Mathematical Biosciences 358 (2023) 108982

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Original Research Article

Optimal classification and generalized prevalence estimates for diagnostic settings with more than two classes

Rayanne A. Luke [a,b,*], Anthony J. Kearsley [b], Paul N. Patrone [b]

[a] Johns Hopkins University, Department of Applied Mathematics and Statistics, Baltimore, 21218, MD, USA
[b] National Institute of Standards and Technology, Information Technology Laboratory, Gaithersburg, 20899, MD, USA

ARTICLE INFO

Keywords:
Antibody testing
Diagnostics
Multiclass classification
Prevalence estimation
SARS-CoV-2

ABSTRACT

An accurate multiclass classification strategy is crucial to interpreting antibody tests. However, traditional methods based on confidence intervals or receiver operating characteristics lack clear extensions to settings with more than two classes. We address this problem by developing a multiclass classification based on probabilistic modeling and optimal decision theory that minimizes the convex combination of false classification rates. The classification process is challenging when the relative fraction of the population in each class, or generalized prevalence, is unknown. Thus, we also develop a method for estimating the generalized prevalence of test data that is independent of classification of the test data. We validate our approach on serological data with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) naïve, previously infected, and vaccinated classes. Synthetic data are used to demonstrate that (i) prevalence estimates are unbiased and converge to true values and (ii) our procedure applies to arbitrary measurement dimensions. In contrast to the binary problem, the multiclass setting offers wide-reaching utility as the most general framework and provides new insight into prevalence estimation best practices.

1. Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has highlighted the importance of accurate classification of antibody test results. Most work has focused on labeling data as previously infected (seropositive) or naïve. Due to the deployment of SARS-CoV-2 vaccines in late 2020, there is a clear need for a classification scheme that correctly distinguishes between naïve, previously infected, and uninfected but vaccinated individuals. However, the traditional diagnostic classification methods of confidence intervals and receiver operating characteristics have no obvious extensions to a multiclass setting.

Current multiclass applications in diagnostic classification are mostly limited to supervised learning and discriminant analysis (DA) and often do not address the central role of mathematical modeling in diagnostics. Example studies in the field of machine learning include the application of support-vector machines to automatically sort endomysial autoantibody tests of celiac disease into one of four classes [1], and another that trained deep neural networks to label resting-state functional magnetic resonance imaging (MRI) results with one of six Alzheimer's disease state [2]. Some research (e.g., see Li et al. [3]) suggests DA approaches may be preferable to machine learning for multiclass classification. One example project used linear DA (LDA) to classify echocardiogram signals into five groups related to cardiac arrhythmia [4]. Some studies use DA and machine learning in combination. One obtained intermediate scores by applying LDA to multimodal data and then used an extreme learning machine-based decision tree to classify these measurements into three or four classes of Alzheimer's stages [5].

However, both supervised learning and common implementations of DA such as LDA and quadratic DA (QDA) may not accurately model training populations or account for the role of prevalence, both of which should inform the classification procedure. In contrast, modeling can overcome these limitations. Binary (two class) examples include two-dimensional (2D) modeling of antigen targets coupled with optimal decision theory [6], statistical modeling applied to either antibody or viral-load tests [7], and an approach to the time-dependent problem for antibody measurements [8]. However, none of these works discussed multiclass extensions.

This paper uses mathematical modeling to fully address the task of multiclass classification in the context of diagnostic testing. We begin by showing that the notion of *generalized prevalence*–the relative fraction of the population in each class–is fundamental for defining our objective function, the convex combination of false classification rates (Section 3). Minimization thereof yields optimal classification. Interestingly, we show that these prevalences can be computed without

* Corresponding author at: Johns Hopkins University, Department of Applied Mathematics and Statistics, Baltimore, 21218, MD, USA.
E-mail address: rluke3@jhu.edu (R.A. Luke).

https://doi.org/10.1016/j.mbs.2023.108982
Received 9 October 2022; Received in revised form 25 January 2023; Accepted 14 February 2023
Available online 17 February 2023

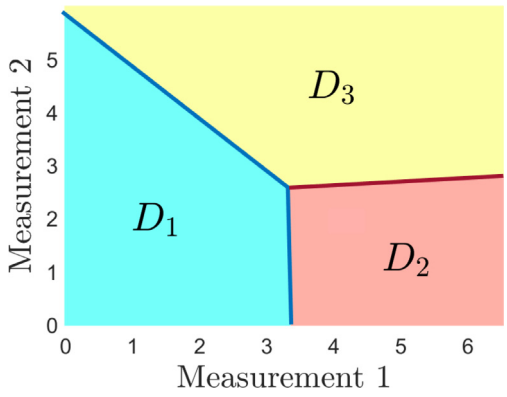


Fig. 1. Illustration of classification domains D_1, D_2, and D_3 in which the sets of equal probability of a measurement belonging to two or more classes are shown as lines separating the optimal regions.

classification by solving a linear system. We validate our methods using a SARS-CoV-2 serological data set with naïve, previously infected, and vaccinated classes [9,10][1] (Section 4). We then computationally validate the convergence of generalized prevalence estimates to the true values in mean square and illustrate a generalization to 2D data (Section 5). Finally, the discussion includes further analysis of prevalence estimation, connections and comparison to DA, extensions, and limitations (Section 6).

2. Notation

This work combines set and measure theory with applied diagnostics; readers will likely not be experts in both. In order for the ideas of this paper to be readily implemented by diagnostics experts and the applications understood by mathematicians, we provide baseline terminology from both fields.

2.1. Definitions from applied diagnostics

- The naïve class comprises individuals that have not been previously infected or vaccinated. In a binary classification, such samples are often referred to as 'negative'.
- The previously infected class comprises individuals with a prior infection but who are unvaccinated. In a binary classification, such samples are often referred to as 'positive'.
- The vaccinated class comprises individuals who have been inoculated against a disease without a prior infection.
- Training data correspond to samples for which the true classes are known. Typically, such data are used to construct conditional probability models.
- Test data correspond to samples for which the true classes are unknown or assumed to be unknown for validation purposes. Typically, a classification procedure is applied to such data.
- Generalized prevalence is the relative fraction of samples in a population that belong to each class.

[1] Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

2.2. Definitions from measure theory

- A set is a collection of objects, e.g. measurement values. A domain is a set in a continuous measurement space; see Fig. 1 for an example.
- The symbol $\mathbb{R}$ denotes the set of all real numbers. The symbol $\mathbb{R}^m$ denotes the real coordinate space of dimension m consisting of all real-valued vectors of length m.
- The symbol $\in$ indicates set inclusion. The expression $\boldsymbol{r} \in A$ means $\boldsymbol{r}$ is in set A.
- The symbol $\subset$ denotes the subset relationship of two sets. The expression $A \subset B$ means that all elements in A are contained in B.
- The use of a superscript C denotes the complement of a set. The set D^C contains all elements in the measurement space not in D.
- The symbol $\emptyset$ denotes the empty set, which contains no elements.
- The operator $\cup$ denotes the union of two sets. The set $C = A \cup B$ contains all elements in either A or B or both.
- The operator $\cap$ denotes the intersection of two sets. The set $C = A \cap B$ contains all elements in both A and B.
- The operator $\setminus$ denotes the set difference. The set $C = A \setminus B$ contains all objects in A that are not also in B. An equivalent interpretation is that $A \setminus B$ is the result of removing the common elements of A and B from A.
- The notation $A = \{\boldsymbol{r} : *\}$ defines the set A as all $\boldsymbol{r}$ that satisfy condition $*$.

2.3. Notation specific to this paper

- The set Ω denotes the entire measurement space.
- The label C_j refers to the jth class.
- The generalized prevalence for class C_j is denoted by q_j.
- The set D_j denotes a domain corresponding to C_j.
- The use of a superscript $\star$ denotes an optimal quantity. For example, $D_j^\star$ could be an optimal classification domain corresponding to class C_j.

3. Generalized prevalence estimation of test data and multiclass classification

Prevalence estimation of test data and classification of both training and test data rely on the same framework of antibody measurements.[2] For each individual or sample, we represent corresponding measurements as a vector $\boldsymbol{r} = (r_1, \ldots, r_m) \in \Omega \subset \mathbb{R}^m$. Here, $\boldsymbol{r}$ could denote m antibody types targeting different parts of a virus as measured in median fluorescence intensity (MFI). Let $P_j(\boldsymbol{r})$ describe the probability that a sample from class C_j yields measurement value $\boldsymbol{r}$. These conditional probability density functions (PDFs) are assumed known in this section; their construction is considered for example serological training data in Section 4.1.

The generalized prevalence q_j is the relative fraction of the population corresponding to class C_j. In what follows we assume there are n classes. The generalized prevalences must sum to one,[3] which implies

$$\sum_{j=1}^{n} q_j = 1, \tag{1a}$$

$$q_k = 1 - \sum_{\substack{j=1 \\ j\neq k}}^{n} q_j, \quad \text{for } k \in \{1, \ldots, n\}. \tag{1b}$$

[2] We assume the prevalences of training data are known or verified separately.

[3] The vector of prevalences $\{q_1, \ldots, q_n\}$ takes values on Δ^1_{n-1}, the $(n-1)$ dimensional simplex with total mass 1, or probability simplex. This is an equivalent geometric formulation of the property given by (1).

The probability density $Q(\boldsymbol{r})$ of a measurement $\boldsymbol{r}$ for a test sample is given by

$$Q(\boldsymbol{r}) = \sum_{j=1}^{n} q_j P_j(\boldsymbol{r}). \tag{2}$$

The product $q_j P_j(\boldsymbol{r})$ is the probability that a random sample both belongs to class C_j and has measurement value $\boldsymbol{r}$; thus, the expression for Q is an instance of the law of total probability. This quantity plays an important role in prevalence estimation of test data and classification of both training and test data.

3.1. Generalized prevalence estimation

To demonstrate the importance of prevalence in diagnostic classification, consider the United States population's SARS-CoV-2 antibody response. In early 2020, most samples should have been classified as naïve because the disease prevalence was small. By February 2022, the disease prevalence was estimated at 57.7 % [11]; a significant fraction of samples should have been classified as previously infected. Crucially, *the same measurement value may be classified differently depending on the disease prevalence*. This example shows that prevalence plays an integral role in classifying diagnostic tests and should be estimated before classification of test data. We address this need by designing unbiased estimators for the prevalences $\{q_j\}$ of test data.

For n classes, consider an arbitrary partition $\{D_j\}$ that separates the measurement space Ω into n nonempty domains. It is important to note that these D_j are not classification domains. Define

$$Q_j = \int_{D_j} Q(\boldsymbol{r})d\boldsymbol{r} = \int_{D_j} \sum_{k=1}^{n} q_k P_k(\boldsymbol{r})d\boldsymbol{r} = \sum_{k=1}^{n} q_k \int_{D_j} P_k(\boldsymbol{r})d\boldsymbol{r} = \sum_{k=1}^{n} P_{j,k} q_k, \tag{3}$$

where

$$P_{j,k} = \int_{D_j} P_k(\boldsymbol{r})d\boldsymbol{r}. \tag{4}$$

Writing this as a linear system yields

$$\begin{bmatrix} Q_1 \\ \vdots \\ Q_n \end{bmatrix} = \begin{bmatrix} P_{1,1} & \cdots & P_{1,n} \\ \vdots & \ddots & \vdots \\ P_{n,1} & \cdots & P_{n,n} \end{bmatrix} \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix}. \tag{5}$$

Taking $k = n$ in (1b) implies

$$\begin{bmatrix} Q_1 \\ \vdots \\ Q_{n-1} \end{bmatrix} - \begin{bmatrix} P_{1,n} \\ \vdots \\ P_{n-1,n} \end{bmatrix} = \left(\begin{bmatrix} P_{1,1} & \cdots & P_{1,n-1} \\ \vdots & \ddots & \vdots \\ P_{n-1,1} & \cdots & P_{n-1,n-1} \end{bmatrix} - \begin{bmatrix} P_{1,n} \\ \vdots \\ P_{n-1,n} \end{bmatrix} \underbrace{[1, \ldots, 1]}_{n-1} \right) \begin{bmatrix} q_1 \\ \vdots \\ q_{n-1} \end{bmatrix}. \tag{6}$$

This yields the prevalences $\boldsymbol{q}$ as the solution to the system

$$\boldsymbol{q} = (\boldsymbol{P} - \boldsymbol{P_n})^{-1} \left(\overline{\boldsymbol{Q}} - \overline{\boldsymbol{P_n}} \right), \tag{7a}$$

$$q_k \geq 0 \quad \text{for } k = 1, 2, \ldots, n-1, \tag{7b}$$

where $\boldsymbol{q}$ is the vector of length $n-1$ whose jth entry is q_j, $\boldsymbol{P}$ is the $(n-1)\times(n-1)$ matrix whose (i,j)th entry is $P_{i,j}$, $\boldsymbol{P_n}$ is the $(n-1)\times(n-1)$ matrix whose (i,j)th entry is $P_{i,n}$, $\overline{\boldsymbol{Q}}$ is the vector of length $n-1$ whose jth entry is Q_j, and $\overline{\boldsymbol{P_n}}$ is the vector of length $n-1$ whose jth entry is $P_{j,n}$. The last prevalence q_n is found via (1b) with $k = n$. We assume that the inverse of the matrix $\boldsymbol{P} - \boldsymbol{P_n}$ exists; Section 6.1 further discusses the matrices $\boldsymbol{P}$ and $\boldsymbol{P} - \boldsymbol{P_n}$.

To estimate the generalized prevalences, estimate the Q_j by $\hat{Q}_j$, where

$$Q_j \approx \hat{Q}_j = \frac{1}{S} \sum_{i=1}^{S} \mathbb{I}(\boldsymbol{r}_i \in D_j). \tag{8}$$

Here, S is the total number of samples and $\mathbb{I}$ denotes the indicator function. This Monte Carlo estimate of Q_j is the empirical relative frequency of points in domain D_j. Substituting $\hat{Q}_j$ for Q_j in (7) yields an estimate $\hat{q}_k$ for q_k:

$$\hat{\boldsymbol{q}} = (\boldsymbol{P} - \boldsymbol{P_n})^{-1} \left(\hat{\boldsymbol{Q}} - \overline{\boldsymbol{P_n}} \right). \tag{9}$$

Then $\hat{q}_n$ is given by $\hat{q}_n = 1 - \sum_{j=1}^{n-1} \hat{q}_j$. When the PDFs $P_j(\boldsymbol{r})$ are known and $(\boldsymbol{P} - \boldsymbol{P_n})^{-1}$ exists, these estimates $\hat{q}_k$ are unbiased, i.e., $E[\hat{q}_k] = q_k$. This follows directly from the fact that $\hat{Q}_j$ is a Monte Carlo estimator of Q_j. Further, the generalized prevalence estimates converge to the true values in mean square as the number of samples is increased [12]. This is illustrated in Section 5.1.

We note that the generalized prevalence estimates are not unique due to the arbitrariness of the $\{D_j\}$. However, the non-uniqueness allows us to select any reasonable partition over which to find the estimators $\{\hat{Q}_j\}$. In this way, the prevalence estimates are also not necessarily optimal (in the sense of minimizing variances). This is discussed further in Section 6.1.

3.2. Optimal classification

Our task is to define a partition $\{D_j\}$ (not necessarily the same as for prevalence estimation) of the measurement space Ω such that each domain corresponds to one and only one class C_j. A measurement $\boldsymbol{r}$ is assigned to class j if $\boldsymbol{r} \in D_j$. We require

$$\mu_j \left(\bigcup_{k=1}^{n} D_k \right) = 1 \quad \forall j \in \{1, \ldots, n\}, \tag{10a}$$

$$\mu_\ell(D_j \cap D_k) = 0 \text{ for } j \neq k, \quad \forall \ell = 1, 2, \ldots, n, \tag{10b}$$

where $\mu_\ell(X) = \int_X P_\ell(\boldsymbol{r})d\boldsymbol{r}$. Here, (10a) ensures that any sample can be classified and (10b) enforces single-label classification (up to sets of measure zero). To identify an optimal partition $\{D_j^\star\}$, we construct the loss function

$$\mathscr{L}(D_1, \ldots, D_n) = \sum_{j=1}^{n} q_j \int_{\Omega \setminus D_j} P_j(\boldsymbol{r})d\boldsymbol{r}. \tag{11}$$

Here, (11) is the generalized prevalence-weighted convex combination of false classification rates as a function of the domains D_j. Intuitively, we expect that a sample with measurement $\boldsymbol{r}$ should be assigned to the class domain D_j to which it has the highest probability of belonging; that is, the highest value of $q_j P_j(\boldsymbol{r})$ for all j. Accordingly, the loss function (11) penalizes misclassified measurements $\boldsymbol{r}$ with high probability values. The optimization problem yields optimal domains as the solution to

$$\{D_1^\star, \ldots, D_n^\star\} = \underset{D_1, \ldots, D_n}{\operatorname{argmin}} \mathscr{L}(D_1, \ldots, D_n). \tag{12}$$

To address situations in which a measurement has an equal highest probability of belonging to two or more classes, we introduce the following set for each class C_j:

$$\mathscr{E}_j = \bigcup_{\substack{k=1 \\ k \neq j}}^{n} \{\boldsymbol{r} : q_k P_k(\boldsymbol{r}) = q_j P_j(\boldsymbol{r}) = \max_i q_i P_i(\boldsymbol{r})\}. \tag{13}$$

In most practical implementations all $\mathscr{E}_j$ have measure zero, and the domains

$$D_j^\star = \{\boldsymbol{r} : q_j P_j(\boldsymbol{r}) > q_k P_k(\boldsymbol{r}) \text{ for } k \neq j\} \tag{14}$$

minimize the loss function $\mathscr{L}$ up to sets of measure zero. The proof is shown in Appendix A and involves a straightforward application of set theory; see also Rasmussen and Williams [13] for similar ideas.

If $\mathscr{E}_j$ has nonzero measure, randomly assigning a measurement in $\mathscr{E}_j$ to one of the classes to which it has equal maximal probability of

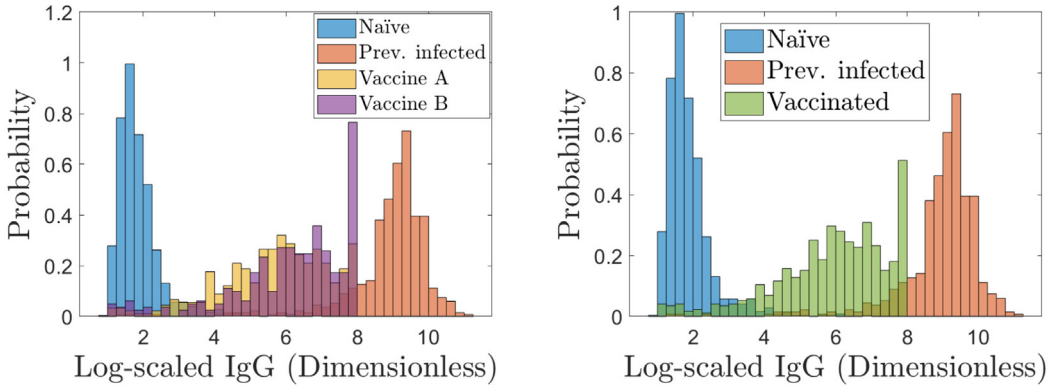


Fig. 2. Histograms of previously infected, naïve, and vaccinated data from Ainsworth et al. (2020) and Wei et al. (2021).

belonging does not effect the loss $\mathscr{L}$. In this case, the optimal domains are generalized to

$$D_j^\star = \{\boldsymbol{r} : q_j P_j(\boldsymbol{r}) > q_k P_k(\boldsymbol{r}) \text{ for } k \neq j\} \cup Z_{\mathscr{E}_j}, \tag{15}$$

where $Z_{\mathscr{E}_j}$ is an element of a partition of $\mathscr{E}_j$ that we define iteratively as follows.

$$Z_{\mathscr{E}_1} = \mathscr{E}_1, \tag{16a}$$

$$Z_{\mathscr{E}_k} = \mathscr{E}_k \setminus \bigcup_{j=1}^{k-1} \mathscr{E}_j, \quad k \in \{2, \dots, n\}. \tag{16b}$$

This ensures that no measurement in a set $\mathscr{E}_j$ is assigned to more than one optimal domain. Note that by construction, $Z_{\mathscr{E}_n}$ is empty.

Fig. 1 shows a 2D conceptual illustration of $\{\mathscr{E}_j\}$, which are the lines delineating the optimal regions. In 2D, line segments have Lebesgue measure zero. Thus, classification follows (14). Note that the "multi-point" at which the lines meet has equal probability of belonging to all three classes. We discuss this further in Section 6.2.

4. Example applied to SARS-CoV-2 antibody data with three classes

To demonstrate the concepts developed in Section 3, we apply our methods to serological data with three classes. Publicly available data sets associated with Ainsworth et al. [9] and Wei et al. [10] provide previously infected, naïve, and vaccinated antibody measurements. The vaccine data [10] are recorded for individuals that were inoculated with one of two vaccines. We refer to these as Vaccine A and Vaccine B and analyze the populations separately and together. The studies provide SARS-CoV-2 anti-spike immunoglobulin G (IgG) antibody measurements; see Appendix B for measurement details. We use one-dimensional (1D) data to illustrate a straightforward multiclass example; Section 5.2 demonstrates that our analysis holds for higher measurement dimensions.

All data are transformed to a logarithmic scale as follows:

$$r = \log_2(\tilde{r} + 2) - 1. \tag{17}$$

Here, $\tilde{r}$ and r represent the original and log-transformed values; $\tilde{r}$ has units of ng/mL and r is nondimensional. This transformation puts the data on the scale of bits and allows for better viewing of measurements that range over several decades of MFI values in the original units. Wei et al. [10] truncated vaccinated samples, with lower and upper transformed limits of 1 and roughly 8.

Fig. 2 shows a histogram of the data with the vaccinated category split by vaccine manufacturer (Fig. 2(a)) and combined (Fig. 2(b)). Previously infected samples have the largest spike IgG antibody levels and naïve samples the smallest; the vaccinated class falls in the middle. The vaccinated class overlaps with some naïve and previously infected samples. Due to the truncation of vaccinated measurements, the corresponding right-most histogram bin contains many samples. We separate the data into randomly generated training (80 % of samples) and test (20 %) populations.

4.1. Conditional probability distributions

We fit the training data to probability distributions to model the naïve, previously infected, and vaccinated antibody responses. For our purposes, we assume these are distinct classes; a sample belongs to one and only one of the three categories. To construct the conditional PDF for each population, we select a parameterized model that empirically characterizes the shape and spread of the samples. We determine parameters separately for the three training populations by maximum likelihood estimation (MLE). Several parameterized models were originally considered and fit by each training population; we select the model with the highest MLE value among those considered.

The naïve training population is fit to a Burr distribution

$$N(r) = \frac{ck}{\lambda}\left(\frac{r}{\lambda}\right)^{c-1}\left[1 + \left(\frac{r}{\lambda}\right)^{c}\right]^{-k-1}, \tag{18}$$

which describes a right-skewed sample population.

The previously infected training population is fit to a stable distribution described by characteristic function

$$\phi(r) = \exp\left\{ir\delta - |\gamma r|^\alpha\left[1 + i\beta\,\mathrm{sgn}(r)\tan\left(\frac{\pi\alpha}{2}\right)(|\gamma r|^{1-\alpha} - 1)\right]\right\} \tag{19}$$

for $\alpha \neq 1$. Here, i is the imaginary unit and sgn is the sign function, which returns +1, -1, or 0. This distribution describes a left-skewed sample population.

We fit the vaccinated training populations to an extreme-value distribution after observing the mostly symmetric shape of the data with a spike at the right truncation limit:

$$V(r) = \frac{1}{\sigma}\exp\left(\frac{r-\mu}{\sigma}\right)\exp\left[-\exp\left(\frac{r-\mu}{\sigma}\right)\right]. \tag{20}$$

We apply data censoring to better fit the truncated data; this is described in Appendix C.

The analysis in this section is identical for all three visualizations of the vaccinated class. In what follows, we report all results but show only the Vaccine A figures as examples. Corresponding figures for the Vaccine B and combined visualizations of the vaccine class are left to the Supplemental Data.

Fig. 3 shows the conditional PDFs, represented as continuous curves, trained on the three-class training data with a Vaccine A vaccinated class. The blue, red, and black curves correspond to the naïve, previously infected, and vaccinated models. The effect of truncating the data at the upper limit is visible in the right-most bin of the vaccinated class

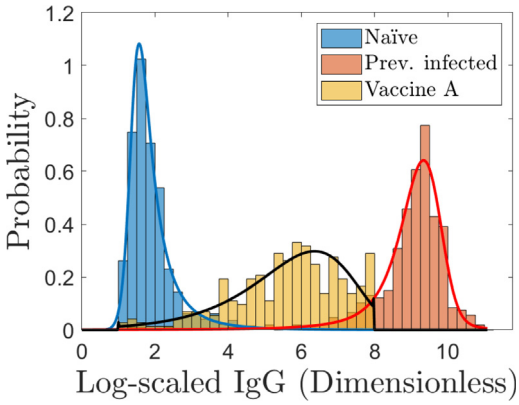


Fig. 3. Conditional PDFs for the naïve, previously infected, and vaccinated classes trained on the training data for a Vaccine A visualization of the vaccinated class. See Supplemental Figure S1 for the PDFs of the Vaccine B and combined visualizations.

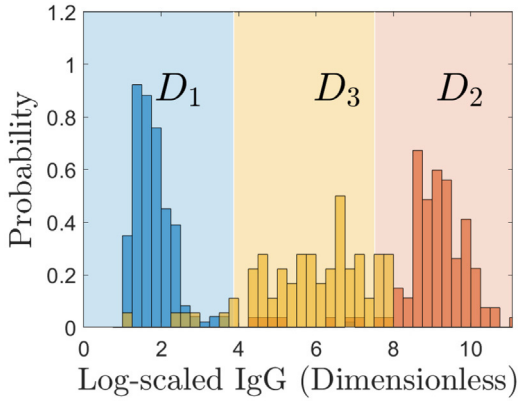


Fig. 4. Test data k-means partitioning with a Vaccine A vaccinated class for generalized prevalence estimates. We use $k = 3$ classes; the clustered domains are labeled as D_1, D_2, and D_3. See Supplemental Figure S2 for the partitions of the Vaccine B and combined visualizations of the vaccinated class.

histogram; this is accounted for by the data-censoring. As a result, the vaccinated class PDF exhibits spikes at the upper and lower truncation values. This spike is an artifact of the original data collection process and not a typical problem.

4.2. Generalized prevalence estimation

Recall that prevalence estimation of test data requires a partition that separates the measurement space Ω into n nonempty domains. Here, the number of classes is $n = 3$. We create a partition using k-means clustering with $k = 3$, which assigns each measurement to the cluster with the closest mean. Fig. 4 shows the partition for our test data set with a Vaccine A vaccinated class. The clustering separates the three populations reasonably well; see Section 6.1 for the importance of this statement. The partition need not perfectly separate the data by class to estimate prevalences with high accuracy. We estimate generalized prevalences for the test data via (9) and record true and estimated values in Table 1.

4.3. Optimal classification

We classify the training data using known generalized prevalences via (14). Fig. 5(a) shows the optimal domains, labeled $D_N^\star$, $D_V^\star$, and $D_P^\star$, for a Vaccine A vaccinated class. For this 1D example with three classes, the optimal classification domain boundaries can be represented by upper and lower threshold levels. Samples with measurements below the smaller level are classified as naïve, samples with measurements between the thresholds as vaccinated, and samples with measurements above the larger level as previously infected. All three populations have overlapping PDFs, which reduces classification accuracy.

Accurate classification of test data is possible with reasonably close prevalence estimates. We classify the test data using estimated generalized prevalences and display the optimal classification domains for a Vaccine A vaccinated class in Fig. 5(b).

Training and test data classification errors are recorded in Table 2. Taken over the three considerations of the vaccinated class, the average error for the training data is 7.61 % and the same for the test data is 5.11 %.

5. Computational validation

We numerically demonstrate two important features of our generalized prevalence estimation and multiclass optimal classification procedures. First, we show the convergence of our prevalence estimates to the true values as the number of samples is increased. Second, we present a 2D tri-class problem to show how the method generalizes to higher dimensional measurement spaces.

5.1. Convergence of prevalence estimates

We use our probability models (18)–(20) to generate synthetic data sets whose relative frequencies match the generalized prevalences of the Ainsworth et al. [9] and Wei et al. [10] data. We systematically increase S, the number of elements in each synthetic data set, from 10^2 to 10^5 while holding generalized prevalences fixed to study the effect of sample size on prevalence convergence. For each S, we randomly generate 1000 synthetic data sets by sampling from the class probability models and compute statistics on this set of 1000 data sets.

Fig. 6 shows our analysis for the Vaccine A vaccinated class. Fig. 6(b) shows a boxchart of the statistics for $S = 10^2, 10^3, 10^4$, and 10^5. The estimates have more outliers and variation when few samples are used, which decreases as S is increased. Even with few samples, the median generalized prevalence estimates are close to the true generalized prevalences. Table 3 records the mean and standard deviations taken over the 1000 data sets for each S. Even for only $S = 10^3$ samples, our estimates agree with the true generalized prevalences with roughly 2 % relative error.

Fig. 6(c) plots the standard deviation of the prevalence estimate error taken over the 1000 data sets on a log–log scale against number of samples S. The standard deviation should decrease with the inverse square root of S [12], which is plotted for comparison. Our empirical convergence rates all agree with the theory through $S = 10^4$ samples; the rate is maintained for the Vaccine A vaccinated class.

5.2. Generalization to higher dimensions

We now explore a 2D synthetic numerical validation of generalized prevalence estimation and multiclass optimal classification. See Luke et al. [14] for a discussion of the implications of higher-dimensional modeling on diagnostic testing accuracy. The synthetic values we use are modeled off the receptor-binding domain (RBD) and nucleocapsid (N) SARS-CoV-2 antibody targets; together these form a measurement double $\boldsymbol{r}$. Details about the models and information about the data are given in Appendix D. Fig. 7(a) shows an example of 2D synthetic antibody measurements with naïve, previously infected, and vaccinated classes with true prevalences of 0.3, 0.2, and 0.5. The conditional PDFs are shown as contour lines of constant probability. We generate a synthetic data set with $S = 10^3$.

To quantify uncertainty in the prevalence estimates, we randomly generate 1000 synthetic data sets using fixed prevalences. We then partition the measurement space via k-means clustering using one synthetic data set (see Fig. 7(b)), fix the partition, and use (9) to

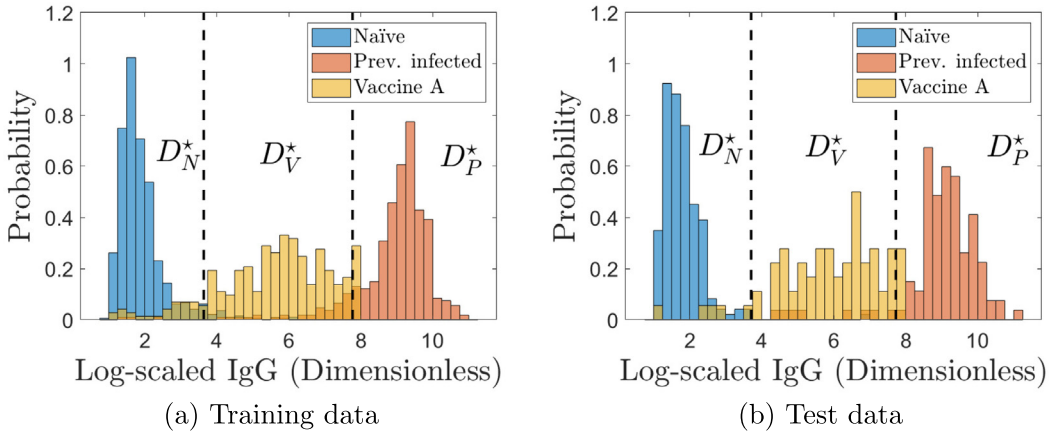


Fig. 5. Training (a) and test (b) data with a Vaccine A vaccinated class with optimal decision thresholds using a known prevalence. Vertical dashed lines indicate the optimal decision boundaries. The optimal naïve, vaccinated, and previously infected domains are labeled $D_N^\star$, $D_V^\star$, and $D_P^\star$. See Supplemental Figures S3 and S4 for the optimal domains for Vaccine B and combined visualizations of the vaccinated class.

Table 1
Estimated and true generalized prevalences for the test data naïve (N), previously infected (P), and vaccinated (V) classes. Vaccine A (A) and Vaccine B (B) are considered separately and together (All).

		Vaccine data set Estimated (true) generalized prevalence			Errors (%)			Avg. errors (%)
		A	*B*	*All*	*A*	*B*	*All*	
Class	*N*	0.523 (0.521)	0.538 (0.531)	0.445 (0.444)	0.377	1.02	0.223	0.540
	P	0.300 (0.286)	0.313 (0.292)	0.285 (0.244)	4.69	7.10	17.0	9.93
	V	0.177 (0.193)	0.149 (0.177)	0.270 (0.312)	7.99	15.0	13.6	12.2
Avg.					4.74	7.67	10.3	7.55

Table 2
Classification errors for training data using known prevalences and test data with estimated prevalences. Vaccine A and Vaccine B are considered separately and together (Combined).

	Training classification error (%)	Test classification error (%)
Vaccine A	7.87	5.08
Vaccine B	7.07	5.46
Combined	7.90	4.78

generate prevalence estimates for all sets. The results are shown in Table 4. Fig. 8 shows histograms of the generalized prevalence estimates and true values, which fall within the middle of each distribution. We classify using these estimated prevalences via (14) and find an average error of 1.58 %. Fig. 7(c) shows example optimal classification domains. The gold region is the previously infected domain, the purple is the vaccinated, and the remainder of the measurement space, colored in light blue, defines the naïve domain. For this example, the false classification rate is 1.8 %.

6. Discussion

In this section, we include a further analysis of prevalence estimation, connections and comparison to discriminant analysis, extensions, and limitations. First, we note that our methods extend the binary prevalence estimation and classification procedures of Patrone and Kearsley [6] and may provide insight into the simpler setting. More precisely, the multiclass framework reduces to the binary problem when $n = 2$.

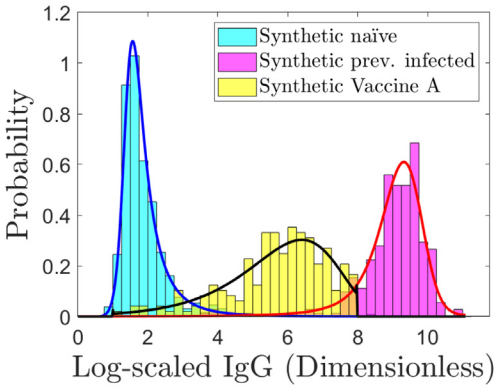
6.1. Limiting cases of prevalence estimation and implications for assay design

An interesting observation of our prevalence estimation scheme is that the structure of the matrix underpinning the linear system encodes information about overlap between populations. As such, the matrix potentially informs best practices for prevalence estimation. Here we examine limiting cases of prevalence estimation and connect characteristics of the matrix $\boldsymbol{P}$ to assay accuracy.

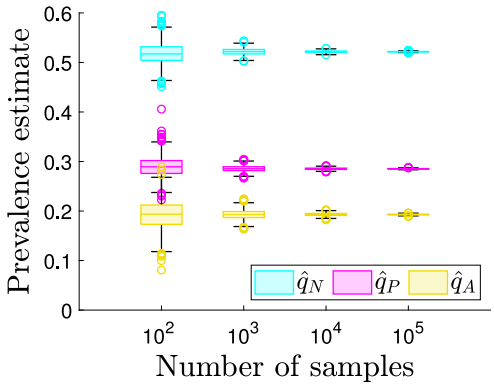
We explore interpretations of equivalent definitions of singularity of the matrix $\boldsymbol{P}$. Recall that the quantity $P_{j,k}$ gives the probability density of class k falling in domain D_j. If all elements of a row (column) of the matrix $\boldsymbol{P}$ are zero, the probability of any measurement value falling in (belonging to) the corresponding domain (class) is zero. If the columns of $\boldsymbol{P}$ are linearly dependent, the probability of a sample belonging to class C_k having a measurement in domain D_j is a linear combination of the probabilities of samples belonging to all other classes having measurements in domain D_j. This occurs for a choice of partition where all points fall in a single domain D_j. In this extreme case, there is an apparent dependence (in the linear algebra sense) of the measurement values of different classes. As a related example, for the 1D SARS-CoV-2 antibody data from Ainsworth et al. [9] and Wei et al. [10], we can construct a partition where one trial domain is empty, both $\boldsymbol{P}$ and $\boldsymbol{P}-\boldsymbol{P_n}$ are singular, and therefore prevalence estimation is not possible.

To avoid this situation, one should select nonempty trial domains, i.e., training data should lie in each element of the partition. In the limiting case that $P_{ij} = 0$, the measurement of a sample in class C_j has zero probability of falling in domain D_i. The most extreme separation of training data occurs when the PDFs have nonzero support only on mutually exclusive elements of the partition. In this setting, the matrix $\boldsymbol{P}$ is a permutation matrix, and the prevalence estimates are merely the relative fractions $\hat{\boldsymbol{Q}}$ of measurements in each domain. If the partition elements are correctly matched to the classes, this extreme separation corresponds to a perfect assay because there are no misclassifications. We note that the matrices that result from a k-means partition of the 1D SARS-CoV-2 tri-class data are close to permutation matrices; one example is

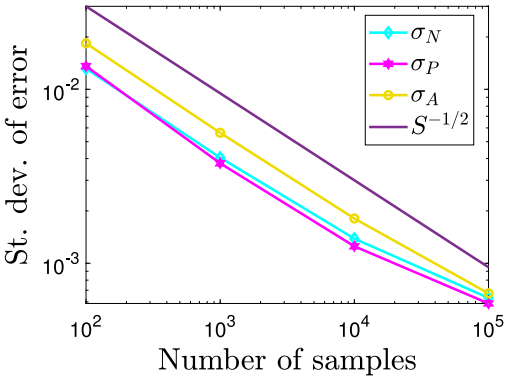
$$\boldsymbol{P} = \begin{bmatrix} 0.9749 & 0.0058 & 0.1055 \\ 0.0237 & 0.0495 & 0.8101 \\ 0.0015 & 0.9447 & 0.0844 \end{bmatrix}. \tag{21}$$



(a) $S = 10^3$ synthetic samples



(b) Boxchart of estimate statistics



(c) Convergence of the error in mean square

Fig. 6. Prevalence estimation convergence for synthetic data using a Vaccine A vaccinated class (1000 simulations). In (b), the boxes display the median and upper and lower quartiles as the line inside the box and its top and bottom edges. The whiskers show non-outlier maximum and minimum values; outliers vary from the median by more that 1.5 times the difference between the upper and lower quartiles, and are shown as circles. In (b) and (c), the subscripts N, P, and A denote naïve, previously infected, and Vaccine A vaccinated. The number of samples is S.

Table 3
Generalized prevalence estimate means and standard deviations taken over 1000 simulations of synthetic data generated from the probability models for increasing number of samples S for a Vaccine A vaccinated class.

		Naïve	Previously infected	Vaccine A
True generalized prevalences		0.521	0.286	0.193
Number of samples	10^2	0.518 ± 0.0212	0.289 ± 0.0211	0.193 ± 0.0299
	10^3	0.522 ± 0.0067	0.286 ± 0.0061	0.193 ± 0.0093
	10^4	0.522 ± 0.0021	0.285 ± 0.0020	0.193 ± 0.0030
	10^5	0.522 ± 0.0007	0.285 ± 0.0007	0.193 ± 0.0011

Table 4
Statistics for the data shown in Fig. 8. The true values of the prevalences are given along with the mean μ, standard deviation σ, and coefficient of variation (CV) of the estimates.

	True val	μ	σ	CV $\left(\frac{\sigma}{\mu}\right)$
q_1	0.3	0.300	9.6×10^{-3}	0.0319
q_2	0.2	0.200	7.4×10^{-3}	0.0371
q_3	0.5	0.500	6.3×10^{-3}	0.0126

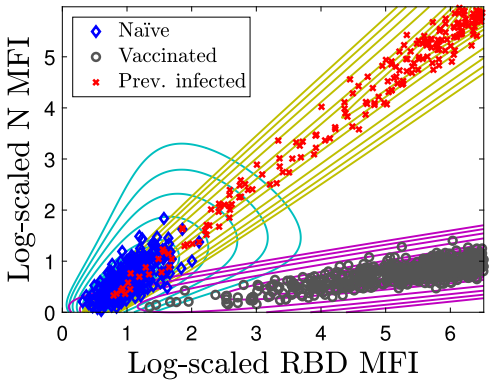
Selection of trial domains with a high degree of class separation may be a key to our low-error prevalence estimates.

We speculate that under certain conditions a matrix $\boldsymbol{P}$ that is a permutation matrix may be optimal in the sense that it minimizes the prevalence estimate error. In 1D, it may be possible to construct an optimization in terms of the samples assigned to each element of the partition, such as

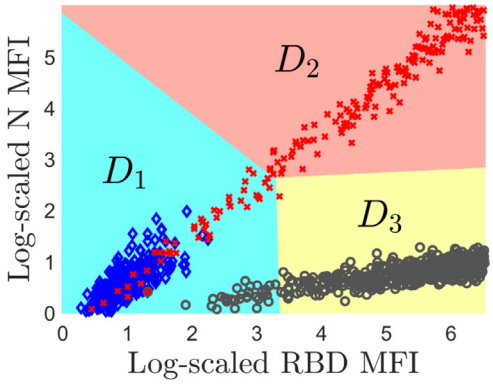
$$\underset{f_1(x),f_2(x),\boldsymbol{P}_\pi}{\operatorname{argmin}} \|\boldsymbol{P}_\pi \boldsymbol{P} - \boldsymbol{I}\|_2^2, \tag{22}$$

where $f_1(x)$ and $f_2(x)$ are indicator functions determining which samples are assigned to elements 1 and 2 of the partition (without loss of generality). Here, $\boldsymbol{P}_\pi \boldsymbol{P}$ is the row permutation of $\boldsymbol{P}$ closest to the identity matrix, $\boldsymbol{I}$. For the matrix $\boldsymbol{P}$ given by (21), for example, $\boldsymbol{P}_\pi = [\boldsymbol{e}_1; \boldsymbol{e}_3; \boldsymbol{e}_2]$, where $\boldsymbol{e}_j$ is the jth standard basis vector.

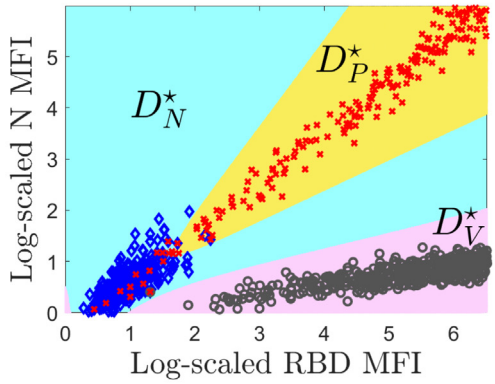
We leave a search for the minimum prevalence error estimate to future work; see Patrone and Kearsley [15] for an approach to the binary case. The extension of their work to the multiclass setting is not obvious because the objective function to minimize can be generalized in many different ways.



(a) Probability model contours and synthetic data



(b) k-means clustering



(c) Optimal classification domains

Fig. 7. (a) Level sets of conditional PDFs with example synthetic data, (b) k-means clustering, (c) optimal classification domains with estimated generalized prevalences. In (c), the subscripts N, P, and V denote naïve, previously infected, and vaccinated.

As a final note on extreme cases of prevalence estimation, in expectation, the problem is unconstrained. The constrained problem may be a viable alternative when it is known that one prevalence is close to zero.

6.2. Local accuracy

Recall that in Section 3.2 we needed to consider sets of measurements with equal probability of belonging to more than one class. This concept is related to *local accuracy*, Z, which compares the probability that a test sample belongs to a particular class and has measurement $\boldsymbol{r}$ to the measurement density of a test sample with measurement $\boldsymbol{r}$. We generalize the binary version from Patrone et al. [16] to the multiclass setting:

$$Z(\boldsymbol{r}, D_1, \dots, D_n) = \frac{q_k P_k(\boldsymbol{r})}{Q(\boldsymbol{r})} = \frac{q_k P_k(\boldsymbol{r})}{\sum_{j=1}^n q_j P_j(\boldsymbol{r})}, \quad \boldsymbol{r} \in D_k, \tag{23}$$

where $\{D_k\}$ partitions Ω. Let $Z^\star(\boldsymbol{r}) = Z(\boldsymbol{r}, D_1^\star, \dots, D_n^\star)$ be the local accuracy of the optimal solution to the multiclass problem. It is straightforward to show that $1/n \le Z^\star \le 1$. Due to optimality, if $\boldsymbol{r} \in D_k^\star$, we have $q_k P_k(\boldsymbol{r}) \ge q_j P_j(\boldsymbol{r})$ for $j \ne k$. Then

$$nq_k P_k(\boldsymbol{r}) \ge \sum_{j=1}^n q_j P_j(\boldsymbol{r}) = Q(\boldsymbol{r}), \tag{24}$$

and so $q_k P_k(\boldsymbol{r})/Q(\boldsymbol{r}) = Z^\star(\boldsymbol{r}) \ge 1/n$ for $r \in D_k^\star$. $Z^\star$ is maximized at 1 when $Q(\boldsymbol{r}) = q_k P_k(\boldsymbol{r})$ for $\boldsymbol{r} \in D_k^\star$. In the multiclass setting, we have $Z^\star = 1/n$ when

$$q_1 P_1(\boldsymbol{r}) = \cdots = q_n P_n(\boldsymbol{r}). \tag{25}$$

We will refer to such an $\boldsymbol{r}$, if it exists, as a multipoint of the optimal domains. The lower bound on $Z^\star$ is only attained at a multipoint. To see this, consider some measurement $\boldsymbol{v}$ that is not a multipoint. Then there exist $j, k \in \{1, \dots, n\}$, $j \ne k$, such that $q_j P_j(\boldsymbol{v}) < q_k P_k(\boldsymbol{v})$. Then, since the classification is optimal, $\boldsymbol{v} \notin D_j^\star$ and $\boldsymbol{v} \in D_m^\star$ for some m (it may be that $m = k$). Clearly, $q_j P_j(\boldsymbol{v}) < q_m P_m(\boldsymbol{v})$. Further, $q_\ell P_\ell(\boldsymbol{v}) \le q_m P_m(\boldsymbol{v})$ for $\ell \ne j$. It follows that

$$Q(\boldsymbol{v}) = \left[\sum_{\substack{i=1 \\ i \ne j}}^n q_i P_i(\boldsymbol{v})\right] + q_j P_j(\boldsymbol{v}) < (n-1) q_m P_m(\boldsymbol{v}) + q_m P_m(\boldsymbol{v}) \tag{26}$$

and so $Q(\boldsymbol{v}) < nq_m P_m(\boldsymbol{v})$, which gives $1/n < Z^\star(\boldsymbol{v})$ for a non-multipoint.

The concept of local accuracy could be used to decide which values to hold out in an indeterminate class in order to meet a global accuracy target (see [16]). For any measurement, we can compute what the local accuracy would be if we chose to assign it to each class in turn. Using the SARS-CoV-2 tri-class example as an illustration, conducting this procedure on a measurement $\boldsymbol{r}$ may result in, say, similarly high local accuracies for the previously infected and vaccinated classes, but a low local accuracy for the naïve class. In this situation and without an optimal classification scheme, we may not feel confident labeling the sample as previously infected or as vaccinated, since their probabilities are similar, but we can say the sample is almost certainly not naïve. This leads naturally to the observation that any subset of classes can be combined to make a new class. In particular, we can reduce the problem to a binary classifier. For our serological example, perhaps it is desirable to consider previously infected and vaccinated samples together, or equally possible that it is difficult to tell them apart for a

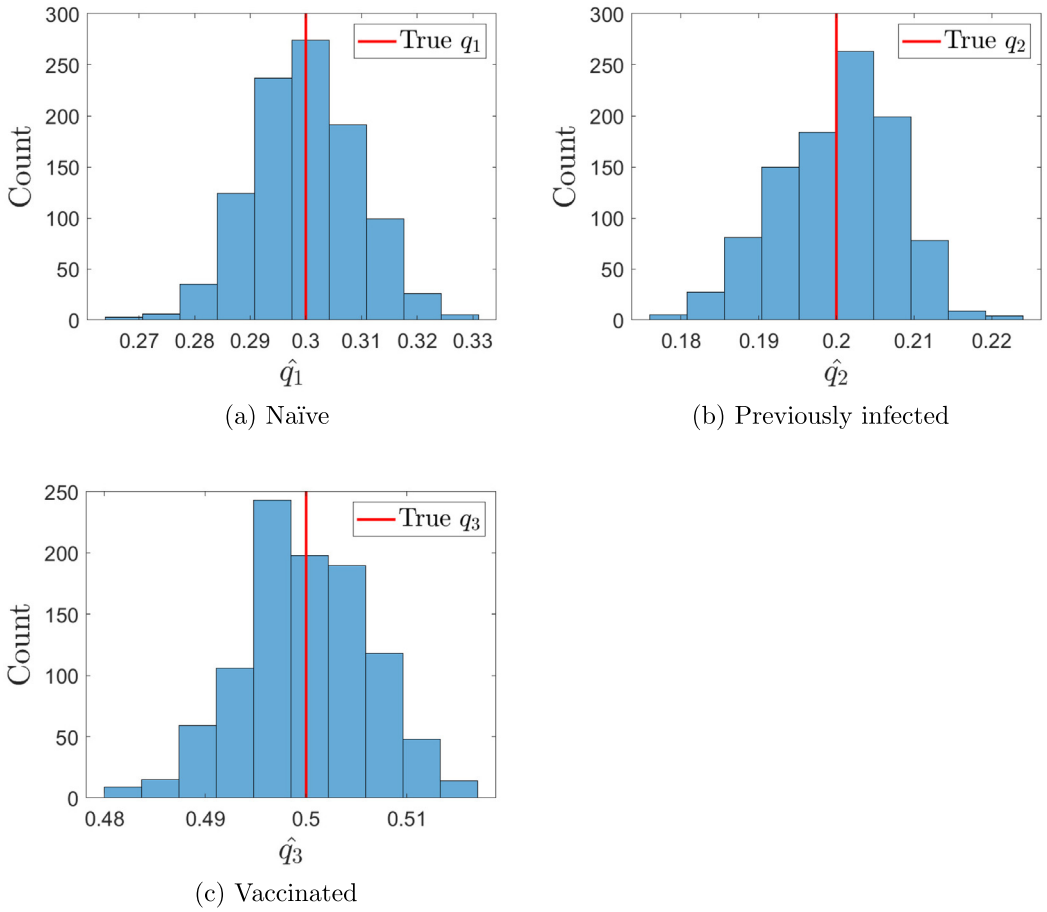


Fig. 8. (a–c): Histograms of generalized prevalence estimates from 1000 synthetic data sets. Ten bins are used for each histogram and the true prevalence is shown as a vertical red line.

particular assay, and so our goal becomes to classify them separately from naïves. Conversely, we may want to reduce a multiclass problem to a binary classification in which our "negative" to prior infection class includes samples from vaccinated individuals.

This reduction of the problem size by combining classes is in a sense a projection onto a lower class space. Specifically, consider

$$Q(\boldsymbol{r}) = \sum_{j=1}^{n} q_j P_j(\boldsymbol{r}) = \underbrace{\sum_{j=1}^{k} q_j P_j(\boldsymbol{r})}_{\tilde{q}_1 \tilde{P}_1(\boldsymbol{r})} + \underbrace{\sum_{j=k+1}^{n} q_j P_j(\boldsymbol{r})}_{\tilde{q}_2 \tilde{P}_2(\boldsymbol{r})}, \tag{27}$$

where $\tilde{P}_1(\boldsymbol{r})$ and $\tilde{P}_2(\boldsymbol{r})$ are newly-created PDFs with associated prevalences $\tilde{q}_1$ and $\tilde{q}_2 = 1 - \tilde{q}_1$. The task becomes to find $\tilde{q}_1$, for which there may be an optimal strategy.

The diagnostic community's current analog to local accuracy is the concept of a likelihood ratio (LR), calculated as $s_e/(1 - s_p)$ for a previously infected test result, where s_e and s_p represent sensitivity and specificity. The previously infected LR can be interpreted as the ratio of probabilities of correctly to incorrectly predicting a previously infected result [17]. These values use average population information through s_e and s_p values, which may not always be available or representative. In contrast, local accuracy uses local information, since the latter is conditioned on knowing individual measurement values.

6.3. Connections and comparison to discriminant analysis

An alternative to our classification method is DA, which classifies observations into known groups [13]. Most realizations of DA do not optimize an accuracy-related objective relevant to most diagnostic applications [3]. In contrast, we ground our work in probability theory and modeling by directly minimizing the misclassification rate adapted to a given assay. Diagnostic classification benefits from uncertainty quantification. Although DA supports this, it also makes assumptions that increase model form errors [18]. Our method addresses this shortcoming by using the data to inform the underlying probability distributions and prevalences used for classification.

In more detail, our approach, which generalizes DA, connects several strategies together, including Bayesian classification methods. The latter, which are most closely aligned with our approach, still assume a prior probability of being in each class without reference to the test data. However, the decision rule is essentially identical to ours and corresponds to the Bayes 0–1 loss [19], except that it does not address boundary cases associated with (13), (15), and (16). Thus, our method extends the Bayes rule to cover this situation.

LDA, a common implementation, assumes all training populations are well-approximated by normal distributions with the same covariance [19]. QDA relaxes the equal covariance assumption. In contrast to LDA and QDA, the first step of our procedure is to select appropriate, and not necessarily normal, distributions for each training class. The modeling exercise is important because many biological populations are not normally distributed. Rasmussen and Williams [13] note, and we find, that if the normality assumption is inappropriate, this may lead to poor performance of the classification method. This assumption is also a characteristic of traditional binary methods such as using confidence intervals, which have been outperformed in related work [6,14,16].

We perform LDA and QDA using `fitcdiscr` and `predict` in MATLAB with default options (MathWorks, Natick, MA, USA). In Table 5 we compare classification errors for the example data in Figs. 5(b)

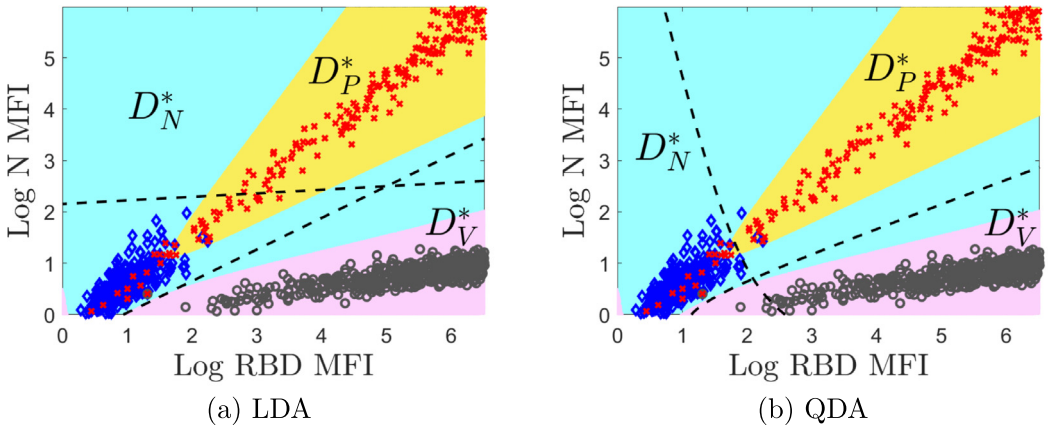


Fig. 9. Comparison of our optimal classification method to linear (LDA) and quadratic (QDA) discriminant analysis. The optimal classification domains are colored whereas the discriminant analysis thresholds are shown as dashed lines.

Table 5
Classification errors for our optimal method compared to linear (LDA) and quadratic (QDA) discriminant analysis. The examples correspond to the data from Figs. 5(b) and 7(c). We show the example 2D comparison in Fig. 9.

Classification method	Classification error (%)	
	1D example	2D example
Optimal classification	5.08	1.80
LDA	7.22	5.20
QDA	7.22	2.20

and 7(c) for our method against those from LDA and QDA. For both examples, our method outperforms the discriminant analyses. Our results range from an 18.2 % to 65.4 % reduction in classification error. We show the 2D comparison in Fig. 9. An observation about Fig. 9 is that the previously infected and vaccinated populations are not normally distributed, as LDA and QDA assume. Thus, it is not surprising that a discriminant threshold cuts directly through a population in the linear case. In contrast to both DA implementations, the power of our modeling is evident in the form of the optimal classification boundaries, thereby reducing the error rate.

6.4. Extensions

Multiclass methods are readily equipped to handle further stratification of antibody data, such as by age group, biological sex, or coronavirus disease of 2019 (COVID-19) booster status. An additional class could be added for individuals who are both vaccinated and previously infected. Studies have demonstrated a greater antibody response post-vaccination for previously-infected versus COVID-19 naïve recipients [20,21]; this could allow for these populations to be distinguished by our classification scheme. Further, we minimize the prevalence-weighted combination of misclassifications, but the optimization problem can be rewritten for any desired objective function. Reformulations include "rule-in" or "rule-out" tests that meet desired sensitivity or specificity targets [22]. Our methods may even be generalizable to multi-label classification, in which a sample can be assigned to more than one class; we anticipate challenges designing the corresponding optimization problem. Finally, the methods presented here can be applied to any setting where class size estimation and population labeling are required; an example is cell sorting in flow cytometry.

6.5. Limitations

Model selection is inherently subjective; Schwartz [23] showed that the error goes to zero as more data points are added. As the number of antibody measurements increases, corresponding to viewing the data in higher dimensions, additional modeling choices become available. Patrone and Kearsley [6] suggest the possibility of minimizing misclassifications over a family of models; see also Smith [18] for a discussion of model form errors. Classification accuracy and prevalence estimation of the 1D data sets from Ainsworth et al. [9] and Wei et al. [10] suffer from overlap between their spike IgG values. If more measurements were available per sample, modeling the data in a higher dimension could improve class separation and thereby lower error rates (see [14]). Further, our models do not account for time-dependence. This concept is important when classifying antibody tests, which are known to have a half life on the order of several months post infection or vaccination [24,25]. See Bedekar et al. [8] for a time-dependent approach to the binary setting.

6.6. Implications for assay developers

We have solved the multiclass diagnostic classification problem, which was previously unresolved. Antibody measurements from vaccinated individuals can now be distinguished from previously infected and naïve samples.

Our work is the first to obtain unbiased predictions of the relative fractions of vaccinated, previously infected, and naïve individuals in a population. These estimates are improved as more samples are added. Best practices for conducting these predictions include dividing the range of all possible measurement values into nonempty regions that create separation between samples of neighboring regions. This can be easily achieved using pre-defined clustering algorithms. Our procedure hinges on selecting probability distributions to model training populations, which can be conducted automatically for measurements of a single antibody target in several open-source programming languages. Our classification scheme is also easily implementable, and can be modified to prioritize specificity if desired. Regardless of the reformulation, the error is minimized by construction.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Analysis scripts and data developed as a part of this work are available upon reasonable request. Original data are provided in Ainsworth et al. [9] and Wei et al. [10].

Acknowledgments

This work is a contribution of the National Institute of Standards and Technology, United States and is not subject to copyright in the United States. R.L. was funded through the NIST PREP grant 70NANB18H162. The aforementioned funder had no role in study design, data analysis, decision to publish, or preparation of the manuscript. Use of data provided in this paper has been approved by the NIST Research Protections Office (IRB no. ITL 2020 2057). The authors wish to thank Drs. Daniel Anderson and Eric Shirley for useful discussions during preparation of this manuscript.

Appendix A. Optimality of classification domains

Lemma 1. *Let the PDFs $P_j(\boldsymbol{r})$ be bounded and summable functions on Ω and suppose that the measure of a single measurement $\boldsymbol{r}$ is zero with respect to all distributions. Further, suppose the boundary set $\mathscr{E}_j$ has measure zero for all j. Define the loss function $\mathscr{L}$ as in (11). Then the domains given by (14) minimize the loss function $\mathscr{L}$.*

Proof. Consider a partition of Ω given by $\{D_j^\star\}$ for which each $D_j^\star$ satisfies the criteria given in (10) and (14). Without loss of generality, absorb each q_j into $P_j(\boldsymbol{r})$, and consider the loss function

$$\mathscr{L}(D_1,\dots,D_n)=\sum_{j=1}^n\int_{\Omega\setminus D_j}P_j(\boldsymbol{r})d\boldsymbol{r}. \tag{A.1}$$

Now consider a different partition of the measurement space $\{\hat{D}_j\}$, where at least two elements $\hat{D}_k$ and $\hat{D}_i$ differ from $D_k^\star$ and $D_i^\star$ by more than a set of measure zero.[4] Without loss of generality, suppose the elements of the partitions $\{\hat{D}_j\}$ and $\{D_j^\star\}$ are ordered such that $\hat{D}_1,\dots\hat{D}_M$ differ from $D_1^\star,\dots,D_M^\star$ by more than a set of measure zero and $\hat{D}_{M+1},\dots\hat{D}_n$ and $D_{M+1}^\star,\dots,D_n^\star$ do not. Note that $\Omega\setminus\hat{D}_j=(\hat{D}_j)^C$. As a slight abuse of notation, we will rewrite $(\hat{D}_j)^C$ and subsequent, similar expressions as

$$(\hat{D}_j)^C=\bigcup_{\substack{k=1\\k\neq j}}^n\hat{D}_k \tag{A.2}$$

with the understanding that this is true up to sets of measure zero and that the discrepancy does not affect integration. Then we may decompose the loss function (A.1) as

$$\mathscr{L}(\hat{D}_1,\dots,\hat{D}_n)=\sum_{j=1}^n\int_{\cup_{k,k\neq j}\hat{D}_k}P_j(\boldsymbol{r})d\boldsymbol{r}. \tag{A.3}$$

Likewise,

$$\mathscr{L}(D_1^\star,\dots,D_n^\star)=\sum_{j=1}^n\int_{\cup_{k,k\neq j}D_k^\star}P_j(\boldsymbol{r})d\boldsymbol{r}. \tag{A.4}$$

Decomposing $(\hat{D}_j)^C$ further gives

$$\begin{aligned}(\hat{D}_j)^C&=\bigcup_{\substack{k=1\\k\neq j}}^n\hat{D}_k=\bigcup_{\substack{k=1\\k\neq j}}^n\left[(\hat{D}_k\cap D_k^\star)\cup(\hat{D}_k\setminus D_k^\star)\right]\\&=\left[\bigcup_{\substack{k=1\\k\neq j}}^n(\hat{D}_k\cap D_k^\star)\right]\cup\left[\bigcup_{\substack{k=1\\k\neq j}}^n(\hat{D}_k\setminus D_k^\star)\right].\end{aligned} \tag{A.5}$$

[4] If only one element $\hat{D}_k$ differed by more than a set of measure zero from $\hat{D}_k^\star$, the set $\{\hat{D}_j\}$ would no longer form a partition of Ω.

Here, (A.5) can be further decomposed using our assumptions on which elements of the partitions are equivalent up to sets of measure zero:

$$\begin{aligned}(\hat{D}_j)^C=&\left[\bigcup_{\substack{k=1\\k\neq j}}^M(\hat{D}_k\cap D_k^\star)\right]\cup\left[\bigcup_{\substack{k=1\\k\neq j}}^M(\hat{D}_k\setminus D_k^\star)\right]\cup\left[\bigcup_{\substack{k=M+1\\k\neq j}}^n(\hat{D}_k\cap D_k^\star)\right]\\&\cup\left[\bigcup_{\substack{k=M+1\\k\neq j}}^n(\hat{D}_k\setminus D_k^\star)\right],\end{aligned} \tag{A.6}$$

The third term in square brackets in (A.6) has the same measure as $\cup_{k=M+1,k\neq j}^n\hat{D}_k$ and the last term has measure zero by our assumptions. We let the first term be denoted by $\hat{D}_{j\cap}$, the second by $\hat{D}_{j\setminus}$, and the third by $\hat{D}_{j\cup}$. Define $D_{j\cap}^\star$, $D_{j\setminus}^\star$, and $D_{j\cup}^\star$ similarly as the first, second, and third terms in (A.6) but with $\wedge$ and $\star$ swapped. Note that all terms in (A.6) are disjoint. Thus, subtracting (A.4) from (A.3) gives

$$\begin{aligned}\Delta\mathscr{L}&=\mathscr{L}(\hat{D}_1,\dots,\hat{D}_n)-\mathscr{L}(D_1^\star,\dots,D_n^\star)\\&=\sum_{j=1}^n\int_{\hat{D}_{j\cap}}P_j(\boldsymbol{r})d\boldsymbol{r}+\sum_{j=1}^n\int_{\hat{D}_{j\setminus}}P_j(\boldsymbol{r})d\boldsymbol{r}+\sum_{j=1}^n\int_{\hat{D}_{j\cup}}P_j(\boldsymbol{r})d\boldsymbol{r}\\&\quad-\sum_{j=1}^n\int_{D_{j\cap}^\star}P_j(\boldsymbol{r})d\boldsymbol{r}-\sum_{j=1}^n\int_{D_{j\setminus}^\star}P_j(\boldsymbol{r})d\boldsymbol{r}-\sum_{j=1}^n\int_{D_{j\cup}^\star}P_j(\boldsymbol{r})d\boldsymbol{r}\\&=\sum_{j=1}^n\int_{\hat{D}_{j\setminus}}P_j(\boldsymbol{r})d\boldsymbol{r}-\sum_{j=1}^n\int_{D_{j\setminus}^\star}P_j(\boldsymbol{r})d\boldsymbol{r},\end{aligned} \tag{A.7}$$

where we have used our assumption that $\hat{D}_{j\cup}=D_{j\cup}^\star$ up to sets of measure zero and

$$\hat{D}_{j\cap}=\bigcup_{\substack{k=1\\k\neq j}}(\hat{D}_k\cap D_k^\star)=\bigcup_{\substack{k=1\\k\neq j}}(D_k^\star\cap\hat{D}_k)=D_{j\cap}^\star. \tag{A.8}$$

Since $\{\hat{D}_j\}$ and $\{D_j^\star\}$ partition Ω, we may write

$$\Delta\mathscr{L}=\sum_{j=1}^n\sum_{\substack{k=1\\k\neq j}}^M\left[\int_{\hat{D}_k\setminus D_k^\star}P_j(\boldsymbol{r})d\boldsymbol{r}-\int_{D_k^\star\setminus\hat{D}_k}P_j(\boldsymbol{r})d\boldsymbol{r}\right]. \tag{A.9}$$

From our assumption that $\hat{D}_j$ differs from $D_j^\star$ by a measurable set for all $j\in\{1,\dots,M\}$, each set $\hat{D}_k\setminus D_k^\star$ in $\hat{D}_{j\setminus}$ and $D_i^\star\setminus\hat{D}_i$ in $D_{j\setminus}^\star$ have nonzero measure. Next, we may write

$$\begin{aligned}\hat{D}_{j\setminus}=\bigcup_{\substack{k=1\\k\neq j}}^M\hat{D}_k\setminus D_k^\star&=\bigcup_{\substack{k=1\\k\neq j}}^M\left\{\left[(\hat{D}_k\setminus D_k^\star)\cap D_j^\star\right]\cup\left[(\hat{D}_k\setminus D_k^\star)\setminus D_j^\star\right]\right\}\\&=\left\{\bigcup_{\substack{k=1\\k\neq j}}^M\left[(\hat{D}_k\setminus D_k^\star)\cap D_j^\star\right]\right\}\cup\left\{\bigcup_{\substack{k=1\\k\neq j}}^M\left[(\hat{D}_k\setminus D_k^\star)\setminus D_j^\star\right]\right\},\end{aligned} \tag{A.10}$$

and similarly for $D_{j\setminus}^\star=\bigcup_{\substack{k=1\\k\neq j}}D_k^\star\setminus\hat{D}_k$. Call the first term in curly braces in (A.10) $\hat{D}_{j\setminus\cap}$ and the second $\hat{D}_{j\setminus\setminus}$, and let $D_{j\setminus\cap}^\star$ and $D_{j\setminus\setminus}^\star$ be similarly defined. Note that $\hat{D}_{j\cap\setminus}$ and $D_{j\setminus\setminus}^\star$ are disjoint by construction. This allows us to write

$$\begin{aligned}\Delta\mathscr{L}=\sum_{j=1}^n\sum_{\substack{k=1\\k\neq j}}^M\Bigg[&\int_{\hat{D}_{j\setminus\cap}}P_j(\boldsymbol{r})d\boldsymbol{r}+\int_{\hat{D}_{j\setminus\setminus}}P_j(\boldsymbol{r})d\boldsymbol{r}\\&-\int_{D_{j\setminus\cap}^\star}P_j(\boldsymbol{r})d\boldsymbol{r}-\int_{D_{j\setminus\setminus}^\star}P_j(\boldsymbol{r})d\boldsymbol{r}\Bigg].\end{aligned} \tag{A.11}$$

We rewrite $(D_j^\star)^C$ to find

$$(\hat{D}_k \setminus D_k^\star) \setminus D_j^\star = (\hat{D}_k \setminus D_k^\star) \cap \left[\bigcup_{\substack{i=1\\ i\neq j}}^{n} D_i^\star\right] = \hat{D}_k \cap \left[\bigcup_{\substack{\ell=1\\ \ell\neq k}}^{n} D_\ell^\star\right] \cap \left[\bigcup_{\substack{i=1\\ i\neq j}}^{n} D_i^\star\right]$$
$$= \bigcup_{\substack{i=1\\ i\neq j,k}}^{n} \hat{D}_k \cap D_i^\star. \tag{A.12}$$

We rewrite $(\hat{D}_j)^C$ similarly. Here, (A.12) and the symmetry of the intersection operator give

$$\hat{D}_{j\setminus\setminus} = \bigcup_{\substack{i=1\\ i\neq j,k}}^{n} \hat{D}_k \cap D_i^\star = \bigcup_{\substack{i=1\\ i\neq j,k}}^{n} D_k^\star \cap \hat{D}_i = D_{j\setminus\setminus}^\star. \tag{A.13}$$

Thus, these terms in (A.11) are equal and of opposite signs, and cancel. This leaves

$$\Delta\mathscr{L} = \sum_{j=1}^{n}\sum_{\substack{k=1\\ k\neq j}}^{M}\left[\int_{\hat{D}_{j\setminus\cap}} P_j(\boldsymbol{r})d\boldsymbol{r} - \int_{D_{j\setminus\cap}^\star} P_j(\boldsymbol{r})d\boldsymbol{r}\right]. \tag{A.14}$$

Since

$$D_{j\setminus\cap}^\star = \bigcup_{\substack{k=1\\ k\neq j}}(D_k^\star \setminus \hat{D}_k)\cap \hat{D}_j \subset (D_j^\star)^C, \tag{A.15}$$

we have $\mu(D_{j\setminus\cap}^\star \cap D_j^\star) \leq \mu((D_j^\star)^C \cap D_j^\star) = 0$. Thus, since there is no measurable overlap with the maximal set $D_j^\star$, we have $P_j(\boldsymbol{r}) \leq P_k(\boldsymbol{r})$ on $D_{j\setminus\cap}^\star$ for $k \neq j$. In contrast, $\hat{D}_{j\setminus\cap}\cap D_j^\star = (\hat{D}_k \setminus D_k^\star)\cap D_j^\star$ is by construction a subset of $D_j^\star$, and in totality $\bigcup_j \hat{D}_{j\setminus\cap}\cap D_j^\star \neq \emptyset$ since we have assumed that $\hat{D}_1,\ldots,\hat{D}_M$ differ from $D_1^\star,\ldots,D_M^\star$ by a set of nonzero measure. Thus $P_j(\boldsymbol{r}) \geq P_k(\boldsymbol{r})$ on $\hat{D}_{j\setminus\cap}$ for $k \neq j$.

Since P_j is maximal on each $\hat{D}_{j\setminus\cap}$ and submaximal on each $D_{j\setminus\cap}^\star$, and $\bigcup_j \hat{D}_{j\setminus\cap}$ has nonzero measure, we have

$$\Delta\mathscr{L} = \sum_{j=1}^{n}\sum_{\substack{k=1\\ k\neq j}}^{M}\left[\int_{\hat{D}_{j\setminus\cap}} P_j(\boldsymbol{r})d\boldsymbol{r} - \int_{D_{j\setminus\cap}^\star} P_j(\boldsymbol{r})d\boldsymbol{r}\right] \geq 0, \tag{A.16}$$

which proves that $\{D_j^\star\}$ as defined minimizes the loss function $\mathscr{L}$. □

Appendix B. Measurement details

Ainsworth et al. [9] provide SARS-CoV-2 serological anti-spike IgG antibody measurements for previously infected and naïve samples (which they refer to as positive and negative). Ainsworth et al. [9] used a true negative rate of 99 % to set a positivity threshold of 8 million MFI units. Previously infected samples were taken at least 20 days post symptom onset from individuals whose infections were confirmed via reverse transcription polymerase chain reaction (RT-PCR) by a nose or throat swab. The naïve samples were collected pre-pandemic.

Wei et al. [10] provide SARS-CoV-2 serological anti-spike IgG antibody measurements for vaccinated samples. Measurements were recorded from individuals that had been inoculated with either the Oxford–AstraZeneca ChAdOx1 nCoV-19 (Vaccine A) or the Pfizer-BioNTech BNT162b2 (Vaccine B). The vaccinated samples were collected at various time points relative to the first inoculation, ranging from 14 days prior to 66 days after. To control for variation resulting from the length of time after inoculation, we use only data from 28 days post first dose.

Ainsworth et al. [9] recorded antibody measurements in MFI. For comparison, Wei et al. [10] convert from MFI to ng/mL following:

$$\log_{10}(m) = A + Bf + C\mathbb{I}_{f>D}(f - D), \tag{B1}$$

where the constants A, B, C, and D are given by $A = 0.221738, B = 1.751889 \times 10^{-7}, C = 5.416675 \times 10^{-7}$, and $D = 9.19031 \times 10^6$. Here, m is the measurement in ng/mL, f is the measurement in MFI, and $\mathbb{I}$ denotes the indicator function on all MFI values, returning one if the measurement is above D and zero otherwise. Wei et al. [10] truncated vaccinated measurements below 2 ng/mL (0.4 % of total data) and above 500 ng/mL (7 % of total data). The truncation was not applied to the previously infected and naïve samples because they were first reported in Ainsworth et al. [9].

In total, there are 976 naïve, 536 previously infected, and 686 vaccinated samples (362 Vaccine A and 324 Vaccine B samples taken at the 28 day mark).

As a note on the specific choice of antibody studied, immunoglobulin M (IgM) and immunoglobulin A (IgA) could also be suitable candidates for such an analysis. IgG and IgM have been shown to exhibit high sensitivity and specificity in binary serological tests [26] and blood-based samples of all three of IgG, IgM, and IgA showed high separation of values for known positive and negative samples [27]. Separation of training data populations is important to prevalence estimation because it can result in nonoverlapping class conditional probability distributions, which admits the possibility of zero-error prevalence estimation if the partition is suitably chosen (see Section 6.1).

Appendix C. Data censoring

Given a general distribution $f(x;\boldsymbol{\varphi})$ of parameters $\boldsymbol{\varphi}$ that fits the data between the truncation limits $x_{\min} < x < x_{\max}$, the probability that a measurement is above the upper truncation limit is

$$p_{\max} = \int_{x_{\max}}^{\infty} f(x)dx, \tag{C1}$$

and the probability that a measurement is below the lower truncation limit is

$$p_{\min} = \int_{-\infty}^{x_{\min}} f(x)dx. \tag{C2}$$

These definitions of $p_{\max}$ and $p_{\min}$ are used in the updated likelihood function

$$L(x) = \begin{cases} p_{\min}, & x = x_{\min},\\ f(x), & x_{\min} < x < x_{\max},\\ p_{\max}, & x = x_{\max}. \end{cases} \tag{C3}$$

The PDF can be written as

$$\hat{f}(x) = f(x)\mathbb{I}(x \in (x_{\min}, x_{\max})) + \delta(x - x_{\min})p_{\min} + \delta(x - x_{\max})p_{\max}. \tag{C4}$$

Appendix D. 2D synthetic probability models

Our 2D exploration builds off work by Patrone and Kearsley [6], which used data from an assay developed in Liu et al. [26]. The two dimensions correspond to measurements taken at the receptor-binding domain (RBD), a substructure of the spike protein of the SARS-CoV-2 molecule, and the nucleocapsid (N) that stabilizes the ribonucleic acid (RNA). Values are recorded as MFIs but follow the same logarithmic scale given by (17). The log-transformed RBD values are x and N values are y. All model parameters are determined via MLE. The naïve samples should have relatively low values of both RBD and N since they have no specific immune response to SARS-CoV-2. However, we still expect small naïve signals due to cross-reactivity with other coronaviruses such as NL63 and HKU1, which are versions of the common cold. In contrast, previously infected samples should have produced a relatively high immune response to both RBD and N.

It is natural to expect some correlation between the signals, which motivates a change of variables $z = (x + y)/\sqrt{2}, w = (x - y)/\sqrt{2}$. This will cause the data to be distributed along the diagonal. Our naïve distribution N is given by

$$N(z,w;k,\alpha,\theta,\beta,\mu) = \frac{z^{k-1}}{\sqrt{2\pi}\alpha\Gamma(k)\theta^k}\exp\left(-\frac{z}{\theta} - \frac{z}{\beta} - \frac{(w-\mu)^2}{2\alpha^2\exp(2z/\beta)}\right). \tag{D1}$$

This is a hybrid gamma-normal distribution where the variance of the variable corresponding to the direction perpendicular to the diagonal, w, depends on the variance of the variable corresponding to the diagonal, z: $\sigma_w = \alpha \exp(z/\beta)$. This allows for the data to fan out slightly away from the origin along the diagonal.

The previously infected distribution P is given by

$$P(\zeta, w; \alpha, \beta, \theta, \mu) = \frac{\Gamma(\alpha+\beta)}{\theta\sqrt{2\pi\zeta}\Gamma(\alpha)\Gamma(\beta)} \zeta^{\alpha-1}(1-\zeta)^{\beta-1} \exp\left(-\frac{(w-\mu)^2}{2\theta^2\zeta}\right). \tag{D2}$$

This is a hybrid beta-normal distribution. Here, we use a different change of variables for the diagonal direction to rescale the beta distribution to its domain $[0, 1]$: $\zeta = (x+y)/(9\sqrt{2})$. Again, the variance of the second variable w depends on the first, ζ: $\sigma_w = \theta\sqrt{\zeta}$. This is a modeling choice that allows for a slightly wider distribution for large N and RBD values. The beta distribution is selected because we recognize that the previously infected samples should span the entire diagonal line $y = x$.

When creating a synthetic vaccinated category, we consider characteristics that set vaccinated and previously infected individuals' antibody measurements apart. Since the vaccines target the binding site on the spike protein, we expect both vaccinated and previously infected individuals to have high RBD values. However, only naturally infected individuals should exhibit high N values. Thus, we create a vaccinated category that is clustered in the bottom right of an N vs. RBD plot.

For the vaccinated (V) population, we use a hybrid Weibull–normal distribution:

$$V(z, w; \alpha, \beta, \delta, \lambda, \mu) = \frac{\delta}{\lambda\alpha\sqrt{2\pi}} \left(\frac{z}{\lambda}\right)^{\delta-1} \exp\left[-\left(\frac{z}{\lambda}\right)^{\delta} - \frac{z}{\beta} - \frac{(w-\mu)^2}{2\alpha^2\exp(2z/\beta)}\right]. \tag{D3}$$

The variance of the second variable w depends on the first, z, and is the same form as that for the naïve distribution.

Appendix E. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mbs.2023.108982.

References

[1] F.L. Caetano dos Santos, I.M. Michalek, K. Laurila, K. Kaukinen, J. Hyttinen, K. Lindfors, Automatic classification of IgA endomysial antibody test for celiac disease: a new method deploying machine learning, Sci. Rep. 9 (1) (2019) 1–7, http://dx.doi.org/10.1038/s41598-019-45679-x.

[2] F. Ramzan, M.U.G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, Z. Mehmood, A deep learning approach for automated diagnosis and multi-class classification of Alzheimer's disease stages using resting-state fMRI and residual neural networks, J. Med. Syst. 44 (2) (2020) 1–16, http://dx.doi.org/10.1007/s10916-019-1475-2.

[3] T. Li, S. Zhu, M. Ogihara, Using discriminant analysis for multi-class classification: an experimental investigation, Knowl. Inf. Syst. 10 (4) (2006) 453–472, http://dx.doi.org/10.1007/s10115-006-0013-y.

[4] Y.-C. Yeh, W.-J. Wang, C.W. Chiou, Cardiac arrhythmia diagnosis method using linear discriminant analysis on ECG signals, Measurement 42 (5) (2009) 778–789, http://dx.doi.org/10.1016/j.measurement.2009.01.004.

[5] W. Lin, Q. Gao, M. Du, W. Chen, T. Tong, Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data, Comput. Biol. Med. 134 (2021) 104478, http://dx.doi.org/10.1016/j.compbiomed.2021.104478.

[6] P.N. Patrone, A.J. Kearsley, Classification under uncertainty: data analysis for diagnostic antibody testing, Math. Med. Biol. 38 (3) (2021) 396–416, http://dx.doi.org/10.1093/imammb/dqab007.

[7] L. Böttcher, M.R. D'Orsogna, T. Chou, A statistical model of COVID-19 testing in populations: effects of sampling bias and testing errors, Phil. Trans. R. Soc. A 380 (2214) (2022) 20210121, http://dx.doi.org/10.1098/rsta.2021.0121.

[8] P. Bedekar, A.J. Kearsley, P.N. Patrone, Prevalence estimation and optimal classification methods to account for time dependence in antibody levels, J. Theoret. Biol. (2022) 111375, http://dx.doi.org/10.1016/j.jtbi.2022.111375.

[9] M. Ainsworth, M. Andersson, K. Auckland, J.K. Baillie, E. Barnes, S. Beer, A. Beveridge, S. Bibi, L. Blackwell, M. Borak, et al., Performance characteristics of five immunoassays for SARS-CoV-2: a head-to-head benchmark comparison, Lancet Infect. Dis. 20 (12) (2020) 1390–1400, http://dx.doi.org/10.1016/s1473-3099(20)30634-4.

[10] J. Wei, N. Stoesser, P.C. Matthews, D. Ayoubkhani, R. Studley, I. Bell, J.I. Bell, J.N. Newton, J. Farrar, I. Diamond, et al., Antibody responses to SARS-CoV-2 vaccines in 45,965 adults from the general population of the United Kingdom, Nat. Microbiol. 6 (9) (2021) 1140–1149, http://dx.doi.org/10.1038/s41564-021-00947-3.

[11] K.E. Clarke, J. Jones, Y. Deng, E. Nycz, A. Lee, I. R, G. AV, H. AJ, M. A, Seroprevalence of infection-induced SARS-CoV-2 antibodies–United States, September 2021–February 2022, MMWR 71 (2022) http://dx.doi.org/10.15585/mmwr.mm7117e3.

[12] R.E. Caflisch, Monte Carlo and quasi-Monte Carlo methods, Acta Numer. 7 (1998) 1–49, http://dx.doi.org/10.1017/s0962492900002804.

[13] C.E. Rasmussen, C.K. Williams, Gaussian Processes for Machine Learning, Vol. 2, No. 3, MIT Press, 2006, http://dx.doi.org/10.7551/mitpress/3206.001.0001.

[14] R.A. Luke, A.J. Kearsley, N. Pisanic, Y.C. Manabe, D.L. Thomas, P.N. Patrone, Modeling in higher dimensions to improve diagnostic testing accuracy: theory and examples for multiplex saliva-based SARS-CoV-2 antibody assays, PLoS One (2023) https://arxiv.org/pdf/2206.14316.pdf, in press.

[15] P. Patrone, A. Kearsley, Minimizing uncertainty in prevalence estimates, 2022, arXiv preprint arXiv:2203.12792.

[16] P.N. Patrone, P. Bedekar, N. Pisanic, Y.C. Manabe, D.L. Thomas, C.D. Heaney, A.J. Kearsley, Optimal decision theory for diagnostic testing: Minimizing indeterminate classes with applications to saliva-based SARS-CoV-2 antibody assays, Math. Biosci. 351 (2022) 108858, http://dx.doi.org/10.1016/j.mbs.2022.108858.

[17] R.H. Riffenburgh, Statistics in Medicine, Associated Press, 2012, http://dx.doi.org/10.1016/b978-0-12-384864-2.00010-x.

[18] R.C. Smith, Uncertainty Quantification: Theory, Implementation, and Applications, Vol. 12, SIAM, 2013.

[19] W.K. Härdle, L. Simar, Applied Multivariate Statistical Analysis, Springer Nature, 2019, http://dx.doi.org/10.1007/978-3-030-26006-4.

[20] L. Dalle Carbonare, M.T. Valenti, Z. Bisoffi, C. Piubelli, M. Pizzato, S. Accordini, S. Mariotto, S. Ferrari, A. Minoia, J. Bertacco, et al., Serology study after BTN162b2 vaccination in participants previously infected with SARS-CoV-2 in two different waves versus naïve, Commun. Med. 1 (1) (2021) 1–11, http://dx.doi.org/10.1038/s43856-021-00039-7.

[21] T.M. Narowski, K. Raphel, L.E. Adams, J. Huang, N.A. Vielot, R. Jadi, A.M. de Silva, R.S. Baric, J.E. Lafleur, L. Premkumar, SARS-CoV-2 mRNA vaccine induces robust specific and cross-reactive IgG and unequal neutralizing antibodies in naïve and previously infected people, Cell Rep. 38 (5) (2022) 110336, http://dx.doi.org/10.1016/j.celrep.2022.110336.

[22] C.M. Florkowski, Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests, Clin. Biochem. Rev. 29 (Suppl 1) (2008) S83, https://pubmed.ncbi.nlm.nih.gov/18852864/.

[23] S.C. Schwartz, Estimation of probability density by an orthogonal series, Ann. Math. Stat. (1967) 1261–1265, http://dx.doi.org/10.1214/aoms/1177698795.

[24] W. Xia, M. Li, Y. Wang, L.E. Kazis, K. Berlo, N. Melikechi, G.R. Chiklis, Longitudinal analysis of antibody decay in convalescent COVID-19 patients, Sci. Rep. 11 (1) (2021) 1–9, http://dx.doi.org/10.1038/s41598-021-96171-4.

[25] S.L. Kwok, S.M. Cheng, J.N. Leung, K. Leung, C.K. Lee, J.M. Peiris, J.T. Wu, Waning antibody levels after COVID-19 vaccination with mRNA Comirnaty and inactivated CoronaVac vaccines in blood donors, Hong Kong, April 2020 to October 2021, Eurosurveillance 27 (2) (2022) 2101197, http://dx.doi.org/10.2807/1560-7917.es.2022.27.2.2101197.

[26] T. Liu, J. Hsiung, S. Zhao, J. Kost, D. Sreedhar, C.V. Hanson, K. Olson, D. Keare, S.T. Chang, K.P. Bliden, et al., Quantification of antibody avidities and accurate detection of SARS-CoV-2 antibodies in serum and saliva on plasmonic substrates, Nat. Biomed. Eng. 4 (12) (2020) 1188–1196, http://dx.doi.org/10.1038/s41551-020-00642-4.

[27] N. Pisanic, P.R. Randad, K. Kruczynski, Y.C. Manabe, D.L. Thomas, A. Pekosz, S.L. Klein, M.J. Betenbaugh, W.A. Clarke, O. Laeyendecker, et al., COVID-19 serology at population scale: SARS-CoV-2-specific antibody responses in saliva, J. Clin. Microbiol. 59 (1) (2020) e02204–e02220, http://dx.doi.org/10.1128/jcm.02204-20.