Confidence intervals for prevalence estimates from complex surveys with imperfect assays

Damon M. Bayer, Michael P. Fay, Barry I. Graubard

1Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, USA
2Department of Statistics, University of California, Irvine, Irvine, California, USA
3Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA

Correspondence
Michael P. Fay, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20892, USA.
Email: mfay@niaid.nih.gov

There are established methods for estimating disease prevalence with associated confidence intervals for complex surveys with perfect assays, or simple random sample surveys with imperfect assays. We develop and study methods for the complicated case of complex surveys with imperfect assays. The new methods use the melding method to combine gamma intervals for directly standardized rates and established adjustments for imperfect assays by estimating sensitivity and specificity. One of the new methods appears to have at least nominal coverage in all simulated scenarios. We compare our new methods to established methods in special cases (complex surveys with perfect assays or simple surveys with imperfect assays). In some simulations, our methods appear to guarantee coverage, while competing methods have much lower than nominal coverage, especially when overall prevalence is very low. In other settings, our methods are shown to have higher than nominal coverage. We apply our method to a seroprevalence survey of SARS-CoV-2 in undiagnosed adults in the United States between May and July 2020.

KEYWORDS
confidence distributions, melding confidence intervals, seroprevalence, survey sampling, weighted sampling

1 INTRODUCTION

Estimating and quantifying uncertainty for disease prevalence is a standard task in epidemiology. For rare events, these estimates are highly sensitive to misclassification, making adjustments for sensitivity and specificity critically important. While estimating prevalence (or any event proportion in a population) in complex surveys and adjusting estimates for misclassification have been well studied separately, performing both of these tasks simultaneously remains relatively unexplored. This article fills that gap by studying estimates and confidence intervals for prevalence from complex surveys with misclassification. We develop a new tractable confidence interval procedure designed to guarantee coverage. Our confidence interval combines three statistical methods: (1) the gamma confidence interval for directly standardized rates is applied to the apparent prevalence and (2) standard misclassification adjustments to that apparent prevalence for sensitivity and specificity are used together with (3) the melding method, which allows combination of confidence intervals for different parameters. We show by simulation only (there are no proofs of guaranteed coverage) that, with a complex survey with low prevalence and misclassification, our method appears to have at least nominal coverage. Even in special cases (eg, simple random samples with imperfect assays, or weighted samples with perfect assays) where there

Abbreviations: ANA, anti-nuclear antibodies; APC, antigen-presenting cells; IRF, interferon regulatory factor.

© 2023 John Wiley & Sons Ltd. This article has been contributed to by U.S. Government employees and their work is in the public domain in the USA.
are existing methods, our new method has at least nominal simulated coverage, while some of the tractable existing methods may under-cover. The cost for apparently achieving nominal coverage is that our intervals may be quite wide in some situations.

Recent overviews of confidence interval procedures for prevalence in surveys without misclassification are provided by Dean and Pagano\(^5\) and Franco et al.\(^6\) For simple random sample surveys with imperfect sensitivity and specificity, Lang and Reiczigel\(^7\) proposed an approximate confidence interval that performed well in simulations. Recent work by DiCiccio et al.\(^8\) and Cai et al.\(^9\) study both valid (ie, exact) and approximate intervals. Their valid intervals use test inversion and the adjustment of Berger and Boos,\(^10\) while their approximation intervals use the bootstrap with the test inversion approach. Fewer methods are available for constructing frequentist confidence intervals for prevalence estimates from complex surveys while adjusting for sensitivity and specificity. Kalish et al.\(^11\) developed one such method that is closely related to one of the methods presented here, but that method’s properties were not studied. Cai et al.\(^9\) (see also the discussion in DiCiccio et al.\(^8\)) modify their approximation approach to allow sample weights for strata or individual specific weights. Another recent advancement is the method developed by Rosin et al.\(^12\) that makes use of asymptotic normal approximations which reduce to the Wald interval when sensitivity and specificity are perfect. This problem has also previously been addressed in Bayesian literature, recently by Gelman and Carpenter.\(^13\)

We work up to our ultimate goal in stages. First, in Section 2.2, we propose confidence intervals for simple random samples where prevalence is assessed with an assay with imperfect sensitivity and/or specificity. Next, in Section 2.3, we present confidence intervals for weighted samples where prevalence is assessed with an assay without misclassification. In Section 2.4, we combine these methods to create confidence intervals for weighted samples where prevalence is assessed with an assay with imperfect sensitivity and specificity. Because the combined method reduces to one of the first two methods as a special case, we can think of the first two stages as testing the combined method in those cases. Finally, in Section 2.5 we show how certain complex surveys may fit into the format for our new method.

In simulations we compare our method to established frequentist competitors and show through simulations that it beats the best of those in each of the three stages with respect to guaranteeing coverage. However, in some simulated settings, our proposed methods are overly conservative, meaning that they demonstrate higher than nominal coverage, while competitor methods maintain closer to nominal coverage. We did not include in our simulations some recent methods that have been developed in response to the COVID-19 pandemic.\(^8,9,12\) The exact method of DiCiccio et al.\(^8\) would guarantee coverage, although applying it to a survey with a large number of strata would be “computationally expensive,” and it has not been applied to surveys using post-stratification weighting. In contrast, our new method can very tractable in those situations.

## 2 CONFIDENCE INTERVAL METHODS

### 2.1 Notation and problem set-up

To introduce notation, consider first the stratified simple random sample. Suppose we have a population partitioned into \(K\) strata, with \(N_1, N_2, \ldots , N_K\) individuals in the \(K\) strata of the population. We sample \(n_1, n_2, \ldots , n_K\) individuals via a simplerandom sample from each of the \(K\) strata to have an assay performed on each individual to determine who has the disease. Let \(X_i\) be the number of positive results from an assay performed on the \(n_i\) individuals from stratum \(i\) and assume \(X_i \sim \text{Binomial}(n_i, \theta_i)\), where \(\theta_i\) is the population frequency of positive results for assays performed on individuals from stratum \(i\). Similarly, let \(X_i^*\) be the unobserved true number of people with the disease among the \(n_i\) individuals from stratum \(i\) and assume \(X_i^* \sim \text{Binomial}(n_i, \theta^*_i)\), where \(\theta^*_i\) is the population frequency of cases in stratum \(i\). In the case of a perfect assay, \(\theta_i = \theta^*_i\).

Therefore, the population prevalence is

\[
\beta^* = \frac{\sum_{i=1}^{K} N_i \theta^*_i}{\sum_{i=1}^{K} N_j} = \sum_{i=1}^{K} w_i \theta^*_i, \tag{1}
\]

and the apparent prevalence is

\[
\beta = \frac{\sum_{i=1}^{K} N_i \theta_i}{\sum_{i=1}^{K} N_j} = \sum_{i=1}^{K} w_i \theta_i, \tag{2}
\]
where $w_i = N_i / \sum_{j=1}^{K} N_j$ and, therefore, $\sum_{i=1}^{K} w_i = 1$. This set-up will approximately work for other complex survey samples, where we can estimate survey weights such that the complex survey sample may be treated as a multinomial sample with probabilities proportional to those weights (see Section 2.5).

We can relate $\theta_i$ and $\theta^*_i$ using the sensitivity ($\phi_p$) and specificity ($1-\phi_n$) of the assay, where $\phi_p$ and $\phi_n$ are the proportion of positive assays from a population of positive controls (ie, individuals known to have the disease) and negative controls (ie, individuals known to be without the disease), respectively. Then $\theta_i = \phi_p \theta^*_i + \phi_n (1 - \theta^*_i)$, or equivalently, 3

$$\theta^*_i = \frac{\theta_i - \phi_n}{\phi_p - \phi_n},$$

and we have

$$\beta^* = \sum_{i=1}^{K} w_i \theta^*_i = \sum_{i=1}^{K} w_i \left( \frac{\theta_i - \phi_n}{\phi_p - \phi_n} \right)$$

$$= \frac{\sum_{i=1}^{K} w_i \theta_i}{\phi_p - \phi_n} - \frac{\phi_n \sum_{i=1}^{K} w_i}{\phi_p - \phi_n} = \frac{\sum_{i=1}^{K} w_i \theta_i}{\phi_p - \phi_n} - \frac{\phi_n}{\phi_p - \phi_n}.$$  

(3)

Suppose the assay is measured on $m_n$ individuals known to not have the disease and on $m_p$ individuals known to have the disease. Let $C_n$ and $C_p$ be the number who test positive from the respective samples. Assume that the negative and positive controls act like simple random samples from their respective populations. Thus, $C_n \sim \text{Binomial}(m_n, \phi_n)$ where $1 - \phi_n$ is the specificity of the assay, and $C_p \sim \text{Binomial}(m_p, \phi_p)$, where $\phi_p$ is the sensitivity of the assay. Let $\hat{\theta}_i = \sum_{i=1}^{K}$, $\hat{\phi}_n = \frac{C_n}{m_n}$, and $\hat{\phi}_p = \frac{C_p}{m_p}$. Then a plug-in estimator for $\beta^*$ is

$$\hat{\beta}^* = \frac{\sum_{i=1}^{K} w_i \hat{\theta}_i}{\hat{\phi}_p - \hat{\phi}_n} - \frac{\hat{\phi}_n}{\hat{\phi}_p - \hat{\phi}_n}.$$  

(4)

This estimator serves as an important basis for developing confidence intervals in this work. Section 2.2 is concerned with confidence intervals for $\beta^*$ in the case where $K = 1$, $\phi_n > 0$, $\phi_p < 1$, that is, estimating prevalence from a simple random sample with an imperfect assay. Section 2.3 is concerned with confidence intervals for $\beta^*$ in the case where $K > 1$, $\phi_n = 0$, $\phi_p = 1$, that is, estimating prevalence from a weighted sample with a perfect assay. Section 2.4 is concerned with confidence intervals for $\beta^*$ in the case where $K > 1$, $\phi_n > 0$, $\phi_p < 1$, that is, estimating prevalence from a weighted sample with an imperfect assay.

### 2.2 Estimating prevalence from a simple random sample with an imperfect assay

First we consider the scenario where $K = 1$, $\phi_n > 0$, and $\phi_p < 1$. We develop a confidence interval for the population prevalence, $\beta^*$. When $K = 1$, the estimand in Equation (3) becomes $\beta^* = (\theta_1 - \phi_n)/(\phi_p - \phi_n)$. We have $\phi_p > \phi_n$ for any useful assay, and since the sample is a mixture of individuals with and without the disease of interest, $\phi_p \geq \theta_1 \geq \phi_n$. The estimator of $\beta^*$ is

$$\hat{\beta}^* \equiv g(\hat{\theta}_1, \hat{\phi}_n, \hat{\phi}_p) \equiv \begin{cases} 
1 & \text{if } \hat{\phi}_n < \hat{\phi}_p \leq \hat{\theta}_1 \\
\frac{\hat{\theta}_1 - \hat{\phi}_n}{\hat{\phi}_p - \hat{\phi}_n} & \text{if } \hat{\phi}_p > \hat{\theta}_1 \geq \hat{\phi}_n \\
0 & \text{otherwise.} 
\end{cases}$$  

(5)

To create a confidence interval for $\hat{\beta}^*$, we use a generalization of the melding method, 4 which makes use of lower and upper confidence distributions on functions of independent estimators to account for variability in $\hat{\theta}_1$, $\hat{\phi}_n$, and $\hat{\phi}_p$. Before describing the melding method, we briefly review confidence distributions.
Confidence distributions are frequentist estimators of a parameter, which can be used similarly to the bootstrap or posterior distributions. For continuous responses (and asymptotically for discrete responses), one can formally define confidence distributions without relying on previously derived confidence interval procedures, but for intuition with discrete responses, it is easier to derive confidence distributions the classical way, that is, directly from confidence interval procedures. For discrete responses there are two confidence distributions (the lower and upper ones), used to ensure validity of the resulting inferences. Under the classical derivation, we can create random variables associated with the lower and upper confidence distributions from two one-sided, nested, confidence interval procedures, where for a nested confidence interval procedure, the \((1 - \alpha_1)\) interval is a subset of the \((1 - \alpha_2)\) interval whenever \((1 - \alpha_1) \leq (1 - \alpha_2)\). Denote those one-sided intervals by \([L(x, 1 - a), \infty)\) and \((-\infty, U(x, 1 - a))\), which are defined for any observed data \(x\), and \((1 - a) \in (0, 1)\). For fixed \(x\), define the lower and upper confidence distribution random variables as \(L(x, A)\) and \(U(x, B)\), where \(A\) and \(B\) are independent uniform random variables. Then a \(100(1 - \alpha)\%\) central confidence interval has the \((\alpha/2)\)th quantile of \(L(x, A)\) as the lower limit, and the \((1 - \alpha/2)\)th quantile of \(U(x, B)\) as the upper limit.

Because each estimated component in Equation (5) is a binomial probability parameter, we now focus on the confidence distributions associated with the exact binomial confidence interval. For a binomial experiment with \(x\) successes out of \(n\) trials, the lower confidence distribution is \(Beta(x, n - x + 1)\) with associated random variable \(B_L\), and the upper confidence distribution is \(Beta(x + 1, n - x)\) with random variable \(B_U\), where for \(\alpha > 0\) we let \(Beta(0, \alpha)\) and \(Beta(\alpha, 0)\) be point masses at 0 and 1, respectively. This result comes from the relationship between the binomial and beta distributions, namely \(Pr[X \leq x] = Pr[B > \theta]\), where \(B \sim Beta(x + 1, n - x)\) for \(x \in \{-1, 0, 1, \ldots, n\}\). Additionally, in the binomial case, the lower and upper confidence distributions are equivalent to the posterior distributions that result from using well-calibrated null preference priors. Let \(q(a, W)\) be the \(a\)th quantile of a random variable \(W\). Then the exact \(1 - \alpha\%\) central confidence interval of Clopper-Pearson for the binomial parameter is

\[
\left\{ q\left(\frac{\alpha}{2}, B_L^*\right), q\left(1 - \frac{\alpha}{2}, B_U^*\right) \right\}.
\]

Fay et al. proposed the melding method for obtaining confidence intervals for functions of two parameters that are monotonic within the allowable range for each parameter given the other is fixed. The melding method uses quantiles of those functions except replacing lower or upper confidence distribution random variables for their associated parameter in the function. Here we generalize that method to \(\beta^*\), which is a function of three parameters. When \(1 \geq \phi > \theta > \phi_n \geq 0\) then \(\beta^*\) is monotonically increasing in \(\theta\), monotonically decreasing in \(\phi\), and monotonically decreasing in \(\phi_n\). For an assessment of monotonicity in other scenarios see Appendix A. Then the \(1 - \alpha\%\) confidence interval for \(\beta^*\) is

\[
\left\{ q\left(\frac{\alpha}{2}, g\left(B_{\theta_1}^*, B_{\phi_1}^*, B_{\phi_n}^*\right)\right), q\left(1 - \frac{\alpha}{2}, g\left(B_{\theta_1}^*, B_{\phi_1}^*, B_{\phi_n}^*\right)\right) \right\}.
\]

where \(g(\cdot)\) is defined in Equation (5). In Equation (7), the choice of confidence distribution random variable (lower or upper) is determined by the associated direction of the monotonicity to attempt to ensure coverage. As the sample sizes \((e.g., n, m_p, m_n)\) increase, there is little difference between the lower and upper confidence distributions.

The quantiles of these melded distributions are calculated by Monte Carlo sampling from each of the component distributions. We compare this method to one described in Lang and Reiczigel as implemented in prevSeSp function in the asht R package, which provides approximate confidence intervals for true prevalence when sensitivity and specificity are estimated from independent samples, as they are in this section. The Lang-Reiczigel interval is given by

\[
\beta_1^{**} + d\beta \pm q\left(1 - \frac{\alpha}{2}, \varphi\right) \cdot \sqrt{Var(\beta_1^{**})},
\]

where \(q_{\varphi} \equiv q\left(1 - \frac{\alpha}{2}, \varphi\right)\), \(\varphi \sim N(0, 1)\),

\[
d\beta = 2 \cdot q_{\varphi}^2 \cdot \left\{ \begin{array}{c}
\phi_n'(1 - \phi_n) - (1 - \beta_1^{**}) \cdot \frac{(1 - \phi_n') \phi_n'}{m_n'}
\end{array} \right\},
\]

\[
Var(\beta_1^{**}) = \frac{\phi_n'(1 - \beta_1^{**})}{m_n'} + (\beta_1^{**})^2 \frac{\phi_n'(1 - \phi_n')}{m_n'} + (1 + \beta_1^{**})^2 \frac{(1 - \phi_n') \phi_n'}{m_n'},
\]
2.3 Estimating prevalence from a weighted sample with a perfect assay

Next, we present a confidence interval for the population prevalence, $\beta^*$, in the scenario $K > 1$, $\phi_n = 0$, $\phi_p = 1$. Our method is a straightforward adaptation of the gamma confidence interval presented in Fay and Feuer, which was developed to create confidence intervals for a standardized population rate which is assumed to be a weighted sum of Poisson rate parameters. We note that for sufficiently large sample size $n$ and small rate $\lambda$, a Poisson $(n \lambda)$ distribution is approximately equal in distribution to a Binomial $(n, \lambda)$ distribution. Under this Poisson assumption, we suggest the 100$(1 - \alpha)\%$ gamma confidence interval for $\beta^*$:

$$
\left( q \left( \frac{\alpha}{2}, G^L_{\beta^*} \right), q \left( 1 - \frac{\alpha}{2}, G^U_{\beta^*} \right) \right),
$$

where

$$
G^L_{\beta^*} \sim \text{Gamma} \left( \frac{y^2}{v} \cdot \frac{y}{y^*} \right), \quad G^U_{\beta^*} \sim \text{Gamma} \left( \frac{y^*}{v^2} \cdot \frac{y^*}{y^*} \right),
$$

$$
y = \sum_{i=1}^{K} w_i \hat{\theta}_i, \quad v = \sum_{i=1}^{K} \left( \frac{w_i}{n_i} \right)^2 x_i,
$$

$$
y^* = y + \max \left( \frac{w_1}{n_1}, \ldots, \frac{w_K}{n_K} \right), \quad v^* = v + \left\{ \max \left( \frac{w_1}{n_1}, \ldots, \frac{w_K}{n_K} \right) \right\}^2.
$$

We call this the wsPoisson method, since it assumes a weighted sum of Poissons. We compare the wsPoisson confidence interval to two methods presented in Dean and Pagano, which were recommended for scenarios with low prevalence. Dean and Pagano showed in simulations that the standard Wald interval had poor coverage with low prevalence (e.g., fig. 1 of that paper showed 95% confidence intervals with coverage of less than 85% for prevalence values less than 2%). Since the confidence interval of Rosin et al reduces to the Wald interval with perfect assays, we will not include that method in the simulation comparisons.

The first recommended method of Dean and Pagano is an adaptation of the method of Agresti and Coull for the survey setting. The interval for $\beta^*$ is given by:

$$
\hat{p} \pm q_Z \sqrt{\hat{p}(1 - \hat{p})/\hat{n}},
$$

where

$$
\hat{\bar{x}} = \left( \sum_{i=1}^{k} w_i \hat{\theta}_i \right) n_{eff} + c, \quad \hat{n} = n_{eff} + 2c,
$$

$$
\hat{p} = \frac{\hat{\bar{x}}}{\hat{n}}, \quad c = q_Z^2/2,
$$

$$
n_{eff} = \frac{\left( \sum_{i=1}^{k} w_i \hat{\theta}_i \right) \left( 1 - \sum_{i=1}^{k} w_i \hat{\theta}_i \right)}{\sum_{i=1}^{k} \frac{w_i^2}{n_i} \hat{\theta}_i}. \quad (11)
$$

In the case where $\sum_{i=1}^{k} \frac{w_i^2}{n_i} \hat{\theta}_i = 0$, we instead let $n_{eff} = \sum_{i=1}^{k} n_i$. 

$$
\phi_p' = \frac{m_p + \phi_p + 1}{m_p + 2}, \quad \phi_p' = \frac{m_n \cdot (1 - \phi_n) + 1}{m_n + 2}, \quad m_p' = m_p + 2, \quad m_n' = m_n + 2.
$$

$$
\beta_1' = \frac{\beta_1' - \phi_n'}{\phi_p' - \phi_n'},
$$

$$
\beta_1' = \frac{n_1 \cdot \hat{\theta}_1 + q_Z^2/2}{n_1 + q_Z^2}.$$
We also compare our suggested method to Dean and Pagano’s modification of the method of Korn and Graubard.\textsuperscript{5,20} This interval is given by

\[
\left( q \left( \frac{\alpha}{2}, B^L_{\text{KG}} \right), q \left( 1 - \frac{\alpha}{2}, B^U_{\text{KG}} \right) \right),
\]

where analogously to the Clopper-Pearson interval (see Equation 6),

\[
B^L_{\text{KG}} \sim \text{Beta} \left( x_{\text{eff}}, n_{\text{eff}} - x_{\text{eff}} + 1 \right), \quad B^U_{\text{KG}} \sim \text{Beta} \left( x_{\text{eff}} + 1, n_{\text{eff}} - x_{\text{eff}} \right),
\]

with \( x_{\text{eff}} = n_{\text{eff}} \sum_{i=1}^{k} w_i \hat{\theta}_i \), and \( n_{\text{eff}} \) defined in Equation (11). Although Dean and Pagano\textsuperscript{5} expressed this in terms of \( F \)-distributions, the beta distribution representation is equivalent.

\subsection*{2.4 Estimating prevalence from a weighted sample with an imperfect assay}

Lastly, we develop a confidence interval for the population prevalence, \( \beta^* \), in the case where \( K > 1, \phi_n > 0, \phi_p < 1 \).

The two methods we discuss are closely related to each other and the methods discussed in Sections 2.2 and 2.3. As in Section 2.2, we use the melding method\textsuperscript{4} to create \( 1 - \alpha \)\% confidence interval very similar to Equation (7). The confidence distributions for \( \phi_p \) and \( \phi_n \) are the same beta distributions as in Section 2.2. The two methods differ in their confidence distributions for the apparent prevalence \( \beta \).

In the first case, we use the adaptation of the gamma confidence interval\textsuperscript{2} presented in Section 2.3 to derive the \( 1 - \alpha \)\% confidence interval for \( \beta^* 

\[
\left\{ q \left( \frac{\alpha}{2}, a_{p}, B^L_{\hat{\phi}_p}, B^U_{\hat{\phi}_p} \right), q \left( 1 - \frac{\alpha}{2}, a_{p}, B^L_{\hat{\phi}_p}, B^U_{\hat{\phi}_p} \right) \right\}.
\]

where \( a_{p} \) are defined in Section 2.3. We refer to this method as the WprevSeSp Poisson—weighted prevalence with sensitivity and specificity, where the prevalence confidence distribution is based on the weighted sum of Poissons.

The alternative method is very similar to that used in Kalish et al.\textsuperscript{11} We use the modification of Korn and Graubard’s method presented by Dean and Pagano,\textsuperscript{5} as in Section 2.3, to derive the \( 100(1 - \alpha) \)\% confidence interval for \( \beta^* 

\[
\left\{ q \left( \frac{\alpha}{2}, a_{n}, B^L_{\hat{\phi}_n}, B^U_{\hat{\phi}_n} \right), q \left( 1 - \frac{\alpha}{2}, a_{n}, B^L_{\hat{\phi}_n}, B^U_{\hat{\phi}_n} \right) \right\}.
\]

where \( B^L_{\hat{\phi}_n} \) and \( B^U_{\hat{\phi}_n} \) are as defined in Section 2.3. Although equivalent, this expression looks different than the one used in Kalish et al\textsuperscript{11} because they used a parameter for specificity, rather than \( \phi_n \), which is 1 minus specificity. We refer to this as WprevSeSp Binomial—weighted prevalence with sensitivity and specificity, where the prevalence confidence distributions are based on a binomial assumption. The WprevSeSp Binomial and WprevSeSp Poisson methods are implemented in the asht R package.\textsuperscript{18}

\subsection*{2.5 Applications to more complex surveys}

\subsection*{2.5.1 When we can use these methods}

In Section 2.1, we assumed that the apparent prevalence was a weighted sum of binomial random variables, \( \beta = \sum_{i=1}^{K} w_i \hat{X}_i / n_i \), where \( X_i \sim \text{Binomial}(n_i, \theta_i) \). Then in Section 2.3, we used the fact that for small \( \theta_i \) and large \( n_i \), the binomial can be approximated by the Poisson, giving \( X_i \sim \text{Poisson}(n_i \theta_i) \). Thus, whenever we can model a complex survey estimator of apparent prevalence as a weighted sum of Poisson variates, then we can apply the methods of this article.

In the upcoming Section 2.5.2, we give a detailed review relating the multinomial sampling model to a weighted sum of Poisson variates model. The multinomial sampling model treats the survey sample as if it is a sampling with replacement from the entire population of \( N \) individuals, where each of the \( N \) individuals has a probability of \( P_j \) of being sampled for each of the \( n \) samples from the survey, with \( \sum_{j=1}^{N} P_j = 1 \). Under this model the number of times each of the \( N \) individuals
is included in the sample is a multinomial with parameters \( n \) and \( \{P_1, \ldots, P_N\} \). The multinomial model describes sampling with replacement, but it is nevertheless used to approximate a sampling design where the \( j \)th individual is sampled without replacement with probability \( P_j \), even though under that design (unlike the multinomial model) no individual is included in the sample more than once. The multinomial model is a common approximation for other complex survey designs. See, for example, Reference 21 (p. 14). For example, in the Kalish et al.\(^\text{11}\) analysis of Section 4, each individual in the sample is assigned a pseudo-weight approximating one over their sampling probability from a multinomial model. The actual sample was not a probability sample. In fact, it was a quota sample from a very large pool of self-selected volunteers, and the pseudo-weights were calculated using a different large survey that was a probability weighted survey. The pseudo-weights were calculated such that if they were analyzed under the multinomial model, they would adjust for selection bias due to self-selection of the volunteers and the imperfection of the quota sampling.

### 2.5.2 Multinomial sampling model

Let \( Y_1, \ldots, Y_N \) be the binary indicators of event in the \( N \) individuals in the population of interest, so the prevalence is \( \beta = N^{-1} \sum_{j=1}^{N} Y_j \). There are many ways to design a complex survey sample, and it is often useful to analyze them as if individuals were sampled with replacement with the sampling probability of the \( j \)th individual equal to \( P_j \), with \( \sum_{j=1}^{N} P_j = 1 \). In other words, we treat the sample as if it were \( n \) independent multinomial samples each with one trial and selection probability vector \( \{P_1, \ldots, P_N\} \). Let \( I_{ij} = 1 \) if the \( i \)th draw for the sample is individual \( j \) in the population, and 0 otherwise. Then let \( y_i = Y_j \) and \( p_i = P_j \) when \( I_{ij} = 1 \). Here, following the tradition in the survey literature, we use capital letters for the population of interest (e.g., \( N, Y, P \)), and lower case letters for the sample (e.g., \( n, y, p \)). In this notation, both \( y_i \) and \( Y_j \) are fixed, and only the variables representing the sampling (i.e., the \( I_{ij} \) variables) are random. Under this independent multinomial model, since \( E(I_{ij}) = P_j \), an unbiased estimator of \( \beta \) is

\[
\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{n} \sum_{j=1}^{N} \frac{I_{ij} Y_j}{N P_j},
\]

(15)

and an unbiased estimator of \( \text{var}(\hat{\beta}) \) under the multinomial model is

\[
\text{var}_M(\hat{\beta}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left( \frac{y_i}{nP_i} - \hat{\beta} \right)^2.
\]

(16)

We can write \( \hat{\beta} \) as a weighted sum.\(^{21}(\text{Problem 2.2-10})\) Traditional survey weighting defines the weights so that the weight for the \( i \)th sampled individual can be interpreted as the number of individuals in the population that the \( i \)th sampled individual represents. Following that tradition, let \( w_i^{\text{(rad)}} = 1/(nP_i) \) and \( W_j^{\text{(rad)}} = 1/(nP_j) \), then the expected sum of the sampled weights is \( N \),

\[
E \left( \sum_{i=1}^{n} w_i^{\text{(rad)}} \right) = E \left( \sum_{i=1}^{n} \sum_{j=1}^{N} I_{ij} W_j^{\text{(rad)}} \right) = \sum_{i=1}^{n} \sum_{j=1}^{N} \frac{E(I_{ij})}{nP_j} = \sum_{i=1}^{n} \sum_{j=1}^{N} \frac{1}{n} = N.
\]

Sometimes the weights are scaled after selection so that the scaled weights are \( w_i^{\text{(strad)}} = \frac{N w_i^{\text{(rad)}}}{\sum_{i=1}^{n} w_i^{\text{(rad)}}} \) and are forced to sum to \( N \). For example, in Kalish et al.\(^{11}\) rescaling (sometimes called post-stratification) was done in a more complicated manner to ensure that the weights summed to the US census population within age group, sex, race, ethnicity, and region.

For this article we define the weights differently, because we want to model our estimator as a weighted sum of Poisson random variables. Thus, we use \( w_i = 1/(nNP_i) \) and \( W_j = 1/(nNP_j) \) so that the sums have expectation 1. In the complex survey case, we start with the independent multinomial model as in Equation (15), then we use the relationship between the multinomial and Poisson distributions. Using the “multinomial-Poisson transformation,” the maximum likelihood estimates (MLE) for a multinnomial random variable are equivalent to the MLEs for independent Poisson random variables, and the variances are asymptotically equivalent.\(^{22}\) Even though we model \( \hat{\beta} \) using multinomial random variables, where there are many missing values (which occurs in our situation whenever \( I_{ij} = 0 \)), the multinomial-Poisson relationship holds even when there are missing variables.\(^{22}(\text{Section 3})\) For both the Poisson and multinomial models, \( E(I_{ij}) = P_j \), and
\( \hat{\beta} \) is unbiased under either model. For the Poisson model, all the \( I_{ij} \) are independent and each mean equals its variance, so that the variance of \( \hat{\beta} \) under this model is

\[
\text{var}_P (\hat{\beta}) = \text{var}_P \left( \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} I_{ij} Y_j \right) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{N} \text{var}_P (I_{ij}) Y_j^2}{n^2 N^2 P_j^2} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{N} Y_j}{n^2 N^2 P_j^2}.
\]

We estimate \( \text{var}_P (\hat{\beta}) \) by multiplying each term in the sum by \( I_{ij} / P_j \), which has an expectation of 1 and eliminates terms of non-selected individuals, giving

\[
\hat{\text{var}}_P (\hat{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{N} \frac{I_{ij} Y_j}{n^2 N^2 P_j^2} = \sum_{i=1}^{n} \frac{y_i}{n^2 N^2 P_i^2}.
\]

Under the Poisson model \( \hat{\text{var}}_P (\hat{\beta}) \) is an unbiased estimator of \( \text{var}_P (\hat{\beta}) \).

3 | SIMULATIONS

We explore simulations in a variety of scenarios; however, despite this variety, we limit the scenarios to cases similar to the application of Section 4, which has low prevalence (when the Poisson assumption is adequate), and high sensitivity and specificity (ie, good assays).

3.1 | Estimating prevalence from a simple random sample with an imperfect assay

We assess and compare our new method (melding, ie, Equation (7)) to that of Lang and Reiczigel (LR) in a variety of simulated settings. In each simulation, 100 subjects are tested to estimate prevalence, 60 are tested to estimate sensitivity, and 300 are tested to estimate specificity. Several combinations of prevalences (0.5%-2%), sensitivities (75%-100%) and specificities (75%-100%) are assessed. Each simulated scenario is replicated 10 000 times. Figure 1 compares the two methods based on coverage, while Figures 2 and 3 present the lower and upper error frequencies for these scenarios, respectively.

Figure 1 shows that, when specificity is less than perfect, the Lang-Reiczigel method achieves approximately nominal coverage, while the melding method is slightly more conservative, generally demonstrating 96%-97% coverage with 95% confidence intervals. When specificity is 100%, both methods are very conservative, achieving nearly 100% coverage. For a more nuanced depiction of each method’s properties, we separate the overall coverage into lower and upper errors. Figure 3 shows that both methods make upper errors with roughly the same frequency. Figure 2 demonstrates that while the melding procedure bounds the lower error frequency below 2.5%, the Lang-Reiczigel method generally has lower error frequency above 2.5%. Each of these behaviors may be undesirable, depending on the context in which the methods are applied. For applications in which there is a need to bound the lower errors, the melding method appears to be superior.

3.2 | Estimating prevalence from a weighted sample with a perfect assay

We compare the wsPoisson method to the more traditional Dean-Pagano modification of the Agresti-Coull (DPAC) method and the Korn-Graubard (KG) method for survey proportions in a variety of settings. Our simulations examine varying levels of disease prevalence (0.5% or 5%), different types survey designs (50 sampling strata with 200 subjects each or 8000 individuals, each with their own weight), distributions of weights among the sampling strata or individuals (coefficient of variation from approximately 0% to nearly 600%), and the number and weights of sampling strata with non-zero prevalence. For each combination of prevalence \( p \), and group type, up to 500 sets of weights are simulated. These 500 sets of weights are designed to span a range of coefficients of variation. For a target coefficient of variation, \( v \), \( n \) weights (\( w_i \))
**FIGURE 1** Coverage properties of 95% confidence intervals for our new method (melding) and the Lang-Reiczigel method in a variety of settings, each simulated 10,000 times.

**FIGURE 2** Lower error properties of 95% confidence intervals for our new method (melding) and the Lang-Reiczigel method in a variety of settings, each simulated 10,000 times.
FIGURE 3 Upper error properties of 95% confidence intervals for our new method (melding) and the Lang-Reiczigel method in a variety of settings, each simulated 10,000 times.

are simulated by generating \( n \) samples from a Beta \( \left( \frac{1}{n^2} - \frac{1}{m^2} - \frac{1}{n^2} + \frac{n-1}{m^2} + \frac{n-1}{n} \right) \) distribution and normalizing so that \( \sum_{i=1}^{n} w_i = 1 \). This assures that the coefficient of variation among these weights is approximately \( v \). Then, certain weights are chosen to have non-zero prevalence (5%, 25%, or 75% distributed either among the highest weights, lowest weights, or distributed uniformly). These weights with non-zero prevalence are given a prevalence such that \( \sum_{i=1}^{n} w_i \hat{\theta}_i = p \). For each simulated set of parameters and weights, 10,000 data sets are simulated and assessed.

The coverage properties for these simulations are presented in Figures 4–7. Additional properties for these simulations are presented in Figures B1–B12.

From Figures 4–7, we note that the two competitor methods generally exhibit lower coverage as the coefficient of variation among the weights increases. In Figure 4, this coverage falls below 60% when the prevalence is very low and is concentrated among the highest weights, and the coefficient of variation among the weights exceeds 4. Uniform distribution of prevalence among the weights, increased overall prevalence, and larger sample sizes among fewer groups all appear to lessen the severity of this problem. In contrast, the wsPoisson method appears to guarantee coverage in all scenarios. The wsPoisson method tends to become more conservative when the coefficient of variation among the weights increases, when the other methods can have problems guaranteeing coverage. In all cases, the wsPoisson method is more conservative than the competitor methods. In scenarios with higher overall prevalence, the Agresti-Coul and Korn-Graubard methods achieve close to nominal coverage, while the wsPoisson method strongly over-covers. This is similar to the behavior observed in Fay and Feuer,2 where, in simulations, the overall error rate for the gamma intervals decreased as the variance of the weights increased. Because our methods appear to be very conservative, with coverage near 100% in some cases, we present the widths of the confidence intervals in Figures B9–B12. In scenarios where coefficient of variation among the survey weights is high, the wsPoisson intervals are often two or three times wider than intervals produced by competing methods.

3.3 Estimating prevalence from a weighted sample with an imperfect assay

We compare properties of our melded confidence interval, WprevSeSp Poisson, to another melded confidence interval method, WprevSeSp Binomial, for cases with weighted samples and imperfect assays. The methods are assessed in a several simulated scenarios with varying levels of disease prevalence (0.5% or 5%), types of groups surveyed...
Coverage properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 50 groups of 200 people are sampled. The horizontal dashed line indicates the nominal coverage, 95%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.

50 groups of 200 subjects each or 8000 groups of 1 subject), distributions of weights among the groups (coefficient of variation from approximately 0% to nearly 600%), the number of groups with non-zero prevalence, and the specificity of the assay (80% - 100%). In each scenario, the assay has 95% sensitivity. Each scenario creates up to 500 new sets of weights and \( \theta_i \) parameters (as in Section 3.2), and each of those is simulated 10,000 times, with new prevalence, sensitivity, and specificity surveys generated and 95% confidence intervals are created. Modeled after the study of Kalish et al, the simulated sensitivity is assessed based on 60 tests, while specificity is based on 300 tests.
FIGURE 5 Coverage properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 8,000 individuals are sampled. The horizontal dashed line indicates the nominal coverage, 95%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.

The coverage properties for these simulations are presented in Figures 8–11. Additional properties for these simulations are presented in Figures B13–B24.

Based on Figures B13–B20, we note that, in most settings, the two melding procedures result in conservative confidence intervals, often nearing 100% coverage. With perfect specificity, the WprevSeSp Binomial method fails to maintain nominal coverage when the coefficient of variation among the weights is high and specificity is perfect. In the other scenarios tested, the WprevSeSp Binomial generally achieves closer to nominal coverage than the WprevSeSp Poisson, which strongly over-covers. Our proposed WprevSeSp Poisson method maintains or exceeds the desired coverage in all
Coverage Properties for Simulations with 5.0% Prevalence Among 50 Groups of 200

Each Point = 10,000 Replications

Weights with Non–0 Prev.
5%

Weights with Non–0 Prev.
25%

Weights with Non–0 Prev.
75%

FIGURE 6 Coverage properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 50 groups of 200 people are sampled. The horizontal dashed line indicates the nominal coverage, 95%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.

scenarios. In these scenarios, we also assess properties of the wsPoisson procedure that does not account for the imperfect assay. This method results in approximately 0% coverage in any scenarios where specificity is less than perfect. For this reason, results from that method are omitted in the figures. Because both WprevSeSp methods appear to be very conservative, with coverage near 100% in some cases, we present the widths of the confidence intervals in Figures B21–B24. The WprevSeSp Binomial and WprevSeSp Poisson methods typically produce wide intervals of approximately the same size - sometimes as wide as 12%, even when true prevalence is 0.05%. One notable exception to this is presented in Figure B23, which shows that for tests with perfect specificity, the WprevSeSP Binomial method produces much narrower confidence intervals than the other method.
FIGURE 7 Coverage properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of Agresti-Coull and Korn-Graubard. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 8000 individuals are sampled. The horizontal dashed line indicates the nominal coverage, 95%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.

4 | APPLICATION

We apply these two methods to a real data set from Kalish et al. This data set was collected to estimate seroprevalence of SARS-CoV-2 in undiagnosed adults in the United States between May and July 2020. The assay used in this data is estimated to have perfect sensitivity, based on 56 tests on individuals with confirmed SARS-CoV-2 and perfect specificity based on 300 tests on individuals confirmed to not have SARS-CoV-2. First we apply the methods to the full data set ($n = 8058$, weight coefficient of variation = 252%). The seroprevalence in Kalish et al was 4.6%
Coverage Properties for Simulations with 0.5% Prevalence Among 50 Groups of 200

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

FIGURE 8 Coverage properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 50 groups of 200 people are sampled. Each dataset also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal coverage, 95%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
**Coverage Properties for Simulations with 0.5% Prevalence Among 8,000 Groups of 1**

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

---

**FIGURE 9** Coverage properties for the confidence interval procedures, \( W_{\text{prevSeSp Binomial}} \) and \( W_{\text{prevSeSp Poisson}} \). Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 8000 individuals are sampled. Each dataset also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal coverage, 95%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Coverage Properties for Simulations with 5.0% Prevalence Among 50 Groups of 200

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

FIGURE 10 Coverage properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 50 groups of 200 people are sampled. Each datasets also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal coverage, 95%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Coverage Properties for Simulations with 5.0% Prevalence Among 8,000 Groups of 1
Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

FIGURE 11 Coverage properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 8000 individuals are sampled. Each datasets also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal coverage, 95%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
with (95% CI: 2.6% to 6.5%), using a confidence interval method that was nearly the same as the WprevSeSp Binomial method (the method Kalish et al included a calculation of the variability of the weights due to the estimation of the weights, whereas in this article we treat the weights as fixed constants). The Korn and Graubard type melded confidence interval with imperfect assay adjustments (WprevSeSp Binomial) studied in this article produced a 95% confidence interval for population prevalence of (2.53%, 6.68%), very similar to that of Kalish et al, while the wsPoisson type melded confidence interval with imperfect assay adjustments (WprevSeSp Poisson) produced the 95% confidence interval (2.56%, 7.54%). We also apply the wsPoisson method from Section 2.3, which does not account for imperfections in the assay, resulting in a 95% confidence interval of (3.04%, 7.39%). While all three intervals overlap to a large degree, the WprevSeSp Poisson interval is the widest. Our simulations show that in this situation, the WprevSeSp Binomial interval may be the best, because with coefficient of variance about 250% (see Figures B15 and B20, top right panel) the error on both sides of the confidence interval is bounded at 2.5% and the width of the intervals are better (Figure B23).

We also apply the methods to the subset of only Hispanic participants (n = 1281, weight coefficient of variation = 306%), where Kalish et al estimated the undiagnosed adult seroprevalence estimate as 6.1% (95% CI: 2.4% to 11.5%). The WprevSeSp Binomial method produces a 95% confidence interval for population prevalence (2.35%, 11.75%), while the WprevSeSp Poisson method produces a 95% confidence interval (2.40%, 20.02%). The wsPoisson method produces a 95% confidence interval of (2.80%, 19.63%). In this case, the two Poisson-based confidence intervals are much wider than the WprevSeSp Binomial interval, which is as expected since the melding method is designed to guarantee coverage, although the simulations show that the WprevSeSp Binomial interval may be reasonable (see eg, Figure B20). The smaller WprevSeSp Binomial interval is also unsurprising and is similar to the results observed in our simulation study.

5 DISCUSSION

We presented several methods for creating confidence intervals to assess disease prevalence in a variety of settings, including simple random samples with imperfect tests, weighted sampling with perfect tests, and weighted sampling with imperfect tests. A main point of this article was to develop and study methods for the last setting, where there has been little previous work. Two methods were studied for that setting, WprevSeSp Poisson and WprevSeSp Binomial, the latter of which was very similar to the method used by Kalish et al.11 In simulations based on the Kalish et al application, the WprevSeSp Poisson method had greater than nominal coverage in all cases but could be very conservative, while the WprevSeSp Binomial was slightly less conservative but had less than nominal coverage in a few scenarios.

In less complicated scenarios, we could compare the new methods to established methods. These new confidence intervals appear to guarantee coverage in most simulated settings, including some scenarios where competitor methods dramatically under-cover. In general, our methods demonstrate higher coverage than competitor methods in most scenarios, sometimes being very conservative while competitor methods demonstrate closer to nominal coverage. In the case of the simple random sample with an imperfect test, our new methods are able to bound the lower error rate for a 95% confidence interval at 2.5%, while the Lang-Reiczigel7 method maintains 95% coverage by allowing a higher lower error rate.

A big advantage of the new methods is that they may be applied with complex survey methods, where each individual has their own weight, such as in Kalish et al.11 However, in this article we have only studied fixed weights and not when the weights are estimated as in Kalish et al. Further, we have not included estimates of the variability of the weights as was done in Kalish et al; however, recalculating the confidence intervals on the same data shows that ignoring variability of the weights made little difference in that case. Further work is needed to address this issue. In addition, further work is needed to consider other complex sample designs such as multistage cluster designs that are used in household and institutional surveys such as hospital and medical practice surveys.

Although the new methods have coverage that is at least nominal in nearly all simulated situations, the cost is that they can be overly conservative. Future work could create less conservative methods by, for example, modifying the new methods by replacing the lower or upper confidence distributions with a mixture of the two (as in Veronese and Melilli23), or using a mid-p version of the gamma intervals (as in Fay and Kim24).

Although this article has done extensive simulations, the focus of the simulations was for low prevalence and high sensitivity and specificity. For other situations, especially high prevalence situations, more work is needed to more fully explore the properties of these methods.
Our methods’ conservative properties are especially advantageous in settings where the competitor methods exhibit much lower than nominal coverage. For example, when there is high variance among the sample weights and prevalence is concentrated among the highest-weighted samples, competitor coverage of 95% confidence intervals can fall to 60%, while our method exhibits >95% coverage. Thus, we suggest that our melding methods be employed when working with survey settings which involve high variance among the weights or lower errors are particularly undesirable.

ACKNOWLEDGEMENTS
For sharing the data from the Kalish et al study, we thank Matthew Memoli and the LID Clinical Studies Unit of the National Institute of Allergy and Infectious Diseases, NIH, Kaitlyn Sadler from the National Institute of Biomedical Imaging and Bioengineering, NIH, Matthew Hall from the National Center for Advancing Translational Sciences, NIH, and Dominic Esposito from the Fredrick National Laboratory for Cancer Research, NCI, NIH.

DATA AVAILABILITY STATEMENT
Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID
Damon M. Bayer https://orcid.org/0000-0003-2819-7645
Michael P. Fay https://orcid.org/0000-0002-8643-9625
Barry I. Graubard https://orcid.org/0000-0002-6787-1105

REFERENCES
1. Hemenway D. The myth of millions of annual self-defense gun uses: a case study of survey overestimates of rare events. Chance. 1997;10(3):6-10. doi:10.1080/09332480.1997.10542033
2. Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. Stat Med. 1997;16(7):791-801.
3. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. Am J Epidemiol. 1978;107(1):71-76.
4. Fay MP, Proschan MA, Brittain E. Combining one-sample confidence procedures for inference in the two-sample case. Biometrics. 2015;71(1):146-156.
5. Dean N, Pagano M. Evaluating confidence interval methods for binomial proportions in clustered surveys. J Surv Stat Methodol. 2015;3(4):484-503.
6. Franco C, Little RJ, Louis TA, Slud EV. Comparative study of confidence intervals for proportions in complex sample surveys. J Surv Stat Methodol. 2019;7(3):334-364.
7. Lang Z, Reiczigel J. Confidence limits for prevalence of disease adjusted for estimated sensitivity and specificity. Prev Vet Med. 2014;113(1):13-22. doi:10.1016/j.prevetmed.2013.09.015
8. DiCiccio TJ, Ritzwoller DM, Romano JP, Shaikh AM. Confidence intervals for seroprevalence. Stat Sci. 2022;37:306-321.
9. Cai B, Ioannidis JPA, Bendavid E, Tian L. Exact inference for disease prevalence based on a test with unknown specificity and sensitivity. J Appl Stat. 2022;1:25. doi:10.1080/02664763.2021.2019687
10. Berger RL, Boos DD. P values maximized over a confidence set for the nuisance parameter. J Am Stat Assoc. 1994;89(427):1012-1016. doi:10.1080/01621459.1994.10476836
11. Kalish H, Klumpp-Thomas C, Hunsberger S, et al. Undiagnosed SARS-CoV-2 seropositivity during the first 6 months of the COVID-19 pandemic in the United States. Sci Transl Med. 2021;13:eabh3826.
12. Rosin S, Shook-Sa BE, Cole SR, Hudgens MG. Estimating SARS-CoV-2 seroprevalence. arXiv preprint arXiv:2111.02910, 2021.
13. Mignone X, Singh K. Confidence distribution, the frequentist distribution estimator of a parameter: a review. Int Stat Rev. 2013;81(1):3-39. doi:10.1111/insr.12000
APPENDIX A. MONOTONICITY OF $g$

It is clear that $g$ is monotonic within each of its piecewise-defined functions. In the following sections we consider if monotonicity holds at the change points between piecewise functions.

A.1 Monotonicity in $\hat{\phi}_p$
- Case 1: $\hat{\phi}_n < \hat{\theta}_1$. Not monotonic.

\[
g(\hat{\theta}_1, \hat{\phi}_n, \hat{\phi}_p) = \begin{cases} 
0 & \hat{\phi}_p < \hat{\phi}_n < \hat{\theta}_1 \\
0 & \hat{\phi}_n = \hat{\phi}_p < \hat{\theta}_1 \\
1 & \hat{\phi}_n < \hat{\phi}_p < \hat{\theta}_1 \\
\frac{\hat{\theta}_1 - \hat{\phi}_n}{\hat{\phi}_p - \hat{\phi}_n} & \hat{\phi}_n < \hat{\theta}_1 < \hat{\phi}_p \\
\frac{\hat{\theta}_1 - \hat{\phi}_n}{\hat{\phi}_p - \hat{\phi}_n} & \hat{\phi}_n < \hat{\theta}_1 < \hat{\phi}_p.
\end{cases} 
\]

(A1)

- Case 2: $\hat{\theta}_1 < \hat{\phi}_n$. Monotonic.

\[
g(\hat{\theta}_1, \hat{\phi}_n, \hat{\phi}_p) = \begin{cases} 
0 & \hat{\phi}_p < \hat{\theta}_1 < \hat{\phi}_n \\
0 & \hat{\theta}_1 = \hat{\phi}_p < \hat{\phi}_n \\
0 & \hat{\phi}_p < \hat{\theta}_1 < \hat{\phi}_n \\
0 & \hat{\theta}_1 < \hat{\phi}_n < \hat{\phi}_p.
\end{cases} 
\]

(A2)

- Case 3: $\hat{\theta}_1 = \hat{\phi}_n$. Monotonic.

\[
g(\hat{\theta}_1, \hat{\phi}_n, \hat{\phi}_p) = \begin{cases} 
0 & \hat{\phi}_p < \hat{\theta}_1 = \hat{\phi}_n \\
0 & \hat{\theta}_1 = \hat{\phi}_p = \hat{\phi}_n \\
0 & \hat{\phi}_p < \hat{\theta}_1 < \hat{\phi}_n \\
\frac{\hat{\theta}_1 - \hat{\phi}_n}{\hat{\phi}_p - \hat{\phi}_n} & \hat{\phi}_n < \hat{\theta}_1 < \hat{\phi}_p.
\end{cases} 
\]

(A3)

A.2 Monotonicity in $\hat{\theta}_1$
- Case 1: $\hat{\phi}_n < \hat{\phi}_p$. Monotonic.

\[
g(\hat{\theta}_1, \hat{\phi}_n, \hat{\phi}_p) = \begin{cases} 
0 & \hat{\theta}_1 < \hat{\phi}_n < \hat{\phi}_p \\
\frac{\hat{\theta}_1 - \hat{\phi}_n}{\hat{\phi}_p - \hat{\phi}_n} = 0 & \hat{\phi}_n = \hat{\phi}_p < \hat{\theta}_1 \\
\frac{\hat{\theta}_1 - \hat{\phi}_n}{\hat{\phi}_p - \hat{\phi}_n} & \hat{\phi}_n < \hat{\theta}_1 < \hat{\phi}_p \\
0 & \hat{\phi}_n < \hat{\phi}_p < \hat{\theta}_1.
\end{cases} 
\]

(A4)
• Case 2: $\hat{\phi}_p < \hat{\phi}_n$. Monotonic.

$$
g(\hat{\theta}_1, \hat{\phi}_n, \hat{\phi}_p) = 
\begin{cases} 
0 & \hat{\theta}_1 < \hat{\phi}_p < \hat{\phi}_n \\
0 & \hat{\phi}_p = \hat{\theta}_1 < \hat{\phi}_n \\
0 & \hat{\phi}_p < \hat{\theta}_1 < \hat{\phi}_n \\
0 & \hat{\phi}_p < \hat{\phi}_n < \hat{\theta}_1.
\end{cases} \quad \text{(A5)}
$$

• Case 3: $\hat{\phi}_p = \hat{\phi}_n$. Monotonic.

$$
g(\hat{\theta}_1, \hat{\phi}_n, \hat{\phi}_p) = 
\begin{cases} 
\hat{\theta}_1 < \hat{\phi}_p = \hat{\phi}_n \\
0 & \hat{\phi}_p = \hat{\phi}_n < \hat{\theta}_1
\end{cases} \quad \text{(A6)}
$$

A.3 Monotonicity in $\hat{\phi}_n$

• Case 1: $\hat{\phi}_p < \hat{\theta}_1$. Monotonic.

$$
g(\hat{\theta}_1, \hat{\phi}_n, \hat{\phi}_p) = 
\begin{cases} 
1 & \hat{\phi}_n < \hat{\phi}_p < \hat{\theta}_1 \\
0 & \hat{\phi}_p = \hat{\phi}_n < \hat{\theta}_1 \\
0 & \hat{\phi}_p < \hat{\phi}_n < \hat{\theta}_1 \\
0 & \hat{\phi}_p < \hat{\phi}_n = \hat{\theta}_1 \\
0 & \hat{\phi}_p < \hat{\theta}_1 < \hat{\phi}_n.
\end{cases} \quad \text{(A7)}
$$

• Case 2: $\hat{\theta}_1 < \hat{\phi}_p$. Monotonic.

$$
g(\hat{\theta}_1, \hat{\phi}_n, \hat{\phi}_p) = 
\begin{cases} 
\hat{\phi}_n < \hat{\phi}_p < \hat{\theta}_1 \\
\hat{\phi}_p = \hat{\phi}_n < \hat{\theta}_1 \\
\hat{\phi}_p < \hat{\phi}_n < \hat{\theta}_1 \\
\hat{\phi}_p < \hat{\phi}_n = \hat{\theta}_1 \\
\hat{\phi}_p < \hat{\theta}_1 < \hat{\phi}_n.
\end{cases} \quad \text{(A8)}
$$

• Case 3: $\hat{\theta}_1 = \hat{\phi}_p$. Monotonic.

$$
g(\hat{\theta}_1, \hat{\phi}_n, \hat{\phi}_p) = 
\begin{cases} 
\hat{\phi}_n < \hat{\theta}_1 < \hat{\phi}_p \\
\hat{\phi}_n = \hat{\theta}_1 < \hat{\phi}_p \\
\hat{\phi}_n < \hat{\phi}_p < \hat{\theta}_1
\end{cases} \quad \text{(A9)}
$$
APPENDIX B. ADDITIONAL FIGURES

Lower Error Frequency Properties for Simulations with 0.5% Prevalence Among 50 Groups of 200
Each Point = 10,000 Replications

Weights with Non-0 Prev. 5%
Weights with Non-0 Prev. 25%
Weights with Non-0 Prev. 75%

Weights with Prev. High Weights
Weights with Prev. Uniform Weights
Weights with Prev. Low Weights

Lower Error Frequency

0.0% 1.0% 2.0% 3.0%
0% 200% 400% 600%

0.0% 1.0% 2.0% 3.0%
0% 200% 400% 600%

0.0% 1.0% 2.0% 3.0%
0% 200% 400% 600%

Method
- Agresti–Coull
- Korn–Graubard
- wsPoisson

FIGURE B1 Lower error properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 50 groups of 200 people are sampled. The horizontal dashed line indicates the nominal lower error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Lower Error Frequency Properties for Simulations with 0.5% Prevalence Among 8,000 Groups of 1

Each Point = 10,000 Replications

FIGURE B2  Lower error properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 8000 individuals are sampled. The horizontal dashed line indicates the nominal lower error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
**Lower Error Frequency Properties for Simulations with 5.0% Prevalence Among 50 Groups of 200**

Each Point = 10,000 Replications

| Weights with Non-0 Prev. | Weights with Non-0 Prev. | Weights with Non-0 Prev. |
|--------------------------|--------------------------|--------------------------|
| 5%                       | 25%                      | 75%                      |

| Weight Coefficient of Variation | Lower Error Frequency |
|---------------------------------|-----------------------|
| 0%                              | 2%                    |
| 200%                            | 0%                    |
| 400%                            | 1%                    |
| 600%                            | 2%                    |
| 800%                            | 3%                    |
| 100%                            | 4%                    |

Method: Agresti-Coull, Korn-Graubard, wsPoisson

**FIGURE B3** Lower error properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 50 groups of 200 people are sampled. The horizontal dashed line indicates the nominal lower error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
**FIGURE B4** Lower error properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 8000 individuals are sampled. The horizontal dashed line indicates the nominal lower error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Upper Error Frequency Properties for Simulations with 0.5% Prevalence Among 50 Groups of 200

Each Point = 10,000 Replications

Method
- Agresti–Coull
- Korn–Graubard
- wsPoisson

FIGURE B5 Upper error properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 50 groups of 200 people are sampled. The horizontal dashed line indicates the nominal upper error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Upper Error Frequency Properties for Simulations with 0.5% Prevalence Among 8,000 Groups of 1

Each Point = 10,000 Replications

Weights with Non-0 Prev.
5%

Weights with Non-0 Prev.
25%

Weights with Non-0 Prev.
75%

Weights with Prev.
High Weights

Weights with Prev.
Uniform Weights

Weights with Prev.
Low Weights

FIGURE B6  Upper error properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10 000 simulations of datasets from a population with 0.5% prevalence where 8000 individuals are sampled. The horizontal dashed line indicates the nominal upper error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Upper Error Frequency Properties for Simulations with 5.0% Prevalence Among 50 Groups of 200
Each Point = 10,000 Replications

FIGURE B7  Upper error properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 50 groups of 200 people are sampled. The horizontal dashed line indicates the nominal upper error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines
**Upper Error Frequency Properties for Simulations with 5.0% Prevalence Among 8,000 Groups of 1**

Each Point = 10,000 Replications

**FIGURE B8** Upper error properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 8000 individuals are sampled. The horizontal dashed line indicates the nominal upper error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Confidence Interval Width Properties for Simulations with 0.5% Prevalence Among 50 Groups of 200
Each Point = 10,000 Replications

FIGURE B9  Confidence interval width properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 50 groups of 200 people are sampled. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Confidence Interval Width Properties for Simulations with 5.0% Prevalence Among 50 Groups of 200

Each Point = 10,000 Replications

F I G U R E  B10  Confidence interval width properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 50 groups of 200 people are sampled. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
FIGURE B11  Confidence interval width properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 8000 individuals are sampled. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
 FIGURE B12  Confidence interval width properties for the wsPoisson model and two standard methods, the Dean-Pagano modification of the Agresti-Coull method and of the Korn-Graubard method. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 8000 individuals are sampled. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Lower Error Frequency Properties for Simulations with 5.0% Prevalence Among 50 Groups of 200

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

FIGURE B13  Lower error properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 50 groups of 200 people are sampled. Each dataset also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal lower error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines. If the WprevSeSp Binomial line is not visible, then it is covered by the WprevSeSp Poisson line.
Lower Error Frequency Properties for Simulations with 0.5% Prevalence Among 50 Groups of 200

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

**FIGURE B14** Lower error properties for the confidence interval procedures, \( W_{\text{prevSeSp Binomial}} \) and \( W_{\text{prevSeSp Poisson}} \). Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 8000 individuals are sampled. Each datasets also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal lower error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
**FIGURE B15** Lower error properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 50 groups of 200 people are sampled. Each dataset also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal lower error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Lower Error Frequency Properties for Simulations with 0.5% Prevalence Among 8,000 Groups of 1

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

**FIGURE B16**  Lower error properties for the confidence interval procedures, $W_{\text{prevSeSp Binomial}}$ and $W_{\text{prevSeSp Poisson}}$. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 8,000 individuals are sampled. Each dataset also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal lower error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Upper Error Frequency Properties for Simulations with 0.5% Prevalence Among 50 Groups of 200
Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

FIGURE B17  Upper error properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 50 groups of 200 people are sampled. Each dataset also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal upper error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
**Upper Error Frequency Properties for Simulations with 0.5% Prevalence Among 8,000 Groups of 1**

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

---

**FIGURE B18**  Upper error properties for the confidence interval procedures, \(W_{prevSeSp\text{Binomial}}\) and \(W_{prevSeSp\text{Poisson}}\). Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 8000 individuals are sampled. Each datasets also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal upper error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Upper Error Frequency Properties for Simulations with 5.0% Prevalence Among 50 Groups of 200

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

FIGURE B19 Upper error properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 50 groups of 200 people are sampled. Each datasets also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal upper error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
FIGURE B20 Upper error properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 8000 individuals are sampled. Each dataset also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. The horizontal dashed line indicates the nominal upper error rate, 2.5%. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Confidence Interval Width Properties for Simulations with 0.5% Prevalence Among 50 Groups of 200
Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

FIGURE B21  Confidence interval width properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 50 groups of 200 people are sampled. Each datasets also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. Colored dashed lines are estimates from a logistic regression model using quadratic splines
Confidence Interval Width Properties for Simulations with 5.0% Prevalence Among 50 Groups of 200

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

FIGURE B22  Confidence interval width properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 50 groups of 200 people are sampled. Each datasets also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
**Confidence Interval Width Properties for Simulations with 0.5% Prevalence Among 8,000 Groups of 1**

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

*FIGURE B23* Confidence interval width properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 0.5% prevalence where 8000 individuals are sampled. Each dataset also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. Colored dashed lines are estimates from a logistic regression model using quadratic splines.
Confidence Interval Width Properties for Simulations with 5.0% Prevalence Among 8,000 Groups of 1

Each Facet = 95% Sensitivity, Each Point = 10,000 Replications

FIGURE B24  Confidence interval width properties for the confidence interval procedures, WprevSeSp Binomial and WprevSeSp Poisson. Each point represents 10,000 simulations of datasets from a population with 5% prevalence where 8000 individuals are sampled. Each dataset also includes simulated results of tests to evaluate the sensitivity and specificity of the assay performed on 60 and 300 individuals, respectively. Colored dashed lines are estimates from a logistic regression model using quadratic splines.