomly sampled from the full data and sent for chart review. For each phenotype, the PA was obtained by fitting PheNorm without denoising using the total number of (i) phenotype-related PheCodes and (ii) positive mentions of the phenotype-related clinical concepts as the silver-standard labels and the number of notes in a patient's EHR as the measure of healthcare utilization. The phenotypes represent different levels of PA accuracy, labeled and unlabeled dataset sizes, and prevalence (P). A summary of the 5 phenotypes is presented in Table 2.

Benchmark method and reported metrics

We compared the PA evaluation results from ssROC and the benchmark, supROC, using the simulated, semi-synthetic, and real EHR data. We transformed the PA scores by their respective empirical cumulative distribution functions prior to ROC analysis. This transformation improves the performance of the imputation step, particularly when the distribution of S is skewed.[50] For the kernel regression, we used a Gaussian kernel with bandwidth determined by the standard deviation of the transformed PA scores divided by $n^{0.45}$.[51] Additional detail related to the imputation step is provided in Section S1. We obtained variance estimates for the ROC parameters using perturbation resampling with 500 replications and weights from a scaled beta distribution, $4 \cdot \text{Beta}(1/2, 3/2)$, to improve finite-sample performance.[52] We focused on logit-based CIs, described in Section S2.2, due to their improved coverage relative to standard Wald intervals.[53]

Table 2. Summary of the MGB phenotypes.

Phenotype	n	N	P	PheCode	CUI
Cerebral aneurysm (CA)	134	18 679	.68	433.5	C0917996
Congestive heart failure (CHF)	140	155 112	.18	428	C0018801
Parkinson's disease (PD)	97	17 752	.62	332	C0030567
Systemic sclerosis (SS)	189	4272	.43	709.3	C0036421
Type 1 diabetes (T1DM)	121	46 013	.17	250.1	C0011854

The labeled dataset size (n), the unlabeled dataset size (N), and the prevalence (P) of the 5 phenotypes as well as the main PheCode and concept unique identifier (CUI) used to train PheNorm. The underlying full data for each phenotype included all participants who passed the filter of ≥ 1 PheCode for the phenotype of interest.

We assessed percent bias for both supROC and ssROC by computing the mean of [(point estimate−ground truth)/(ground truth)×100%] across the replicated datasets. The ground truth values of the ROC parameters for the simulated and semi-synthetic data are provided in Tables S1 and S6. The empirical standard error (ESE) was computed as the standard deviation of the estimates from these datasets. The asymptotic standard error (ASE) was computed as the mean of the standard error estimates derived from the perturbation resampling procedure across the replicated datasets. Using mean squared error (MSE) as an aggregate measure of bias and variance, we evaluated the relative efficiency (RE) as the ratio of the MSE of supROC to the MSE of ssROC. The performance of our resampling procedure was assessed with the coverage probability (CP) of the 95% CIs for both estimation procedures. In the real data analysis, we present point estimates from both supROC and ssROC and the RE defined as the ratio of the variance of supROC to ssROC. We evaluated the performance of the PAs at an FPR of 10% and report the results for the AUC, classification threshold (Threshold), TPR, PPV, and NPV for all analyses.

Results

Simulation study

Figures 3 and 4 show the percent bias and RE in the high and low accuracy settings, respectively. Both ssROC and supROC generally exhibit low bias across all settings and ssROC often has lower bias than supROC. Additionally, ssROC has lower variance than supROC in all settings, as indicated by REs that consistently exceed 1. In the high accuracy setting, the median REs across all calibration patterns and labeled sizes is between 1.3 (AUC) and 1.9 (Threshold). For the low accuracy setting, the median REs range from 1.1 (AUC) to 1.6 (Threshold). Practically, these results imply that ssROC is more precise for a fixed amount of labeled than supROC. Alternatively, this reduction in variance can also be interpreted as a reduction in sample size required for ssROC to achieve the same variance as supROC. For example, the RE for PPV with $n = 250$ under the setting of high PA accuracy and perfect calibration is 2, which suggests that ssROC can achieve the same variance as supROC with half the amount of labeled data.

When S is independent of Y, Figure S3 shows that ssROC has negligible bias, yields precision similar to supROC for the ROC parameters, and has improved precision for the

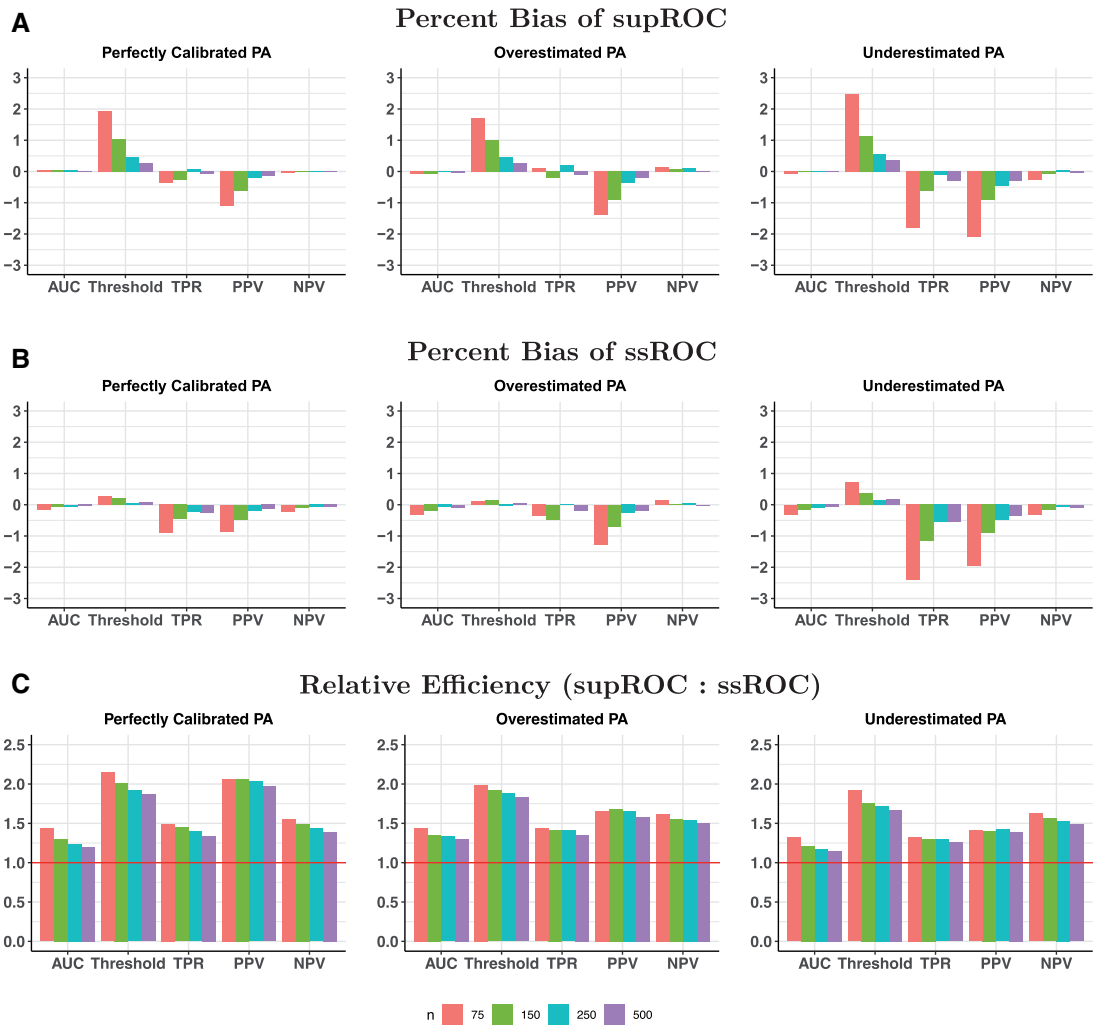

Figure 3. Percent bias and relative efficiency (RE) for high phenotyping algorithm (PA) accuracy settings at a false positive rate (FPR) of 10%. RE is defined as the mean squared error of supROC compared to the mean squared error of ssROC. For all scenarios, the size of the unlabeled was N=10 000.

threshold. These empirical findings demonstrate the robustness of ssROC to a wide range of PA scores and are further supported by our theoretical analysis in Section S3. Specifically, our analysis verifies that ssROC is guaranteed to perform on par supROC for the ROC parameters and yield more precise estimation for the Threshold when S is independent of Y.

The ESE, ASE, and CP for the 95% CIs of both supROC and ssROC are presented in Tables S2 and S3. The proposed logit-based CI consistently achieves reasonable coverage for both methods. The estimated variance for ssROC is also generally more accurate than that from supROC. Additionally, our results underscore the advantages of employing the logit-based interval over the standard Wald interval, particularly when n is small and/or the point estimate is near the boundary.

Semi-synthetic data analysis

The findings from the semi-synthetic EHR data analysis align closely with the results of our simulations, further demonstrating the robustness of ssROC to the PA score. Generally, ssROC has smaller bias than supROC and both methods have small bias across all settings as highlighted in Figure 5A and B. ssROC again demonstrates improved precision relative to supROC. The median RE across labeled data sizes in the setting with high PA accuracy is between 1.3 (AUC) and 1.9 (Threshold) and between 1.1 (AUC) and 2.1 (Threshold) for the low accuracy setting. Additionally, Tables S7 and S8 show that the logit-based CIs for both methods yield reasonable coverage.

Analysis of 5 PAs from MGB

Table 3 presents the point estimates for the 5 phenotypes from MGB, ordered by the AUC estimates from ssROC, at a FPR of 10%. As our primary focus is to compare ssROC with supROC, a single FPR was chosen for consistency across the phenotypes. However, this does lead to low TPRs for some phenotypes, such as CHF. Generally, the point estimates from ssROC are similar to those from supROC. There are some differences in the Threshold estimates for CA and SS, which leads to some discrepancies in the other estimates. As supROC is only evaluated at the unique PA scores in the labeled dataset, the Threshold estimate can be unstable at some FPRs. In contrast, ssROC is evaluated across a broader

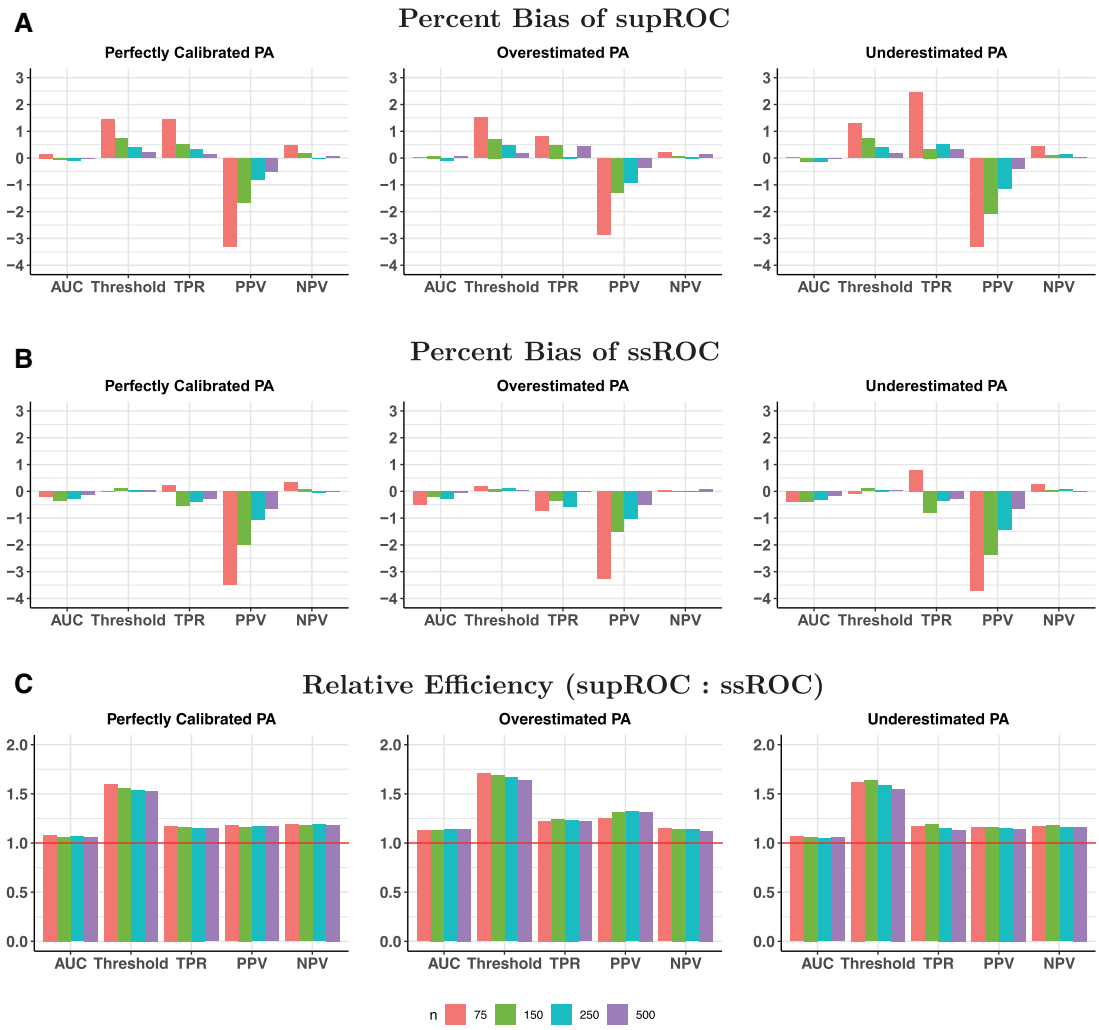

Figure 4. Percent bias and relative efficiency (RE) for low phenotyping algorithm (PA) accuracy settings at a false positive rate (FPR) of 10%. RE is defined as the mean squared error of supROC compared to the mean squared error of ssROC. For all scenarios, the size of the unlabeled was N=10 000.

range of PA scores in the unlabeled dataset and results in more stable estimation.

Figure 6 shows the RE of supROC to ssROC across the 5 phenotypes at a FPR of 10%. The median RE gain across phenotypes ranges from approximately 1.5 (AUC, TPR) to 2.7 (Threshold), implying that the estimates from ssROC are approximately 30%-60% less variable than supROC on average. It is worth noting the RE for the Threshold estimate for CHF is quite high. Figure S5 illustrates that this behavior can be explained by the empirical distribution of the resampled estimates. The distribution of the estimates from supROC are multimodal, while those from ssROC are approximately normal as expected. This behavior further emphasizes the stability of ssROC relative to supROC in real data.

Consistent with our simulation and theoretical results, we also observe that RE is linked to PA accuracy. For example, a phenotype with high PA accuracy, such as T1DM, exhibits a higher RE compared to CA, which has the lowest PA accuracy. Overall, these findings underscore the advantages of our proposed ssROC method compared to supROC in yielding more precise ROC analysis.

Discussion

Although high-throughput phenotyping is the backbone of EHR-based research, there is a paucity of methods for reliably evaluating the predictive performance of a PA with limited labeled data. The proposed ssROC method fills this gap. ssROC is a simple 2-step estimation procedure that leverages large volumes of unlabeled data by imputing missing gold-standard labels with a nonparametric recalibration of a PA score. Unlike existing procedures for PA evaluation in the informatics literature, ssROC eliminates the requirement that the PA be correctly specified to yield unbiased estimation of the ROC parameters and may be utilized for ML-based PAs.[25,37] While we focus specifically on weakly-supervised PAs in our theoretical analysis and data examples given their increasing popularity and ability to automate PA estimation, ssROC can also be used to evaluate rule-based or other ML-based PAs. Moreover, by harnessing unlabeled data, ssROC yields substantially less variable estimates than supROC in simulated, semi-synthetic, and real data. Practically, this translates into a significant reduction in the amount of chart review required to obtain a precise understanding of PA performance.

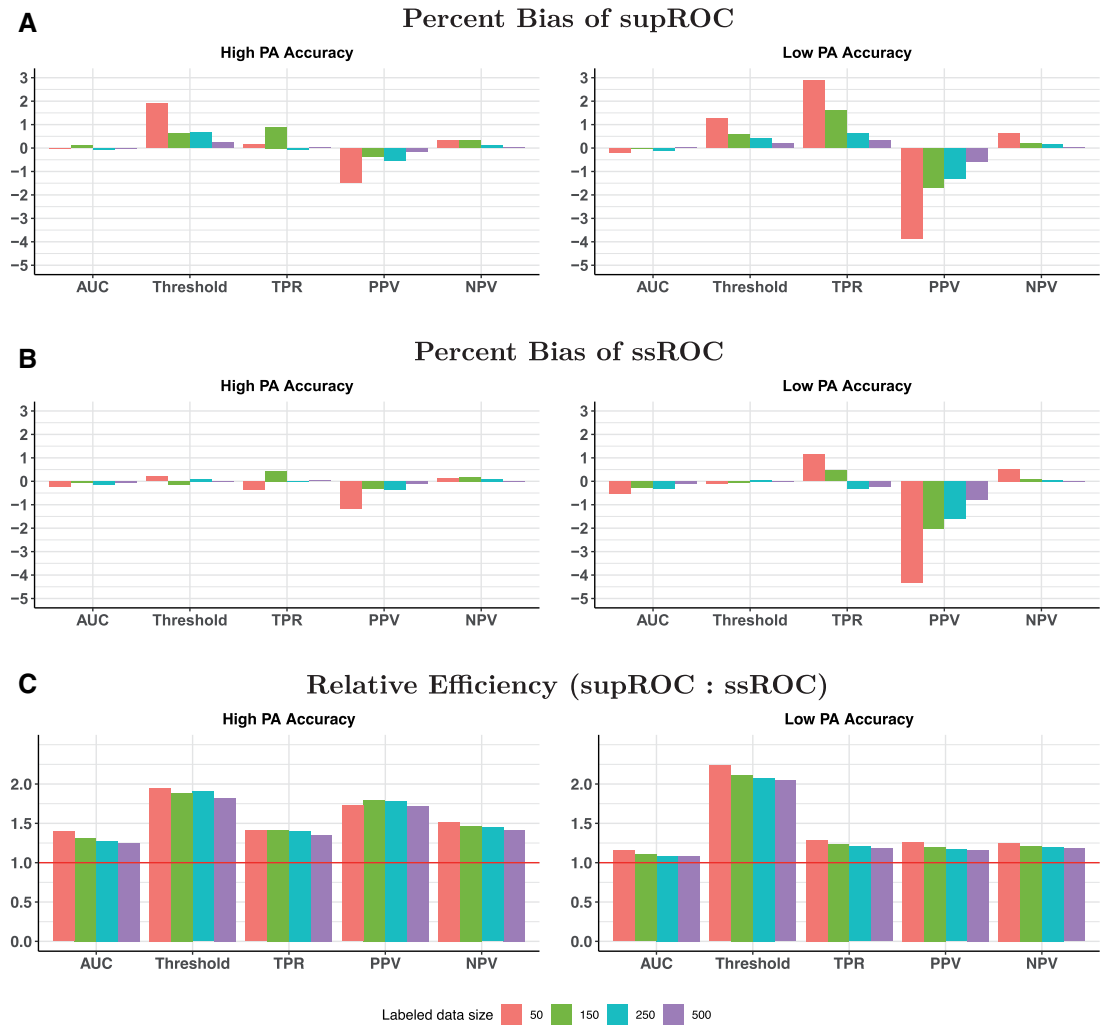

Figure 5. Percent bias and relative efficiency (RE) for the semi-synthetic data analysis at a false positive rate (FPR) of 10%. RE is defined as the mean squared error of supROC compared to the mean squared error of ssROC. For both settings, the size of total data was 32 172.

Table 3. Point estimates for the 5 phenotypes from MGB at a FPR of 10%.

Phenotype	Method	AUC	Threshold	TPR	PPV	NPV
CA	ssROC	81.3	64.0	51.0	89.8	51.5
	supROC	80.4	73.7	35.2	87.4	40.1
CHF	ssROC	83.2	84.9	42.0	44.1	89.2
	supROC	79.3	86.1	36.0	43.9	86.6
PD	ssROC	85.9	75.3	49.1	74.1	75.2
	supROC	81.6	80.1	34.1	72.4	64.1
SS	ssROC	89.4	59.8	60.6	89.9	60.7
	supROC	87.5	64.6	52.3	89.7	53.2
T1DM	ssROC	90.5	80.4	68.8	57.3	93.7
	supROC	91.5	80.1	75.0	59.8	94.8

Abbreviations: CA, cerebral aneurysm; CHF, congestive heart failure; FPR, false positive rate; MGB, Mass General Brigham; NPV, negative predictive value; PD, Parkinson's disease; PPV, positive predictive value; SS, systemic sclerosis; ssROC, semi-supervised receiver operating characteristic analysis; supROC, supervised receiver operating characteristic analysis; T1DM, type 1 diabetes; TPR, true positive rate.

Although our work is a first step toward streamlining PA evaluation, there are several avenues that warrant future research. First, ssROC assumes that the labeled examples are randomly sampled from the underlying full data. In situations where the goal is to phenotype multiple conditions or comorbidities, more effective sampling strategies such as stratified random sampling have the potential to further enhance the efficiency of ssROC.[54] However, due to the large discrepancy in size of the labeled and unlabeled data, developing procedures to accommodate non-random sampling is non-trivial.[55] Second, the nonparametric recalibration step demands a sufficient amount of labeled data for the kernel regression to be well estimated. While our extensive simulation studies across a wide variety of PAs and sample sizes illustrate the robustness of ssROC, our future work will develop a parametric recalibration procedure that accommodates smaller labeled data sizes. Third, ssROC can also be extended for model comparisons and evaluation of fairness metrics, which are urgently needed given the increasing recognition of unfairness in informatics applications. The calibration step would need to be augmented in both settings to utilize additional information in multiple PA scores or protected attributes, respectively. This augmentation could potentially lead to a more efficient procedure as ssROC only uses information from one PA score for imputation. Lastly, our results demonstrate the ability of ssROC to provide accurate ROC evaluation for 5 phenotypes with variable prevalence, labeled and unlabeled dataset sizes, and PA accuracy within one health system. Further work is needed to understand the

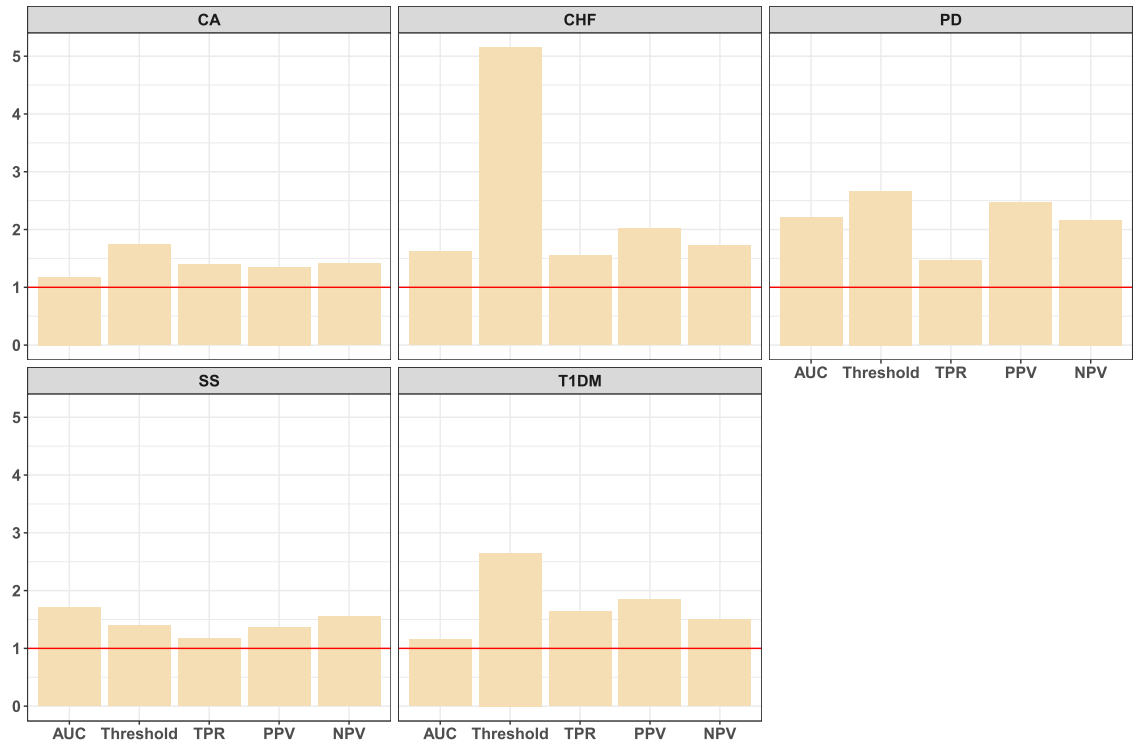

Figure 6. Relative efficiency (RE) for the 5 phenotypes from Mass General Brigham (MGB) at an false positive rate (FPR) of 10%. RE is defined as the ratio of the variance of supROC to that of ssROC.

performance of our method across a diverse range of phenotypes and to extend our approach to accommodate federated analyses across multiple healthcare systems.

Conclusion

In this article, we introduced a semi-supervised approach, ssROC, that leverages a large volume of unlabeled data together with a small subset of gold-standard labeled data to precisely estimate the ROC parameters of PAs. PA development involves 2 key steps: (i) algorithm estimation and (ii) algorithm evaluation. While a considerable amount of effort has been placed on algorithm estimation, ssROC fills the current gap in robust and efficient methodology for predictive performance evaluation. Additionally, ssROC is simple to implement and is available in open-source R software to encourage use in practice. When used in conjunction with weakly-supervised PAs, ssROC demonstrates the potential to facilitate the reliable and streamlined phenotyping that is necessary for a wide variety of translational EHR applications.

Author contributions

Je.G. conceived and designed the study. Ji.G. conducted simulation and semi-synthetic data analyses. C.B. and C.H. conducted real data analyses. Je.G., Ji.G., C.B., P.V., and K.Z. analyzed and interpreted the results. Je.G. and Ji.G. drafted and revised the manuscript. All authors reviewed and approved the final manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

The project was supported by the Natural Sciences and Engineering Research Council of Canada grant (RGPIN-2021-03734), the University of Toronto Connaught New Researcher Award, and the University of Toronto Seed Funding for Methodologists Grant (to Je.G.).

Conflict of interest

The authors have no conflicts of interest to declare.

Data availability

Our proposed method is implemented as an R software package, `ssROC`, which is available at https://github.com/jlgrons/ssROC.

References

1. McGinnis JM, Olsen L, Goolsby WA, et al. *Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary*. National Academies Press; 2011.
2. Boockvar KS, Livote EE, Goldstein N, et al. Electronic health records and adverse drug events after patient transfer. *Qual Saf Health Care*. 2010;19(5):e16.
3. Kurreeman F, Liao K, Chibnik L, et al. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multiethnic cohort derived from electronic health records. *Am J Hum Genet*. 2011;88(1):57-69.
4. Liao KP, Kurreeman F, Li G, et al. Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non-rheumatoid arthritis controls. *Arthritis Rheum*. 2013;65(3):571-581.

5. Chen C-Y, Lee PH, Castro VM, et al. Genetic validation of bipolar disorder identified by automated phenotyping using electronic health records. *Transl Psychiatry*. 2018;8(1):86.
6. Li R, Chen Y, Ritchie MD, et al. Electronic health records and polygenic risk scores for predicting disease risk. *Nat Rev Genet*. 2020;21(8):493-502.
7. Brat GA, Weber GM, Gehlenborg N, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *NPJ Digit Med*. 2020;3:109.
8. Bastarache L. Using phecodes for research with the electronic health record: from PheWAS to PheRS. *Annu Rev Biomed Data Sci*. 2021;4:1-19.
9. Prieto-Alhambra D, Kostka K, Duarte-Salles T, et al. Unraveling COVID-19: a large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS. *Res Square*. 2021;14:369-384.
10. Henry KE, Adams R, Parent C, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nature Med*. 2022;28(7):1447-1454.
11. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2014;21(2):221-230.
12. Banda JM, Seneviratne M, Hernandez-Boussard T, et al. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annu Rev Biomed Data Sci*. 2018;1:53-68.
13. Alzoubi H, Alzubi R, Ramzan N, et al. A review of automatic phenotyping approaches using electronic health records. *Electronics*. 2019;8(11):1235.
14. Yang S, Varghese P, Stephenson E, et al. Machine learning approaches for electronic health records phenotyping: a methodical review. *J Am Med Inform Assoc*. 2023;30(2):367-381.
15. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat Protoc*. 2019;14(12):3426-3444.
16. Murphy S, Churchill S, Bry L, et al. Instrumenting the health care enterprise for discovery research in the genomic era. *Genome Res*. 2009;19(9):1675-1681.
17. Castro V, Shen Y, Yu S, et al. Identification of subjects with polycystic ovary syndrome using electronic health records. *Reprod Biol Endocrinol*. 2015;13:116.
18. Teixeira PL, Wei W-Q, Cronin RM, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *J Am Med Inform Assoc*. 2017;24 (1):162-171.
19. Geva A, Gronsbell JL, Cai T, et al.; Pediatric Pulmonary Hypertension Network and National Heart, Lung, and Blood Institute Pediatric Pulmonary Vascular Disease Outcomes Bioinformatics Clinical Coordinating Center Investigators. A computable phenotype improves cohort ascertainment in a pediatric pulmonary hypertension registry. *J Pediatr*. 2017;188:224-231.e5.
20. Meaney C, Widdifield J, Jaakkimainen L, et al. Using biomedical text as data and representation learning for identifying patients with an osteoarthritis phenotype in the electronic medical record. *Int J Popul Data Sci*. 2018;3(4):168.
21. Gehrmann S, Dernoncourt F, Li Y, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One*. 2018;13(2):e0192360.
22. Liao KP, Sun J, Cai TA, et al. High-throughput multimodal automated phenotyping (MAP) with application to PheWAS. *J Am Med Inform Assoc*. 2019;26(11):1255-1262.
23. Nori VS, Hane CA, Sun Y, et al. Deep neural network models for identifying incident dementia using claims and EHR datasets. *PLoS One*. 2020;15(9):e0236400.
24. Ni Y, Bachtel A, Nause K, et al. Automated detection of substance use information from electronic health records for a pediatric population. *J Am Med Inform Assoc*. 2021;28(10):2116-2127.
25. Swerdel JN, Hripcsak G, Ryan PB. PheValuator: development and evaluation of a phenotype algorithm evaluator. *J Biomed Inform*. 2019;97:103258.
26. Chartier C, Gfrerer L, Austen WG. ChartSweep: a HIPAA-compliant tool to automate chart review for plastic surgery research. *Plast Reconstr Surg Global Open*. 2021;9(6):e3633.
27. Yu S, Liao KP, Shaw SY, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc*. 2015;22(5):993-1000.
28. Yu S, Chakrabortty A, Liao KP, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. *J Am Med Inform Assoc*. 2017;24(e1):e143-e149.
29. Nogues I-E, Wen J, Lin Y, et al. Weakly semi-supervised phenotyping using electronic health records. *J Biomed Inform*. 2022;134:104175.
30. Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. *J Biomed Inform*. 2010;43(6):891-901.
31. Wright A, Pang J, Feblowitz JC, et al. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. *J Am Med Inform Assoc*. 2011;18(6):859-867.
32. Agarwal V, Podchiyska T, Banda JM, et al. Learning statistical models of phenotypes using noisy labeled training data. *J Am Med Inform Assoc*. 2016;23(6):1166-1173.
33. Banda JM, Halpern Y, Sontag D, et al. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. *AMIA Jt Summits Transl Sci Proc*. 2017;2017:48-57.
34. Huang J, Duan R, Hubbard RA, et al. PIE: a prior knowledge guided integrated likelihood estimation method for bias reduction in association studies using electronic health records data. *J Am Med Inform Assoc*. 2018;25(3):345-352.
35. Tong J, Huang J, Chubak J, et al. An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. *J Am Med Inform Assoc*. 2020;27(2):244-253.
36. Yin Z, Tong J, Chen Y, et al. A cost-effective chart review sampling design to account for phenotyping error in electronic health records (EHR) data. *J Am Med Inform Assoc*. 2022;29(1):52-61.
37. Swerdel J, N, Schuemie M, Murray G, et al. PheValuator 2.0: methodological improvements for the PheValuator approach to semi-automated phenotype algorithm evaluation. *J Biomed Inform*. 2022;135:104177.
38. Gronsbell JL, Cai T. Semi-supervised approaches to efficient evaluation of model prediction performance. *J R Stat Soc B* 2018;80 (3):579-594.
39. Gronsbell J, Liu M, Tian LU, Cai T. Efficient evaluation of prediction rules in semi-supervised settings under stratified sampling. *J R Stat Soc Series B Stat Methodol*. 2022;84(4):1353-1391.
40. Van Calster B, McLernon DJ, Van Smeden M, et al.; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230-237.
41. Huang Y, Li W, Macheret F, et al. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc*. 2020;27(4):621-633.
42. Banda JM, Halpern Y, Sontag D, et al. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. *AMIA Jt Summits Transl Sci Proc*. 2017;2017:48-57.
43. Yu S, Ma Y, Gronsbell J, et al. Enabling phenotypic big data with PheNorm. *J Am Med Inform Assoc*. 2018;25(1):54-60.
44. Gronsbell J, Minnier J, Yu S, et al. Automated feature selection of predictors in electronic medical records data. *Biometrics*. 2019;75 (1):268-277.
45. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; 2003.
46. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
47. Minnier J, Tian L, Cai T. A perturbation method for inference on regularized regression estimates. *J Am Stat Assoc*. 2011;106 (496):1371-1382.

48. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.
49. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31(12):1102-1110.
50. Wand MP, Marron JS, Ruppert D. Transformations in density estimation. *J Am Stat Assoc*. 1991;86(414):343-353.
51. Silverman BW. *Density Estimation for Statistics and Data Analysis*. Routledge; 2018.
52. Sinnott JA, Cai T. Inference for survival prediction under the regularized Cox model. *Biostatistics*. 2016;17(4):692-707.
53. Agresti A. *Categorical Data Analysis*. John Wiley & Sons; 2012.
54. Tan WK, Heagerty PJ. Surrogate-guided sampling designs for classification of rare outcomes from electronic medical records data. *Biostatistics*. 2022;23(2):345-361.
55. Zhang Y, Chakrabortty A, Bradic J. Double robust semi-supervised inference for the mean: selection bias under MAR labeling with decaying overlap. *Information and Inference: A Journal of the IMA*. 2023;12(3):2066-2159.