Title

Constrained binary classification using ensemble learning: an application to cost-efficient targeted PrEP strategies

Journal

Statistics in Medicine, 37(2)

ISSN

0277-6715

Authors

Zheng, Wenjing
Balzer, Laura
van der Laan, Mark
et al.

Publication Date

2018-01-30

DOI

10.1002/sim.7296

Peer reviewed

HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2019 January 30.

Published in final edited form as:

Stat Med. 2018 January 30; 37(2): 261–279. doi:10.1002/sim.7296.

Constrained binary classification using ensemble learning: an application to cost-efficient targeted PrEP strategies

Wenjing Zheng[*,a], **Laura Balzer**[b], **Mark van der Laan**[a], **Maya Petersen**[a], and **the SEARCH Collaboration**

[a]Division of Biostatistics, School of Public Health, University of Calfornia, Berkeley

[b]Dept of Biostatistics, Havard T.H. Chan School of Public Health

Abstract

Binary classifications problems are ubiquitous in health and social sciences. In many cases, one wishes to balance two competing optimality considerations for a binary classifier. For instance, in resource-limited settings, an HIV prevention program based on offering Pre-Exposure Prophylaxis (PrEP) to select high-risk individuals must balance the sensitivity of the binary classifier in detecting future seroconverters (and hence offering them PrEP regimens) with the total number of PrEP regimens that is financially and logistically feasible for the program. In this article, we consider a general class of constrained binary classification problems wherein the objective function and the constraint are both monotonic with respect to a threshold. These include the minimization of the rate of positive predictions subject to a minimum sensitivity, the maximization of sensitivity subject to a maximum rate of positive predictions, and the Neyman-Pearson paradigm, which minimizes the type II error subject to an upper bound on the type I error. We propose an ensemble approach to these binary classification problems based on the Super Learner methodology. This approach linearly combines a user-supplied library of scoring algorithms, with combination weights and a discriminating threshold chosen to minimize the constrained optimality criterion. We then illustrate the application of the proposed classifier to develop an individualized PrEP targeting strategy in a resource-limited setting, with the goal of minimizing the number of PrEP offerings while achieving a minimum required sensitivity. This proof of concept data analysis uses baseline data from the ongoing Sustainable East Africa Research in Community Health study.

Keywords

Super Learner; constrained binary classification; Neyman-Pearson; sensitivity; Rate of Positive Predictions; PrEP; ensemble classification; cross-validation

1. Introduction

Binary classifications problems often arise in health and social science applications, wherein individuals classified into the ‘positive’ class are to receive an intervention of interest, which

[*]Correspondence to: wenjing.zheng@berkeley.edu.

caries with it an associated resource cost. Therefore, it is often desirable, especially in resource-limited settings, to strike a balance between capacity constraints and the sensitivity of the classification algorithm. For example, consider a targeted HIV prevention strategy that prescribes a Pre-Exposure Prophylaxis (PrEP) regimen to individuals with substantial risk of infection. Delivery of PrEP requires a meaningful resource expenditure per individual treated, including ongoing medication and monitoring costs [1]. WHO Guidelines advocate targeting PrEP to subpopulations known to be at high risk for HIV infection [2]. However, within a generalized epidemic, the optimal demographic subgroups to target may not be self-evident, and simply offering PrEP to known high-risk subgroups, such as young women, or mobile populations, may be inefficient. In other words, a strategy that targets PrEP based on a more sophisticated use of individual characteristics may be able to reduce the resource spending per new HIV infection prevented. A natural question, therefore, is 'how can individual characteristics be used to offer targeted PrEP in order to prevent as many new HIV infections as possible, given some fixed constraint on the total number of PrEP regimens offered?'. This questions translates into a binary classification problem that aims to maximize sensitivity, subject to a constraint on the Rate of Positive Predictions (RPP). Alternatively, one might ask 'how should PrEP be targeted at the individual-level in order to minimize the number of PrEP regimens offered while preventing a desired percentage of new infections?' This question translates into a binary classification problem that minimizes the RPP, subject to a sensitivity constraint.

These two problems are ubiquitous in devising cost-efficient intervention or prevention strategies. In fact, in many real-world applications, the cost of misclassification may be much higher in one class than the other, or one may wish to balance two competing optimality considerations for a binary classifier. To this end, we propose in this article a general group of Super Learner-based binary classifiers that aim to satisfy a wide class of performance-constrained optimality criteria. Super Learner [3] is an ensemble learning method in which a user-supplied library of algorithms are combined through a convex weighted combination, with the optimal weights selected to minimize a cross-validated empirical risk specified by the user. It can accommodate large classes of user-specified objective functions; standard implementations include optimizing the squared error loss or the log-likelihood loss. Theoretical results [4–6] exist to guarantee that the ensemble algorithm improves upon any of its constituent algorithms asymptotically. We first consider the binary classification problem of minimizing the Rate of Positive Predictions, subject to achieving a minimum sensitivity requirement. The proposed Super Learner-based binary classifier is characterized by combination weights and a discriminating threshold for classification that together aim to minimize a sensitivity-constrained RPP. Next, we describe how the proposed method can be adapted to the converse problem of maximizing a RPP-constrained sensitivity. We then further extend the proposed Super Learner to a larger group of performance-constrained binary classification problems where the objective function and the constraint function are monotonic in the same direction with respect to the threshold function. This type of classification problem includes the Neyman-Pearson paradigm [7] which minimizes the type II error subject to an upper bound on the type I error.

As an illustration of the proposed method, we develop and evaluate a hypothetical HIV prevention strategy that uses a Super Learner-based binary classifier to offer PrEP to

selected individuals, with the goal of minimizing the number of PrEP offerings while achieving a minimum target sensitivity. We use baseline data from the Sustainable East Africa Research in Community Health (SEARCH, NCT01864603) study, an ongoing cluster randomized HIV "test and treat" trial in rural Kenya and Uganda, to illustrate the development and evaluation of this targeted PrEP algorithm. We compare its projected performance to standard subgroup-based PrEP strategies, which rely on broad demographic categories (e.g. young women, fishermen). In this example, classifiers are trained to predict baseline (prevalent) HIV status using individual-level demographics and other risk factor variables collected at baseline. We note that in real-world development of such a targeted PrEP algorithm, one would instead train the classifier to predict HIV seroconversions among baseline HIV uninfected individuals. In the SEARCH Study, the method was applied to interim seroconversion data from the intervention arm of the trial to develop an empirically evaluated classifier. This classifier is being used in the second phase of the study to offer PrEP to those who do not self-recommend or who are not in a serodiscordant relationship. However, as these seroconversions are interim primary outcomes of the ongoing SEARCH study, they will not be used in this example. We also employ in this example a second-level cross-validation evaluation scheme to assess and compare the performance (in terms of sensitivity and capacity savings) of different classifiers. This scheme seeks to mimic, to the extent possible, an intervention in which the classifier is trained on a random subsample of the population and applied to the remaining individuals. In this sense, we believe it to be a more pragmatic approach to evaluating the performance of a classifier developed with this objective than the standard area under the ROC curve [8].

1.1. Literature overview

A general solution to binary classification with performance constraints has been proposed by Bounsiar et al. [9] within the context of statistical hypothesis testing, and encompasses the problems considered in the current paper. While the solutions developed by Bounsiar et al. [9] have universal applicability, their implementations are, to the best of our understanding, with respect to specific classification or prediction algorithms, and therefore may not be immediately translatable to ensemble learning, which allows one to combine several algorithms, and may have a higher technical barrier for implementation.

Of the class of performance-constrained binary classification problems we consider here, the Neyman-Pearson paradigm is perhaps the most common one. The theoretical properties of single classifiers that solve the corresponding constrained optimization problem with biased versions of the empirical False Negative Rate and empirical False Positive Rate were studied by Cannon et al. [10] and Scott and Nowak [11]. Theoretical properties of an ensemble classifier based on convex-weighted majority vote of the constituent classifiers, with weights solving the corresponding convex optimization problem, were studied in Rigollet and Tong [12]. In the current paper, we show that the performance-constrained problems considered, including the Neyman-Pearson paradigm, can be recast as optimization of the objective function evaluated at an appropriate threshold, and therefore applicable beyond problems with convex objective and performance functions. We also approach the ensemble differently, by employing cross-validated versions of the objective functions and

performance constraints to reduce overfitting, and by developing both a scoring function and a discriminating threshold to obtain a final classifier, instead of combining base classifiers.

In applications in HIV treatment, the use of individualized rules to offer selective HIV viral load testing to detect treatment failure in resource-limited settings had been proposed by Liu et al. [13] and Petersen et al. [14], among others. Liu et al. [13] models the distribution of the risk score (based on a user-supplied scoring scheme) through a nonparametric or semi-parametric approach, and seeks a tripartite rule that minimizes a user-specified weighted combination of False Negative Rate and False Positive Rate, subject to a RPP constraint. In this sense, this program aims to satisfy a different goal than the RPP-constrained sensitivity or the Neyman-Pearson paradigm. While this constrained optimality criterion does not fall into the class we study here, it is an optimization objective that is common to many applications. The synergy between this work and the current paper would be a promising direction of research.

Petersen et al. [14] proposed a Super Learner-based binary classifier to identify patients for selective viral load testing based on routinely collected data. This classifier first obtains risk prediction using the standard Super Learner which optimizes the log-likelihood loss (described in section 2.2.2). A second-level cross-validation scheme is used to evaluate the performance of classifiers (our proposed evaluation scheme models after this one). The general performance of a classifier is summarized using the cross-validated area under the ROC curve across a range of discriminating thresholds. For a given lower bound on sensitivity, the cross-validated ideal RPP of a classifier is obtained by first computing on each validation set the RPP under the largest threshold for which the sensitivity criterion is satisfied, and then averaging this ideal RPP across the validation sets. This is the ‘ideal’ RPP in that it uses the threshold one would have chosen if given the data-generating distribution of the evaluation data, not a threshold estimated from the learning data. The methods proposed in the current paper build upon and extend those in [14] in that the Super Learner weights are now optimized for the target constrained classification criterion, construction of the discriminating threshold is built into the classifier development, and the evaluation scheme assesses the empirical RPP under the score function–threshold duo.

1.2. Organization

This article is organized as follows. In section 2.1 we formulate the binary classification problem of minimizing RPP subject to a sensitivity constraint. In section 2.2 we propose a cross-validated objective function and the implementation of a Super Learner-based classifier that aims to optimize this objective function. In sections 3.1 and 3.2, we describe how the proposed formulation can be extended to the converse problem of maximizing sensitivity subject to a RPP constraint, and to a general class of binary classification problem with monotonic objective function and performance constraints. The corresponding Super Learner classifier is described in section 3.3. In section 4, we illustrate the development and evaluation of a targeted PrEP strategy based on the Super Learner classifier proposed in section 2. We conclude the article with a summary.

2. Sensitivity-constrained minimization of the rate of positive predictions

2.1. Problem formulation

Consider the observed data structure $O = (Y, W) \sim P_0$, with $Y \in \{0,1\}$ a binary class of interest and W a set of covariates. For a score function $\psi : \mathcal{W} \rightarrow [0,1]$, and a threshold c, the pair (ψ, c) defines a binary classification algorithm on W, wherein $\psi(W) \geq c$ is classified to the class $Y = 1$. Our goal is to learn a classification procedure that achieves a sensitivity of at least s_0, for some user-specified $s_0 \in (0,1)$, with a minimal Rate of Positive Predictions.

The sensitivity of (ψ, c) under a data-generating distribution P is given by

$$s(P;\psi,c) \equiv P(\psi(W) \geq c | Y=1). \quad (1)$$

Note that $s(P;\psi, c)$ is monotonically non-increasing in c. In particular, for every ψ, we can define a unique *sensitivity threshold for* ψ *under* P as:

$$c(P;\psi) \equiv max\{c : s(P;\psi,c) \geq s_0\}. \quad (2)$$

In other words, $c(P,\psi)$ is the largest threshold for ψ under distribution P at which the sensitivity is at least s_0.

Consider an objective function for ψ, denoted $r(P;\psi, c)$, that is monotonically non-increasing in c. In this section, we take r to be the Rate of Positive Predictions:

$$r(P;\psi,c) \equiv P(\psi(W) \geq c).$$

For a fixed data-generating P_0, our goal is a binary classification algorithm (ψ, c) that satisfies the *sensitivity-constrained minimization*

$$min_{\psi,c} r(P_0;\psi,c) \text{ such that } s(P_0;\psi,c) \geq s_0. \quad (3)$$

Using the sensitivity threshold defined in (2), we can define a *sensitivity-constrained objective function* as

$$r(P_0;\psi) \equiv r(P_0;\psi, c(P_0,\psi)). \quad (4)$$

In words, this is the RPP of a classification procedure that combines the score function ψ with its sensitivity threshold under P_0. Our optimal binary classifier is thus given by $(\psi_0, c(P_0, \psi_0))$, where the optimal score function is

$$\psi_0 \equiv \arg\min_{\psi} r(P_0;\psi). \quad (5)$$

It is easy to see that the constrained minimization problem in (3) can be solved by (ψ_0, $c(P_0,\psi_0)$). Indeed, firstly, we know that (ψ_0, $c(P_0,\psi_0)$) satisfies the sensitivity constraint of (3). Secondly, suppose (ψ', c') also satisfies the sensitivity constraint. Since for fixed P_0 and ψ, s is a non-increasing function in c, the definition of $c(P_0,\psi')$ given in (2) implies that $c(P_0,\psi') \geq c'$. Since r is non-increasing in c, this inequality implies that $r(P_0;\psi', c') \geq r(P_0;\psi', c(P_0,\psi')) \equiv r(P_0;\psi')$. By definition of ψ_0 as a solution of (5), we know that $r(P_0;\psi') \geq r(P_0;\psi_0)$. Therefore $r(P_0;\psi', c') \geq r(P_0;\psi') \geq r(P_0;\psi_0) \equiv r(P_0;\psi_0, c(P_0,\psi_0))$. In other words, ($\psi_0$, $c(P_0,\psi_0)$) achieves the minimum of $r(P;\psi, c)$ under the constraint.

Consequently, we can solve the constrained minimization problem in (3) by minimizing the sensitivity-constrained objective function in (5). The latter problem seeks a score function ψ that minimizes the objective function when evaluated at its sensitivity threshold, compared to other score functions at their respective sensitivity thresholds. The formulation in (5) is more amenable to application under the existing Super Learner framework, and to asymptotic studies of a cross-validated sensitivity-constrained objective function. We will devote our attention to estimating this optimal classifier (ψ_0, $c(P_0,\psi_0)$).

2.2. Super Learner classifier to minimize the sensitivity-constrained RPP

In this section, we consider a Super Learner-based classifier that estimates the unknown optimal classifier defined in (5). Let $\mathcal{M}$ denote the set of all distributions for O, including the true unknown P_0, and $\mathcal{W}$ denote the outcome space of W. A scoring procedure $\Psi : \mathcal{M} \to \mathcal{W}^{[0,1]}$ inputs a distribution P and outputs a score function $\psi = \Psi(P) : \mathcal{W} \to [0,1]$. In most applications, it is often difficult to specify precisely how a large number of risk factors W interact to influence the outcome of interest. Therefore, we use a nonparametric model for $\mathcal{M}$. In such cases, an ensemble learning method such as Super Learner would allow one to invoke a wide array of scoring procedures, both parametric and nonparametric.

For a measurable function $f(O)$ of the data, and a distribution P, we will use the notation $Pf \equiv E_P(f(O))$.

2.2.1. Cross-validated sensitivity-constrained RPP—We described an objective function (4) for our classification problem, and appointed its minimizer (5) to be our unknown optimal binary classifier. Therefore, estimating this objective function is central to our tasks of assessing the performance of candidate algorithms and selecting the optimal among them. To provide protection against overfitting, we will accomplish these tasks using cross-validation.

Consider a split of a sample of n independent and identically distributed (i.i.d.) copies of O into a *validation set* and a *training set*. This can be represented by a random vector $B \in \{v, t\}^n$, indicating whether each of the n observations is in the validation set (v) or the training set (t). We use P_n to denote the empirical distribution of the n i.i.d. observations, $P^v_{n,B}$ the

empirical distribution of the validation set, and $P^t_{n,B}$ the empirical distribution of the training set. Note that in our notation for B, we suppressed the fact that B depends on n. The particular choice of cross-validation procedure is characterized by the outcome space and distribution for B. For instance, in an M-fold cross-validation, the distribution would place weight $1/M$ to each of the M vectors corresponding to each of the M folds.

We define the *empirical cross-validated sensitivity-constrained RPP* of Ψ as

$$r_n(P_n, \Psi) \equiv E_B \, r\left(P^v_{n,B}; \Psi(P^t_{n,B})\right). \quad (6)$$

In words, for a sample spilt B, we obtain the constrained objective $r\left(P^v_{n,B}; \Psi(P^t_{n,B})\right)$ as follows:

1. Fit Ψ on the training set $P^t_{n,B}$ to obtain a score function $\psi_{n,B} \equiv \Psi(P^t_{n,B}): \mathscr{W} \to [0, 1]$.
2. Obtain the sensitivity threshold $c_{n,B} \equiv c(P^v_{n,B}, \psi_{n,B})$ of this score function under the empirical distribution of the validation set. That is, we apply $\psi_{n,B}$ to obtain scores for the validation set observations, and find the largest threshold c for which the sensitivity constraint is satisfied, i.e. $P^v_{n,B} I(\psi_{n,B}(W) \geq c, Y{=}1) / P^v_{n,B} I(Y{=}1) \geq s_0$. This can be implemented using the quantile function on the observations in the validation set with $Y = 1$.
3. The constrained objective $r\left(P^v_{n,B}; \Psi(P^t_{n,B})\right)$ is given by the $\mathrm{RPP}\, P^v_{n,B} I(\psi_{n,b}(W) \geq c_{n,B})$, i.e. the proportion of the observations in the validation set whose score under $\psi_{n,B}$ surpasses the corresponding threshold $c_{n,B}$.

Note that this empirical cross-validated sensitivity-constrained RPP in (6) is an estimator for the *oracle cross-validated sensitivity-constrained RPP*

$$r_0(P_n, \Psi) \equiv E_B \, r(P_0; \Psi(P^t_{n,B})). \quad (7)$$

In words, if we knew P_0, we would fit Ψ on the training set to obtain the score function $\psi_{n,B}$, and then determine the sensitivity-constrained threshold and corresponding RPP for this score function $\psi_{n,B}$ under the true P_0. This is the true conditional sensitivity-constrained RPP of the procedure Ψ, conditional on being fitted on the training sets under the specified cross-validation procedure on a sample of size n.

2.2.2. Super Learner: a general overview—Super Learner is a generalized stacking learning method that finds the best convex combination of a given set of constituent

procedures for a user-specified optimality criteria. Suppose we have J constituent scoring procedures $\Psi^1, \ldots, \Psi^J$. A constituent procedure may be a pre-specified parametric regression model, as well as machine learning approaches such as neural networks and random forests. It can also be augmented with a screening algorithm (e.g. only using variables that pass a correlation criterion).

For α in the (J−1)-simplex Δ^J, we define

$$\Psi_\alpha(P) \equiv \sum_{j=1}^{J} \alpha^j \Psi^j(P).$$

Each Ψ_α is thus an algorithm that takes J independent variables, which are the scores from the J constituent algorithms, and combines them through the linear combination given by α.

To apply the framework from the previous section, we can consider a representation Δ_n^J of Δ^J by partition into $K(n)$ many grids with size converging to 0 (e.g. size $1/n^q$ for $q > 0$). As discussed in van der Laan et al. [3], minimization over Δ_n^J vs Δ^J would produce asymptotically equivalent procedures.

Consider an M-fold sample split, with $P_{n,m}^v$ and $P_{n,m}^t$ denoting the m-th empirical distributions of the validation and training sets, respectively. Standard implementations of Super Learner [15] use the minus log-likelihood loss or a squared error loss as optimality criteria. Specifically, they produce predictor Ψ_{α_n}, where α_n minimizes $\frac{1}{M}\sum_{m=1}^{M} P_{n,m}^v L(\Psi_\alpha(P_{n,m}^t))(O)$, with $L(\Psi_\alpha(P_{n,m}^t))(O)$ being the minus log-likelihood loss

$$-\{Y \log \Psi_\alpha(P_{n,m}^t)(A,W) + (1-Y)\log(1-\Psi_\alpha(P_{n,m}^t)(A,W))\}, \quad (8)$$

or the squared-error loss $(Y-\Psi_\alpha(P_{n,m}^t)(A,W))^2$. A Super Learner that maximizes the area under the ROC curve is presented in LeDell et al. [16].

2.2.3. Super Learner classifier for the proposed problem—Now we are ready to present a Super Learner for the binary classification problem under consideration. The goal is to find the optimal weight α, and a corresponding threshold c.

The proposed Super Learner scoring function, which optimizes the constrained criterion in (3), is given by Ψ_{α_n}, where α_n minimizes the empirical cross-validated objective function in (6):

$$\alpha_n \equiv \arg\min_{\alpha \in \Delta_n^J} r_n(P_n, \Psi_\alpha) = \arg\min_{\alpha \in \Delta_n^J} \frac{1}{M}\sum_{m=1}^{M} r\left(P_{n,m}^v; \sum_j \alpha^j \Psi^j(P_{n,m}^t)\right) \quad (9)$$

In words, for each α, we implement the function $r_n(P_n, \Psi_\alpha)$ as follows:

1. At m-th fold, fit each Ψ^j on the training set to obtain a score function $\Psi^j(P^t_{n,m}):\mathscr{W} \to [0,1]$, and then use α to combine these to produce an ensemble score function $\sum_j \alpha^j \Psi^j(P^t_{n,m}) \equiv \Psi_\alpha(P^t_{n,m}):\mathscr{W} \to [0,1]$.
2. Then use the validation set $P^v_{n,m}$ to obtain the sensitivity threshold $c\left(P^v_{n,m}, \Psi_\alpha(P^t_{n,m})\right)$ and the corresponding sensitivity-constrained RPP $P\left(\Psi_\alpha(P^t_{n,m}) \geq c(P^v_{n,m}, \Psi_\alpha(P^t_{n,m}))\right)$ for this combined score function.
3. The desired $r_n(P_n, \Psi_\alpha)$ is given by the average of such fold-specified sensitivity constrained RPPs across the M folds.

To obtain the desired optimal α_n, we can use a nonlinear optimization algorithm such as the `nloptr` package in in `R` [17].

To complete the classifier, we now require a threshold. The score function Ψ_{α_n} is one that has minimal (cross-validated) RPP at its sensitivity threshold. Therefore, we now focus our efforts on estimating its sensitivity threshold. Following analogous procedure, consider the *empirical cross-validated sensitivity* of a classification procedure based on a scoring procedure Ψ and threshold c:

$$s_n(P_n;\Psi,c)=\frac{1}{M}\sum_{m=1}^{M} s\left(P^v_{n,m};\Psi(P^t_{n,m}),c\right)=\frac{1}{M}\sum_{m=1}^{M} P^v_{n,m}(\Psi(P^t_{n,m})(W)\geq c|Y=1). \quad (10)$$

This is an estimator of the *oracle cross-validated sensitivity*

$$s_0(P_n;\Psi,c)=\frac{1}{M}\sum_{m=1}^{M} s\left(P_0;\Psi(P^t_{n,m}),c\right).$$

This latter is the true conditional sensitivity of Ψ under threshold c, conditional on the training sets used to fit the scoring procedure. The sensitivity threshold for Ψ_{α_n} can then be estimated by finding a threshold that satisfies the constraint on the empirical cross-validated sensitivity:

$$c_n \equiv \max\{c\in(0,1):s_n(P_n;\Psi_{\alpha_n},c)\geq s_0\}. \quad (11)$$

The final classifier is given by the pair $(\Psi_{\alpha_n}(P_n), c_n)$, where the score function $\Psi_{\alpha_n}(P_n)=\sum_j \alpha^j_n \Psi^j(P_n)$ is obtained by using α_n to combine the constituent score functions fitted on the full dataset. It classifies a given W as $I(\Psi_{\alpha_n}(P_n)(W)\geq c_n)$.

2.2.4. Case-control sampling in applications with rare outcomes—In the HIV example considered in this paper, as well as in other applications, the outcomes of interest

may be rare. In such cases, irrespective of the objective function considered, instead of using the full sample, the Super Learner can use a case-control subsample [18, 19] that consists of all the H cases in the full sample plus a random sample of $(C-1)\times H$ controls, for a user-specified C. Each observation in the subsample will be weighted by the inverse of its probability of being sampled from the learning data: cases will have weights 1, controls will have weights given by the number of controls in the full data divided by the number of controls in the subsample. Subsequently, the algorithm fits on the training set, as well as the fold-specific evaluations of the constraint and objective function, will use weighted observations. Moreover, we can implement the Super Learner using a M-fold sample split that is stratified by outcome case, and thus ensuring that the validation sets have similar number of cases.

3. More general performance-constrained binary classification problems

In section 2, we considered a Super Learner-based binary classifier that minimizes the RPP subject to achieving a minimum sensitivity. In this section, we first consider the converse to this problem: maximizing the sensitivity subject to an upper bound on the RPP. We then unify these two under a larger class of constrained binary classification problems.

3.1. RPP-constrained maximization of sensitivity

Suppose our goal now is to learn a classification procedure that can achieve maximal sensitivity subject to an upper bound s_0 on the RPP, for some user-specified $s_0 \in (0,1)$. To keep the language and notations parallel, we will formulate this problem in terms of minimizing the False Negative Rate (FNR), subject to a minimum Rate of Negative Predictions (RNP).

The RNP of a classifier (ψ, c) under a data-generating distribution P is given by

$$s(P;\psi,c) \equiv P(\psi(W)<c). \quad (12)$$

This is the cumulative distribution of $\psi(W)$, and hence is monotonically non-decreasing in c. In particular, for every ψ, we can define a unique *RNP threshold for ψ under P* as:

$$c(P;\psi) \equiv min\{c:s(P;\psi,c) \geq s_0\}. \quad (13)$$

In other words, $c(P,\psi)$ is the smallest threshold for ψ under distribution P at which the RNP is at least s_0.

Consider the objective function for ψ, denoted $r(P;\psi, c)$, to be the False Negative Rate:

$$r(P;\psi,c) \equiv P(\psi(W)<c|Y=1).$$

Like $s(P;\psi, c)$, $r(P;\psi, c)$ is also non-decreasing in c.

For a fixed data-generating P_0, our goal is a binary classification algorithm (ψ, c) that satisfies the *RNP-constrained minimization*

$$min_{\psi,c} r(P_0;\psi,c) \text{ such that } s(P_0;\psi,c)>s_0. \quad (14)$$

Using the RNP threshold defined in (13), we can define a *RNP-constrained objective function* as

$$r(P_0;\psi) \equiv r(P_0;\psi,c(P_0,\psi)). \quad (15)$$

In words, this is the FNR of a classification procedure that combines the score function ψ with its RNP threshold under P_0. Our optimal binary classifier is thus given by (ψ_0, $c(P_0,\psi_0)$), where

$$\psi_0 \equiv \arg\min_{\psi} r(P_0;\psi). \quad (16)$$

It is easy to see that the constrained minimization problem in (14) can be solved by (ψ_0, $c(P_0,\psi_0)$). Indeed, firstly, we know that (ψ_0, $c(P_0,\psi_0)$) satisfies the RNP constraint of (16). Secondly, suppose (ψ', c') also satisfies the RNP constraint. Since for fixed P_0 and ψ, s is a non-decreasing function in c, the definition of $c(P_0,\psi')$ given in (13) implies that $c(P_0,\psi') \leq c'$. Since r is non-decreasing in c, this inequality implies that $r(P_0;\psi', c') \geq r(P_0;\psi', c(P_0,\psi')) \equiv r(P_0;\psi')$. By definition of ψ_0 as a solution of (16), we know that $r(P_0;\psi') \geq r(P_0;\psi_0)$. Therefore $r(P_0;\psi', c') \geq r(P_0;\psi') \geq r(P_0;\psi_0) \equiv r(P_0;\psi_0, c(P_0,\psi_0))$. In other words, ($\psi_0$, $c(P_0,\psi_0)$) achieves the minimum of $r(P;\psi, c)$ under the constraint.

3.2. A general class of performance-constrained binary classification problems

The two constrained binary classification problems we considered in section 2 and 3.1 can be generalized to a larger class of constrained binary classification problems where the objective function and the constraint are monotonic with respect to the threshold.

Specifically, for a binary classifier characterized by a score function ψ and a threshold c, we wish to minimize an objective function $r(P_0;\psi, c)$ that is monotonic in c, subject to a constraint $s(P_0;\psi, c) \geq 0$, where the *constraint function* $s(P_0;\psi, c)$ is also monotonic in c. Suppose the constraint function s is monotonic in c in the same direction of the objective function r — that is, either both are non-decreasing in c or both are non-increasing in c. Then, we can define $c(P_0,\psi) \equiv \max\{c : s(P_0;\psi, c) \geq 0\}$, in the non-increasing case, and $c(P_0;\psi) \equiv \min\{c : s(P_0;\psi, c) \geq 0\}$, in the non-decreasing case. In the two problems we considered previously, the RPP and the minimal sensitivity requirement correspond to non-increasing objective and constraint, and the FNR and the minimal RNP requirement corresponds to a non-decreasing objective and constraint.

The constrained binary classification problem of

$$\min_{\psi,c} r(P_0;\psi,c) \text{ such that } \tilde{s}(P_0;\psi,c) \geq 0$$

can thus be solved by $(\psi_0, c(P_0,\psi_0))$ where

$$\psi_0 \equiv \arg\min_{\psi} r(P_0;\psi,c(P_0,\psi)).$$

Indeed, if a pair (ψ', c') satisfies the constraint, then either $c' \leq c(P_0,\psi')$ and $r(P_0;\psi', c') \geq r(P_0;\psi', c(P_0,\psi'))$ in the non-increasing case, or $c' \geq c(P_0,\psi')$ and $r(P_0;\psi', c') \geq r(P_0;\psi', c(P_0,\psi'))$ in the non-decreasing case. Hence, in both cases, $r(P_0;\psi', c') \geq r(P_0;\psi', c(P_0,\psi')) \geq r(P_0;\psi_0, c(P_0,\psi_0))$, by definition of ψ_0.

This group of classification problems includes most constraint and objective functions that are the traditional performance metrics, and addresses many applications where one must balance competing performance criteria. In particular, it includes the commonly known Neyman-Pearson criterion, which aims to minimize type II error (i.e. minimize False Negative Rate) with an upper bound on type I error (i.e. lower bound on True Negative Rate).

3.3. Super Learner

Once the parallel formulation to the problem considered in section 2 is established, the corresponding Super Learner-based classifier can be obtained in a similar manner. We will not repeat the entire description here, but only highlight the relevant modifications.

The empirical cross-validated objective value $r_n(P_n,\Psi_\alpha)$ of each potential weight α is obtained as follows. At fold m, fit each constituent algorithm Ψ_j on the training set to produce the combined score function $\Psi_\alpha(P^t_{n,m}) \equiv \sum_j \alpha^j \Psi^j(P^t_{n,m})$. To compute the threshold $c(P^v_{n,m}, \Psi_\alpha(P^t_{n,m}))$ of this score function under the empirical distribution of the validation set, we apply $\Psi_\alpha(P^t_{n,m})$ to obtain scores for the validation set observations, and either find the largest threshold c, in the case of non-increasing objective and constraint, or find the smallest threshold c, in the case of non-decreasing objective and constraint, among those satisfying the constraint, i.e. among the set $\{c:\tilde{s}(P^v_{n,m};\Psi_\alpha(P^l_{n,m}),c) \geq 0\}$. The corresponding constrained objective value of Ψ_α on this fold is thus $r(P^v_{n,m};\Psi_\alpha(P^t_{n,m}), c(P^v_{n,m}, \Psi_\alpha(P^t_{n,m})))$, i.e. the objective function evaluated at the score function $\Psi_\alpha(P^t_{n,m})$ and its corresponding constraint threshold $c(P^v_{n,m}, \Psi_\alpha(P^t_{n,m}))$. The desired cross-validated objective value of Ψ_α is thus given by the average of such fold-specific objective values:

$$r_n(P_n, \Psi_\alpha) \equiv \frac{1}{M}\sum_m r\left(P^v_{n,m}; \Psi_\alpha(P^t_{n,m}), c\left(P^v_{n,m}, \Psi_\alpha(P^t_{n,m})\right)\right).$$

The Super Learner weights α_n is the weight vector that minimizes $r_n(P_n, \Psi_\alpha)$. In the RNP-constrained minimization of FNR considered in section 3.1, the threshold for the fold m would be the smallest c such that $P^v_{n,m} I(\Psi_\alpha(P^t_{n,m})(W) < c) - s_0 \geq 0$, and the corresponding objective value on this fold is the $\mathrm{FNR}\ P^v_{n,m} I(\Psi_\alpha(P^t_{n,m})(W) < c, Y = 1)/P^v_{n,m} I(Y = 1)$.

Correspondingly, the *empirical cross-validated constraint function* $\tilde{s}$ of a classification procedure based on scoring procedure Ψ and threshold c is

$$\tilde{s}_n(P_n; \Psi, c) = \frac{1}{M}\sum_{m=1}^{M} \tilde{s}(P^v_{n,m}; \Psi(P^t_{n,m}), c).$$

Consequently, the threshold for our score function Ψ_{α_n} can be estimated by finding a threshold that satisfies the empirical cross-validated constraint:

$$c_n \equiv \max\{c : \tilde{s}_n(P_n; \Psi_{\alpha_n}, c) \geq 0\} \text{ and } c_n \equiv \min\{c : \tilde{s}_n(P_n; \Psi_{\alpha_n}, c) \geq 0\},$$

in the non-increasing and the non-decreasing cases, respectively.

The final classifier is given by the pair $(\Psi_{\alpha_n}(P_n), c_n)$, where the score function $\Psi_{\alpha_n}(P_n) = \sum_j \alpha^j_n \Psi^j(P_n)$ is obtained by using α_n to combine the constituent scoring procedures fitted on the full dataset. It classifies a given W as $I(\Psi_{\alpha_n}(P_n)(W) \geq c_n)$.

The comments in section 2.2.4 on case control sampling in applications with rare outcomes naturally apply here.

Note that our proposed method focuses on how to combine a given set of constituent scoring procedures and how to find a discriminating threshold to satisfy the constrained optimality criterion of interest. We have not discussed how each of these constituent scoring procedures should be fitted. Standard fits for these procedures are not tailored for the desired constrained optimality criterion. One can derive for each constituent procedure the optimal fit for the target criterion under consideration, but it may make certain algorithms difficult to implement and raise the technical barrier for the use of these classification tools. Using the standard fits for the constituent procedures could be a practical tradeoff, at the expense of potentially shortchanging performance compared to using fits tailored for the target criterion. In the summary section we will discuss future work to investigate this tradeoff.

4. Application to an individualized targeted PrEP strategy

4.1. Background

We now consider an example from HIV prevention. Pre-exposure prophylaxis (PrEP) is an HIV prevention method in which uninfected individuals follow a regimen of antiretroviral medication to reduce their risk of infection. As of September 2015, the World Health Organization recommends that individuals with high risk of HIV infection be offered PrEP as part of a comprehensive prevention strategy [2]. The success of this prevention tool relies on consistent use of the medication and regular monitoring, leading to considerable resource expenditure associated with each PrEP regimen. Therefore, for long-term sustainability, prevention programs need strategies for identifying high risk individuals for PrEP eligibility that optimize population level impact within resource constraints. In regions with generalized epidemics, offering PrEP to known demographic risk groups may be neither optimally effective nor optimally efficient. The highest risk subgroups, such as individuals in a serodiscordant relationship, may represent only a minority of total new infections in the general population, while broader demographic groups, such as young women, that include a larger proportion of new infections may have too low an incidence to form the basis of a cost-efficient targeting strategy. Flexible machine learning methods for building individual risk scores that appropriately tradeoff sensitivity and constrained roll out therefore have the potential to improve the impact and sustainability of PrEP as an HIV prevention tool.

In this example, we consider a hypothetical PrEP-based prevention program in Eastern Uganda. The goal of this program is to offer PrEP to select HIV uninfected individuals in the target population in order to prevent 80% of new infections, while keeping the number of such offerings to a minimum. To this end, we would like an algorithm that uses individual-level data to identify prospective seroconverters with a sensitivity of at least 80% while minimizing the number of positive predictions. To further illustrate strategy development, we consider an implementation scenario where, while the algorithm training has at its disposal a large array of variables, at the program rollout only a limited number of variables can be collected at real-time on the prospective individuals. Consequently, the constrained optimality criterion will also be used to select a small subset of the variables to be used in the implemented algorithm. We will compare the performance of the targeted Super Learner-based strategy to a conventional subgroup-based strategy wherein one offers PrEP to everyone in a pre-specified subgroup defined by strata of demographic factors. In this example, we could also use a standard implementation of the Super Learner with a minus log-likelihood loss function, as carried out in [14] for predicting viral load failure among HIV patients on treatment (more detail in section 1.1). This standard implementation is not designed to optimize the constrained criterion under consideration, but it is still of interest for our application. It will be included in our example for comparison.

4.2. Methods

4.2.1. Data, target population and outcome of interest—In this example, we will use baseline data from the SEARCH study to illustrate the development and demonstrate applicability of such an targeted PrEP algorithm. The SEARCH study is a cluster-randomized trial that includes 32 communities of roughly 5000 adults (age ≥ 15) each, in

rural Uganda and Kenya. The first phase of this study tests a community-level intervention that consists of annual community-based HIV and multi-disease testing, with immediate linkage to care, antiretroviral therapy (ART) eligibility for all HIV-infected individuals, and streamlined ART delivery using a patient-centered model. At baseline, the population of each community was enumerated through a door-to-door household census, and basic demographics (age, sex, marital status and occupation) were collected on all household members. Then, baseline HIV testing and other baseline data collection were performed during a community health campaign and subsequent home-based tracking for those that did not attend the campaign. We refer to Chamie et al. [20, 21] for a detailed exposition on the census and the community-based HIV and multi-disease testing campaign. In this example, we use baseline data from 10 communities in Eastern Uganda.

Our target population is adult community residents with a conclusive baseline HIV test result from these 10 communities. Our classifier will be trained to predict the baseline prevalent HIV status with the goal of achieving at least 80% sensitivity while minimizing the number of positive predictions. Importantly, this baseline data analysis is intended solely as a proof of concept; in designing a classifier for use in the actual targeted PrEP strategy deployed in the second phase of the SEARCH study, we instead trained the classifier to predict seroconversion outcomes among baseline HIV uninfected individuals. However, as these seroconversions are interim primary outcomes of the ongoing SEARCH study, this seroconversion analysis is not described here. We chose Eastern Uganda as an illustration of the method because it has the lowest baseline HIV prevalence, and is thus more comparable to a seroconversion outcome, which is expected to be rare.

4.2.2. Candidate predictors and models—In this example, we consider an implementation scenario where only a limited number of risk factors can be collected on the prospective individuals during the rollout of the program. Therefore, as part of the algorithm development the investigator must decide which subsets of the variables should be used. Suppose also that variables within the same domain can often be found in the same data source. Therefore to minimize the number of data sources needed at the program rollout, one would group the variables by domain:

- **Demographics**: age, gender, occupation, marital status, polygamy, educational attainment, and circumcision (for males).
- **Mobility**: number of months a resident had lived outside the community in the past year, number of nights spent in one’s residence in the past month.
- **Reproductive Health**: pregnancy in the past 12 months (females), whether self or partner is currently using contraception.
- **Drinking**: whether drink alcohol, frequency of binge drinking (defined as 6 or more drinks at once), number of days in a months drink alcohol, number of drinks in a typical day.
- **Depression**: Patient Health Questionnaire-2 score [22], Generalized Anxiety-2 score [23].

- **Work Productivity**: days worked in the past month, hours worked in a normal day in the past week.

From here onward, by a 'Model' we mean a combination of risk factor variables from these domains. For instance the model **Demographics.Mobility** would use the variables under the domains Demographics and Mobility. We will be considering models that combine Demographics with each one of the other domains. These make up a total of 6 models under consideration.

4.2.3. Building the Super Learner-based classification algorithm—For each of the models considered, we apply the Super Learner classifier described in section 2 to classify the baseline HIV status, with the goal of minimizing the Rate of Positive Predictions while achieving a sensitivity of at least 80%. The constituent algorithms consist of screening-scoring pairs. The scoring algorithms include Lasso regression [24], main term logistic regression, generalized additive model [25, 26], random forest [27], Bayes logistic regression [28], and recursive partitioning regression [29]. We will use the standard fits of these algorithms as implemented in `R`. Each of these candidate scoring algorithms is augmented with screening algorithms that either use a) all the variables, b) only the top 10% most correlated variables, or c) only variables with a T-test p-value of less than 0.1. We implement a Super Learner-based classifier that constructs a score function through a linear combination of the constituent algorithms, with weights minimizing the sensitivity-constrained RPP, and uses as its threshold the cross-validated sensitivity threshold in (11). The optimal weights are computed using an algorithm for finding global optima in the `nloptr` package ([17]) in `R`. Besides the proposed Super Learner, we can also use a standard implementation of the Super Learner prediction (with weights minimizing the standard risk associated with minus log-likelihood loss), coupled with the cross-validated sensitivity threshold in (11). We will call the former the *constrained RPP Super Learner*, and the latter the *log-likelihood* Super Learner. We will apply both Super Learner-based classifiers in this example for comparison.

As we described in section 2.2.4, to mitigate the low prevalence outcome, the Super Learner uses a case-control subsample from the input data that consists of all the H baseline HIV positive cases and a random sample (with replacement) of $(C-1)\times H$ controls, with $C = 10$. Each of these observations are inversely weighted by the probability of being sampled from the input dataset. We implement the Super Learner using a 10-fold sample split that is stratified by outcome case. This stratification means that each validation set will have approximately the same number of cases.

4.2.4. Performance assessment—We assess the performance of each classifier in terms of empirical sensitivity, as measured by the true positive rate, and the number needed to treat (NNT), as measured by the total number of positive predictions divided by the total number of cases identified. If a case consisted of a seroconversion (rather than, as here, a prevalent HIV case), NNT conveys the number of individuals offered PrEP per infection potentially prevented. Actual infections prevented would of course also depend on uptake and adherence to PrEP among those individuals to whom it was offered. NNT allows for capacity-spendings comparison across individual-based and subgroup-based strategies. The

empirical sensitivity and NNT are assessed through the average of 10 repetitions of a 10-fold split of the baseline target population into a learning dataset and an evaluation dataset. Specifically, we split the sample into 10 folds; on each fold, we use the learning dataset to learn the Super Learner classifier (characterized by weights a_n and threshold c_n, with 'full data' P_n being the learning dataset), and then apply it to classify the individuals in the evaluation set and obtain the fold-specific sensitivity and NNT measures of the classifier. We then average each performance measure across the 10 folds to obtain the cross-validated sensitivity and the cross-validated NNT of this classifier under the 10-fold split. Lastly, we repeat this 10-fold splitting and cross-validation evaluation scheme 10 times, and then average the resulting cross-validated sensitivity and cross-validated NNT. We call these the **average cross-validated sensitivity** (aCV-sensitivity) and the **average cross-validated NNT** (aCV-NNT), respectively. They would assess the average sensitivity and NNT of a strategy where we use a random subset of individuals in the population to train the classifier and apply the learned strategy to an independent sample from the same population.

These average cross-validated sensitivity and NNT measures can also be applied to evaluate the performance of subgroup-based strategies, wherein one only recommends PrEP to individuals in a pre-defined subgroup prescribed by baseline variable strata. In these cases, as there is no algorithm fitting in the learning set, the fold-specific sensitivity is the number of cases in the stratum in the validation set divided by the number of cases in the validation set, and the fold-specific NNT is the size of the stratum in the validation set divided by the number of cases in the stratum in the validation set. We believe the average cross-validated measures are more realistic assessments compared to the absolute sensitivity and NNT based on entire population stratum, since they mimic a real-world implementation where one learns, from a random sample, strata with highest risk of infection, and then subsequently offer PrEP to others in the population within those strata.

4.3. Results

The dataset consists of 44,762 adult (age 15 or older) residents from the 10 Eastern Ugandan communities enumerated in the SEARCH baseline survey, with conclusive baseline HIV test results. Of these, 1493 had a positive baseline HIV test (3.3% prevalence). In Table 1, we describe the baseline HIV status per stratum of key baseline variables. We reiterate here that since only baseline data is used in this example for illustration and proof of concept for the proposed classifier, the reader must not interpret the subject matter-specific results in this analysis as directly translatable to risk factors in seroconversion, nor the performance assessments as indicative of actual results expected from such a targeted PrEP strategy.

4.3.1. Subgroup-based strategies—We first considered more conventional subgroup-based PrEP strategies. A subgroup-based strategy recommends PrEP to everyone in a broad subgroup defined by specific strata of one or few demographic or risk factors. By contrast, the proposed Super Learner-based strategy (results in section 4.3.2) provides individualized PrEP recommendations based on a wide array of demographic and risk factor values on the individual. In this illustration, we considered all the subgroups that can be defined by using common demographic and risk factors variables fed into the Super Learner strategies. In the Table 1, each row in the table represents a subgroup given by a stratum of a demographic or

risk factor (examples of subgroups are all males, or all individuals aged 15–19). If we were to recommend PrEP to all those in the subgroup, then the average sensitivity and NNT one would achieve are depicted in the two right columns of the table.

Specifically, a strategy to offer PrEP to everyone in the population would have a sensitivity of 100%, at the cost of 30 individuals offered PrEP per infection potentially prevented; this should serve as a benchmark for the upper-bound cost of a PrEP prevention program. By way of comparison, if we were to offer PrEP to all those employed in the farming sector, we would achieve a sensitivity of 74% at the cost of 25.33 NNT. In general, a subgroup-based strategy using any one stratum in this table would have a cost of 30 NNT or greater in order to achieve a sensitivity of at least 80%. For an NNT less than 30, the highest sensitivity achieved is less than 75%.

Based on the above observation, an ad-hoc data-adaptive approach to building a targeted PrEP strategy might simply combine the most promising pre-specified subgroups; for example those with a sensitivity above 60% and an NNT less than 30. In our example, such an approach would offer PrEP to all women as well as men that are married and/or employed in farming. This subgroup has a total of 38,321 individuals (85% of the total population), with 1,457 positives. This strategy would have an average cross-validated performance of 98% sensitivity with a cost of 26.86 NNT. This ad-hoc strategy illustrates that the more variables we combine, the greater gain in capacity savings (less NNT for a given sensitivity level).

4.3.2. Super Learner-based strategies—Now, we turn to the performance of the proposed Super Learner-based PrEP strategy, calibrated to achieve at least 80% sensitivity while minimizing the rate of positive prediction.

The empirical performance of the constrained RPP Super Learner using each of the models considered in section 4.2.2, as assessed by the average cross-validated sensitivity and NNT, is depicted in Figure 1. The empirical sensitivities were about 80–81%, above the nominal 80% and thus satisfying the required constraint, with a cost of only 17–18 NNT. In other words, the proposed constrained RPP Super Learner-based strategies are less costly than the subgroup-based strategies in Table 1 that could yield over 70% sensitivity, and are more sensitive than subgroup-based strategies of similar cost.

We further contrast the performance of the constrained RPP Super Learner proposed in this paper with the standard log-likelihood Super Learner. The performance of the log-likelihood Super Learner-based classifier is depicted in Figure 2. The cross-validated sensitivity threshold again ensured that the sensitivity constraint is achieved in a new dataset. However, as this Super Learner predictor was optimized for the log-likelihood loss, not the RPP, the resulting classifier tends to overshoot the required sensitivity level, resulting in a higher NNT than that achieved by the constrained RPP Super Leaner.

We have seen in section 4.3.1 that a composite subgroup strategy (all women as well as men who are married and/or employed in farming) could yield a classifier that achieves 98% sensitivity with about 27 NNT. We also saw in Figure 2 that an individual strategy using a

log-likelihood Super Learner classifier could achieve a 98% sensitivity with about 29 NNT. Let us now consider the proposed constrained RPP Super Learner classifier calibrated to achieve at least 98% sensitivity. Its performance is depicted in Figure 3. To achieve 98% empirical sensitivity, such strategy would use about 25 NNT. To translate these performance metrics into implementation logistics, in a population with about 1500 cases, a strategy with 98% sensitivity at 25 NNT would result in $1500 \times .98 \times 25 = 36{,}750$ individuals offered PrEP in the population, and one at 27 NNT would result in about 39,690 individuals offered PrEP. In this case, an NNT difference of merely 2 points results in 3,000 more PrEP regimens being offered.

4.3.3. Interpretation—From this data analysis, we saw that, at least for rare outcome applications, principled individual-based strategies were generally more sensitive and less costly (for a given sensitivity level) than strategies based on pre-specified demographic subgroups. Composite subgroup-based strategies that use several predictor strata yielded larger gains in sensitivity and capacity savings than single subgroup-based strategies alone. However, such approaches remained more costly (i.e. required higher NNT for a given sensitivity) than an approach that used the proposed constrained RPP Super Learner to build a flexible individual based targeting strategy. In short, in this application at least, the use of a state-of the art machine learning approach (Super Learner) that employs an optimality criteria specifically aligned with the implementation objective of optimizing efficient and effective PrEP offerings can result in substantial performance improvements.

5. Summary

In this article, we proposed a Super Learner-based classifier for a class of constrained binary classification problems. As an illustration, we developed and evaluated a hypothetical HIV prevention strategy that uses this Super Learner-based binary classifier to offer PrEP on an individual basis, with the goal of minimizing the number of PrEP offerings while achieving the required proportion of new infections prevented.

Super Learner is an ensemble machine learning algorithm that combines its constituent algorithms linearly using weights that minimize a cross-validated user-supplied objective function. The constrained binary classification problems under consideration are the ones where the objective and constraint functions have the same monotonicity with respect to the discriminating threshold. As specific examples, we examined the minimization of the rate of positive predictions subject to a lower bound on the sensitivity, and the maximization of the sensitivity subject to an upper bound on the rate of positive predictions. To construct these classifiers, we first expressed the constrained optimization problem as the minimization of a constrained objective function. Then, we obtained a Super Learner score function with weights minimizing the cross-validated version of said function; the discriminating threshold of the corresponding binary classifier is one that satisfies the cross-validated version of the constraint.

In our targeted PrEP example, we used baseline data from the SEARCH study and trained the classifiers to predict baseline (prevalent) HIV status using individual-level demographics and other risk factor variables collected at baseline. The performance of this and other

standard subgroup-based classifiers was assessed in terms of sensitivity and NNT. These measures were obtained under a 10-fold sample-split evaluation scheme, wherein the classifiers were trained in the learning set, and their sensitivity and NNT were evaluated based on their performance in classifying the evaluation set. Averaging these performance measures across the 10 folds, we obtained a cross-validated sensitivity and NNT of each strategy. We conducted 10 repetitions of such 10-fold sample split evaluation to obtain as our final performance assessment an average cross-validated sensitivity and NNT for each classifier. For this application, we believe this empirical performance assessment to be a more pragmatic evaluation scheme than the standard area under the ROC curve, as deriving an appropriate threshold is part of the classifier development. In the results of this data analysis, we saw that Super Learner-based classifiers are generally more sensitive and less costly than subgroup-based strategies. Moreover, a Super Learner-based classifier that targets the desired constrained RPP may outperform (in terms of the desired capacity savings optimization), or at least perform as well as, a Super Learner-based classifier that targets the log-likelihood loss. In summary, such individualized classifiers targeting the desired optimality criterion offer great promise to applications in a heterogeneous population in which the desired strategy must balance complex logistics and scientific needs that may not be fully captured by standard loss functions.

In addition to using the empirical objective and constraint metrics described here as an evaluation scheme, we could also adopt an inferential approach, in which the oracle cross-validated sensitivity-constrained RPP (7) of a scoring procedure Ψ is considered a (data-adaptive) target parameter of interest (see Hubbard et al. [30] on data-adaptive target parameters). One can use a non-parametric MLE estimator (6) for this target parameter, and use bootstrap to obtain a confidence interval. However, bootstrap procedures may be prohibitively time-consuming when using machine learning algorithms on large datasets. Alternatively, we note that conditional on a fitted score function, this target parameter is path-wise differentiable and thus its efficient influence curve can be derived, providing basis for influence curve-based confidence intervals. This approach has been proposed in LeDell et al. [31] with the area under the ROC as performance metric and target parameter. Besides the nonparametric MLE estimator, for finite sample gain, we can also use Targeted Maximum Likelihood Estimator [32] or its cross-validated version [33] to estimate this target parameter. The latter may help reduce second order terms in the linear expansion, as the target parameter is not linear in P_0. This research topic is currently under development and will be presented in a separate work.

A limitation in the implementation of the proposed Super Learner classifiers is that we have not extended our constrained optimality criterion, which guided our selection of the Super Learner weights, to the fitting of the constituent algorithms themselves. We used the default implementation in each constituent algorithm, which aims to estimate the true conditional outcome probability, and then combined the fitted predictors in a way that optimizes the proposed constrained criterion. This was a practical consideration, in an effort to allow for the inclusion of ready-to-use algorithms in the most general settings. In a future work, we will evaluate analytically how far the optimal ψ_0 is to the true conditional outcome probability. We will also investigate in which cases and to what extent, using parametric constituent algorithms that are each fitted to satisfy the constrained optimality criterion

would be more advantageous than using more data-adaptive constituent algorithms that are each fitted to estimate the true conditional outcome probability, in the implementation of the proposed Super Learner classifier.

Acknowledgments

Contract/grant sponsor: NIH 5 R01 AI074345-08, NIAID U01AI099959, PEPFAR

References

1. World Health Organization. Technical report. World Health Organization; Jun. 2016 Consolidated guidelines on the use of antiretroviral drugs for treating and preventing hiv infection.
2. World Health Organization. Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis for hiv. Sep. 2015
3. van der Laan M, Polley E, Hubbard A. Super learner. Statistical Applications in Genetics and Molecular Biology. 2007; 6(25) ISSN 1.
4. van der Laan, M., Dudoit, S. Technical report. Division of Biostatistics, University of California; Berkeley: Nov. 2003 Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples.
5. Dudoit S, van der Laan M. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. Statistical Methodology. 2005; 2(2):131–154.
6. van der Vaart A, Dudoit S, van der Laan M. Oracle inequalities for multi-fold cross-validation. Statistics and Decisions. 2006; 24(3):351–371.
7. Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypothesis. Phil Trans Royal Soc A. 1933; 231(9):289–337.
8. Bradley A. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997; 30:1145–1159.
9. Bounsiar A, Beauseroy P, Grall-Maes E. General solution and learning method for binary classification with performance constraints. Pattern Recognition Letters. 2008; 29(10):1455–1465.
10. Cannon A, Howse J, Hush D, Scovel C. Learning with the neyman-pearson and min-max criteria. Technical report, LA-UR-02-2951. 2002
11. Scott C, Nowak R. A neyman-pearson approach to statistical learning. IEEE Transactions on Information Theory. 2005; 51:3806–3819.
12. Rigollet P, Tong X. Neyman-pearson classification, convexity and stochastic constraints. Journal of Machine Learning Research. 2011; 12:2831–2855.
13. Liu T, Hogan JW, Wang L, Zhang S, Kantor R. Optimal allocation of gold standard testing under constrained availability: application to assessment of hiv treatment failure. Journal of the American Statistical Association. 2013; 108(504):1173–1188. [PubMed: 24672142]
14. Petersen ML, LeDell E, Schwab J, Sarovar V, Gross R, Reynolds N, Haberer JE, Goggin K, Golin C, Arnsten J, et al. Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective hiv rna monitoring. Journal of acquired immune deficiency syndromes. 2015; 69(1):109–118. [PubMed: 25942462]
15. Polley, E., LeDell, E., van der Laan, M. Package 'SuperLearner'. 2016. https://cran.r-project.org/web/packages/SuperLearner/SuperLearner.pdf
16. LeDell E, van der Laan MJ, Peterson M. Auc-maximizing ensembles through metalearning. The international journal of biostatistics. 2016; 12(1):203–218. [PubMed: 27227721]
17. Johnson, SG. The nlopt nonlinear-optimization package. 2013. http://ab-initio.mit.edu/nlopt
18. van der Laan, M., Rose, S. Springer Series in Statistics. 1. Springer; 2011. Targeted Learning: Causal Inference for Observational and Experimental Data.
19. LeDell, E. Package 'casecontrolsl'. 2014. http://www.stat.berkeley.edu/~ledell/R/casecontrolSL.pdf

20. Chamie G, Clark TD, Kabami J, Kadede K, Ssemmondo E, Steinfeld R, Lavoy G, Kwarisiima D, Sang N, Jain V, et al. A hybrid mobile approach for population-wide hiv testing in rural east africa: an observational study. The Lancet HIV. 2016; 3(3):e111–e119. [PubMed: 26939734]
21. Chamie G, Kwarisiima D, Clark TD, Kabami J, Jain V, Geng E, Balzer LB, Petersen ML, Thirumurthy H, Charlebois ED, et al. Uptake of community-based hiv testing during a multi-disease health campaign in rural uganda. PLoS One. 2014; 9(1):e84317. [PubMed: 24392124]
22. Kroenke K, Spitzer RL, Williams JB. The patient health questionnaire-2: validity of a two-item depression screener. Medical care. 2003; 41(11):1284–1292. [PubMed: 14583691]
23. Kroenke K, Spitzer R, Williams J, et al. The 2-item generalized anxiety disorder scale had high sensitivity and specificity for detecting gad in primary care. J Intern Med. 2007; 146:317–25.
24. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996:267–288.
25. Hastie T, Tibshirani R. Generalized additive models. Statistical science. 1986:297–310.
26. Hastie, TJ., Tibshirani, RJ. Generalized additive models. Vol. 43. CRC Press; 1990.
27. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32.
28. Gelman A, Jakulin A, Pittau M, Su YS. A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics. 2009; 2(3):1360–1383.
29. Breiman, L., Friedman, J., Olshen, R., Stone, C. The Wadsworth statistics/probability series. Wadsworth International Group; 1984. Classification and regression trees.
30. Hubbard AE, Kherad-Pajouh S, van der Laan MJ. Statistical inference for data adaptive target parameters. The international journal of biostatistics. 2016; 12(1):3–19. [PubMed: 27227715]
31. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the roc curve estimates. Electronic journal of statistics. 2015; 9(1):1583. [PubMed: 26279737]
32. van der Laan M, Rubin D. Targeted maximum likelihood learning. The International Journal of Biostatistics. 2006; 2(1):1–40.
33. Zheng, W., van der Laan, M. Technical report 273. Division of Biostatistics, University of California; Berkeley: Nov. 2010 Asymptotic theory for cross-validated targeted maximum likelihood estimation. http://www.bepress.com/ucbbiostat/paper273

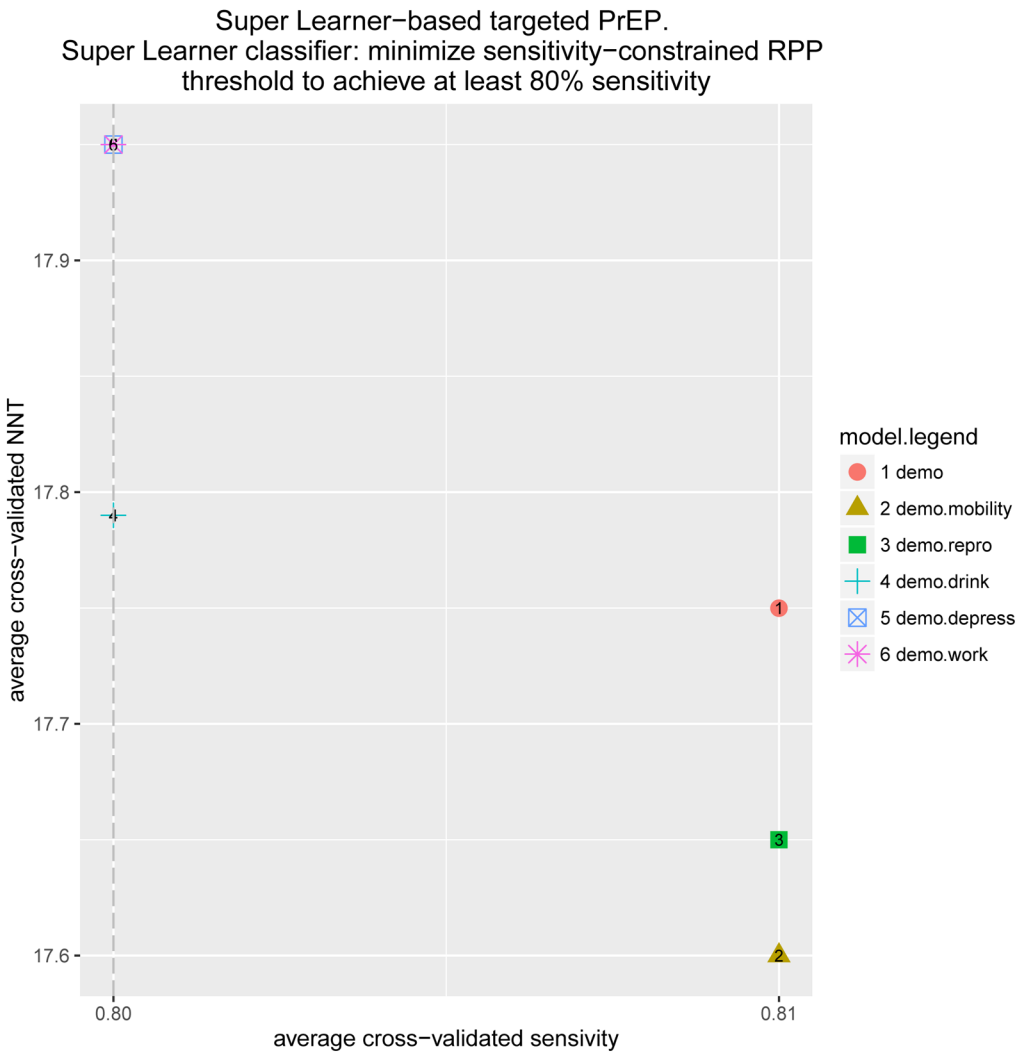

Figure 1.

Empirical performance of a Super Learner classifier that minimizes RPP under the nominal constraint of achieiving at least 80% sensitivity. Performance measures are given by average cross-validated sensitivity, and average cross-validated number needed to treat (NNT).

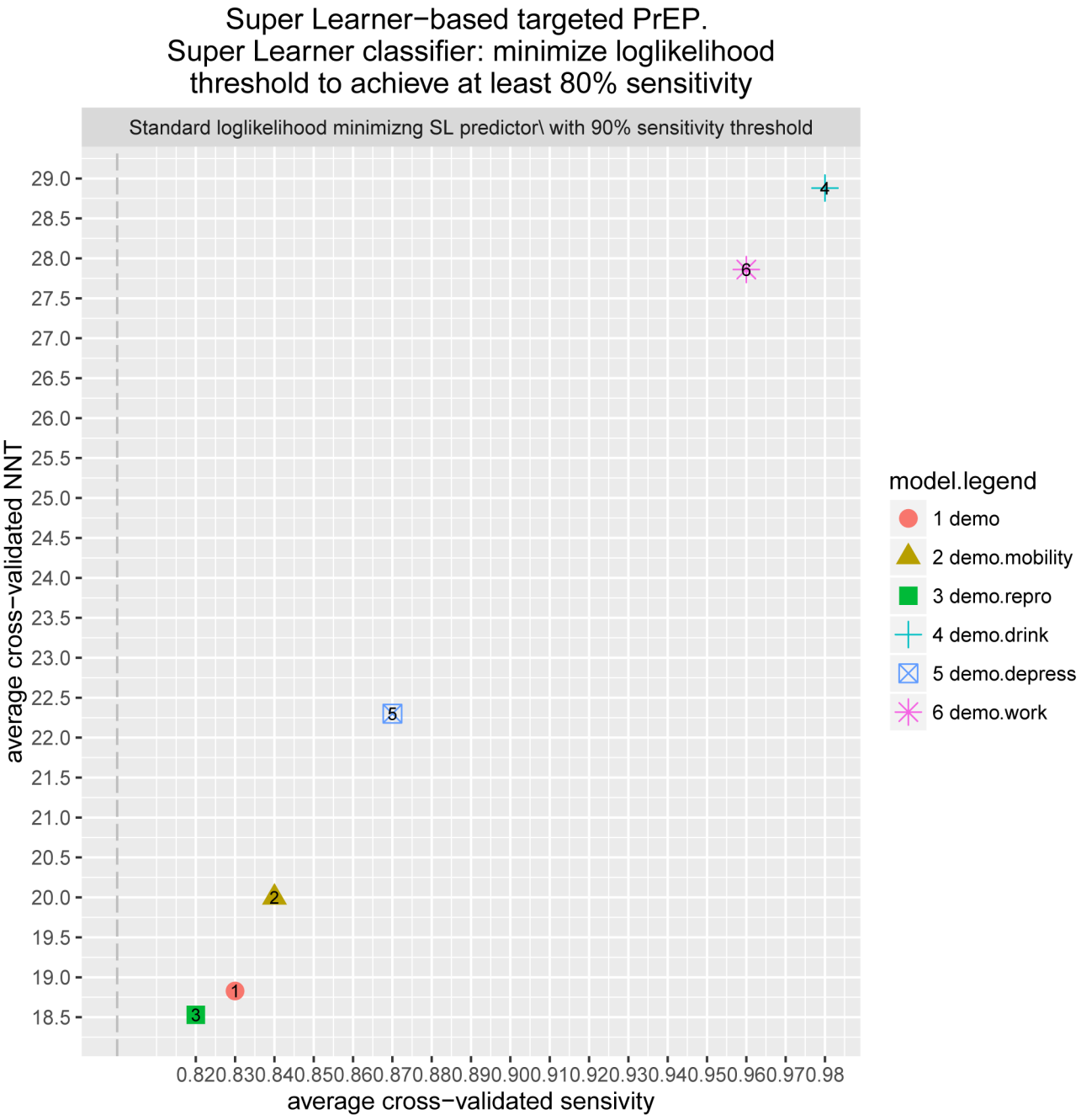

Figure 2.

Empirical performance of a Super Learner predictor that minimizes the minus log-likelihood, coupled with a cross-validated 80% sensitivity threshold. Performance measures are given by average cross-validated sensitivity, and average cross-validated number needed to treat (NNT).

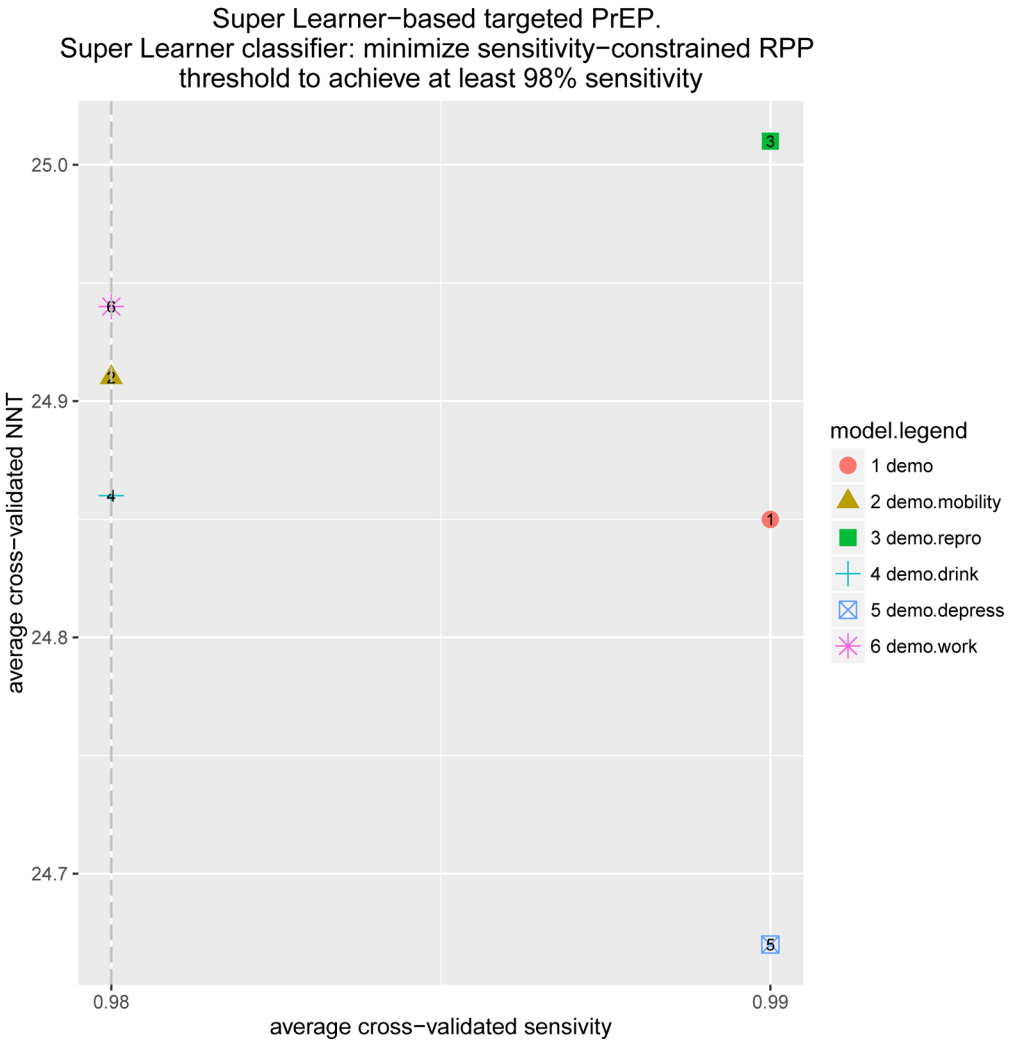

Figure 3.

Empirical performance of a Super Learner classifier that minimizes RPP under the nominal constraint of achieiving at least 98% sensitivity. Performance measures are given by average cross-validated sensitivity, and average cross-validated number needed to treat (NNT).

Stat Med. Author manuscript; available in PMC 2019 January 30.

Table 1

Baseline HIV status by baseline variables. For a subgroup-based strategy defined by a stratum, we assess a) the average cross-validated true positive rate (aCV-sensitivity), and b) the average cross-validated number needed to treat (aCV-NNT)

	negative	**positive**	**total**	**aCV-sensitivity**	**aCV-NNT**
Pop'n total	43,269	1,493	44,762	1	30
Gender					
male	19,646	527	20,173	0.350	39.19
female	23,623	966	24,589	0.650	26.07
Age group					
15–19	10,154	57	10,211	0.040	209.55
20–29	12,066	280	12,346	0.190	45.91
30–39	7,485	452	7,937	0.300	18,30
40–49	5,453	415	5,868	0.280	14.86
50–59	3,592	199	3,791	0.130	20.43
60 or older	4,519	90	4,609	0.060	59.63
Marital Status					
no answer	183	6	189	0.00	*NA*
single	12,694	117	12,811	0.080	117.81
married	25,763	958	26,721	0.640	28.44
widowed	2,582	219	2,801	0.150	13.77
divorced	380	44	424	0.030	12.38
separated	1,667	149	1,816	0.100	13.67
Polygamy					
no answer	17510	536	18,046	0.360	34.65
no	19,468	647	20,115	0.430	31.89
yes	6,291	310	6,601	0.210	22.35
Occupation					
No answer	188	6	194	0	*NA*
farm	26,290	1,107	27,397	0.740	25.33

	negative	positive	total	aCV-sensitivity	aCV-NNT
fish	87	10	97	0.010	5.79
food/tourism	295	41	336	0.030	10.13
household worker	1,355	46	1,401	0.030	38.58
industrial	577	22	599	0.02	27.87
market/shopkeeper	1,132	55	1,187	0.040	26.36
no job/other	2,124	78	2,202	0.050	33.15
public sector	520	35	555	0.020	18.87
student	9,503	29	9,532	0.020	427.67
teacher/clerk	794	38	832	0.030	25.87
transport	404	26	430	0.020	20.98
Education					
No School	6,562	276	6,838	0.180	25.80
Primary	25,730	885	26,615	0.590	30.69
Secondary	10,977	332	11,309	0.220	35.38
Stable Resident					
not stable	1,660	39	1,699	0.030	54.16
stable	41,609	1,454	43,063	0.970	30.13
Contraception Use					
no answer	8,638	143	8,781	0.100	67
no	27,113	914	28,027	0.610	31.28
yes	7,518	436	7,954	0.290	19
Drink Alcohol					
no answer	52	0	52	0	*NA*
no	36,064	1,116	37,180	0.750	33.84
yes	7,153	377	7,530	0.250	20.83
Binge drink					
no answer	36,116	1,116	37,232	0.750	33.88
never	4,717	263	4,980	0.180	19.95
less than monthly	800	41	841	0.030	26.56

	negative	positive	total	aCV-sensitivity	aCV-NNT
monthly	632	27	659	0.020	27.34
weekly	586	23	609	0.020	27.85
daily	418	23	441	0.020	21.52
Days in a month drinking					
no answer	36,116	1,116	37,232	0.750	33.88
0–3	1,906	83	1,989	0.060	28.14
4–7	1,200	55	1,255	0.040	30.22
8–11	669	45	714	0.030	20.13
12–15	629	45	674	0.030	19.05
16–19	202	7	209	0.010	9.40
20–23	513	26	539	0.020	24.19
24 or more	2,034	116	2,150	0.080	21.25
Number of drinks in a day					
no answer	36,116	1,116	37,232	0.750	33.88
1	2,789	158	2,947	0.110	20.25
2	2,365	115	2,480	0.080	23.95
3	1,114	66	1,180	0.040	22.22
4	442	15	457	0.010	23.53
5 or more	443	23	466	0.020	22.77
Days worked in past month					
no answer	71	1	72	0	*NA*
0–12	6,873	163	7,036	0.110	45.90
13–19	3,055	95	3,150	0.060	36.67
20–23	8,049	246	8,295	0.160	35.45
24–27	17,316	645	17,961	0.430	28.56
28 or more	7,905	343	8,248	0.230	25.08
Hours/Day worked in past week					
no answer	73	1	74	0	*NA*
0–4	15,440	493	15,933	0.330	33.15

	negative	positive	total	aCV-sensitivity	aCV-NNT
5–7	12,774	460	13,234	0.310	29.53
8–10	8,788	297	9,085	0.200	31.98
11 or more	6,194	242	6,436	0.160	27.88
PHQ-2 score					
no answer	6,523	30	6,553	0.020	260.64
0	13,289	422	13,711	0.280	33.39
1	6,681	280	6,961	0.190	25.95
2	11,582	510	12,092	0.340	24.46
3	2,082	102	2,184	0.070	24.55
4	1,962	85	2,047	0.060	27.85
5	274	13	287	0.010	15.30
6	876	51	927	0.030	23.86
GAD-2 score					
no answer	6,525	30	6,555	0.020	260.76
0	14,727	478	15,205	0.320	32.65
1	5,792	245	6,037	0.160	25.95
2	10,788	487	11,275	0.330	23.86
3	1,870	85	1,955	0.060	26.12
4	2,269	98	2,367	0.070	27.38
5	411	18	429	0.010	23.52
6	887	52	939	0.030	21.52
Composite group woman or married or farming	36864	1457	38321	0.976	26.86