The Population and Personalized AUCs

Journal:	*Biostatistics*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Michael, Haben; University of Massachusetts, Department of Mathematics and Statistics Tian, Lu; Stanford University, Health Research and Policy
Keywords:	Case-Control studies, classification and prediction, Longitudinal data analysis, Non-parametric methods

The Population and Personalized AUCs

Haben Michael[1] and Lu Tian[2]

[1]Department of Mathematics and Statistics, University of Massachusetts

[2]Department of Biomedical Data Science, Stanford University

Abstract: We consider two generalizations of the AUC to accommodate clustered data. We describe situations in which the two cluster AUCs diverge and other situations in which they coincide. We apply the results to data collected on urban policing behavior.

Keywords: AUC, Confounding, Clustered data, Simpson's paradox

1 Introduction

The AUC is a widely used measure of how well a scalar predictor discriminates between two outcomes. As a population parameter, the AUC is the probability that the value of a randomly sampled predictor from one of the outcome classes is less than an independently sampled predictor from the other outcome class. There are several ways to generalize the AUC to accommodate clustered data. What we refer to as the "population AUC" appears to be the most commonly studied. The population AUC evaluates the predictor's typical effect on an entire population, as further discussed below.

While the population AUC is an important part of understanding the usefulness of a predictor, the medical field has lately focused on personalizing treatment. For example, in 2018 the National Academy of Medicine concluded: "The individuality of the patient should

be at the core of every treatment decision. One-size-fits-all approaches to treating medical conditions are inadequate; instead, treatments should be tailored to individuals based on heterogeneity of clinical characteristics and their personal preferences."

We examine a "personalized AUC" in conjunction with the population AUC. These two evaluations may give different accounts of the usefulness of a marker. In the extreme case, the phenomenon known as Simpson's paradox may occur: The personalized AUC may be nearly uninformative while the population AUC is nearly perfectly predictive, or vice versa. Modern accounts of Simpson's paradox, working in the framework of causal inference, delineate situations in which the personalized AUC is appropriate, and other situations in which the population AUC is appropriate.

Obuchowski (1997) proposes a nonparametric estimator for the variance of an estimator for the population AUC. We give an alternate derivation here. We also clarify the statistical model and target of inference. Michael et al. (2019) analyzes population and personalized versions of the ROC curve which may, in principle, be used to obtain estimates of the population and personalized AUCs discussed here under certain distributional assumptions. The analysis here is nonparametric and further avoids the inefficiency introduced by first estimating the entire ROC curve, which may not be feasible for many smaller data sets.

In Section 2 we introduce the AUC and the two generalizations to clustered data considered here, the population and personalized AUCs. In Section 3 we discuss several examples of data-generating processes that highlight the differences between the two AUCs. In Section 4 we describe two results available under the assumption of independence between the cluster sizes and the values of the predictor of interest. In Section 5 we give the asymptotic joint distribution of estimators for the two AUCs. The performance of these estimators are then analyzed using synthetic data generated under two models in Section 6. In Section 7 we analyze collected data on urban policing from the standpoint of the two cluster AUCs, and we conclude in Section 8 with directions for future work.

2 Setting and Notation

Let $\psi : \mathbb{R}^2 \to \mathbb{R}$ denote the function $(x, y) \mapsto \{x < y\} + \frac{1}{2}\{x = y\}$, using $\{\cdot\}$ to denote the indicator function. Given independent draws X and Y from two distributions F_X and F_Y, the AUC is defined as

$$\theta = E(\psi(X, Y)) = P(X < Y) + \frac{1}{2}P(X = Y).$$

Given samples $X_1, X_2, \ldots, X_M$, IID as F_X and $Y_1, Y_2, \ldots, Y_N$, IID as F_Y, an unbiased estimator of the AUC of F_X and F_Y is

$$\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\psi(X_i, Y_j).$$

The function ψ is referred to as the "kernel." The AUC is often used to evaluate how effectively the data distinguish the two distributions. The AUC is close to 1/2 when the distinction is poor, and equals 1/2 in the extreme case that $F_X = F_Y$. The AUC is close to 1 when the distinction is better. In the extreme, there is a number $c \in \mathbb{R}$ such that always $X < c$ and $Y > c$, and then $\theta = 1$. We informally refer to these two classes given by the two distributions as "control" and "case," and the scalar predictor as "marker." Switching the observations designated "control" and "case" reflects the AUC across 1/2, $AUC \mapsto 1 - AUC$, so $|AUC - \frac{1}{2}|$ is often of greater interest than the AUC itself.

We extend the AUC to accommodate 1) vectors of case and control observations and 2) dependence between case and control observations. Examples of data of this type are:

1. The predictors are longitudinal measurements of tumor antigens (CEA, CA15-3, TPS), and the outcomes are progression or non-progression of breast cancer (Emir et al., 2000).

2. The predictors are longitudinal measurements of levels of vascular enothelial growth factor and a soluble fragment of Cytokeratin 19, and the outcomes are progression or

non-progression of non-small cell lung cancer (Wu and Wang, 2011).

3. The predictors are longitudinal measurements of an HIV positive patient's CD4 counts, and the outcome is "blip" status, a binary measurement representing a transient spike in viral load (Michael et al., 2019).

Let (X, Y, M, N) be a random vector with joint distribution P such that X and Y are sequences and M and N are counting numbers.

$$\begin{aligned} &(X, Y, M, N) \sim P \\ &X = (X_1, X_2, \ldots) \in \mathbb{R}^{\mathbb{N}}, Y = (Y_1, Y_2, \ldots) \in \mathbb{R}^{\mathbb{N}} \\ &M, N \in 1, 2, 3, \ldots . \end{aligned} \tag{1}$$

Informally, we regard X and Y as vectors of length M and N, ignoring the rest of the sequences. The formulation (1) lets us avoid working with vectors of variable length.

Extend the AUC kernel $\psi(\cdot, \cdot)$ to vector arguments as

$$\psi(x, y) = \psi((x_1, \ldots, x_m), (y_1, \ldots, y_n)) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\{x_i < y_j\} + \frac{1}{2}\{x_i = y_j\} \right). \tag{2}$$

We define the personalized AUC as

$$\theta_{11}(P) = E\left(\frac{\psi(X, Y)}{MN} \right). \tag{3}$$

With (X_1, Y_1, M_1, N_1) and (X_2, Y_2, M_2, N_2), being two independent draws from P, we define the population AUC as

$$\begin{aligned} &\theta_{12}(P) = \frac{E\psi(X_1, Y_2)}{E(M_1)E(N_2)} \\ &(X_1, Y_1, M_1, N_1), (X_2, Y_2, M_2, N_2) \overset{\text{IID}}{\sim} P. \end{aligned} \tag{4}$$

The personalized AUC may be undefined if M or N can take the value 0 with positive

probability, which is the reason for restricting them to counting numbers. The population AUC may still be well-defined and when analyzed without regard to the personalized AUC the possibility of $M = 0$ or $N = 0$ may be allowed (Obuchowski, 1997). In applications where $M = 0$ or $N = 0$ is possible, our analysis is therefore conditional on $M > 0, N > 0$, a sub-population in which all clusters have at least 1 case and 1 control observation.

For estimation, suppose a sample is given,

$$(X_1, Y_1, M_1, N_1), \ldots, (X_I, Y_I, M_I, N_I) \overset{\text{IID}}{\sim} P.$$

An unbiased estimator of θ_{11} is

$$\hat{\theta}_{11} = \frac{1}{I} \sum_{i=1}^{I} \frac{\psi(X_i, Y_i)}{M_i N_i}.$$

A consistent estimator of θ_{12} is

$$\hat{\theta}_{12} = \frac{\sum \sum_{i \neq j} \psi(X_i, Y_j)}{\sum_i M_i \sum_i N_i}. \tag{5}$$

Both the population and personalized AUC, like the usual AUC, are bounded between 0 and 1, $\frac{1}{2}$ represents poor discrimination, and distance from $\frac{1}{2}$ represents increasing discrimination. However, they describe distinct measures of discrimination. It is possible for one to be informative and therefore far from 1/2, while the other is non-informative, or close to 1/2. Whereas the personalized AUC is the average AUC of a typical cluster, the population AUC is, setting aside ties in the data, the probability that a typical control observation in the population is less than a typical case observation. The following proposition makes this description precise. The consistency of $\hat{\theta}_{12}$ follows from Corollary 2.

Proposition 1. *1. Let $(X_1, Y_1, M_1, N_1), \ldots, (X_I, Y_I, M_I, N_I)$, be a random sample of size I IID according to P. Let P_I be the joint distribution of independent random selections from among the elements of $X_1, \ldots, X_I$, and $Y_1, \ldots, Y_I$, and let $(\xi_I, \eta_I) \sim P_I$. Then*

$\theta(P_I) = Pr(\xi_I < \eta_I) + \frac{1}{2}Pr(\xi_I = \eta_I) \to \theta_{12}(P)$ *as* $I \to \infty$.

2. *Let* $(X_1, Y_1, M_1, N_1), \ldots,$ *be an infinite random sequence sampled IID according to* P. *Let* P_∞ *be the joint distribution of independent random selections from among the elements of* $X_1, \ldots,$ *and* $Y_1, \ldots,$ *and let* $(\xi_\infty, \eta_\infty) \sim P_\infty$. *Then* $\theta(P_\infty) = Pr(\xi_\infty < \eta_\infty) + \frac{1}{2}Pr(\xi_\infty = \eta_\infty) = \theta_{12}(P)$.

The definition of the population AUC (4) allows for dependence between (M, N) and (X, Y) in capturing a population-level AUC in the sense of Proposition 1. Practical reasons to avoid assuming $(X, Y) \perp\!\!\!\perp (M, N)$ include informative censoring, imbalanced designs, and confounding by indication. As an alternative definition of the population AUC, consider

$$\theta'_{12} = E\left(\frac{\psi(X_1, Y_2)}{M_1 N_2}\right). \tag{6}$$

This parameter is formally a closer counterpart to the personalized AUC (3), but does not take into account different cluster sizes, with a small cluster contributing as much as a large cluster. This estimator would not therefore represent discrimination at the population level, except in case $(X, Y) \perp\!\!\!\perp (M, N)$.

Similar to the population AUC estimator (5), Obuchowski (1997) presents the estimator

$$\frac{\sum\sum_{i,j}\psi(x_i, y_j)}{\sum_i m_i \sum_i n_i} = \hat{\theta}_{12} + \frac{\sum_i \psi(x_i, y_i)}{\sum_i m_i \sum_i n_i}. \tag{7}$$

This estimator may be obtained as $\theta_{12}(P_I)$, where P_I is the empirical distribution given a sample of size I. It differs from $\hat{\theta}_{12}$ only in including the diagonal terms, an asymptotically negligible $O(1/I)$ bias. The definition (4) was chosen in part as the probability limit of (7). Though Obuchowski (1997) does not enunciate a clear statistical model, the analysis of (7) rather than the simpler (6) perhaps suggests that Obuchowski (1997) too contemplates $(X, Y) \not\perp\!\!\!\perp (M, N)$.

The population AUC, which appears more prominently in past research, may lay a claim

to being the more natural generalization of the usual AUC since it equals the usual AUC when $M = N = 1$. Below we argue that in general the population and personalized AUCs are both important, complementary tools in evaluating an estimator. In the other direction, we give inequalities that may be used in some situations to relate the two cluster AUCs.

3 Examples

3.1 Random effects model

We illustrate the population and personalized AUCs and their differences using a generic random effects model with a location shift parameter. We show that the location shift can be used to control the personalized AUC while separately the random effect can be used to control the population AUC.

Let the distribution of (X, Y, M, N) given M, N be

$$\begin{aligned} X \mid M, N &\sim Z(M, N) + \xi_i^x, i = 1, \ldots, M \\ Y \mid M, N &\sim Z(M, N) + \xi_j^y + \Delta, j = 1, \ldots, N \end{aligned} \tag{8}$$

Here, $\Delta > 0$ is a non-random location shift between the control and case values, Z is a random, cluster-level effect, and $\xi_i^x, \xi_j^y, i = 1, \ldots, M, j = 1, \ldots, N$, are IID individual effects. The within-cluster dependence is induced by Z. The individual effects ξ_i^x, ξ_j^y are assumed to be independent of (M, N), but Z is not assumed to be so. To keep things simple, we assume continuous densities are available, and so $\psi(x, y) = \{x < y\}$.

The personalized AUC is

$$\begin{aligned}\theta_{11} &= E\left(\frac{\psi_{11}}{M_1 N_1}\right) = E\left(\frac{1}{M_1 N_1}\sum_{i=1}^{M_1}\sum_{j=1}^{N_1}\{X_{1i} < Y_{1j}\}\right) \\ &= E\left(\frac{1}{M_1 N_1}\sum_{i=1}^{M_1}\sum_{j=1}^{N_1}\{Z_1 + \xi_i^x < Z_1 + \xi_j^y + \Delta\}\right) \\ &= E\left(\frac{1}{M_1 N_1}\sum_{i=1}^{M_1}\sum_{j=1}^{N_1} P(\xi_i^x - \xi_j^y < \Delta \mid M_1, N_1)\right) \\ &= P(\xi_1 - \xi_2 < \Delta). \end{aligned} \tag{9}$$

Lemma 3 was used to pull the conditional expectation inside the double sum.

The population AUC is

$$\begin{aligned}\theta_{12} &= \frac{1}{E(M)E(N)} E\left(\sum_{i=1}^{M_1}\sum_{j=1}^{N_2}\{X_{1i} < Y_{2j}\}\right) \\ &= \frac{1}{E(M)E(N)} E\left(\sum_{i=1}^{M_1}\sum_{j=1}^{N_2} P(Z_1 + \xi^x < Z_2 + \xi^y + \Delta \mid M_1, N_1, M_2, N_2)\right) \\ &= \frac{1}{E(M)E(N)} E\left(M_1 N_2 P(Z_1 + \xi^x < Z_2 + \xi^y + \Delta \mid M_1, N_1, M_2, N_2)\right) \\ &= E\left(\frac{M_1 N_2}{E(M)E(N)}\{Z_1 - Z_2 + (\xi^x - \xi^y) < \Delta\}\right) \end{aligned} \tag{10}$$

The last expression is a covariance-like term lying between 0 and 1.

3.1.1 Informative personalized AUC, uninformative population AUC

From (9), $\theta_{11} \to 1$ as $\Delta \to \infty$.

Letting Z be independent of (M, N), $\theta_{12} = P(Z_1 + \xi^x - (Z_2 + \xi^y) < \Delta)$. As a difference of two IID random variables, $Z_1 + \xi^x - (Z_2 + \xi^y)$ is symmetric about 0, and $\theta_{12} = P(Z_1 + \xi^x - (Z_2 + \xi^y) < \Delta) = 1 - P(Z_1 + \xi^x - (Z_2 + \xi^y) \geq \Delta) = 1 - 1/2P(|Z_1 + \xi^x - (Z_2 + \xi^y)| \geq \Delta)$. As $P(|Z_1 + \xi^x - (Z_2 + \xi^y)| \geq \Delta) \geq 1 - 2\Delta|f_{Z+\xi}|_\infty \geq 1 - 2\Delta|f_Z|_\infty$, a sufficient condition for $\theta_{12} \to 1/2$ is $|f_Z|_\infty \to 0$. For example, suppose Z belongs to a scale family, $f_Z =$

$f_{Z_0}(Z/\sqrt{\text{Var}(Z)})/\sqrt{\text{Var}(Z)}$ for a fixed density $f_{Z_0}, |f_{Z_0}|_\infty < \infty$, and $\text{Var}(Z) \to \infty$.

Therefore, for $\Delta = E(Y_{11}) - E(X_{11})$ large enough, θ_{11} is arbitrarily close to 1, while for any fixed Δ, for $\text{Var}(Z)$ large enough, θ_{12} approaches $1/2$.

3.1.2 Informative population AUC, uninformative personalized AUC

From (9), $\theta_{11} \to 1/2$ as $\Delta \to 0$, $\xi^x - \xi^y$ being symmetric about 0.

Let Δ be fixed. The covariance-like term (10) approaches 1 when there is a large negative covariance between M, N, and $Z_1 - Z_2$, i.e., a large negative covariance between M and Z or large positive covariance between N and Z, or both.

Figure 1 presents a simulation using gaussian data to demonstrate the discussed differences between the population and personalized AUCs. The normal model is an example of the random effects model (8) and is discussed further in Section 6. Though a large location shift can push the personalized AUC close to 1, large inter-cluster variance relative to intra-cluster variance keeps the population AUC uninformative. Similarly, if the number of case observations relative to control is positively associated with the observation values, the population AUC may approach 1 irrespective of the personalized AUC.

3.2 Binary response model

Models for case and control data are often given by specifying the status conditional on the marker, rather than vice versa as in Example 3.1. Let σ denote a monotone link such as the probit or logistic function. Fixing the cluster size $M + N$, let continuous cluster effects Z, continuous within-cluster effects ξ, and within-cluster status indicators D specify the

distribution of a cluster as follows:

$$\vec{\xi} = (\xi_1, \ldots, \xi_{M+N}) \text{ IID}$$

$$Z \perp\!\!\!\perp (\xi_1, \ldots, \xi_{M+N})$$

$$B_i = Z + \xi_i, i = 1, \ldots, M+N$$

$$D_i \mid \vec{Z}, \vec{\xi} \sim \text{bernoulli with parameter } \sigma(\beta_0 Z + \beta_1 \xi_i), i = 1, \ldots, M+N$$

$$M = \sum_{i=1}^{M+N} (1 - D_i), \qquad N = \sum_{i=1}^{M+N} D_i.$$

The control and case observations in a cluster, X_i and Y_i, are then those B_i such that $D_i = 0$ and $D_i = 1$, respectively. Here the cluster allocations M and N and the markers $\vec{B}$ are dependent, both being functions of Z and $\vec{\xi}$, though they are conditionally independent given the statuses $\vec{D}$.

Suppose first that $\beta_0 > 0$ and $\beta_1 = 0$, so $P(D_i = 1 \mid Z, \overline{\xi}) = \sigma(\beta_0 Z)$. The population AUC is

$$\begin{aligned}
\theta_{12} &= \frac{1}{E(M)E(N)} E\left(\sum_{i=1}^{M+N} \sum_{j=1}^{M+N} \{B_{1i} < B_{2j}\}\{D_{1i} = 0\}\{D_{2j} = 1\} \right) \\
&= \frac{1}{E(M)E(N)} E\left(\sum_{i=1}^{M+N} \sum_{j=1}^{M+N} P(B_{1i} < B_{2j} \mid D_{1i} = 0, D_{2j} = 1)\{D_{1i} = 0\}\{D_{2j} = 1\} \right) \\
&= P(B_{11} < B_{21} \mid D_{11} = 0, D_{21} = 1) \\
&= P(Z_{11} - Z_{21} < \xi_{21} - \xi_{11} \mid D_{11} = 0, D_{21} = 1).
\end{aligned}$$

Since $Z_{11} \mid D_{11} = 0$ is stochastically less than $Z_{21} \mid D_{21} = 1$, with the difference increasing in β_0, and since the ξs are independent of the Zs and Ds, the last line is $> \frac{1}{2}$, with the difference increasing in β_0.

On the other hand, $\vec{\xi} \perp\!\!\!\perp (\vec{D}, M, N)$, so

$$\begin{aligned}\theta_{11} &= E\left(\frac{1}{MN}\sum_{i=1}^{M+N}\sum_{j=1}^{M+N}\{B_{1i} < B_{1j}\}\{D_{i1} = 0 \text{ and } D_{ij} = 1\} \mid M > 0, N > 0\right) \\ &= E\left(\frac{1}{MN}\sum_{i=1}^{M+N}\sum_{j=1}^{M+N}\{\xi_{1i} < \xi_{1j}\}\{D_{i1} = 0 \text{ and } D_{ij} = 1\} \mid M > 0, N > 0\right) \\ &= P(\xi_{11} < \xi_{12}) = 1/2.\end{aligned}$$

Two possible instances of the model:

(a) The cluster effect Z represents a genuine signal of disease status D, such as viral load does for HIV status, and ξ represents non-systematic measurement error on instruments measuring Z. In this case, the population AUC better matches expectations of an AUC measurement than the personalized AUC. The biomarker B isn't completely uninformative, as θ_{11} suggests.

(b) The cluster effect Z is a subject's dose of a possibly ineffective drug, and larger doses are administered to sicker patients. The subject-specific measurements ξ represent non-systematic measurement error again. Here the association between the marker and disease status implied by the population AUC is spurious, and may or may not be of value to the analyst. It is possible that the personalized AUC, which does not convey any association, is preferable.

Reversing the roles of the cluster-level effect Z and within-cluster effects ξ, suppose $\beta_0 = 0$ and $\beta_1 > 0$, so that $\theta_{12} \approx 1/2$ and $\theta_{11} > 1/2$. Two instances of this model:

(c) The markers B are measurements on a patient, D indicates the presence of a disease that depends little or not at all on a baseline measure Z but is indicated by the deviations ξ from the baseline. As a second example, the markers B are post-test measurements on a population that has been stratified by pre-test measurement Z. The subject effects $\xi_i = B_i - Z$ represent the difference between post-test and pre-

test measurements, and the status indicators D represent an effective or ineffective intervention. Here the personalized AUC probably carries the correct interpretation.

(d) A population clustered along any given dimension Z, and, analogous to (b), uptake of a possibly ineffective drug is confounded by indication. That is, sicker individuals, those for which D_i is more likely to be 1, take higher doses ξ_i of the drug. Here again a causal analysis would suggest the population AUC as less misleading than the personalized AUC, though a non-causal analysis, e.g., an intention-to-treat analysis, may point to the personalized AUC.

3.3 Relationship to Simpson's paradox

Simpson's paradox, understood broadly, refers to situations where data is clustered and exhibits a consistent trend at each cluster, but exhibits a contrary trend when the unclustered data is analyzed. The examples in Section 3.1 are instance of this phenomenon. The individual and population AUCs are clustered and unclustered analyses that can yield opposite conclusions about the quality of the predictor. Contemporary analyses of Simpson's paradox show the importance of considering both the individual and population AUCs.

Working in the framework of causal inference, Pearl (2014) argues that the paradox arises from the subtle relationship between causal intervention and statistical conditioning. Human judgments, which align more closely with causal relations, may be contradicted by one of the analyses when it represents a non-causal association. Resolution of the paradox therefore amounts to formally identifying which of the two analyses represents causal relationships, if either. The correct analysis in any given situation, whether the clustered or unclustered analysis, requires information about the underlying causal relationships between the treatment, outcome, and clustering variable. See Fig. 2 for schematics that, suitably interpreted under the rules of Pearl (2014) and the references cited there, represent the causal structure of the examples given above.

The AUC, however, is often used for predictive and not necessarily causal reasons. In

interpretation (b) above, for example, the spurious association may be of use in identifying diseased individuals or estimating their number when only knowledge of prescribed dosages is available. In fact, the marker is not a cause of case status, in the sense of Pearl (2014), in any of the interpretations given above. In each case, either the marker and disease status are both downstream effects of Z and $\vec{\xi}$, or the marker is an effect of the status. See Fig. 2.

4 Simplifications when $(X, Y) \perp\!\!\!\perp (M, N)$

Under some conditions, the cluster AUC parameters θ_{12} and θ_{11} may simplify to the $M = N = 1$ case. An example is given in Section 3, where the exchangeable cluster structure enables the simplification.

Proposition 2. *Given $(X, Y, M, N) \sim P$, suppose that $E(\psi(X_{1k}, Y_{1l}) \mid M, N)$ and $E(\psi(X_{1k}, Y_{2l}) \mid M, N)$ do not depend on k, l. Then $\theta_{11}(P) = E\psi(X_{11}, Y_{11})$ and $\theta_{12}(P) = E\psi(X_{11}, Y_{21})$.*

In order for $\hat{\theta}_{12} \to 1$ while $\hat{\theta}_{11} \not\to 1$ in the random effects model discussed in Section 3, it was necessary that $(X, Y) \not\perp\!\!\!\perp (M, N)$. Theorem 3 below bounds θ_{12} by θ_{11} under one case of $(X, Y) \perp\!\!\!\perp (M, N)$, namely, when M and N are each constant.

We introduce the bound in a simple case. Each cluster contributes just one control and one case observation each, and their joint distribution P is supported on finitely many points in the plane: $P = \sum_{i=1}^{B} p_i \delta_{(x_i, y_i)}$ where $(x_i, y_i) \in \mathbb{R}^2$, $0 \le p_i \le 1, i = 1, \ldots, B$, and $p_1 + \ldots + p_B = 1$. For this simple example, assume further that all the x_i and y_i are distinct, so $\psi(x, y) = \{x < y\}$.

The personalized AUC is $\theta_{11}(P) = P(X < Y) = \sum_{i: x_i < y_i} p_i$. The population AUC depends on the product of the marginals of X and Y, say: $P_{\perp\!\!\!\perp}$, $\theta_{12}(P) = P_{\perp\!\!\!\perp}(X < Y)$. Since all the x-coordinates of the support points are distinct, the marginal distribution of X is simply $P_{\perp\!\!\!\perp}(X = x) = \sum_i p_i \delta_{x_i}(x)$. Similarly, $P_{\perp\!\!\!\perp}(Y = y) = \sum_i p_i \delta_{y_i}(y)$. The product measure is therefore a weighted sum of B^2 atoms, $P_{\perp\!\!\!\perp}(X = x, Y = y) = \sum_{i,j} p_i p_j \delta_{(x_i, y_j)}(x, y)$. We give a lower bound for the population AUC $P_{\perp\!\!\!\perp}(X < Y)$. An atom of P lying in $\{x < y\}$ of

mass p contributes p^2 to the mass given by $P_{\perp\!\!\!\perp}(X < Y)$. Each pair of atoms of P lying in $\{x < y\}$ of mass p and q contributes, beyond p^2 and q^2, at least pq and possibly $2pq$ to the mass given by $P_{\perp\!\!\!\perp}(X < Y)$. See Figure 3. Therefore

$$\begin{aligned}
\theta_{12}(P) = P_{\perp\!\!\!\perp}(X < Y) &\geq \sum_{i:x_i<y_i} p_i^2 + \sum_{i:x_i<y_i} \sum_{\substack{j:x_j<y_j \\ i<j}} p_i p_j \\
&= \frac{1}{2}\left(\sum_{i:x_i<y_i} p_i\right)^2 + \frac{1}{2}\sum_{i:x_i<y_i} p_i^2 \\
&\geq \frac{1}{2}\left(\sum_{i:x_i<y_i} p_i\right)^2 + \frac{1}{2|\{i : x_i < y_i\}|}\left(\sum_{i:x_i<y_i} p_i\right)^2 \\
&= \frac{1}{2}(1 + |\{i : x_i < y_i\}|^{-1})\theta_{11}(P)^2.
\end{aligned}$$

The first inequality is tight when each each pair i, j such that $x_i < y_i$ and $x_j < y_j$ contributes exactly $p_i p_j$, i.e., when the square given by x_i, x_j and y_i, y_j has exactly one corner in $\{x < y\}$, so that $y_i - x_i < x_j - x_i$ whenever $x_i < x_j$. The second inequality is Cauchy-Schwarz, and is tight when all the atoms in $\{x < y\}$ have the same mass.

By symmetry,

$$P_{\perp\!\!\!\perp}(X > Y) \geq \frac{1}{2}(1 + |\{i : x_i > y_i\}|^{-1})P(X > Y)^2,$$

leading to an upper bound

$$\theta_{12} \leq 1 - \frac{1}{2}(1 + |\{i : x_i > y_i\}|^{-1})(1 - \theta_{11})^2.$$

Combining these bounds, $1 - \sqrt{2(1-\theta_{12})} \leq \theta_{11} \leq \sqrt{2\theta_{12}}$.

When the personalized AUC is completely uninformative, $\theta_{11} = 1/2$, the informativity of the population AUC is limited, $1/8 \leq \theta_{12} \leq 7/8$. However, when the population AUC is

completely uninformative, $\theta_{12} = 1/2$, the above bounds on the personalized AUC, which are tight, are vacuous, $0 \leq \theta_{11} \leq 1$. Situations as described in Section 3, where the population AUC $\to 1$ while the personalized AUC $\to 1/2$, appear to require some dependence between M, N and X, Y.

Theorem 3 extends the inequality to an arbitrary P so long as M and N are constant.

Theorem 3. *Let $(X, Y, M, N) \sim P$ be given as in (1). Assume further that $M = m$ and $N = n$ are constant. Then*

$$\frac{1}{2}\left(\theta_{11} + \frac{\sum_{k,l} P(X_{1k} = Y_{1l})}{2mn}\right)^2 \leq \theta_{12} \leq 1 - \frac{1}{2}\left(1 - \theta_{11} + \frac{\sum_{k,l} P(X_{1k} = Y_{1l})}{2mn}\right)^2$$

The theorem follows from the lemma,

Lemma 4. *Given a pair of scalar random variables (X, Y) with joint distribution P, let $P_{\perp\!\!\!\perp}$ be the product measure of the marginals, i.e., for all real a, b,*

$$P_{\perp\!\!\!\perp}(\{x < a\} \cap \{y < b\}) = P(\{x < a\})P(\{y < b\}).$$

Then

$$\frac{1}{2}(P(X < Y) + P(X = Y))^2 \leq P_{\perp\!\!\!\perp}(X < Y) + \frac{1}{2}(X = Y) \leq 1 - \frac{1}{2}(1 - P(X < Y))^2.$$

With the random vector $(X, Y, M, N) \sim P$, with constant $M = N = 1$ so that P may be regarded as the joint distribution of (X, Y), the conclusion of the Lemma is

$$\frac{1}{2}(\theta_{11}(P) + \frac{1}{2}P(X = Y))^2 \leq \theta_{12}(P) \leq 1 - \frac{1}{2}(1 - \theta_{11}(P) + \frac{1}{2}P(X = Y))^2.$$

5 Asymptotic Distribution of $(\theta_{12}, \theta_{11})$

Theorem 5 gives the asymptotic joint distribution of the individual and population AUCs. It is stated in somewhat greater generality for any square-integrable kernel, not just the AUC kernel (2). The proof is the same for any random variables M, N, such that $EM \neq 0, EN \neq 0$, $EM^{-2} < \infty, EN^{-2} < \infty$, i.e., M and N need not be the lengths of X and Y. Let $W_i = (X_i, Y_i, M_i, N_i), i = 1, \ldots, I$, and let V denote the space of finite sequences.

Theorem 5. *Let $\psi : V \times V \to \mathbb{R}$, $(X, Y, M, N) \sim P$ with $(X, Y) \in V \times V$, $\psi \in L^2(P)$, M and N counting numbers > 0 with finite means. Then:*

$$\sqrt{I}(\hat{\theta}_{12} - \theta_{12}, \hat{\theta}_{11} - \theta_{11}) \rightsquigarrow \mathcal{N}(0, \Sigma)$$

with

$$\Sigma_{11} = \lim_{I\to\infty} I \operatorname{Var}(\hat{\theta}_{12}) = E\left(\frac{E(\psi_{12} \mid W_1) + E(\psi_{21} \mid W_1)}{E(M)E(N)} - \theta_{12}\left(\frac{M_1}{E(M)} + \frac{N_1}{E(N)}\right)\right)^2$$

$$\Sigma_{22} = \lim_{I\to\infty} I \operatorname{Var}(\hat{\theta}_{11}) = \operatorname{Var}(\psi_{11}/(M_1 N_1))$$

$$\Sigma_{12} = \lim_{I\to\infty} I \operatorname{Cov}(\hat{\theta}_{12}, \hat{\theta}_{11}) = \theta_{12} E\left(\frac{\psi_{11}}{M_1 N_1}\left(\frac{\psi_{12} + \psi_{21}}{E\psi_{12}} - \frac{M_1}{E(M)} - \frac{N_1}{E(N)}\right)\right)$$

Corollary 6. *Under the assumptions of Theorem 5, let $(X_1, Y_1, M_1, N_1), \ldots, (X_I, Y_I, M_I, N_I)$, be IID according to P. For $1 \leq i \leq I$ define*

$$\psi_{i\cdot} = I^{-1} \sum_{j=1}^{I} \psi(X_i, Y_j),$$

$$\psi_{\cdot i} = I^{-1} \sum_{j=1}^{I} \psi(X_j, Y_i),$$

$$\phi_i = \frac{\psi(X_i, Y_i)}{M_i N_i},$$

and analogously for $M_\cdot, N_\cdot$, and $\psi_{\cdot\cdot}$. The asymptotic covariance matrix Σ of $(\hat{\theta}_{12}, \hat{\theta}_{11})$ may

be consistently estimated by $\hat{\Sigma}$ *given by:*

$$\hat{\Sigma}_{11} = \frac{1}{I-1}\sum_{i=1}^{I}\left(\frac{\psi_{i\cdot}+\psi_{\cdot i}}{M_{\cdot}N_{\cdot}} - \hat{\theta}_{12}\left(\frac{M_i}{M_{\cdot}}+\frac{N_i}{N_{\cdot}}\right)\right)^2$$

$$\hat{\Sigma}_{22} = \frac{1}{I-1}\sum_{i=1}^{I}(\phi_i - \phi_{\cdot})^2$$

$$\hat{\Sigma}_{12} = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\phi_i}{\phi_{\cdot}}\left(\frac{\psi_{i\cdot}+\psi_{\cdot i}}{\psi_{\cdot\cdot}} - \frac{M_i}{M_{\cdot}} - \frac{N_i}{N_{\cdot}}\right)\right)$$

Proof. See Sen (1960) for convergence results for random variables $\psi_{i\cdot}$ and $\psi_{\cdot i}$. □

The estimator $\hat{\Sigma}_{11}$ of the asymptotic variance of $\hat{\theta}_{12}$ is the same as given by Obuchowski (1997), derived by a different method. The finite-sample performance of this estimator is examined in Section 6.

6 Simulation

We examine estimation and inference on the population and personalized AUCs jointly. Many of the choices and parameters follow the simulation in Obuchowski (1997) examining what is here referred to as the population AUC. Key differences include: 1) In our model $M > 0, N > 0$, to ensure that the personalized AUC is well-defined; 2) Whereas Obuchowski (1997) take $I = 100$, we take the number of clusters to be $I = 60$ in the coverage simulation, Section 6.2, and $I = 10$ in the power simulation, Section 6.3.

6.1 Data models

To generate (M, N), first a preliminary number $\overline{M} + \overline{N}$ of combined case and control observations belonging in a sample is randomly selected from among $k \in \{2, 3, 4, 5\}$. Next, to obtain the allocation to case and control observations, $\overline{M} + \overline{N}$ normal variables are sampled with unit variance and common pairwise correlation $\rho_{MN} \in \{0, 0.1, 0.4, 0.8\}$. A preliminary number $\overline{M}$ of control observations is taken to be those greater than 0, and the remainder

the preliminary number $\overline{N}$ of case observations. Finally, 1 is added to each to obtain the final number of control and case observations, $M = \overline{M} + 1, N = \overline{N} + 1$. The greater the correlation ρ_{MN}, the greater the imbalance between case and control observations within the clusters.

Two related models were considered for $(X, Y) \mid (M, N)$.

6.1.1 Bivariate normal model

A popular parametric model for the AUC is the bivariate normal model, where the case and control observations are both assumed to follow a normal distribution (Hanley, 1988). Following Obuchowski (1997) we extend this model to accommodate clustered data by modeling the observations as multivariate normal vectors with an exchangeable correlation structure.

$$(X, Y) \mid (M, N) \sim \mathcal{N}_{M+N} \left(\begin{pmatrix} 0 \cdot \mathbb{1}_M \\ \Delta \cdot \mathbb{1}_N \end{pmatrix}, \rho \mathbb{1}_{M+N} \mathbb{1}_{M+N}^T + (1-\rho) Id_{M+N} \right) \tag{11}$$

That is, the case and control observations of a given cluster all have unit variance and share the same pairwise correlation ρ, all the case observations have mean $\Delta > 0$, and all the control observations mean 0. The bivariate normal model is in fact an example of the random effect model described in Section 3, though the random effect is not given explicitly in (11). As the impact of the random effect discussed there is only to change the intra-cluster correlation or mean in (11), it is redundant to the usual multivariate normal parameters. Moreover, further parameters such as for for a non-zero control mean $E(X_{11})$ or non-unit variances $\mathrm{Var}(X_{11})$ and $\mathrm{Var}(Y_{11})$ are redundant for our purpose of modeling AUCs.

Using Proposition 2,

$$\begin{aligned} \theta_{12}(P) &= \Phi\left(\frac{\Delta}{\sqrt{2}}\right) \\ \theta_{11}(P) &= \Phi\left(\frac{\Delta}{\sqrt{2(1-\rho)}}\right) \end{aligned} \tag{12}$$

The formulas (12) show that $\theta_{11} > \theta_{12}$ and further that θ_{12} and θ_{11} are simultaneously $> 1/2, = 1/2$, or $< 1/2$. We give two benefits. The first is that $(\theta_{12}, \theta_{11})$ can be restricted without loss of generality to $[1/2, 1] \times [1/2, 1]$, switching control and case labels if necessary. The pair $(\theta_{12}, \theta_{11})$ may then serve as a parameterization of the bivariate normal model (11), solving for Δ and ρ in (12). The second involves testing. Though AUCs are often compared by magnitude, e.g., $H_0 : AUC_1 - AUC_2 > 0$, one is usually interested in the discrimination, i.e., $|AUC_1 - 1/2|$ versus $|AUC_2 - 1/2|$. The hypothesis $H_0 : AUC_1 - AUC_2 > 0$ is ambiguous, indicating that AUC_1 than is more discriminating than AUC_2 when both are greater than $1/2$, but less discriminating if both are less than $1/2$. A further complication, which will not be solved by switching the class designations, is that one AUC may be greater than $1/2$ and the other less. These complications are avoided in the bivariate normal model for the personalized and population AUCs. A test of $\theta_{12} = \theta_{11}$ versus $\theta_{12} < \theta_{11}$ is also a test of discrimination, $|\theta_{12} - 1/2| = |\theta_{11} - 1/2|$ versus $|\theta_{12} - 1/2| < |\theta_{11} - 1/2|$.

6.1.2 Censored bivariate normal model

We also examine the bivariate normal model under censoring, a mixed discrete-continuous distribution. Let $a > 0$, let $(\overline{X}, \overline{Y}) \mid (M, N)$ be sampled as in in (11), and let

$$\begin{aligned}(X, Y) \mid (M, N) = (&-a\{\overline{X} \le -a\} + \overline{X}\{-a < \overline{X} < a\} + a\{\overline{X} \ge a\},\\ &-a\{\overline{Y} \le -a\} + \overline{Y}\{-a < \overline{Y} < a\} + a\{\overline{Y} \ge a\}).\end{aligned} \tag{13}$$

That is, observations $(\overline{X}, \overline{Y})$ are generated as in the bivariate normal model (11), and the values are then clipped to $\pm a$. This type of data-generating process is used by Obuchowski (1997) to model radiologists' scores, which lie on a 0—100% scale and often accumulate at 0% and 100%.

Let

$$(X_{11}, Y_{11}) \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ \Delta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Again using Proposition 2 to reduce to the $M = N = 1$ case,

$$\begin{aligned}\theta_{12}(P) &= -\int_{-a}^{a} \Phi(x-\Delta)\phi(x)dx + \frac{1}{2}(\Phi(a) - \Phi(a-\Delta) - \Phi(-a-\Delta) \\ &\quad + \Phi(a)(\Phi(-a-\Delta) + \Phi(a-\Delta)) + 1) \\ \theta_{11}(P) &= \int_{-a}^{a}\int_{x}^{a} f_{X_{11},Y_{11}}(x,y)dydx + \Phi(-a) + 1 - \Phi(a-\Delta) - \frac{1}{2}P(X_{11} < -a, Y_{11} < -a) \\ &\quad - \frac{1}{2}P(X_{11} > a, Y_{11} > a) - P(X_{11} < -a, Y_{11} > a).\end{aligned}$$

Due to the censoring, the AUCs may be bounded below 1 in this model, regardless of the magnitude of the location shift between the underlying control and case observations. As $\Delta \to \infty$, θ_{12} and θ_{11} both tend to $\frac{1}{2}(1 + \Phi(a))$.

6.2 Coverage

The parameters Δ and ρ were set to correspond to a population AUC of $\theta_{12} \in \{0.7, 0.8\}$ and personalized AUCs of $\theta_{11} \in \{0.7, 0.8, 0.9\}$ with $\theta_{11} \geq \theta_{12}$. For each setting of $\rho_{MN}, \theta_{12}, \theta_{11}$, 1,000 replicates of size $I = 60$ were sampled and used to form a confidence ellipse for $(\theta_{12}, \theta_{11})$. Specifically, with $\hat{\theta}_{12}, \hat{\theta}_{11}$ computed as in Section 2 and Σ as in Theorem 5, under P,

$$\left|\Sigma^{-1/2}\left(\begin{pmatrix}\theta_{12}\\ \theta_{11}\end{pmatrix} - \begin{pmatrix}\hat{\theta}_{12}\\ \hat{\theta}_{11}\end{pmatrix}\right)\right|^2 \tag{14}$$

has a chi-squared distribution with 2 degrees of freedom. If q is an upper α quantile of this distribution, then

$$\left\{\begin{pmatrix}x\\ y\end{pmatrix} : \left|\Sigma^{-1/2}\left(\begin{pmatrix}x\\ y\end{pmatrix} - \begin{pmatrix}\hat{\theta}_{12}\\ \hat{\theta}_{11}\end{pmatrix}\right)\right|^2 < q\right\}$$

is a level $1-\alpha$ confidence region for $(\theta_{12}, \theta_{11})$, which then covers $(\theta_{12}, \theta_{11})$ when (14) is $< q$. In the simulation, we substitute for Σ the asymptotic approximation $\hat{\Sigma}$ given in Corollary

6. Results are presented in Table 1. The bias is on the order of a hundredth at this sample size, and the coverage is generally close to .95. There is some degradation in the coverage as $(\theta_{12}, \theta_{11})$ approach $(1, 1)$.

6.3 Power

We examine the power of testing the null hypothesis $H_0 : \theta_{12} = \theta_{11}$ using the proposed variance estimators under the bivariate normal model (11). Restricting to $\rho > 0$ in (11), the set of alternatives to $H_0 : 1/2 < \theta_{12} = \theta_{11}$ is $H_A : 1/2 < \theta_{12} < \theta_{11}$, i.e., where the personalized AUC is more discriminating than the population AUC.

The data is generated under (11) using $(\theta_{12}, \theta_{11})$ selected from points randomly and uniformly selected in $[\frac{1}{2}, 1] \times [\frac{1}{2}, 1], \{\theta_{11} \geq \theta_{12}\}$. Estimates $\hat{\theta}_{12}, \hat{\theta}_{11}$, and $\hat{\Sigma}$ were then obtained as described above. The test is carried out by testing the significance of the z-statistic

$$(\hat{\theta}_{12} - \hat{\theta}_{11})/\sqrt{c^t \hat{\Sigma} c}$$

where the contrast vector c is $(1, -1)^t$.

The observed power functions are plotted in Fig. 4. The number of clusters was chosen to be $I = 10$, few relative to the setting in Obuchowski (1997) or the number encountered in the data in Section 7, since the qualitative behavior of the power surface appears clearer with fewer clusters.

7 Data analysis

We examine data on police behavior and give 3 analyses leading to 3 different relationships between the population and personalized AUCs: the population AUC 1) significantly more than, 2) significantly less than, and 3) not significantly different from the personalized AUC.

The data consists of Terry stops in New York City and Boston. A Terry stop is a policing procedure whereby an officer briefly detains an individual based on a reasonable suspicion

that a crime has been committed, which is a lower evidentiary bar than required to arrest the individual. Terry stops are colloquially referred to as "stop and frisks" though the suspect need not be frisked or searched. The analysis here focuses on the relationship between the duration of the stop and race of the suspect. We cluster the stops according to precinct, in the case of NYC, and according to the officer conducting the stop, in the case of Boston. There is an extensive literature examining the relationship between race and Terry stops. Duration of the stop in particular is examined in, e.g., Ridgeway (2006), clustering at the precinct level in, e.g., Goel et al. (2016), and clustering at the officer level in, e.g., Ridgeway and MacDonald (2009).

The NYC data consists of measurements on 54,587 stops carried out between 2017 and 2021. The Boston data consists of 6,591 stops carried out between 2019 and 2021. The stop durations range between 0 minutes and 1–2 hours, with modes at multiples of 5 minutes, and 15 minutes being the most commonly recorded duration. While data is available for years prior to the cutoffs used here, key covariates used in the analysis were either missing or coded differently in the earlier data. So that the personalized AUC could be estimated, the data was further restricted to those clusters with at least 1 control and 1 case observation, where the interpretation of "control" or "case" depends on the racial classification under analysis below. The final number of clusters and cluster sizes are given in Table 3.

The racial classifications we consider are Black, White, and Hispanic, where Black and White are taken to include Black Hispanic and White Hispanic; see Table 3 for breakdowns.

1. $\theta_{12} < \theta_{11}$. With Black race as the binary classification, the AUC analysis looks for a difference in location between the distribution of stop durations of non-Black ("control") and Black ("case") suspects. For the NYC data, the population AUC estimate is $\hat{\theta}_{12} = 0.46$ with 95% CI 0.45—0.47, significantly different from the null value of 1/2. The personalized AUC estimate is $\hat{\theta}_{11} = 0.50$ with a 95% CI 0.47—0.53. A test of equality $H_0 : \theta_{12} = \theta_{11}$ against $\theta_{12} < \theta_{11}$ returns a p-value of .05%. The Boston data is similar. The population AUC estimate is 0.46 [0.42, 0.50] and the personalized AUC

estimate is 0.52 [0.46, 0.58]. A test of equality $H_0 : \theta_{12} = \theta_{11}$ against $\theta_{12} < \theta_{11}$ returns the p-value .91%. Confidence ellipses are plotted in Figure 5. The data recalls the situation depicted in Fig. 1b, though of course the difference between the two AUCs is less dramatic here than in the artificial example constructed there.

2. $\theta_{11} < \theta_{12}$. We next consider differences in duration of stop between non-White ("control") or White ("case") suspect status. As Table 2 indicates, the vast majority of suspects are either Black or White, when those categories are taken inclusive of Hispanics, so one might expect that the analysis for non-White/White status to be nearly the same as the analysis for Black/non-Black status, therefore simply reversing the direction of the results just given, i.e., reflecting the AUCs across 1/2. That expectation largely holds for the NYC data, where the population and personalized AUCs are 0.53 [0.52, 0.54] and 0.50 [0.48, 0.53], and the population AUC remains the only one of the two significantly different from the null value 1/2. For the Boston estimates, however, the personalized AUC, 0.46 [0.40, 0.53], is more informative than the population AUC, 0.52 [0.48, 0.55], with the test of equality versus $\theta_{11} < \theta_{12}$ returning a p-value of 2.5%. This analysis therefore corresponds to the situation in Fig. 1a.

3. No significant difference between θ_{12} and θ_{11}. Finally, we consider duration of the stop between non-Hispanic ("control") and Hispanic ("case") suspects. For both the NYC and Boston data, neither the population AUC nor personalized AUC is significantly different from the null value 1/2, and the test of equality of the two AUCs fails to reject. As a second example, in Boston, whether one takes the case status to be non-Hispanic Black or non-Hispanic White, the two AUCs are statistically indistinguishable from each other and each is indistinguishable from the null value 1/2.

The decision to cluster at the officer or precinct level, as opposed to, say, the time of day of the stop, age of the suspect, or other partition of the data, is in part arbitrary. For the application of the definitions and results given in the previous sections, the decision

amounts to the idealization that the officers' or precincts' data are drawn independently from a universe of officers or precinct Terry stop data. At the same time, many current analyses, such as cited above, besides this IID assumption further impose modeling assumptions such as linear random effects or logistic links. The approach here has the advantage of being otherwise nonparametric.

8 Discussion

We have compared and contrasted two generalizations of the AUC to accommodate clustered, paired data. Straightforward extensions include allowing for multiple dependent AUCs, clusters that are only exchangeable or otherwise fall short of being IID, and covariate-adjusted AUCs. A more delicate extension would allow for estimation of the personalized AUC when some clusters have no control or no case observations. As the personalized AUC is not currently defined for such clusters either the definition would need to be re-worked or a model would need to be introduced for the missing values corresponding to those clusters. No major changes would be required of the analysis under a strong enough assumption such as ignorability, i.e., the assumption that the behavior of the personalized AUC (or the pair) is the same on $M > 1, N > 1$ as on the entire population.

References

Emir, B., Wieand, S., Jung, S.-H., and Ying, Z. (2000). Comparison of diagnostic markers with repeated measurements: a non-parametric ROC curve approach. *Statistics in Medicine*, 19(4):511–523.

Goel, S., Rao, J. M., and Shroff, R. (2016). Precinct or prejudice? Understanding racial disparities in new york city's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–394.

Hanley, J. A. (1988). The robustness of the "binormal" assumptions used in fitting ROC curves. *Medical decision making*, 8(3):197–203.

Michael, H., Tian, L., and Ghebremichael, M. (2019). The ROC curve for regularly measured longitudinal biomarkers. *Biostatistics*, 20(3):433–451.

Obuchowski, N. A. (1997). Nonparametric analysis of clustered ROC curve data. *Biometrics*, pages 567–578.

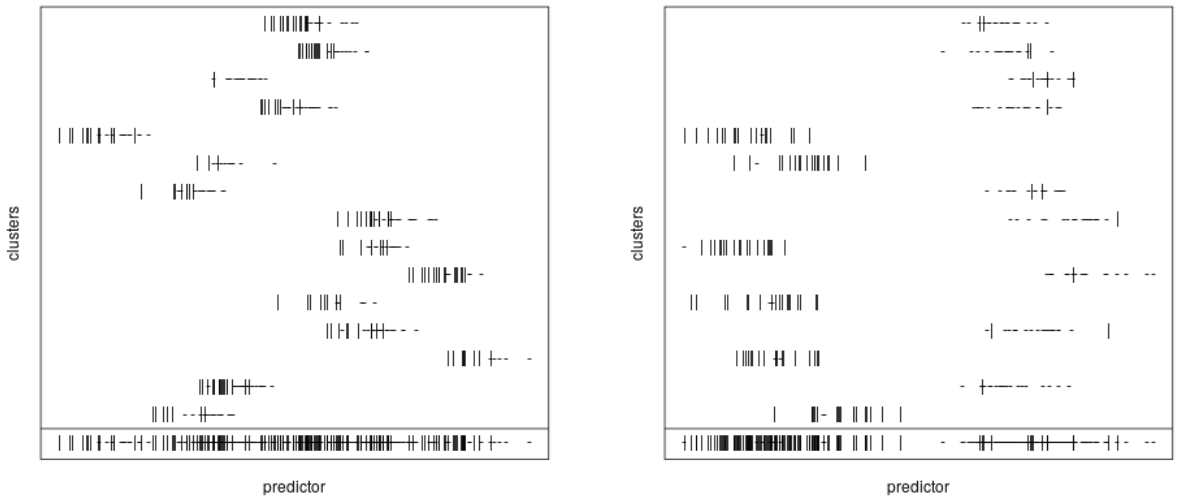
Pearl, J. (2014). Comment: Understanding Simpson's paradox. *The American Statistician*, 68(1):8–13.

Ridgeway, G. (2006). Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *Journal of Quantitative Criminology*, 22(1):1–29.

Ridgeway, G. and MacDonald, J. M. (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association*, 104(486):661–668.

Sen, P. K. (1960). On some convergence properties of U-tatistics. *Calcutta Statistical Association Bulletin*, 10(1-2):1–18.

Wu, Y. and Wang, X. (2011). Optimal weight in estimating and comparing areas under the receiver operating characteristic curve using longitudinal data. *Biometrical journal*, 53(5):764–778.



(a) (b)

Figure 1: Two visualizations contrasting the individual and population AUCs. Each gives rug plots of fifteen clusters of data, each cluster sampled IID according to a bivariate normal model, with the unclustered data combined at the bottom. Case observations are represented with "−" and control observations with "|". On the left, the personalized AUC is informative and the population AUC uninformative. The reverse situation is presented on the right.

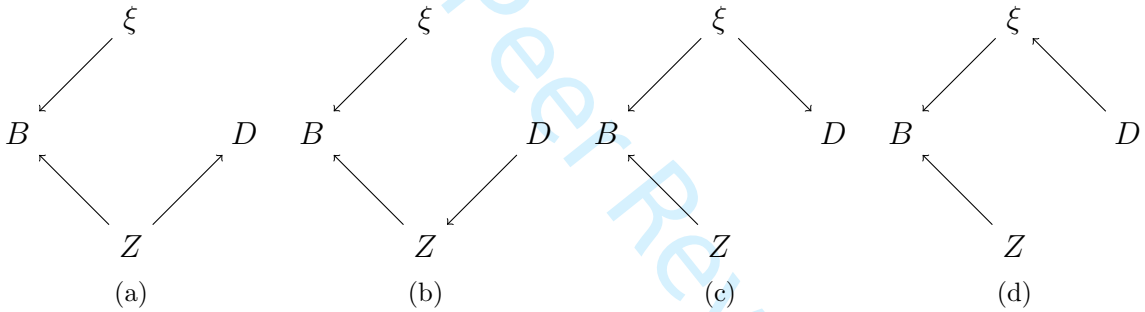


Figure 2: Causal DAGs for the four instances given in Section 3.2.

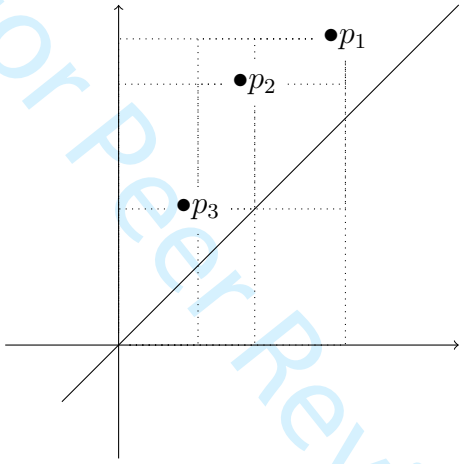
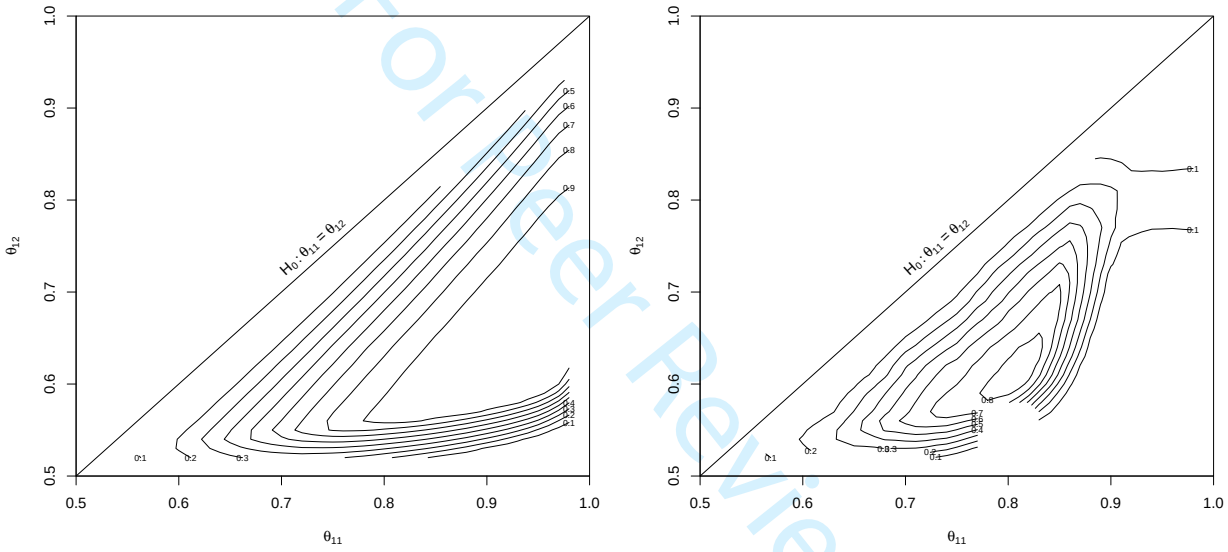
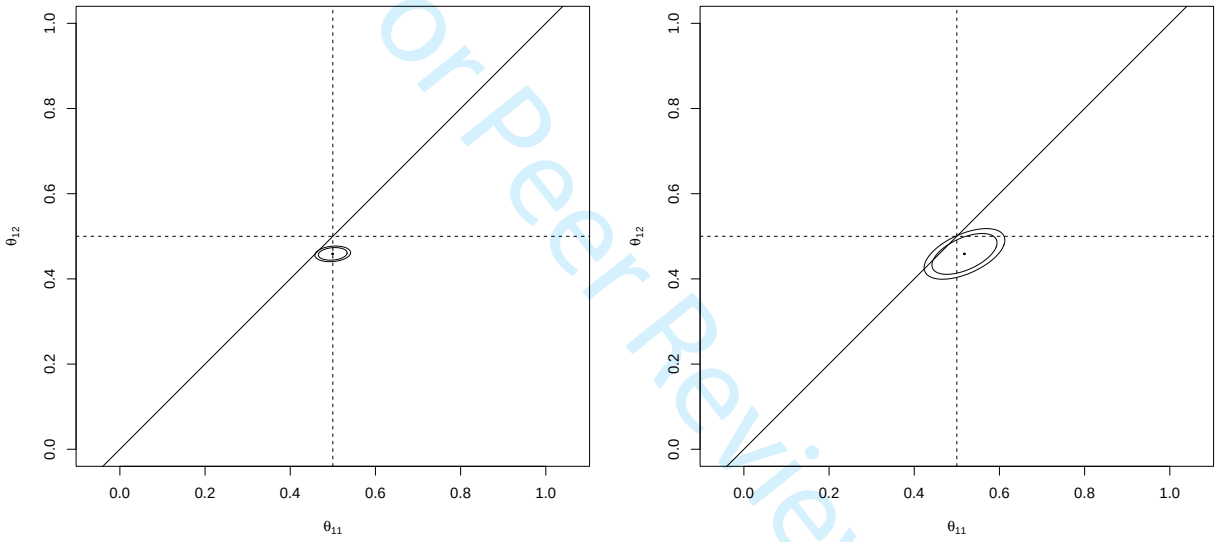


Figure 3: The case $M = N = 1$ and finitely supported (X, Y). The distance between the atoms p_1 and p_2 is small relative to their distances to the line $x = y$, so they contribute $(p_1 + p_2)^2$ to the mass of $\{x < y\}$ under product of the marginals. The distance between p_1 and p_3 is relatively large, so they contribute only $(p_1 + p_3)^2 - p_1 p_3$.



(a) Bivariate normal model (11) (b) Bivariate normal model with censoring (13)

Figure 4: Empirical power function of the test of $H_0 : \theta_{12} = \theta_{11}$ versus $\theta_{12} < \theta_{11}$ using the asymptotic estimator given in Section 5. In the bivariate normal model with or without censoring, the null is equivalent to $H_0 : |\theta_{12} - 1/2| = |\theta_{11} - 1/2|$, equal informativity.



(a) NYC (b) Boston

Figure 5: Level 95% and 99% Confidence ellipses for the estimates of $(\theta_{11}, \theta_{12})$ for duration of Terry stop by non-Black/Black status.

parameters			coverage	bias				
θ_{12}	θ_{11}	ρ_{MN}		θ_{12}	θ_{11}	Σ_{11}	Σ_{12}	Σ_{22}
0.70	0.70	0.00	0.94	0.00	-0.00	0.00	-0.00	-0.00
0.70	0.70	0.10	0.93	0.00	0.00	-0.01	-0.00	0.00
0.70	0.70	0.40	0.94	0.00	0.00	0.01	0.01	0.01
0.70	0.70	0.80	0.94	0.00	0.00	0.00	-0.00	-0.00
0.70	0.80	0.00	0.94	0.00	0.00	0.00	0.00	0.00
0.70	0.80	0.10	0.93	0.00	0.00	0.00	0.00	0.00
0.70	0.80	0.40	0.93	0.00	0.00	-0.01	-0.00	0.00
0.70	0.80	0.80	0.93	-0.00	-0.00	0.01	0.00	0.00
0.80	0.80	0.00	0.92	0.00	-0.00	-0.00	0.00	0.00
0.80	0.80	0.10	0.94	0.00	-0.00	0.00	0.00	0.00
0.80	0.80	0.40	0.93	0.00	0.00	0.00	0.00	0.00
0.80	0.80	0.80	0.93	0.00	-0.00	0.00	-0.00	-0.00
0.80	0.90	0.00	0.92	0.00	0.00	0.00	-0.00	-0.00
0.80	0.90	0.10	0.92	0.00	-0.00	0.00	0.00	0.00
0.80	0.90	0.40	0.91	0.00	-0.00	0.00	0.00	-0.00
0.80	0.90	0.80	0.90	0.01	0.00	-0.00	-0.00	0.00

(a) Binormal model (11)

parameters			coverage	bias				
θ_{12}	θ_{11}	ρ_{MN}		θ_{12}	θ_{11}	Σ_{11}	Σ_{12}	Σ_{22}
0.70	0.70	0.00	0.94	0.00	0.00	0.00	0.00	0.00
0.70	0.70	0.10	0.94	0.00	0.00	-0.00	0.00	0.01
0.70	0.70	0.40	0.93	0.00	0.00	-0.00	-0.01	-0.00
0.70	0.70	0.80	0.94	0.00	0.00	-0.00	-0.00	-0.00
0.70	0.80	0.00	0.94	0.00	0.00	0.00	-0.00	-0.00
0.70	0.80	0.10	0.91	0.00	0.00	-0.01	-0.00	-0.01
0.70	0.80	0.40	0.94	0.00	-0.00	0.00	0.00	0.00
0.70	0.80	0.80	0.93	0.00	0.00	0.00	-0.00	-0.00
0.80	0.80	0.00	0.94	0.00	-0.00	0.00	0.00	0.00
0.80	0.80	0.10	0.93	0.00	-0.00	0.00	0.00	0.00
0.80	0.80	0.40	0.95	-0.00	-0.00	0.00	-0.00	0.00
0.80	0.80	0.80	0.92	0.00	-0.00	-0.01	-0.00	-0.00
0.80	0.90	0.00	0.93	0.00	-0.01	-0.00	-0.00	0.00
0.80	0.90	0.10	0.92	0.00	-0.01	-0.00	-0.00	-0.00
0.80	0.90	0.40	0.92	0.00	-0.01	-0.00	-0.00	0.00
0.80	0.90	0.80	0.93	0.00	-0.01	0.00	0.00	0.01

(b) Binormal model with censoring (13)

Table 1: The results of a simulation examining the coverage of a nominal 95% confidence ellipse obtained using the asymptotic estimator given in Section 5. . For θ_{11} and θ_{12}, the bias is computed as the mean difference between the estimates and the known true values. For the elements of the covariance matrix Σ_{ij}, the bias is the mean difference between the estimates given by Theorem 5 and the empirical covariance.

	NYC			Boston		
group	mean duration (SD)	count	freq.	mean duration (SD)	count	freq.
Asian	14.24 (21.16)	1139	0.02	25.00 (24.22)	53	0.01
Black Hispanic	11.01 (17.12)	4675	0.09	15.28 (18.73)	391	0.06
Black non-Hispanic	10.99 (16.78)	31588	0.58	19.06 (28.93)	3448	0.55
White Hispanic	11.21 (15.15)	11486	0.21	15.63 (15.96)	578	0.09
White non-Hispanic	12.85 (16.18)	4854	0.09	21.74 (33.01)	1760	0.28
other	11.84 (17.70)	261	0.00	20.89 (23.90)	93	0.01

Table 2: Summary estimates on the duration of Terry stops by racial group.

case group	data set	I	ΣM_i	ΣN_i	θ_{12}	θ_{11}	$H_0 : \theta_{12} = \theta_{11}$
Black	NYC	187	17698	36152	0.46 [0.45, 0.47]	0.50 [0.47, 0.53]	0.00
	Boston	112	418	585	0.46 [0.42, 0.50]	0.52 [0.46, 0.58]	0.02
Black non-Hispanic	NYC	185	22348	31490	0.47 [0.46, 0.48]	0.51 [0.48, 0.53]	0.01
	Boston	117	464	569	0.48 [0.44, 0.51]	0.50 [0.44, 0.56]	0.30
Black Hispanic	NYC	154	48847	4672	0.48 [0.47, 0.49]	0.49 [0.47, 0.52]	0.42
	Boston	41	494	62	0.44 [0.37, 0.51]	0.49 [0.40, 0.59]	0.09
White	NYC	185	37547	16298	0.53 [0.52, 0.54]	0.50 [0.48, 0.53]	0.04
	Boston	109	614	385	0.52 [0.48, 0.55]	0.46 [0.40, 0.53]	0.05
White non-Hispanic	NYC	148	48327	4838	0.56 [0.55, 0.58]	0.52 [0.49, 0.55]	0.00
	Boston	106	631	324	0.52 [0.47, 0.56]	0.49 [0.43, 0.56]	0.39
White Hispanic	NYC	176	42333	11463	0.51 [0.50, 0.52]	0.49 [0.47, 0.52]	0.30
	Boston	62	631	89	0.48 [0.41, 0.55]	0.47 [0.39, 0.56]	0.81
Hispanic	NYC	180	37693	16125	0.50 [0.49, 0.51]	0.49 [0.46, 0.52]	0.41
	Boston	85	706	151	0.46 [0.41, 0.50]	0.48 [0.41, 0.55]	0.51

Table 3: Estimates of the population and personalized AUCs of the duration of Terry stops by racial group.