Title

Misclassified group-tested current status data

Journal

Biometrika, 103(4)

ISSN

0006-3444

Authors

Petito, LC

Jewell, NP

Publication Date

2016-12-01

DOI

10.1093/biomet/asw043

Peer reviewed

Biometrika (2016), **103**, 4, *pp.* 801–815
Printed in Great Britain
doi: 10.1093/biomet/asw043

Misclassified group-tested current status data

By L. C. PETITO and N. P. JEWELL

Division of Biostatistics, School of Public Health, 101 Haviland Hall, University of California, Berkeley, California 94720, U.S.A.
lucia.petito@berkeley.edu jewell@berkeley.edu

Summary

Group testing, introduced by Dorfman (1943), has been used to reduce costs when estimating the prevalence of a binary characteristic based on a screening test of k groups that include n independent individuals in total. If the unknown prevalence is low and the screening test suffers from misclassification, it is also possible to obtain more precise prevalence estimates than those obtained from testing all n samples separately (Tu et al., 1994). In some applications, the individual binary response corresponds to whether an underlying time-to-event variable T is less than an observed screening time C, a data structure known as current status data. Given sufficient variation in the observed C values, it is possible to estimate the distribution function F of T nonparametrically, at least at some points in its support, using the pool-adjacent-violators algorithm (Ayer et al., 1955). Here, we consider nonparametric estimation of F based on group-tested current status data for groups of size k where the group tests positive if and only if any individual's unobserved T is less than the corresponding observed C. We investigate the performance of the group-based estimator as compared to the individual test nonparametric maximum likelihood estimator, and show that the former can be more precise in the presence of misclassification for low values of $F(t)$. Potential applications include testing for the presence of various diseases in pooled samples where interest focuses on the age-at-incidence distribution rather than overall prevalence. We apply this estimator to the age-at-incidence curve for hepatitis C infection in a sample of U.S. women who gave birth to a child in 2014, where group assignment is done at random and based on maternal age. We discuss connections to other work in the literature, as well as potential extensions.

Some key words: Current status data; Expectation-maximization algorithm; Group testing; Pool-adjacent-violators algorithm.

1. Introduction

In the past decade, group testing of a binary response has once again become a topic of great interest (Remlinger et al., 2006; Wahed et al., 2006; Dhand et al., 2010). The idea was first introduced in 1943 as a potential cost-saving measure for the detection of syphilis in U.S. army recruits (Dorfman, 1943). Group testing reduces the number of tests by allocating, randomly or otherwise, n individuals into J groups of equal size k and testing each pooled group only once, in order to provide an estimate of the prevalence of a binary characteristic in a population.

More recent work has considered potential issues with group testing, such as dilution effects, non-random group assignment, and misclassification (Hwang, 1976; Wein & Zenios, 1996; Delaigle & Hall, 2012; Liu et al., 2012). Tu et al. (1995) suggested that if the unknown prevalence of a binary characteristic is sufficiently low and the screening test suffers from misclassification,

more precise estimates of the prevalence can be obtained from J group tests than by testing all n individuals separately. The intuition behind this finding is complex. When a test has a rate of misclassification independent of the number of individuals in the pooled sample, performing fewer tests could increase the precision of the prevalence estimate due to fewer tests being performed, thereby leading to less noise in the observations. This is particularly the case when the prevalence is sufficiently small, making it uncommon that two positives will occur in the same group.

The data structure where an individual's binary response corresponds to an underlying time-to-event variable T occurring before an observed screening time C is known as current status data, or interval censoring type I (Jewell & van der Laan, 2003; Jewell & Emerson, 2013). The nonparametric maximum likelihood estimator of the distribution function, F, of T for current status data is the pool-adjacent-violators algorithm, although it is only possible to use this estimator if there is sufficient variation in the observed screening times C (Ayer et al., 1955).

In this paper, we develop a simple algorithm to compute a nonparametric maximum likelihood estimator of F for group-tested current status data, and extend it to settings where the test is subject to misclassification. When misclassification is present, we hypothesize that there will sometimes be substantial gains in precision for values of T at which the prevalence is sufficiently small, as described by Tu et al. (1995) in the case of estimating a single fixed prevalence.

2. NOTATION AND LIKELIHOOD FUNCTION

We assume that the underlying data, prior to grouping, arise from n independent realizations of a bivariate random variable, $\Phi = (1(T < C), C)$, where the survival random variable T and screening random variable C follow distribution functions F and G, respectively. Throughout, we assume that T and C are independent. The observed data are based on grouping these realizations at random into blocks of size k, where for convenience we assume that n/k is an integer. It is trivial to extend all the results below to situations where the block sizes may vary. Thus each original unit corresponds to the jth individual in the ith group, where $i = 1, \ldots, n/k$ and $j = 1, \ldots, k$. The group-tested result from the ith group, Δ_i, is the only test result available, whereas individual screening times, C_{ij}, are observed for all participants. Specifically, $\Delta_i = 0$ if and only if $\Phi_{ij} = 0$ for all $j = 1, \ldots, k$, and $\Delta_i = 1$ otherwise. The group test detects the presence of one or more positives in the group, but cannot distinguish between a single, or several, positive Φ_{ij}. The immediate goal is to estimate the distribution function F.

Owing to the assumed independence of T and C, we can focus on the conditional likelihood of the data given the observed screening times $\{C_{ij} : i = 1, \ldots, n/k; j = 1, \ldots, k\}$. Since $\mathrm{pr}(\Delta_i = 0 \mid C_{ij} : j = 1, \ldots, k) = \prod_{j=1}^{k} \mathrm{pr}(\Phi_{ij} = 0 \mid C_{ij})$, this conditional likelihood is

$$\mathrm{CL} = \prod_{i=1}^{n/k} \left\{ S(c_{i1}) \times \cdots \times S(c_{ik}) \right\}^{1-\delta_i} \left\{ 1 - S(c_{i1}) \times \cdots \times S(c_{ik}) \right\}^{\delta_i}, \tag{1}$$

where $S = 1 - F$ is the survival function of T. This conditional likelihood applies to various methods of selecting the screening times C and assigning the observations to groups for testing. At one extreme, the C values in each group are selected completely at random; at the other end of the spectrum, individuals with a common value of C are assigned to the same group. The latter sampling scheme is only fully feasible if the distribution function G is discrete. While the estimation strategy pursued here applies generally, estimation is much simpler with a common C value in each group, and asymptotic properties of the estimator are more easily derived in

that case. For example, with a common value of C in each grouping of fixed group size k, the likelihood (1) simplifies to that for the standard current status data problem with underlying survival function $S_k(c_i) = S(c_i)^k$. Estimates and inference regarding S_k can then be immediately translated to corresponding statements regarding S itself. In practice, with a continuous G, it may be advantageous to group together individuals with approximately the same value of C.

This development assumes a perfect screening test of whether or not the true group test result was positive, $\Delta_i = 1$. We can extend these ideas to permit misclassification of the test results, and we now use the notation Y to distinguish the potentially misclassified test result from the true result Δ. Assume that the test has known sensitivity and specificity, independent of both the screening time C and the group size, given by $\alpha = \text{pr}(Y = 1 \mid \Delta = 1)$ and $\beta = \text{pr}(Y = 0 \mid \Delta = 0)$ with the assumption that $\alpha + \beta > 1$. Then the conditional likelihood of the potentially misclassified data, given the observed screening times $\{C_{ij} : i = 1, \ldots, n/k; j = 1, \ldots, k\}$, can be written as

$$\text{CL}(\alpha, \beta) = \prod_{i=1}^{n/k} \left\{1 - \alpha + \gamma S(c_{i1}) \times \cdots \times S(c_{ik})\right\}^{1-y_i} \left\{\alpha - \gamma S(c_{i1}) \times \cdots \times S(c_{ik})\right\}^{y_i},$$

where $\gamma = \alpha + \beta - 1$.

3. An expectation-maximization pool-adjacent-violators algorithm

3·1. *Development of the algorithm*

Group-tested current status data can be formulated as a missing data problem. First, consider the setting without misclassification of test results. While the full set of screening times C_{ij} is observed, only group-tested results Δ_i are available, whereas a complete dataset would include all individual test results, Φ_{ij}. This missing information setting naturally allows use of the expectation-maximization algorithm (Dempster et al., 1977).

To implement the expectation-maximization algorithm, we calculate the expected value of the true individual test result, Φ_{ij}, given the observed value of the group-tested result, Δ_i, based on a current estimate of F. These calculations are straightforward when there is no misclassification:

$$E(\Phi_{ij} \mid \Delta_i = 0, C_{i1} = c_{i1}, \ldots, C_{ik} = c_{ik}) = 0, \tag{2}$$

$$E(\Phi_{ij} \mid \Delta_i = 1, C_{i1} = c_{i1}, \ldots, C_{ik} = c_{ik}) = F(c_{ij})\{1 - S(c_{i1}) \times \cdots \times S(c_{ik})\}^{-1}. \tag{3}$$

For misclassified data with sensitivity α and specificity β, computing the expected value of an individual true disease status Φ_{ij} given the potentially misclassified observed group-test result Y_i becomes slightly more complicated; see the Supplementary Material. Letting $\gamma = \alpha + \beta - 1$, this step becomes

$$E(\Phi_{ij} \mid Y_i = 1, C_{i1} = c_{i1}, \ldots, C_{ik} = c_{ik}) = \alpha F(c_{ij})\{\alpha - \gamma S(c_{i1}) \times \cdots \times S(c_{ik})\}^{-1},$$

$$E(\Phi_{ij} \mid Y_i = 0, C_{i1} = c_{i1}, \ldots, C_{ik} = c_{ik}) = \frac{(1 - \alpha)F(c_{ij})}{(1 - \alpha) + \gamma S(c_{i1}) \times \cdots \times S(c_{ik})}.$$

For the maximization step, we simply use a weighted version of the pool-adjacent-violators algorithm on the full dataset $\{\phi_{ij} : i = 1, \ldots, n/k; j = 1, \ldots, k\}$, where $\phi_{ij} = 0$ with weight 1 if $\delta_i = 0$, per (2). On the other hand, according to (3), if $\delta_i = 1$, then $\phi_{ij} = 1$ with weight given by

the right-hand side of (3), and additional observations $\phi_{ij} = 0$ have weight given by 1 minus the right-hand side of (3). The complete algorithm is thus described as follows.

Step 1. Initialize values of $f_{ij}^{(0)} = \hat{F}^{(0)}(c_{ij})$ for each individual and set a threshold τ for convergence.

Step 2 (Expectation). For each individual $j \in \{1, \ldots, k\}$ in group i, calculate the probability f_{ij}^* that the individual tested positive, given their group's test result. For perfectly classified results, δ_i, use

$$f_{ij}^* = \begin{cases} f_{ij}^{(0)} \left\{ 1 - \prod_{J=1}^{k} (1 - f_{iJ}^{(0)}) \right\}^{-1}, & \delta_i = 1, \\ 0, & \delta_i = 0. \end{cases} \tag{4}$$

For group-tested results subject to misclassification, y_i, with sensitivity α and specificity β such that $\gamma = \alpha + \beta - 1$, use

$$f_{ij}^* = \begin{cases} \alpha f_{ij}^{(0)} \left\{ \alpha - \gamma \prod_{J=1}^{k} (1 - f_{iJ}^{(0)}) \right\}^{-1}, & y_i = 1, \\ (1 - \alpha) f_{ij}^{(0)} \left\{ 1 - \alpha + \gamma \prod_{J=1}^{k} (1 - f_{iJ}^{(0)}) \right\}^{-1}, & y_i = 0. \end{cases} \tag{5}$$

Step 3 (Maximization). Use the group-tested results, δ_i or y_i, as the observations for each individual, and the probabilities from Step 2 as the weights in the weighted pool-adjacent-violators algorithm to calculate updated estimates of $f_{ij}^{(1)} = \hat{F}^{(1)}(c_{ij})$.

Step 4. Repeat Steps 2 and 3, using the estimate of $\hat{F}$ from Step 3 as the initial value for Step 2, until convergence, for example until

$$\sum_{i=1}^{n/k} \sum_{j=1}^{k} \{ \hat{F}^{(t+1)}(c_{ij}) - \hat{F}^{(t)}(c_{ij}) \}^2 < \tau.$$

It is important to run the algorithm with several choices of starting values, not only to reduce the possibility of converging to a local extrema, but also to discover possible different nonunique versions of the nonparametric maximum likelihood estimator. We recommend choosing a large set of random starting values of F at the observed set of C_{ij} by generating random Un(0, 1) values ordered so that the starting values are monotonically increasing with C_{ij}.

3·2. *Comments regarding asymptotics*

Asymptotic results for standard current status data are nonstandard. The nonparametric maximum likelihood estimator is known to be consistent, although converging only at the rate $n^{1/3}$, but has a non-Gaussian limiting distribution known as Chernoff's distribution (Groeneboom & Wellner, 1992) in situations where the monitoring time distribution, G, is continuous; Banerjee (2012) provides a concise discussion of this result. Rather than using Wald-type pointwise confidence intervals derived from this limit, Banerjee & Wellner (2001, 2005) suggest using a likelihood ratio approach to construct confidence bands.

On the other hand, when G has finite support, the likelihood is parametric, since F can then be estimated only at this finite number of support points, namely the observed censoring times. As expected from this observation, the nonparametric maximum likelihood estimator now converges to a Gaussian limit at rate $n^{1/2}$, with the asymptotic variance at a specific monitoring time C_0 given simply by $F(C_0)\{1-F(C_0)\}\{g(C_0)\}^{-1}$, which is straightforward to estimate using the obvious plug-in estimators (Yu et al., 1998; Maathuis & Hudgens, 2011). The hybrid problem where the number of support points grows with the sample size is discussed beautifully in Tang et al. (2012). Sal y Rosas & Hughes (2010) proposed the inversion of a likelihood ratio test to obtain pointwise confidence intervals for F when the data are subject to misclassification.

These results can be applied directly to the group-testing scenario only in the simplest situations. For the extreme situation of only one monitoring time, estimation of $F(C_0)$ reduces to the simple estimation of prevalence. This scenario has been studied extensively in the literature on group testing with misclassification; for example, Tu et al. (1994) provided asymptotically normal confidence intervals with convergence rate $n^{1/2}$. Generalizing slightly, the situation with finite support for C, and with no misclassification, simplifies to the case considered by Yu et al. (1998) if individuals within a group all share a common value of C. In this case, $\mathrm{pr}(\Delta = 1 \mid C) = 1-S(C)^k$, so that asymptotic results for the nonparametric maximum likelihood estimator applied to the group-tested data immediately follow through for the plug-in estimator of S, or F, at the finite number of screening times C by using the delta method. We anticipate that this will extend straightforwardly in the presence of misclassification, and we also suggest that use of the bootstrap will be effective here.

Even with a finite number of monitoring times, the situation becomes more complex when screening times are randomly assigned to the groups. This is clear even in the case of only two monitoring times and with pair groupings done at random. Further, there are as yet no known asymptotic results for the nonparametric maximum likelihood estimator of § 3·1 with a continuous screening time distribution, although we anticipate that convergence will remain at a $n^{1/3}$ rate.

4. Elementary example

4·1. *An analytic solution*

For illustration, consider a simple example in a setting without misclassified test results, where there are two groups each containing two individuals; that is, $n = 4$ and $k = 2$. There are twelve possible combinations of group assignments and test results, corresponding to three different possible pair assignments with each pair having two possible test outcomes. Consideration of the conditional likelihood (1) reveals a simple solution in all but one of these cases; we focus on the remaining case, which has the grouping shown in Fig. 1, with $\Delta_1 = 1$ and $\Delta_2 = 0$.

The conditional likelihood (1) in this setting is

$$\mathrm{CL}_4 = \{1 - S(c_1)S(c_3)\}S(c_2)S(c_4).$$

It is immediate that the nonparametric maximum likelihood estimator must have $\hat{S}(c_1) = \hat{S}(c_2)$ and $\hat{S}(c_3) = \hat{S}(c_4)$. Hence, the nonparametric maximum likelihood estimator is not unique but is achieved by any set of $\{\hat{S}(c_1), \ldots, \hat{S}(c_4)\}$ with $\hat{S}(c_1) = \hat{S}(c_2)$, $\hat{S}(c_3) = \hat{S}(c_4)$ and $\hat{S}(c_2)\hat{S}(c_3) = 0{\cdot}5$. We show how the expectation-maximization pool-adjacent-violators algorithm converges to one such solution, with the specific value depending directly on the starting values for $\hat{F}^{(0)}(c_i) = 1 - \hat{S}^{(0)}(c_i)$.

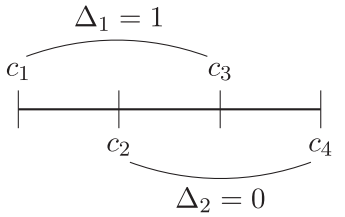

Fig. 1. Elementary example of data configuration with two groups, each of size 2, where the first group has tested positive and the second group has tested negative.

Given an initial set of probabilities $\hat{F}^{(0)}(C_1) = f_1$, $\hat{F}^{(0)}(C_2) = f_2$, $\hat{F}^{(0)}(C_3) = f_3$ and $\hat{F}^{(0)}(C_4) = f_4$ such that $f_1 \leqslant f_2 \leqslant f_3 \leqslant f_4$, the first step of the algorithm calculates the expectation of each of the initial conditional probabilities, f_i^* $(i = 1, \ldots, 4)$, as given in (4) and (5), i.e., the probability that each individual was positive given the known group-tested result. For two of these probabilities, in a setting without misclassification, this calculation is trivial: the pair tested negative, so neither of the individuals was positive. Hence we can set $f_2^* = f_4^* = 0$. For the pair that tested positive, this calculation follows directly from (4):

$$f_1^* = \text{pr}(T_1 \leqslant C_1 \mid \Delta_1 = 1) = f_1\{1 - (1-f_1)(1-f_3)\}^{-1} = f_1\{f_1 + f_3 - f_1 f_3\}^{-1},$$
$$f_3^* = \text{pr}(T_3 \leqslant C_3 \mid \Delta_1 = 1) = f_3\{1 - (1-f_1)(1-f_3)\}^{-1} = f_3\{f_1 + f_3 - f_1 f_3\}^{-1}.$$

The next step of the algorithm is to make these f_j^* monotonic, recalling that $f_2^* = f_4^* = 0$, by using the pool-adjacent-violators algorithm. This yields the following updated estimates of F:

$$\hat{F}^{(1)}(C_1) = \hat{F}^{(1)}(C_2) = f_1^*/2 = f_1\{2(f_1 + f_3 - f_1 f_3)\}^{-1}, \tag{6}$$
$$\hat{F}^{(1)}(C_3) = \hat{F}^{(1)}(C_4) = f_3^*/2 = f_3\{2(f_1 + f_3 - f_1 f_3)\}^{-1}. \tag{7}$$

These steps are then iterated until a determination of convergence based on comparing, say, the sum of the squared differences between $\hat{F}^{(m)}$ and $\hat{F}^{(m+1)}$ at each observed C to a prespecified threshold τ.

4·2. *Multiple convergence values*

As we demonstrated in § 4·1, the initial values for the pair that tested negative, $f_2 = \hat{F}^{(0)}(C_2)$ and $f_4 = \hat{F}^{(0)}(C_4)$, are not relevant to the update step in our expectation-maximization pool-adjacent-violators algorithm. Therefore, when discussing convergence of the algorithm, we will only consider initial values for $\hat{F}^{(0)}(C_1) = f_1$ and $\hat{F}^{(0)}(C_3) = f_3$.

In all settings where $f_1 = f_3 = f$, the update step given by (6) and (7) becomes

$$\hat{F}^{(1)}(C_1) = \hat{F}^{(1)}(C_3) = \hat{f} = 1\{2(2-f)\}^{-1}.$$

Therefore, at convergence, $f = \{2(2-f)\}^{-1}$ so that the algorithm converges to $f = 1 - 2^{-1/2}$, the only solution in $[0, 1]$. This can, of course, also be expressed as $\hat{S}(C_j) = 2^{-1/2}$ for $j = 1, \ldots, 4$.

For any other set of starting values, the ratio f_1/f_2 remains unchanged by the iterations. We can therefore write $f_1 = rf_3$, where $0 < r < 1$ and r stays fixed, as determined by the starting values for f_1 and f_3 . At convergence, (6) then simplifies to

$$f_3 = f_3\{2(rf_3 + f_3 - rf_3^2)\}^{-1}.$$

Thus, convergence occurs when $rf_3 + f_3 - rf_3^2 = 1/2$. After an application of the quadratic formula, this simplifies to

$$f_3 = \left\{r + 1 - (r^2 + 1)^{1/2}\right\}(2r)^{-1},$$

the only feasible solution. It immediately follows that at convergence, $f_1 = \{r+1-(r^2+1)^{1/2}\}/2$ so that the condition $(1 - f_1)(1 - f_3) = 0{\cdot}5$ holds, as noted in § 4·1.

This simple example demonstrates the nonuniqueness of the nonparametric maximum likelihood estimator, with the algorithm converging to a specific solution for $\hat{F}$ determined by the ratio of the starting values of F at C_1 and C_3. When using this algorithm in an applied setting, we suggest repeating it many times, using a different set of randomly drawn starting values each time, and then computing the likelihood function to identify as many different unique solutions to the optimization as possible.

5. Simulations

5·1. *Design of simulations*

We carry out two series of simulations to examine the behaviour of the expectation-maximization pool-adjacent-violators algorithm for group-tested data, as compared to the pool-adjacent-violators algorithm, which is the nonparametric maximum likelihood estimator for individual-level current status data (Barlow et al., 1972). We consider two scenarios, one where the tests are subject to no misclassification, and another where the test is subject to misclassification with known, constant error rates. In the latter case, the comparative estimator for misclassified individual-level current status data was derived by McKeown & Jewell (2010). We consider both continuous and discrete independent screening times. The former are described and discussed below, and the latter in the Supplementary Material.

Each simulation is characterized by a set of fixed parameters: n, the number of individuals; k, the group size; and α and β, the sensitivity and specificity of the screening test, respectively. We set $\alpha = \beta = 1$ in scenarios without misclassification. We first simulate traditional current status data for each individual from the distribution of the true event times, F, and the censoring distribution, G. Each run of the simulations begins with simulating data of sample size n at the individual level, followed by assigning individuals to groups randomly.

The distribution F of the event times T is Weibull with shape and scale parameters 4 and 25, respectively; here F has mean 22·7 and variance 40·4. For the perfectly classified test simulations, the screening distribution G for C is $\mathrm{Un}(0, 36)$, allowing almost all of the distribution F to be identified. The necessary binary datum Φ is then determined from the generated individual values of T and C. The values of Δ, the group-tested results, follow immediately from the values of Φ from each individual in the group, as described in § 2. Each simulation is performed 1000 times in six different settings, given by $n \in \{1000, 5000\}$ and groupings of sizes $k \in \{2, 5, 10\}$.

For misclassified test results, we are most interested in examining performance of the expectation-maximization pool-adjacent-violators estimator in the left tail of F, where false positive test results could have the largest effect on the estimate of F (Tu et al., 1994). Hence, while F remains the same Weibull distribution, we now take C to be $\mathrm{Un}(0, 14)$ to ensure that $F(t) \leqslant 10\%$. Here we select a single sample size $n = 5000$ in 12 different settings with group sizes $k \in \{2, 5, 10\}$ and misclassification rates of $\alpha = \beta \in \{0{\cdot}8, 0{\cdot}9, 0{\cdot}95, 0{\cdot}99\}$. In these simulations, the observed misclassified data are obtained by, first, subjecting each individual test result Φ to misclassification under the specified test characteristics and, second, generating the group-tested outcome Y separately by misclassifying the corresponding group-test result Δ. Here we

have used the same test classification probabilities, assuming independence between the group size and the error rates of the testing procedure.

In each run of the two sets of simulations, for perfectly classified data and misclassified data, we compute both the appropriate expectation-maximization pool-adjacent-violators algorithm for the group-tested data and the appropriate pool-adjacent-violators algorithm for individual data. To select initial values for the expectation-maximization pool-adjacent-violators algorithm, we first draw n values uniformly from the range [0, 1] and sort them from smallest to largest; we then order the observations so that the C values are monotonically increasing, and match the ordered initial probabilities to the ordered data. Although, as noted earlier, for a specific application we recommend choosing multiple starting values, here we opt to randomly select only one set of initial values for each simulated dataset, thereby achieving only one of potentially many possible nonparametric maximum likelihood estimates.

The averages of the estimates of F obtained from each algorithm over the 1000 runs are calculated for each t in the support of G. To calculate the estimate of F at a value of C not observed in a specific simulation, we assume left-continuity of both estimators in situations where this is not imposed by monotonicity. To provide a sense of the variability of each estimator, we also calculate the 2·5th and 97·5th quantiles of the estimates over the 1000 simulations. For the second set of simulations, we use these quantities to compute a measure of pseudo-relative efficiency, the ratio of the widths of these 95% Monte Carlo quantile intervals: $\{(q_{97\cdot5} - q_{2\cdot5})_{\text{group}}\}/\{(q_{97\cdot5} - q_{2\cdot5})_{\text{individual}}\}$. The variances of the simulated estimates are less relevant, since we hypothesize that this estimator does not converge to a Gaussian distribution, nor at a $n^{1/2}$ rate.

The Supplementary Material contains results from two simulations in samples of size $n = 10\,000$, with 10 fixed, equal-frequency screening times C, and with true event probabilities at each screening time fixed at $0{\cdot}005, 0{\cdot}01, \ldots, 0{\cdot}05$. In the first simulation, we randomly group individuals by values of C to allow for the presentation of asymptotically normal confidence intervals, as described in § 3·2; in the second, we group across screening times and again present the widths of the 95% Monte Carlo quantile intervals.

5·2. *Results: perfectly classified data*

Figure 2 displays the results from applying the expectation-maximization pool-adjacent-violators algorithm and the pool-adjacent-violators algorithm to data generated in the six simulations where there is no misclassification of the test results. These simulations show that the finite-sample bias is small, except perhaps when the group size is large, e.g., $k = 10$, and $F(t)$ is small. Even then, this bias declines systematically as the sample size increases. As anticipated, in all situations, the bias is also smaller for the estimator based on individual test results. Similarly, and also to be expected, the latter is more precise, although the gain in precision decreases for larger sample sizes and smaller k. This being said, the group-tested estimator stands up remarkably well given that the screening costs are reduced by 50%, 80% and 90% when $k = 2, 5$ and 10, respectively, assuming that costs are proportional to the number of tests.

Because the asymptotic properties of the expectation-maximization pool-adjacent-violators algorithm are currently unknown, to demonstrate variability in the estimates we delineate the 95% Monte Carlo quantile interval by dashed and dotted lines in Fig. 2. The width of this interval for the pool-adjacent-violators algorithm from individual data is always smaller than that for the expectation-maximization pool-adjacent-violators algorithm applied to group-tested data. This is to be expected, as there is no misclassification in these simulations. Smaller group sizes k in the expectation-maximization pool-adjacent-violators algorithm provide 95% quantile intervals more similar to those estimated from individual data, and as n increases for fixed k, the width of the 95%

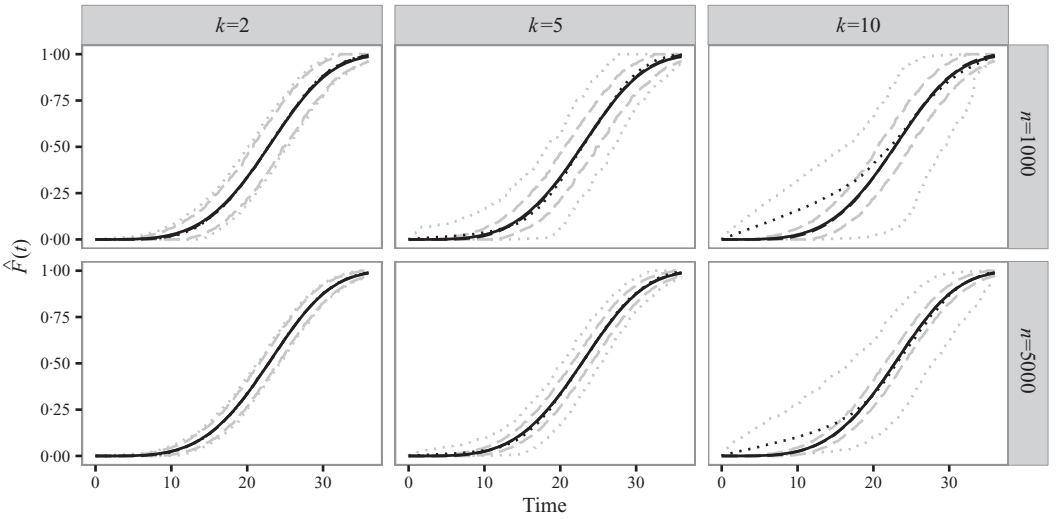

Fig. 2. Results from six simulations of the estimation of F, with 1000 runs each, for different sample sizes n and group sizes k. In each panel, the black lines are the average estimates of $\hat{F}(t)$ over the 1000 simulations, with the solid line representing the true cumulative distribution function Wei(4, 25) and the dashed and dotted lines representing, respectively, the estimates from the pool-adjacent-violators algorithm and the expectation-maximization pool-adjacent-violators algorithm; the grey lines are the 2·5th and 97·5th quantiles from the simulation runs for each estimator, using the same line types.

quantile interval decreases. Overall, Fig. 2 demonstrates that the expectation-maximization pool-adjacent-violators algorithm provides an unbiased estimate of the true underlying distribution, F.

5·3. *Results: misclassified data*

Figures 3 and 4 present results from the twelve simulations in settings with $n = 5000$ individuals and varying group sizes and misclassification rates. Figure 3 shows that the percentage relative bias of both estimators in these finite samples is large, e.g., greater than 100%, for estimates of $F(t)$ that are very small, e.g., less than 0·002, and is very close to zero for estimates of $F(t)$ that are greater than 0·02, even at large group sizes with high misclassification rates. Although the individual-based estimator is less biased at small group sizes and low misclassification rates, we do see similar or lower amounts of bias from the group-testing estimator at higher misclassification rates, e.g., $\alpha = \beta = 0{\cdot}8$ or 0·9, particularly with the larger grouping sizes $k = 5, 10$ and at lower values of T. Ultimately, the shapes of the finite-sample relative bias curves for these two estimators are very similar, so, at the very least, grouping does not introduce substantial amounts of additional bias.

With regard to variability, a comparison of the widths of the 95% Monte Carlo quantile intervals associated with both estimators, as shown in Fig. 4, demonstrates a considerable advantage of our estimator from group-tested data at low t and high levels of misclassification. For example, $T = 10$ corresponds to a true prevalence of 2·5%. If a test is subject to 10% misclassification, i.e., $\alpha = \beta = 0{\cdot}9$, then test results from data grouped into pools of size 10 will provide a more or equally precise estimate of $F(t)$ for $T < 10$ than data from individual tests. This implies that if the cumulative failure rate in question is less than 2·5%, a testing procedure that involves groups of size 10 will cost 90% less than testing everyone individually, and will result in a less biased and more precise estimate of $F(t)$ in this range. In general, the specific threshold t below which

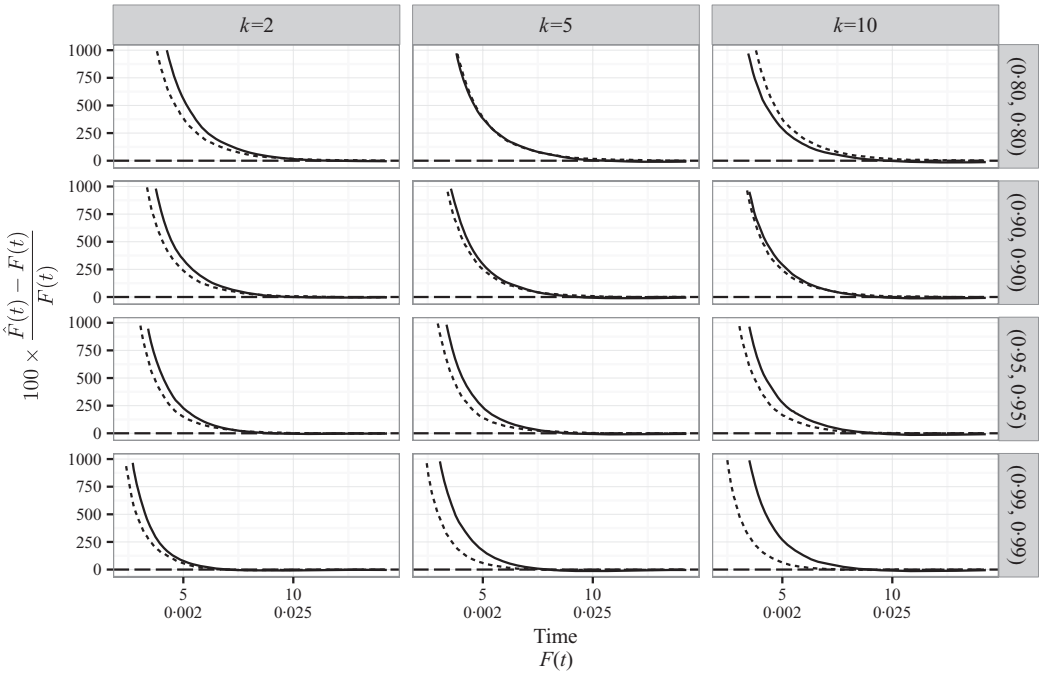

Fig. 3. Graphical representation of the finite-sample percentage relative bias from 12 simulations repeated 1000 times with 5000 individuals each, based on different group sizes k and misclassification rates (α, β), with values of the latter noted along the right-hand side. In each plot, the solid black line represents results obtained from the expectation-maximization pool-adjacent-violators algorithm for group tests, and the short-dashed black line represents results from the pool-adjacent-violators algorithm for misclassified individual test data; the long-dashed black line represents the reference level of 0% bias.

such precision gains can be expected depends on both the group size and the misclassification rate, as suggested by Tu et al. (1994) for estimation of a single fixed prevalence.

The Supplementary Material includes results from simulations of group-tested current status data on a grid, with grouping done solely according to common observation times, which more easily ensures a sufficiently small maximum value of F. As seen in Tu et al. (1994), we observe a reduction in the size of 95% confidence intervals as the group size increases, and separately a reduction in the size of the 95% confidence intervals as the misclassification rates decrease. Additionally, there appears to be no substantial increase in bias as group size increases.

6. Application to hepatitis C data

To investigate the performance of our estimator in a practical setting, we use publicly available data from the 2014 U.S. Birth Data File, created by the National Center for Health Statistics, to investigate the age-at-incidence distribution for hepatitis C in non-Hispanic white women of child-bearing age. The dataset includes all such women of ages 13–40 who gave birth in 2014. We are therefore making the tacit assumption that women who gave birth are a representative sample of women of the same ages that could have given birth in terms of their risk of infection with hepatitis C. This is not exactly correct but seems to be a reasonable approximation, at least for sexually active women. Of the 1 981 521 eligible women, we randomly sampled 10%, creating a sample of $N = 197\,840$ observations, for greater ease of illustration and computation. The

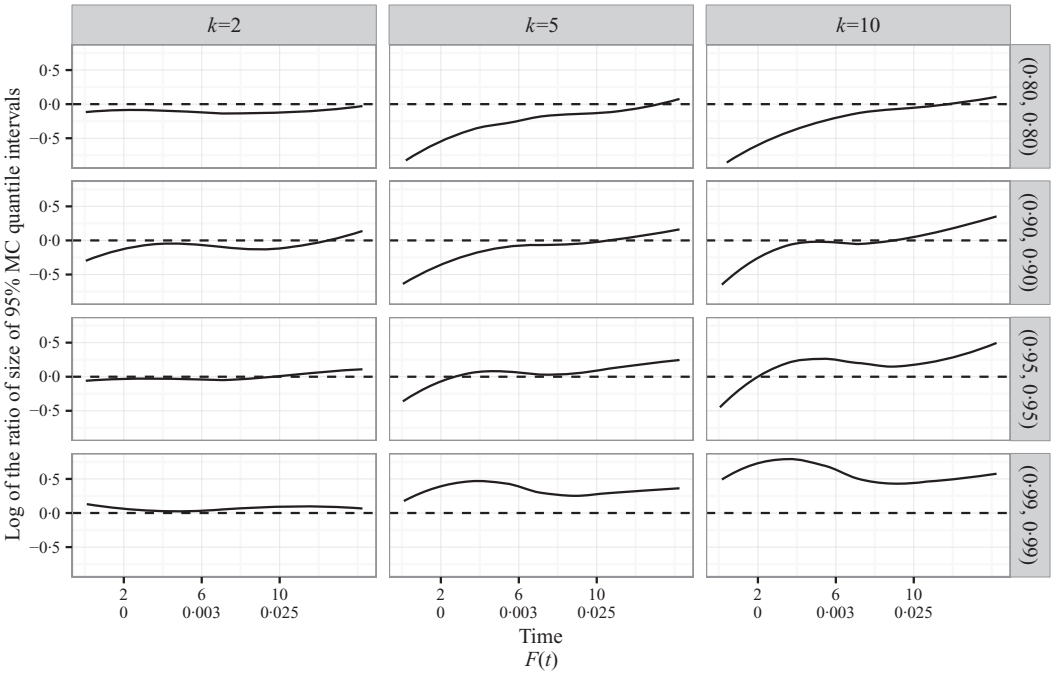

Fig. 4. Logarithm of the pseudo-relative efficiency of the expectation-maximization pool-adjacent-violators algorithm and the adjusted pool-adjacent-violators algorithm from 12 simulations of 1000 runs with 5000 individuals each, based on different group sizes k and misclassification rates (α, β), with values of the latter noted along the right-hand side. In each plot, the solid black line is a lowess curve showing the overall trend in pseudo-relative efficiency as t increases; the dashed black line represents equal-width 95% Monte Carlo quantile intervals for reference: if the solid black line is below zero, the width of the expectation-maximization pool-adjacent-violators 95% Monte Carlo quantile interval is smaller than that obtained from the individual test pool-adjacent-violators algorithm.

data include the mother's age in years and her hepatitis C status at the birth of her child. Of the $N = 197\,840$ women in our investigation, only 901 tested positive for hepatitis C, a cumulative incidence of 0·46%. When accounting for potential misclassification of these test results, we used the sensitivity, $\alpha = 0{\cdot}987$, and specificity, $\beta = 0{\cdot}999$, associated with the most commonly used test for hepatitis C: an enzyme immunoassay test. Although hepatitis C can be spread via sexual contact, it is primarily transmitted through blood, and an increase in the incidence of hepatitis C after age 25 would imply that people are beginning or continuing to engage in risky drug behaviour.

These data are based on individual blood testing for each mother separately. To illustrate our proposed methods, we consider group testing of pooled blood samples, representing potentially enormous savings in test costs depending on the size of the grouping used. These savings persist even if specific infected individuals need to be identified. As discussed above, given the low misclassification rates, we anticipate some loss of accuracy in estimating the prevalence, but this may nonetheless be worth the considerable cost reduction. We created artificial group-test results in two ways: (i) by assigning the data to groups of sizes 2, 5 and 10 according to age (gridded group assignment), and (ii) by randomly assigning the data to groups of sizes 2, 5 and 10. Then, each group test was assigned a positive result if at least one individual test was positive. For gridded group assignments, we computed point estimates and 95% confidence intervals adjusted for misclassification using the method described in § 3·2. For random group assignments, we applied the adjusted pool-adjacent-violators algorithm to the individual test

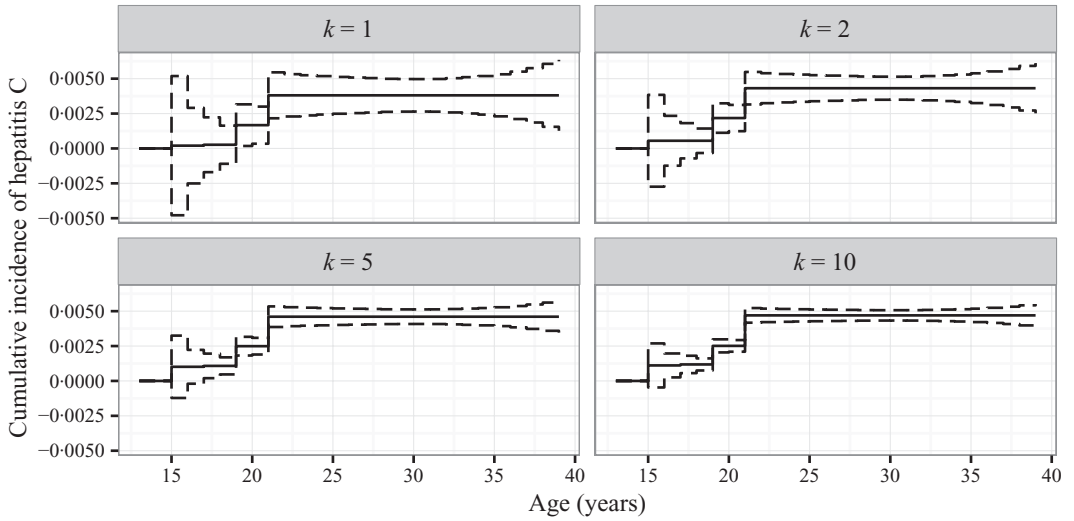

Fig. 5. Four estimates of the cumulative incidence of hepatitis C in non-Hispanic white child-bearing U.S. women of ages 13–40 in 2014 when grouping is assigned according to common values of age. Group sizes considered were $k = 1, 2, 5$ and 10. In each panel, the solid line is the estimate from the individual or group-tested results, and the dashed lines represent the upper and lower bounds of 95% confidence intervals.

results and, for comparison, the expectation-maximization pool-adjacent-violators algorithm to the group-tested results.

Figure 5 displays estimates obtained from individual and group-tested results with groups of sizes 2, 5 and 10 in a setting where group assignment is done by common age. The results are satisfying, as they lead to the same public health implications. Although the estimates are slightly different, they increase with group size, and the major jumps in the estimates occur at ages 19 and 21 for each of the group sizes considered. From these results, we can be fairly certain that any intervention to potentially reduce the public health burden due to hepatitis C infection would best occur during adolescence, ideally before risky behaviours such as drug use and unprotected sexual activity begin. In this example, major cost reductions could be achieved by decreasing the number of tests performed, assuming costs are proportional to the number of tests, without changing the conclusions of the analysis.

Figure 6 displays estimates obtained from individual and group-tested results with groups of sizes 2, 5 and 10 in a setting where group assignment is done completely at random. Unlike the estimates in Fig. 5 obtained from data grouped according to the women's age, here the estimates from data in groups of different sizes yield different implications. The results from the individual tests suggest an essentially flat cumulative incidence of hepatitis C after age 21, having reached a cumulative incidence of approximately 0·38%. This has significant implications for a public health intervention: it potentially indicates, for example, that any future hepatitis C vaccination would be most effective if implemented during late adolescence. No vaccine currently exists, although several candidates are under development. The group-tested results from groups of size 2 support the same conclusion, although they suggest that the cumulative incidence does not increase after age 19. However, the results from groups of sizes 5 and 10 tell a slightly different story: while these estimates increase to a cumulative incidence of roughly 0·4% before age 20, they then both continue to increase with age to somewhere in the range of 0·45–0·55% by age 40, suggesting that a substantial fraction of hepatitis C infections occur post-adolescence.

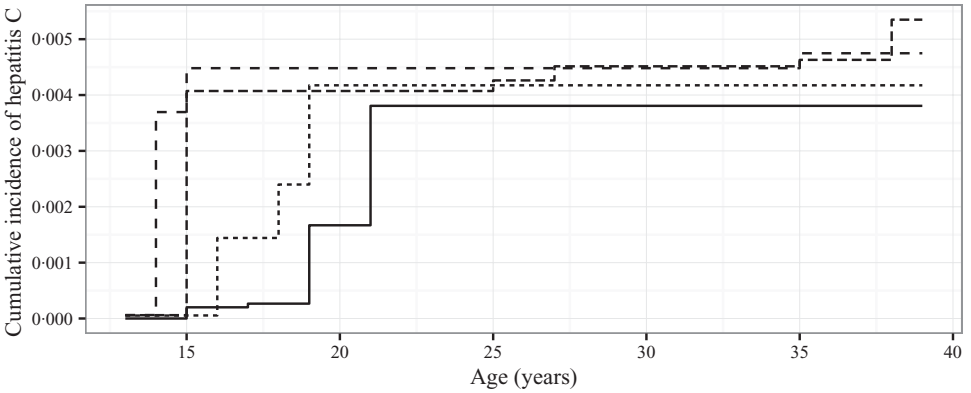

Fig. 6. Four estimates of the cumulative incidence of hepatitis C in non-Hispanic white child-bearing U.S. women of ages 13–40 in 2014 when group testing with random group assignments. The solid line is the pool-adjacent-violators estimate from the individual test results, and the dotted, short-dashed and long-dashed lines are the estimates obtained from the expectation-maximization pool-adjacent-violators algorithm with the individual test results artificially assigned to groups of sizes 2, 5 and 10, respectively.

Because these estimates seem to imply public health interventions at different times in life, it is important to consider which estimate is most reliable in this particular setting. As noted earlier, there is very little misclassification in the testing procedure, so we would expect that the results from the adjusted pool-adjacent-violators algorithm based on individual data would be more accurate, albeit obtained at significantly higher cost. However, the pool-adjacent-violators algorithm adjusted for misclassification has a limitation: it automatically estimates cumulative incidences that are less than $1 - \beta$ as 0; because the cumulative incidences at the early ages are less than 0·5%, if we had set $\beta \leqslant 0{\cdot}995$ in this application, our estimate from the individual data adjusted for misclassification would have been zero at all ages. This suggests a potential issue with individual test results that may not be as much of a problem with group-tested results.

7. Discussion

In this paper we have proposed a modified expectation-maximization algorithm to estimate a distribution function from data obtained by group-tested current status screening with test misclassification. Simulations show that the estimator based on group-tested data adds relatively little extra small-sample bias compared to an estimator based on individual data, but has a far lower cost, although this conclusion necessarily requires a larger n as the grouping size k increases. Additionally, when substantial misclassification is present, and $F(t)$ is low, estimates obtained from the expectation-maximization pool-adjacent-violators algorithm with groups of size 5 or larger may be less biased and have improved precision, although inferential properties for this procedure need further development. This offers the possibility that a significantly less expensive testing procedure might yield a less biased and more precise estimate for the left tail of F.

In the presence of misclassification, these observations suggest possible hybrid grouping strategies that may improve precision at low values of $F(t)$ and maintain performance at higher levels, all in comparison to individual tests whose costs are far greater. That is, where possible, if the screening times are known in advance of pooling, it will likely be advantageous to first group individuals according to the observed C values, and then use larger group sizes at the smaller

values of C and decrease the group size as C increases, even down to individual tests. Simulations to examine variations of these possibilities are currently under way. As noted earlier, when individuals in a group have similar C values, it is possible to also use an approximate individual group-tested current status estimator by treating all C values in the group as being the same.

There are a number of important extensions to these results. As noted, the pool-adjacent-violators estimator for classic current status data converges at a rate of $n^{1/3}$ with a nonstandard asymptotic limit, see a 1987 technical report by P. Groeneboom from the University of Amsterdam. We conjecture that the same asymptotics will hold for the group-tested estimator, although this remains to be established. In practice, in a setting with misclassified individual current status data, the m-out-of-n bootstrap (McKeown & Jewell, 2010) has been shown to provide one method of obtaining valid inference procedures. We look forward to further theoretical progress in this area.

It is natural to anticipate that misclassification rates may depend on group size. This may occur, for example, if the screening test is more sensitive to detecting a positive group when there are more individual positives in the pool, related to the so-called dilution effect (Hwang, 1976; McMahan et al., 2013). Second, covariate-adjusted regression analysis has been a primary focus of the statistical literature on group testing (Vansteelandt et al., 2000; Xie, 2001; Chen et al., 2009; Delaigle & Meister, 2011). In addition, in many applications, interest is focused on regression effects or group comparisons of time-to-event properties rather than on estimation of the underlying distribution function itself, often through use of standard multiplicative or additive regression models. Such regression models have been widely studied for individual current status data (Jewell & Emerson, 2013). Future work will investigate the use of additive hazard regression models for group-tested current status data.

ACKNOWLEDGEMENT

The authors thank the editor, associate editor and reviewers for their insightful feedback. This work was supported by the National Heart, Lung, and Blood Institute, U.S. National Institutes of Health.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online contains a derivation of the expectation step of our expectation-maximization pool-adjacent-violators algorithm in the presence of misclassification, results from both sets of simulations with fixed censoring times, and code needed to replicate the simulations outlined in § 5·1.

REFERENCES

AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. & SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26**, 641–7.

BANERJEE, M. (2012). Current status data in the 21st century: Some interesting developments. In *Interval-Censored Time-to-Event Data: Methods and Applications*. D. G. Chen, J. Sun and K. E. Peace, eds. Boca Raton, Florida: Chapman & Hall/CRC, pp. 45–90.

BANERJEE, M. & WELLNER, J. A. (2001). Likelihood ratio test for monotone functions. *Ann. Statist.* **29**, 1699–731.

BANERJEE, M. & WELLNER, J. A. (2005). Confidence intervals for current status data. *Scand. J. Statist.* **32**, 405–24.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. & BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions*. New York: Wiley.

Chen, P., Tebbs, J. M. & Bilder, C. R. (2009). Group testing regression models with fixed and random effects. *Biometrics* **65**, 1270–8.

Delaigle, A. & Meister, A. (2011). Nonparametric regression analysis for group testing data. *J. Am. Statist. Assoc.* **106**, 640–50.

Delaigle, A. & Hall, P. (2012). Nonparametric regression with homogeneous group testing data. *Ann. Statist.* **40**, 131–58.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* B **39**, 1–38.

Dhand, N. K., Johnson, W. O. & Toribio, J. A. L. (2010). A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size. *J. Agric. Biol. Envir. Statist.* **15**, 452–73.

Dorfman, R. (1943). The detection of defective members of large populations. *Ann. Math. Statist.* **14**, 436–40.

Groeneboom, P. & Wellner, J. A. (1992). *Nonparametric Maximum Likelihood Estimators for Interval Censoring and Deconvolution*. Boston: Birkhäuser.

Hwang, F. K. (1976). Group testing with a dilution effect. *Biometrika* **63**, 671–80.

Jewell, N. P. & Emerson, R. (2013). Current status data: An illustration with data on avalanche victims. In *Handbook of Survival Analysis*. Boca Raton, Florida: Chapman & Hall/CRC, pp. 391–412.

Jewell, N. P. & van der Laan, M. (2003). Current status data: Review, recent developments and open problems. In *Handbook in Statistics*, vol. 23. Amsterdam: Elsevier, pp. 625–42.

Liu, A., Liu, C., Zhang, Z. & Albert, P. S. (2012). Optimality of group testing in the presence of misclassification. *Biometrika* **99**, 245–51.

Maathuis, M. & Hudgens, M. G. (2011). Nonparametric inference for competing risks current status data with continuous, discrete or grouped observation times. *Biometrika* **98**, 325–40.

McKeown, K. & Jewell, N. P. (2010). Misclassification of current status data. *Lifetime Data Anal.* **16**, 215–30.

McMahan, C. S., Tebbs, J. M. & Bilder, C. R. (2013). Regression models for group testing data with pool dilution effects. *Biostatistics* **14**, 284–98.

Remlinger, K. S., Hughes-Oliver, J. M., Young, S. S. & Lam, R. L. (2006). Statistical design of pools using optimal coverage and minimal collision. *Technometrics* **48**, 133–43.

Sal y Rosas, V. G. & Hughes, J. P. (2010). Nonparametric and semiparametric analysis of current status data subject to outcome misclassification. *Statist. Commun. Inf. Dis.* **2010**, article no. 364.

Tang, R., Banerjee, M. & Kosorok, M. R. (2012). Likelihood based inference for current status data on a grid: A boundary phenomenon and an adaptive inference procedure. *Ann. Statist.* **40**, 45–72.

Tu, X. M., Litvak, E. & Pagano, M. (1994). Screening tests: Can we get more by doing less? *Statist. Med.* **13**, 1905–19.

Tu, X. M., Litvak, E. & Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika* **82**, 287–97.

Vansteelandt, S., Goetghebeur, E. & Verstraeten, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* **56**, 1126–33.

Wahed, M. A., Chowdhury, D., Nermell, B., Khan, S. I., Ilias, M., Rahman, M., Persson, L. A. & Vahter, M. (2006). A modified routine analysis of arsenic content in drinking-water in Bangladesh by hydride generation-atomic absorption spectrophotometry. *J. Health Pop. Nutr.* **24**, 36–41.

Wein, L. M. & Zenios, S. A. (1996). Pooled testing for HIV screening: Capturing the dilution effect. *Oper. Res.* **44**, 543–69.

Xie, M. (2001). Regression analysis of group testing samples. *Statist. Med.* **20**, 1957–69.

Yu, G., Schick, A., Li, L. & Wong, G. Y. C. (1998). Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. *Can. J. Statist.* **26**, 619–27.

[*Received December* 2015. *Revised July* 2016]