HHS Public Access

Author manuscript

J Agric Biol Environ Stat. Author manuscript; available in PMC 2024 April 12.

Published in final edited form as:

J Agric Biol Environ Stat. 2021 ; 26: 90–104. doi:10.1007/s13253-020-00411-5.

Bias Correction in Estimating Proportions by Imperfect Pooled Testing

Graham Hepworth,

School of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC 3010, Australia.

Brad J. Biggerstaff

Centers for Disease Control and Prevention, Fort Collins, CO, USA.

Abstract

In the estimation of proportions by pooled testing, the MLE is biased. Hepworth and Biggerstaff (JABES, 22:602–614, 2017) proposed an estimator based on the bias correction method of Firth (Biometrika 80:27–38, 1993) and showed that it is almost unbiased across a range of pooled testing problems involving no misclassification. We now extend their work to allow for imperfect testing. We derive the estimator, provide a Newton–Raphson iterative formula for its computation and test it in situations involving equal or unequal pool sizes, drawing on problems encountered in plant disease assessment and prevalence estimation of mosquito-borne viruses. Our estimator is highly effective at reducing the bias for prevalences consistent with the pooled testing procedure employed.

Keywords

Diagnostic testing; Firth's correction; Group testing; Sensitivity; Specificity

1. INTRODUCTION

Estimation of proportions can sometimes be greatly facilitated by pooled testing, in which individuals are pooled together and tested as a group for the presence of an attribute, usually a marker of disease or infection. Pooled testing (or group testing) is being applied in an increasing number of fields, including HIV prevalence estimation (Zhang et al. 2014), drug discovery (Hughes-Oliver 2006) and chronic illnesses in farm animals (Dhand et al. 2007). Two important areas of application, corresponding to the authors' involvement, are assessment of plant disease levels (e.g., Liu et al. 2011), and estimation of virus prevalence in mosquito vectors (e.g., Komar et al. 2015). Plant disease assessment often arises from sampling a field crop or glasshouse, and pools sizes are generally not large; mosquitoes usually pool more haphazardly in traps, and pool sizes can be very large.

hepworth@unimelb.edu.au .

Disclaimer: The findings and conclusions herein are those of the authors and do not necessarily represent the official position of the US Centers for Disease Control and Prevention.

For pooled testing with a fixed sample size, the maximum likelihood estimator (MLE) of a proportion p is positively biased, except in the trivial case of pools of size 1. Swallow (1985) illustrated the extent of the problem, with tables of bias for a range of pool sizes and numbers of pools. Various approaches have been proposed to address this bias. Some have focused on study design, such as Hepworth and Watson (2009), who proposed sequential testing with pools of decreasing size. They also suggested a numerical correction based on calculating the expectation of the estimator at p equal to the MLE ($\hat{p}$) for each outcome, and adjusting accordingly, with possible iteration of this process. Other studies have searched for an analytical solution, the most successful being Burrows (1987), who derived a bias correction to the MLE for equal pool sizes. His correction, which effectively added approximately half an additional negative pool, resulted in a "shrinkage" estimator with less than 5% of the bias of the MLE. Colon et al. (2001) compared Burrows' estimator to a range of other estimators and found it to have the least bias overall. Hepworth and Watson (2009) investigated various ways of extending Burrows' correction to unequal pool sizes, but they were essentially arbitrary, and none proved to be entirely satisfactory.

Gart (1991) described a general bias correction method for adjusting the MLE. Applying this method to pooled testing with unequal pool sizes mostly works well (Hepworth and Watson 2009). It also removes most of the bias when applied to pooled testing following inverse sampling (Hepworth 2013). One disadvantage of Gart's method is that it cannot be used on the upper boundary of the parameter space ($\hat{p} = 1$). It also tends to over-correct as p increases.

Hepworth and Biggerstaff (2017) showed that the general bias correction method introduced by Firth (1993), and subsequently applied to many statistical problems, is equivalent to Burrows' method for equal pool sizes. Firth's method is therefore an obvious way to extend Burrows' correction to unequal pool sizes. Hepworth and Biggerstaff (2017) examined the resulting bias-corrected estimator and found it to be almost unbiased across a range of pooled testing problems, and less biased overall than the estimator arising from Gart's correction. Rather than finding the MLE and then correcting it using the bias $b(p)$ (as Gart's method does), Firth's method is based on a modification to the score function $S(p)$ and requires the solution to

$$S(p) - I(p)\, b(p) = 0 \tag{1}$$

where $I(p)$ is the expected information function. Rather than being corrective, this method is preventative, which has the advantage of avoiding undefined parameter estimates.

All of the work described so far has assumed perfect testing (no misclassification of positives or negatives). This assumption is appropriate in some situations, but not others. For example, Burkhalter et al. (2014) reported a sensitivity of 97.6% in testing individual mosquitoes for West Nile virus, and a range of values for different pool sizes. The earliest study to quantify the effect of misclassification on pooled testing was that of Tu et al. (1994), who derived the MLE and its asymptotic variance for equal pool sizes. Others have

since derived exact confidence intervals (Reiczigel et al. 2010) and sample size formulae (e.g., Messam et al. 2008). Imperfect testing has been shown not to diminish the usefulness of pooled testing; in fact, Liu et al. (2012) showed that with moderate pool sizes it provides more efficient estimation than fully observed data over a wide range of disease prevalences. However, the testing errors need to be accounted for in estimation of p, including correction for bias.

This paper extends Firth's bias correction to imperfect testing, thus creating a new estimator for pooled testing with misclassification. We firstly derive the expressions needed for the estimator, using the log-likelihood and related quantities and their derivatives. To facilitate computation, we derive a Newton–Raphson iterative formula for calculating the estimator. We then compare the new estimator with the corresponding estimator derived using Gart's method, using pooled testing examples from plant virus testing. Following a similar approach to Hepworth and Biggerstaff (2017), the new estimator is then evaluated for a variety of pooled testing problems, chosen to reflect real situations in either plant disease assessment or mosquito-borne viruses. We show that Firth's method eliminates most of the bias, especially at smaller prevalences.

2. FIRTH'S BIAS CORRECTION APPLIED TO IMPERFECT POOLED TESTING

Suppose that for $i = 1, \ldots, d$, n_i pools of size m_i are tested, of which $X_i = x_i$ pools are positive. The total number of individuals is $N = \sum_{i=1}^{d} n_i m_i$. Let a and b denote the assumed known sensitivity and specificity of the test, respectively, and to ease presentation, assume no loss of either due to pooling; we note that this assumption may be relaxed, as our computations are easily seen to extend to the more general case by subscripting a and b with i. Assuming that the individuals in the pools follow i.i.d. Bernoulli distributions with parameter p, the binomial parameter for the distribution of X_i is

$$\pi_i(p) = a\left[1 - (1-p)^{m_i}\right] + (1-b)(1-p)^{m_i} = a - r(1-p)^{m_i}$$

where $r = a + b - 1$. The quantity r was suggested by Youden (1950) as a measure of overall performance of the test, and by others as a level of "informedness". Here we use r principally to simplify the complexity of mathematical expressions. The log-likelihood is

$$l(p; \boldsymbol{x}) = \sum_{i=1}^{d} \left[\log \binom{n_i}{x_i} + x_i \log \pi_i(p) + (n_i - x_i) \log \left[1 - \pi_i(p)\right] \right]$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)$. For ease of presentation, we temporarily drop the indexing subscript i and the functional notation on $\pi(p)$. Using derivatives of $\pi(p)$ and of the log-likelihood, it can be shown (see "Appendix 1") that the score function is

$$S(p) = \frac{m(a-\pi)(x-n\pi)}{(1-p)\pi(1-\pi)}.$$

In the case that all pools are of equal size m, with x of the n pools positive, the solution requires only that the right bracket of the numerator be equated to 0, in which case the MLE is

$$\hat{p} = 1 - \left(\frac{a - x/n}{r}\right)^{1/m} \tag{2}$$

provided the proportion of positive pools x/n is between $1 - b$ and a (Cowling et al. 1999). For $a = b = 1$ (perfect testing), this simplifies to the well-known formula

$$\hat{p} = 1 - (1 - x/n)^{1/m}.$$

Using the higher-order derivatives of the likelihood (see "Appendix 1"), we find that the information is

$$I(p) = \frac{nm^2(a-\pi)^2}{(1-p)^2\pi(1-\pi)}.$$

The remaining quantity needed to implement Firth's correction is the bias, which is approximately

$$b(p) = -\frac{2\frac{\mathrm{d}I(p)}{\mathrm{d}p} + E\left[\frac{\mathrm{d}^3 l}{\mathrm{d}p^3}\right]}{2[I(p)]^2} \tag{3}$$

(Gart 1991). Each term in (3) was derived for pooled testing by Hepworth (2005) for perfect testing and employed by Hepworth and Biggerstaff (2017) to implement the Firth correction. It can be shown (see "Appendix 1") that the numerator of the bias simplifies to

$$2\frac{\mathrm{d}I(p)}{\mathrm{d}p} + E\left[\frac{\mathrm{d}^3 l}{\mathrm{d}p^3}\right] = -\frac{nm^2(m-1)(a-\pi)^2}{(1-p)^3\pi(1-\pi)}. \tag{4}$$

Recalling that $\pi = a - r(1-p)^m$, and putting together the score (A.1), the information (A.2), and the numerator of the bias (A.4) and adding back indices, we have a relatively simplified expression for the Firth adjusted score:

$$
\begin{aligned}
S^*(p) &= S(p) - I(p)b(p) \\
&= \frac{1}{1-p}\left\{\sum_{i=1}^{d}\frac{m_i(a-\pi_i)(x_i-n_i\pi_i)}{\pi_i(1-\pi_i)} - \frac{1}{2}\frac{\sum_{i=1}^{d}\frac{n_i m_i^2(m_i-1)(a-\pi_i)^2}{\pi_i(1-\pi_i)}}{\sum_{i=1}^{d}\frac{n_i m_i^2(a-\pi_i)^2}{\pi_i(1-\pi_i)}}\right\}. \\
&= \frac{1}{1-p}\sum_{i=1}^{d}\left\{\frac{m_i(a-\pi_i)(x_i-n_i\pi_i)}{\pi_i(1-\pi_i)} - \frac{1}{2}w_i(m_i-1)\right\},
\end{aligned}
\tag{5}
$$

with

$$
v_i = \frac{n_i m_i^2(a-\pi_i)^2}{(1-p)^2\pi_i(1-\pi_i)}
$$

as the individual-pool contribution to the information, and $w_i = v_i \,/\, \sum_i v_i$.

The common pool size Firth adjusted score is thus

$$
S^*(p) = S(p) - \frac{1}{2}\frac{m-1}{1-p} = \frac{1}{1-p}\left[\frac{m(a-\pi)(x-n\pi)}{\pi(1-\pi)} - \frac{1}{2}(m-1)\right].
$$

Setting this expression equal to 0 gives a quadratic equation in π. Writing $\bar{x} = x \,/\, n$, this equation is

$$
[2mn + (m-1)]\,\pi^2 - [2mn\,(\bar{x}+a) + m - 1]\,\pi + 2mna\bar{x} = 0\,.
$$

The solution $\breve{\pi}$ for π in this equation follows directly as the smaller root in the quadratic formula, noting that erroneous solutions should lead to re-evaluation of the specifications for a and b. Then the subsequent Firth corrected estimate for p is $\breve{p} = 1 - [(a-\breve{\pi}) \,/\, r]^{1/m}$.

In computing a solution for p to $S^*(p) = 0$ in the general expression in Eq. (5), the leading factor has no role, so let $\widetilde{S}^*(p)$ be the function $S^*(p)$ without this. A convenient, if iterative, computational approach is to use the Newton–Raphson method. To implement this, we need the derivative $\widetilde{S}^{*'}(p)$, which after some straightforward computation utilizing the chain rule, we have as

$$
\widetilde{S}^{*'}(p) = \sum_{i=1}^{d}\left\{m_i^2\frac{[x_i - n_i(1-a)]\pi_i^3 + a[n_i(1-a) - 3x_i]\pi_i^2 + a(2a+1)x_i\pi_i - a^2x_i}{(1-p)\pi_i^2(1-\pi_i)^2} - \frac{1}{2}w_i'(m_i-1)\right\}
$$

where $v_i' = v_i'(p) = \mathrm{d}I(p) \,/\, \mathrm{d}\,p$ from (A.3) with subscript i, and

$$w_i^{'} = w_i^{'}(p) = \frac{v_i^{'}\sum_j v_j - v_i\sum_j v_j^{'}}{[\sum_j v_j]^2}.$$

Finally, the Newton–Raphson recursion is

$$p_{k+1} = p_k - \frac{\tilde{S}^*(p_k)}{\tilde{S}^{*'}(p_k)}$$

with starting value at $k = 0$ pre-determined, using, for example, the Firth estimate assuming a perfect test. R code (www.r-project.org) for performing the Newton–Raphson iterations is in "Appendix 2".

3. SMALL POOLED TESTING EXAMPLES

We illustrate Firth's bias correction method using two small pooled testing examples. The first example, with equal pool sizes, is described in Hepworth and Watson (2009) and arose in the context of testing carnations for viruses. A collection of 200 carnation plants were tested in 8 pools of 25, for which we adopt the notation $N : m^n = 200 : 25^8$. Consider the outcome $x = 6$ positive pools. Assuming perfect testing, the MLE is 0.0539, and the Firth bias-corrected estimate, which we denote $\breve{p}$, is 0.0480. With sensitivity $a = 0.95$ and specificity $b = 0.99$, applying (2) results in $\hat{p} = 0.0600$. Assuming $a < 1$ has the effect of increasing the estimate, and assuming $b < 1$ has the effect of decreasing it; the net effect here is an increase. Applying the Firth correction to this results in $\breve{p} = 0.0515$. If the sensitivity is assumed much lower at $a = 0.8$ (keeping b at 0.99), the effect is stronger, with $\hat{p} = 0.1045$, but the effect of the Firth correction is also substantial, with $\breve{p} = 0.0658$. If the specificity is assumed much lower at $b = 0.8$ (keeping a at 0.95), the effect is not as great, with $\hat{p} = 0.0515$ and $\breve{p} = 0.0429$.

The second example is described by Hepworth and Biggerstaff (2017) and also arose in the assessment of virus levels in plant populations, with two pool sizes. There were 8 pools of 20 and 8 pools of 5, which in the above notation is written $200 : 20^8\ 5^8$. Before considering the bias of the estimators, it is useful to examine the estimates themselves for a range of outcomes; these are presented in Table 1, with the outcomes selected to give a range of values for the estimates. The MLE and the bias-corrected estimate are shown for perfect testing, and for imperfect testing with $a = 0.95$, $b = 0.99$ for illustration. We use these values of a and b throughout this paper, to reflect the fact that, in practice, loss of sensitivity is usually greater than loss of specificity.

The estimates behave in a predictable way, though for outcomes (5, 7) and (7, 8), the assumption of imperfect testing ($a < 1$ in particular) results in a very large upward adjustment, whether or not the Firth bias correction is applied. These outcomes are of small probability and are somewhat incompatible with the statistical model, because there is a larger proportion of positive pools of size 5 than of size 20. Such incompatibilities are

not necessarily problematical in practice, as the assumed sensitivity or specificity can be re-evaluated and revised as appropriate before final prevalence estimates are made.

For evaluation of the bias of $\check{p}$, rather than examining the entire parameter space [$p \in (0, 1)$], it is more useful to consider values of p consistent with the design of the testing procedure. We use $p \leq \psi$, where ψ is the value of p at which the probability of all positive pools is 0.05 under perfect testing, an approach suggested by Hepworth and Watson (2009). All positive pools ($x_i = n_i$, $i = 1, \ldots, d$) are an uninformative outcome (even though the Firth method admits a solution in this case), and any pooled testing procedure should aim for its probability to be small. For this example, $\psi = 0.211$.

Table 2 shows the expected value of the estimators corrected by either Gart's or Firth's method for $a = 0.95$ and $b = 0.99$, together with the percentage bias and root mean squared error (RMSE), for selected values of p. The corresponding figures for the MLE are also shown. Both corrected estimators are almost unbiased for small p. The bias is still very small for p up to around 0.15 for Gart's method, and for p up to around 0.20 for Firth's method. Above that, the bias for both methods becomes more negative, especially for the Gart estimator, though the larger values of p in Table 2 are unlikely to be encountered in typical applications. The RMSE is very similar for the two estimators, and it essentially represents the standard deviation, due to the very small bias. Both estimators have smaller RMSE than the MLE, especially for larger p.

The negative bias for larger p is greater than that observed for the corresponding estimators for perfect testing [see Table 2 of Hepworth and Biggerstaff (2017)]. For example, at $p = 0.20$ the percentage bias values of −4.94 and −1.25 for Gart and Firth, respectively, are larger than the corresponding values of −0.85 and −0.30 for perfect testing. The RMSE values here are slightly larger for Firth's correction, and about the same for Gart's correction, compared to those for perfect testing.

4. LARGE POOLED TESTING EXAMPLES

We now focus entirely on Firth's bias correction, as we consider a range of larger examples, with $N = 500$, 1000 or 5000, and between 1 and 4 different pool sizes. These are the same examples used by Hepworth and Biggerstaff (2017) for evaluation of estimators with perfect testing and are reflective of procedures used in some applications by the researchers at the US Centers for Disease Control and Prevention in assessing virus prevalence in mosquitoes.

For these larger examples, it was sometimes the case that not all outcomes admitted a solution to $S^*(p) = 0$ for all values of p considered, and occasionally there was more than one solution. This occurred mostly for very small $\mathbf{x}$, due to the specificity being less than 1. This is essentially a multidimensional extension of the restriction $x / n > 1 - b$ for equal group sizes (we comment on this in Discussion). The problematical outcomes were generally of very small probability and, therefore, had a minimal effect on bias calculations. To address the computational issues, outcomes with probability less than 10^{-5} were excluded from the bias calculations, and the probabilities of other outcomes scaled to sum to 1. Remaining

outcomes which failed to converge to a solution were given an estimate of $\check{p} = \psi / 1000$, to be consistent with the small values of **x**.

Table 3 shows the mean absolute percentage bias and RMSE, calculated over 100 equally spaced points in the interval $[\psi / 100, \psi]$, for $a = 0.95$ and $b = 0.99$. Also shown is the bias at $p = \psi$, where its maximum absolute value consistently occurs across the range $[0, \psi]$.

Figure 1 plots the bias of the estimator for six of the procedures listed in Table 3, selected to show a range of $N : m^n$ and the resulting bias patterns. These are the same six procedures for which the bias was plotted for perfect testing by Hepworth and Biggerstaff (2017). One of the plots arises from equal pool sizes, two of them from 2 pool sizes, two from 3 pool sizes and one from 4.

These results confirm that Firth's correction method is effective in keeping the bias small. The mean percentage (absolute) bias is well under 1% for most pooled testing procedures; the only exceptions are those with a very small number of pools (10 or fewer), for which the worst bias (at $p = \psi$) is still very small. These results are not as good as for perfect testing, where the corresponding values were less than 0.6% for all of the procedures shown in Table 3 (Hepworth and Biggerstaff 2017). However, they would likely be satisfactory for most practitioners looking to make a bias correction. The plots show that the bias is virtually zero for p less than about $2\psi / 3$, and it then becomes negative at an increasing rate.

The average RMSE is generally larger than that observed for perfect testing (see Table 3, Hepworth and Biggerstaff 2017), though not markedly so—the largest ratio is 0.0213/0.0135 = 1.58, for $5000 : 5^{1000}$. The average RMSE is generally larger for small pool sizes. However, it is still much smaller than the corresponding RMSE for the MLE.

5. DISCUSSION

We have considered bias correction for the estimation of proportions by pooled testing, in situations where the testing is assumed to be imperfect, with known sensitivity and specificity. We have proposed a new estimator based on the general bias correction method of Firth (1993), following its successful application to pooled testing with no misclassification (Hepworth and Biggerstaff 2017). The new estimator is very effective at reducing the bias of the MLE, particularly for small prevalence, and has the advantage of being preventative rather than corrective.

Our estimator can be applied to problems with any number of pool sizes. The terms in the bias correction formulae are tedious to derive, but they are all based on the log-likelihood, and so can be summed across the different pool sizes. To facilitate computation of estimates, we have derived an easily implemented Newton–Raphson iterative formula. For some pooled testing procedures, it is possible that not all outcomes admit a solution, due to the complicated boundaries imposed by the imperfect testing assumption for multiple pool sizes. In our performance evaluations, we addressed this by excluding highly improbable outcomes, thus producing a very close approximation to the bias. This complexity is unlikely to hinder the use of the estimator in practice, as the specifications for sensitivity

and specificity can be re-evaluated in light of an outcome that appears contradictory to the assumptions. However, it would be useful for future work to address these issues in more detail.

Firth's correction method virtually eliminates the bias for most prevalence values consistent with the design of the pooled testing procedure. It overcorrects for values of p close to ψ, unlike the correction for perfect testing. This may not be entirely detrimental—Colon et al. (2001) found that for equal pool sizes, the MSE was reduced when a negative bias was introduced by allowing greater shrinkage in the estimator.

As stated in Sect. 2, our methods are, in fact, more general than indicated, because they cover the case that sensitivity or specificity varies with the pool tested, in particular that they may vary with pool size, so long as values for these parameters are known. This follows from the fact that throughout our derivations, the parameters a and b could have been given a subscript i, and no computations would have changed; should different a or b be needed for the same pool size, simply adding another subscript to index this situation would suffice, and our computations would carry through. Others have relaxed the assumption of constant a and b in the context of finding optimal designs (Zhang et al. 2014), variance estimation (Cowling et al. 1999) or dealing with the dilution effect (McMahan et al. 2013). The more general applicability of our derivations notwithstanding, our assumptions that sensitivity and specificity are known and not dependent on pool size were described by Tebbs et al. (2013) as "standard in the group testing literature". They added that "proper assay calibration is needed to ensure that this is reasonable in application". One way of achieving non-dependency is to choose a maximum pool size that makes the assumption reasonable; Zhang et al. (2014) gave an example of $m = 15$ in ELISA-based HIV testing. Finally, because of the myriad possibilities in specification, we have evaluated the performance of our new estimator only in the case of common sensitivity and specificity. We encourage researchers considering applying these methods in applications with different sensitivities or specificities by pool to evaluate the methods in their own circumstance, should they be concerned that the Firth correction may not yield results qualitatively comparable to those we present.

Another extension would be to allow uncertainty in sensitivity and specificity. Such uncertainty has been incorporated in interval estimation and sample size determination (Messam et al. 2008), but it is not obvious how it should be incorporated in bias calculations. Mitchell and Pagano (2012) derived a Burrows-type estimator (essentially Firth's correction, as we showed) for equal pool sizes with uncertain sensitivity and known specificity. However, they recognized the circularity of needing the sensitivity to estimate the prevalence unbiasedly, and so recommended an iterative estimation technique, which would be challenging for practitioners to apply.

Our work provides a foundation for interval estimation of p in the context of bias correction and misclassification. As noted in Discussion in Firth (1993), the asymptotic variance of $\breve{p}$ is the inverse of the expected information, $\left[\sum_{i=1}^{d} I_i(p)\right]^{-1}$, so that standard errors and confidence intervals may be computed to first order in the usual way; all that is required is for $I(p)$ from

Eq. A.2 (see "Appendix 1") to be subscripted. Further investigations on interval estimation remain the subject of future work.

Finally, Firth's bias correction can be applied to estimation of other parameters related to prevalence arising from pooled testing with misclassification. Considering only equal sized pools, Roy and Banerjee (2019) did this for the log-odds ratio comparing two subgroups of the population being studied and found the corrected estimator to have much smaller bias and MSE than the MLE.

APPENDIX 1: DERIVATION OF FIRTH'S BIAS CORRECTION APPLIED TO IMPERFECT POOLED TESTING

Recall $\pi(p) = a - r(1-p)^m$. The first two derivatives are

$$\frac{\mathrm{d}\pi(p)}{\mathrm{d}\,p} = mr(1-p)^{m-1} = \frac{m}{1-p}r(1-p)^m = \frac{m}{1-p}[a-\pi(p)],$$
$$\frac{\mathrm{d}^2\pi(p)}{\mathrm{d}\,p^2} = -m(m-1)r(1-p)^{m-2} = -\frac{m(m-1)}{(1-p)^2}[a-\pi(p)].$$

From here on, we drop the functional notation on $\pi(p)$. We need the following derivatives of the log-likelihood:

$$\frac{\mathrm{d}\,l}{\mathrm{d}\pi} = \frac{x}{\pi} - \frac{n-x}{1-\pi},$$
$$\frac{\mathrm{d}^2 l}{\mathrm{d}\pi^2} = -\frac{x}{\pi^2} - \frac{n-x}{(1-\pi)^2},$$
$$\frac{\mathrm{d}^3 l}{\mathrm{d}\pi^3} = \frac{2x}{\pi^3} - \frac{2(n-x)}{(1-\pi)^3}.$$

From these, we have the score function

$$S(p) = \frac{\mathrm{d}\,l}{\mathrm{d}\,p} = \frac{\mathrm{d}\,l}{\mathrm{d}\pi}\frac{\mathrm{d}\pi}{\mathrm{d}\,p} = \frac{m(a-\pi)(x-n\pi)}{(1-p)\pi(1-\pi)}. \tag{A.1}$$

The chain rule and product rule can be used to compute the higher-order derivatives of the likelihood, as follows:

$$\frac{\mathrm{d}^2 l}{\mathrm{d}\,p^2} = \frac{\mathrm{d}^2 l}{\mathrm{d}\pi^2}\left(\frac{\mathrm{d}\pi}{\mathrm{d}\,p}\right)^2 + \frac{\mathrm{d}\,l}{\mathrm{d}\pi}\frac{\mathrm{d}^2\pi}{\mathrm{d}\,p^2},$$

$$\frac{\mathrm{d}^3 l}{\mathrm{d}\,p^3} = \frac{\mathrm{d}^3 l}{\mathrm{d}\pi^3}\left(\frac{\mathrm{d}\pi}{\mathrm{d}\,p}\right)^3 + 3\frac{\mathrm{d}^2 l}{\mathrm{d}\pi^2}\frac{\mathrm{d}^2\pi}{\mathrm{d}\,p^2}\frac{\mathrm{d}\pi}{\mathrm{d}\,p} + \frac{\mathrm{d}\,l}{\mathrm{d}\pi}\frac{\mathrm{d}^3\pi}{\mathrm{d}\,p^3}.$$

We therefore can obtain the information $I(p)$ as:

$$\begin{aligned} I(p) &= E\left[-\frac{\mathrm{d}^2 l}{\mathrm{d}\, p^2}\right] \\ &= E\left[-\left(-\frac{x}{\pi^2}-\frac{n-x}{(1-\pi)^2}\right)\left(\frac{m}{1-p}(a-\pi)\right)^2-\left(\frac{x}{\pi}-\frac{n-x}{1-\pi}\right)\left(-\frac{m(m-1)(a-\pi)}{(1-p)^2}\right)\right] \\ &= \left(\frac{n}{\pi}+\frac{n}{1-\pi}\right)\left(\frac{m}{1-p}(a-\pi)\right)^2 \\ &= \frac{nm^2(a-\pi)^2}{(1-p)^2\pi(1-\pi)}. \end{aligned}$$

Computation of the bias (see Eq. 3) requires $\mathrm{d}I(p)\,/\,\mathrm{d}p$ and $E\left[\mathrm{d}^3 l\ /\ \mathrm{d}\, p^3\right]$ in addition to $I(p)$. The derivation of $\mathrm{d}I(p)\,/\,\mathrm{d}\,p$ is more tedious than for perfect testing, because our expression for $I(p)$ includes both p and π. We remedy this (i.e., put it in terms of π alone and p only implicitly) by writing

$$I(p) = \frac{nm^2(a-\pi)^2}{(1-p)^2\pi(1-\pi)} = \frac{nm^2(a-\pi)^2}{\left(\frac{a-\pi}{r}\right)^{\frac{2}{m}}\pi(1-\pi)}. \tag{A.2}$$

Then

$$\begin{aligned} \frac{\mathrm{d}I(p)}{\mathrm{d}\,p} &= \frac{\mathrm{d}I(p)}{\mathrm{d}\pi}\frac{\mathrm{d}\pi}{\mathrm{d}\,p} \\ &= nm^2\frac{\left(\frac{a-\pi}{r}\right)^{\frac{2}{m}}\pi(1-\pi)[-2(a-\pi)]-(a-\pi)^2\left[\left(\frac{a-\pi}{r}\right)^{\frac{2}{m}}(1-2\pi)+\pi(1-\pi)\frac{2}{m}\left(\frac{a-\pi}{r}\right)^{\frac{2-m}{m}}\left(-\frac{1}{r}\right)\right]}{\left(\frac{a-\pi}{r}\right)^{\frac{4}{m}}\pi^2(1-\pi)^2} \\ &\quad\times\left[\frac{m}{1-p}(a-\pi)\right] \\ &= -\frac{nm^2(a-\pi)^2}{(1-p)^3\pi^2(1-\pi)^2}[2(m-1)\pi(1-\pi)+m(a-\pi)(1-2\pi)] \end{aligned} \tag{A.3}$$

noting that $\frac{r}{a-\pi} = (1-p)^{-m}$ is useful during the simplification. The final quantity for computing the bias $b(p)$ is

$$
\begin{aligned}
E\left[\frac{\mathrm{d}^3 l}{\mathrm{d}\,p^3}\right] &= E\left[\left(\frac{2x}{\pi^3} - \frac{2(n-x)}{(1-\pi)^3}\right)\left(\frac{m}{1-p}(a-\pi)\right)^3\right. \\
&\quad + 3\left(-\frac{x}{\pi^2} - \frac{n-x}{(1-\pi)^2}\right)\left(\frac{m}{1-p}(a-\pi)\right)\left(-\frac{m(m-1)(a-\pi)}{(1-p)^3}\right) \\
&\quad \left. + \left(\frac{x}{\pi} - \frac{n-x}{1-\pi}\right)\left(\frac{m(m-1)(m-2)(a-\pi)}{(1-\pi)^3}\right)\right] \\
&= \left(\frac{2n}{\pi^2} - \frac{2n}{(1-\pi)^2}\right)\left(\frac{m}{1-p}(a-\pi)\right)^3 + 3\left(\frac{n}{\pi} + \frac{n}{1-\pi}\right)\left(\frac{m^2(m-1)(a-\pi)^2}{(1-p)^3}\right) \\
&= \frac{nm^2(a-\pi)^2}{(1-p)^3\pi^2(1-\pi)^2}\left[2m(1-2\pi)(a-\pi) + 3(m-1)\pi(1-\pi)\right].
\end{aligned}
$$

Further simplification results in the following expression for the numerator of the bias (see Eq. 3):

$$
2\frac{\mathrm{d}I(p)}{\mathrm{d}\,p} + E\left[\frac{\mathrm{d}^3 l}{\mathrm{d}\,p^3}\right] = -\frac{nm^2(m-1)(a-\pi)^2}{(1-p)^3\pi(1-\pi)}. \tag{A.4}
$$

APPENDIX 2: R CODE FOR NEWTON–RAPHSON ITERATION TO FIND FIRTH'S BIAS-CORRECTED ESTIMATE OF *P*

```
"ipooledbinom.firth.NR" <- function(x, m, n = rep(1, length(m)),
              a = rep(1, length(m)), b = rep(1, length(m)),
               tol = 1e-8,
              p.start = NULL){
# if just the regular binomial, return the simple proportion
if(all(m == 1) & all(a == 1) & all(b == 1))
 return(sum(x) / sum(n))
if(sum(x) == 0) return(0)

r <- a + b - 1

# Compute a (default) starting value:
# When a starting value is not specified with p.start,
# this code calls this same function using the Recall()
# functionality, but with forced a = 1 and b = 1.
# This approach is to avoid having to reference a different
# function that gives the perfect Firth estimate, noting
# that this function itself gives the correct, perfect-test
# Firth estimate if a = 1 and b = 1.
# So: use p.start if it is explicitly specified, else
#     compute a starting value
if(!is.null(p.start)){
```

```
   p.new <- p.start
} else {
   p.new <- NULL
   if(is.null(p.start) & is.null(p.new)){

      # if all pools are positive, the default starting
      # value given below is 0, so make it something positive
      if(sum(x) == sum(n)){
         p.new <- 1/sum(m*n)
      } else {
         N <- sum(n * m)
         mmw <- N / sum(n) # average pool size, N/sum(n)

         # uses proportion of positive pools, sum(x)/sum(n)
         p.new <- Recall(x,m,n,a=1,b=1,tol=tol,
                   p.start=1-(1-sum(x)/sum(n))^(1/mmw))
      }
   }
}
done <- FALSE
iter <- 0
while(!done){
   iter <- iter+1
   p.old <- p.new
   pip <- a - r * (1-p.old)^m
   vi <- n*m^2*(a-pip)^2 / ((1-p.old)^2 * pip * (1-pip))
   wi <- vi / sum(vi)
   vi.prime <- -(n*m^2*(a-pip)^2*(2*(m-1)*pip*(1-pip) +
       m*(a-pip)*(1-2*pip))) / ((1-p.old)^3*pip^2*(1-pip)^2)
   wi.prime <- (vi.prime * sum(vi) - vi *
    sum(vi.prime)) / sum(vi)^2
   tSs <- sum(m*(a-pip)*(x-n*pip)/(pip*(1-pip)) - 0.5*wi*(m-1))
   tSsp <- sum((m^2/(1-p.old)) * ((x-n*(1-a))*pip^3 +
                    (n*a*(1-a)-3*a*x)*pip^2 + a*(2*a+1)*x*pip
                     - a^2*x) /
                    (pip^2*(1-pip)^2) - 0.5*wi.prime*(m-1))
   p.new <- max(0, p.old - tSs / tSsp)
   # 0 rather than 0.01 gave problems in some cases
   if(iter > 10000){
      stop("Too many iterations")
   }
   if(abs(p.new-p.old)<tol)
      done <- TRUE
 }
```

```
 p.new
}
```

REFERENCES

Burkhalter KL, Horiuchi K, Biggerstaff BJ, Savage HM and Nasci RS (2014) Evaluation of a Rapid Analyte Measurement Platform and real-time reverse-transcriptase polymerase chain reaction assay West Nile Virus detection system in mosquito pools. Journal of the American Mosquito Control Association, 30, 21–30. [PubMed: 24772673]

Burrows PM (1987) Improved estimation of pathogen transmission rates by group testing. Phytopathology, 77, 363–365.

Colon S, Patil GP and Taillie C (2001) Estimating prevalence using composites. Environmental and Ecological Statistics, 8, 213–236.

Cowling DW, Gardner IA and Johnson WO (1999) Comparison of methods for estimation of individual-level prevalence based on pooled samples. Preventive Veterinary Medicine, 39, 211–225. [PubMed: 10327439]

Dhand NK, Eppleston J, Whittington RJ and Toribio JL (2007) Risk factors for ovine Johne's disease in infected sheep flocks in Australia. Preventive Veterinary Medicine, 82, 51–71. [PubMed: 17602766]

Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika, 80, 27–38.

Gart JJ (1991) An application of score methodology: Confidence intervals and tests of fit for one-hit curves. In: Handbook of Statistics, Rao CR, Chakraborty R (eds), 395–406. Amsterdam: Elsevier.

Hepworth G (2005) Confidence intervals for proportions estimated by group testing with groups of unequal size. JABES, 10, 478–497.

Hepworth G and Biggerstaff BJ (2017) Bias correction in estimating proportions by pooled testing. JABES, 22, 602–614. [PubMed: 30636859]

Hepworth G and Watson R (2009) Debiased estimation of proportions in group testing. JRSS-C, 58, 105–121.

Hepworth G (2013) Improved estimation of proportions using inverse binomial group testing. JABES, 18, 102–119.

Hughes-Oliver JM (2006) Pooling experiments for blood screening and drug discovery. In: Dean A, Lewis S (eds) Screening. Springer, New York, NY.

Komar N, Colborn JM, Horiuchi K, Delorey M, Biggerstaff BJ, Damian D, Smith K and Townsend J (2015) Reduced West Nile Virus transmission around communal roosts of Great-Tailed Grackle (*Quiscalus mexicanus*). EcoHealth, 12, 144–151. [PubMed: 25480320]

Liu SC, Chiang KS, Lin CH, Chung WC, Lin SH and Yang TC (2011) Cost analysis in choosing group size when group testing for Potato virus Y in the presence of classification errors. Annals of Applied Biology, 159, 491–502.

Liu A, Liu C, Zhang Z and Albert PS (2012) Optimality of group testing in the presence of misclassification. Biometrika, 99, 245–251. [PubMed: 23049137]

McMahan CS, Tebbs JM and Bilder CR (2013) Regression models for group testing data with pool dilution effects Biostatistics, 14, 284–298. [PubMed: 23197382]

Messam LL, Branscum AJ, Collins MT and Gardner IA (2008) Frequentist and Bayesian approaches to prevalence estimation using examples from Johne's disease. Animal Health Research Reviews, 9, 1–23. [PubMed: 18346298]

Mitchell S and Pagano M (2012) Pooled testing for effective estimation of the prevalence of *Schistosoma mansoni*. American Journal of Tropical Medicine and Hygiene, 87, 850–861. [PubMed: 22964721]

Reiczigel J, Foldi J and Ozsvari L (2010) Exact confidence limits for prevalence of a disease with an imperfect diagnostic test. Epidemiology and Infection, 138, 1674–1678. [PubMed: 20196903]

Roy S and Banerjee T (2019) Estimation of log-odds ratio from group testing data using Firth correction. Biometrical Journal, 61, 714–728. [PubMed: 30645765]

Swallow WH (1985) Group testing for estimating infection rates and probabilities of disease transmission. Phytopathology, 75, 882–889.

Tebbs JM, McMahan CS and Bilder CR (2013) Two-stage hierarchical group testing for multiple infections with application to the Infertility Prevention Project. Biometrics, 69, 1064–1073. [PubMed: 24117173]

Tu XM, Litvak E and Pagano M(1994) Screening tests: can we get more by doing less? Statistics in Medicine, 13, 1905–1919. [PubMed: 7846399]

Youden WJ (1950) Index for rating diagnostic tests. Cancer, 3, 32–35. [PubMed: 15405679]

Zhang Z, Liu C, Kim S and Liu A (2014) Prevalence estimation subject to misclassification: the mis-substitution bias and some remedies. Statistics in Medicine, 33, 4482–4500. [PubMed: 25043925]

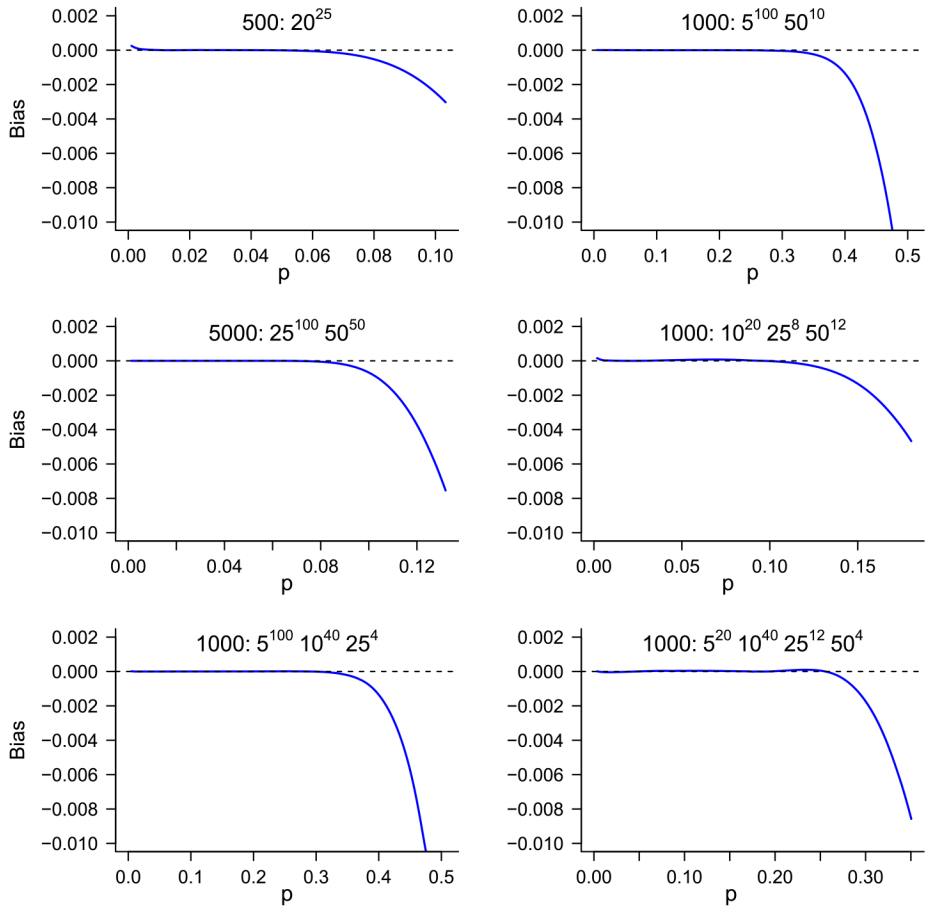

Figure 1.

Bias of estimators over $p < \psi$ corrected by Firth's method, for a range of pooled testing procedures, assuming imperfect testing ($a = 0.95, b = 0.99$).

Table 1.

Estimates of p using the MLE or the Firth bias correction method for perfect testing and imperfect testing (sensitivity = 0.95, specificity = 0.99) for selected outcomes of a procedure with 8 pools of 20 and 8 pools of 5

Testing	**Method**	**Number of positive pools (x_1, x_2)**									
		(1, 2)	**(4, 0)**	**(2, 5)**	**(3, 7)**	**(6, 4)**	**(5, 7)**	**(7, 5)**	**(7, 8)**	**(8, 7)**	**(8, 8)**
Perfect	MLE	0.016	0.025	0.042	0.067	0.085	0.099	0.128	0.205	0.341	1
	Firth	0.016	0.024	0.040	0.064	0.080	0.093	0.118	0.187	0.296	0.455
Imperfect	MLE	0.016	0.026	0.044	0.077	0.097	0.394	0.170	1	0.397	1
	Firth	0.015	0.025	0.042	0.072	0.089	0.124	0.146	0.455	0.327	0.455

Table 2.

Bias of estimators corrected by either Gart's or Firth's method, assuming sensitivity = 0.95 and specificity = 0.99, for 8 pools of 20 and 8 pools of 5

p	MLE			Gart			Firth		
	$\mathbb{E}(\hat{p})$	% Bias	RMSE	$\mathbb{E}(\tilde{p})$	% Bias	RMSE	$\mathbb{E}(\check{p})$	% Bias	RMSE
0.01	0.0106	6.3	0.0083	0.0101	0.70	0.0078	0.0101	0.95	0.0078
0.02	0.0212	5.8	0.0123	0.0200	−0.14	0.0115	0.0200	0.16	0.0115
0.03	0.0319	6.2	0.0161	0.0300	−0.26	0.0147	0.0300	0.11	0.0148
0.04	0.0427	6.9	0.0200	0.0399	−0.34	0.0178	0.0400	0.11	0.0180
0.05	0.0538	7.6	0.0243	0.0498	−0.40	0.0212	0.0501	0.13	0.0213
0.07	0.0766	9.4	0.0352	0.0698	−0.33	0.0290	0.0702	0.31	0.0292
0.10	0.1124	12.4	0.0564	0.0999	−0.05	0.0421	0.1008	0.81	0.0433
0.15	0.1738	15.9	0.1019	0.1477	−1.55	0.0602	0.1509	0.54	0.0655
0.20	0.2402	20.1	0.1613	0.1901	−4.94	0.0745	0.1975	−1.25	0.0823
0.25	0.3165	26.6	0.2266	0.2288	−8.49	0.0879	0.2409	−3.64	0.0939
0.30	0.4010	33.7	0.2643	0.2643	−11.89	0.1008	0.2803	−6.58	0.1007

Table 3.

Mean percentage absolute bias, RMSE and bias at $p = \psi$, for the estimator of p corrected by Firth's method, for a range of pooled testing procedures, assuming sensitivity = 0.95 and specificity = 0.99

N	m^n	ψ	Mean $\lvert\%bias\rvert$	Mean RMSE	Bias at $p = \psi$
500	5^{100}	0.506	0.40	0.0338	−0.0189
500	10^{50}	0.248	0.54	0.0245	−0.0088
500	20^{25}	0.103	0.82	0.0147	−0.0030
500	50^{10}	0.027	1.97	0.0065	−0.0006
500	100^{5}	0.008	4.57	0.0032	−0.0001
1000	20^{50}	0.133	0.61	0.0137	−0.0053
1000	100^{10}	0.013	1.97	0.0033	−0.0003
1000	5^{200}	0.569	0.46	0.0289	−0.0279
5000	5^{1000}	0.687	0.66	0.0213	−0.0528
1000	$5^{100}50^{10}$	0.506	0.38	0.0338	−0.0189
1000	$25^{20}\ 50^{10}$	0.078	0.60	0.0110	−0.0023
5000	$5^{500}\ 50^{50}$	0.641	0.57	0.0252	−0.0419
5000	$25^{100}\ 50^{50}$	0.132	0.73	0.0101	−0.0075
1000	$10^{20}\ 25^{8}\ 50^{12}$	0.180	0.43	0.0251	−0.0047
1000	$10^{50}\ 25^{12}\ 50^{4}$	0.248	0.47	0.0238	−0.0089
1000	$5^{100}\ 10^{40}\ 25^{4}$	0.507	0.39	0.0334	−0.0193
5000	$10^{200}\ 25^{60}\ 50^{30}$	0.344	0.73	0.0229	−0.0222
1000	$5^{10}\ 10^{10}\ 25^{10}\ 50^{12}$	0.261	0.29	0.0400	−0.0043
1000	$5^{20}\ 10^{40}\ 25^{12}\ 50^{4}$	0.350	0.28	0.0367	−0.0086
5000	$10^{50}\ 25^{40}\ 50^{30}\ 100^{20}$	0.248	0.52	0.0262	−0.0089