Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty

BARBRA A. RICHARDSON*
Department of Biostatistics, University of Washington, Box 359909, Seattle, WA 98195, USA
barbrar@u.washington.edu

JAMES P. HUGHES
Department of Biostatistics, University of Washington, Box 359931, Seattle, WA 98195, USA

SUMMARY

Low sensitivity and/or specificity of a diagnostic test for outcome results in biased estimates of the time to first event using product limit estimation. For example, if a test has low specificity, estimates of the cumulative distribution function (cdf) are biased towards time zero, while estimates of the cdf are biased away from time zero if a test has low sensitivity. In the context of discrete time survival analysis for infectious disease data, we develop self-consistent algorithms to obtain unbiased estimates of the time to first event when the sensitivity and/or specificity of the diagnostic test for the outcome is less than 100%. Two examples are presented. The first involves estimating time to first detection of HIV-1 infection in infants in a randomized clinical trial, and the second involves estimating time to first Neisseria gonorrhoeae infection in a cohort of Kenyan prostitutes.

Keywords: Diagnostic testing; Infectious diseases; Measurement error; Product limit estimation.

1. INTRODUCTION

Product limit estimation is an important statistical method for determining both the timing and cumulative incidence of a disease or condition (Kaplan and Meier, 1958). In many studies where product limit estimation is implemented, the outcome of interest can be measured with certainty. For example, in many clinical studies, the final outcome measure is time of death which, in most instances, can be measured exactly. However, in the setting of infectious disease, the outcome is often the result of a diagnostic test which is known to be less than exact. That is, the sensitivity and/or specificity of the test is less than 100%.

Diagnostic tests with less than 100% specificity will result in product limit estimates of the cumulative distribution function (cdf) that are biased towards time zero (i.e. the mass of the distribution will be shifted towards zero). As an example, in the setting of mother-to-child transmission (vertical transmission) of HIV-1, DNA polymerase chain reaction (PCR) testing is the main virologic test used to determine if vertical transmission has occurred. In many studies, the first sample for an infant is blood drawn from the
umbilical cord. This sample is very useful in determining whether virus transmission occurred in-utero, versus during or after delivery. However, because of presence of maternal blood during delivery, cord blood samples may be contaminated and thus test positive for HIV-1 even if the infant is uninfected at birth. One study estimated the specificity of PCR testing on samples of cord blood to be 95% (111/117) (95% confidence interval: 89%–98%) (Simonon et al., 1994). If the specificity of PCR testing at birth is less than 100%, then estimating the timing of HIV-1 infection becomes difficult due to the possibility of a false positive test at birth. This lower specificity of PCR testing will bias estimates of the cdf of time of infection towards the day of birth.

In contrast, diagnostic tests with less than 100% sensitivity will bias product limit estimates of the cdf away from time zero necessitating alternative statistical methods to correct the bias. For example, the sensitivity of culture of endocervical swab specimens for certain sexually transmitted diseases (STDs), specifically Neisseria gonorrhoeae (GC) infection, has very low sensitivity when compared to a ‘gold standard’ of ligase chain reaction (LCR) testing (Buimer et al., 1996; Carroll et al., 1998). In studies investigating time to first gonorrhea infection using swab culture as the diagnostic test, estimates of the cumulative incidence rate for earlier points in time will be too low because of the occurrence of false negative test results. Thus, if one were planning a study using LCR for GC diagnosis, sample size estimates based on point estimates from the cdf estimated using the less sensitive swab culture diagnostic test would be inappropriately inflated.

In the context of infectious diseases, assumptions about the natural history of the disease will affect the type of data that are available as well as the type of adjustment for misclassification that is appropriate. We concentrate on two special cases. The first is a treatable self-limiting infectious disease. In this case, the disease resolves with treatment. Examples are sexually transmitted diseases like gonorrhea and chlamydia. A key feature of data from this type of infectious disease is that no additional follow-up testing to confirm disease status is possible after disease detection, since the disease is treated and cured after the first positive test. The second case is a lifelong infectious disease. In this case, the symptoms of the disease may be treatable, but the disease itself does not resolve. An example is HIV-1 infection. Additional follow-up after first disease detection to verify disease status is possible in this case. Other situations arise in the study of infectious diseases. For example, misclassification may occur when trying to estimate the duration or inter-arrival time of recurrent or intermittent events such as herpes virus shedding. However, in this paper we concentrate on the two special cases described above.

This motivates the development of statistical methods to obtain unbiased estimates of the distribution of event times when the diagnostic test for the event is less than 100% specific and/or less than 100% sensitive. Magder and Hughes addressed the problem for logistic regression (Magder and Hughes, 1997). In this paper we develop methods for the case in which all subjects in a study are recurrently tested at discrete points in time until their first positive test. The censoring pattern of subjects is limited to right censoring, and test sensitivity and/or test specificity are less than 100%. We provide the notation and assumptions used throughout the paper in Section two. In the third, and fourth sections we outline two Expectation Maximization (EM) algorithms: first for the case without follow-up after event detection (e.g. treatable self-limiting infectious disease), and second, for the case with follow-up after event detection (e.g. a lifelong infectious disease). In Section five we explain how to obtain variance estimates of the cdf parameter estimates. We provide examples in Section six and discuss the strengths and limitations of our approach and areas for future research in Section seven. Finally, in the Appendix we provide details on the derivations of the algorithms and estimation of the variance of the parameter estimates.
2. NOTATION AND ASSUMPTIONS

For simplicity, we assume that the diagnostic test for the outcome of interest is performed at discrete points in time. This assumption is valid in the setting of infectious disease research since many prospective studies investigating infectious diseases have set testing schedules for subjects. We assume that all subjects are tested at all time points until they are either right censored (lost to follow-up, or the study ends) or have a positive diagnostic test result, and that the sensitivity and specificity of the diagnostic test do not change with time. In addition, we assume that the censoring mechanism is independent of both failure mechanism and the misclassification mechanism.

Throughout, let \( p_j \) denote the conditional probability of disease (the true event) at time \( j (j = 1, 2, \ldots, T) \) given no disease prior to time \( j \) (i.e. the hazard at time \( j \)). Let \( t_i^{obs} (i = 1, 2, \ldots, N) \) denote the observed time of outcome (event or censoring) for subject \( i \), and let \( d_i^{obs} (i = 1, 2, \ldots, N) \) indicate the observed outcome for subject \( i \) (0 = observed censored, 1 = tested positive for disease/condition). Then \( D_j = \sum_{i=1}^{N} I(t_i^{obs} = j \land d_i^{obs} = 1) \) represents the number of subjects observed to have (first) positive test results at time \( j \), \( M_j = \sum_{i=1}^{N} I(t_i^{obs} = j \land d_i^{obs} = 0) \) represents the number of subjects observed to be censored immediately after time \( j \), and \( N_j = N - \sum_{i=1}^{j-1}(D_i + M_i) \) represents the number of subjects observed to be at risk at time \( j \). Finally, let \( \phi \) represent the specificity (0.0 \leq \phi \leq 1.0) and \( \gamma \) represent the sensitivity (0.0 \leq \gamma \leq 1.0) of the diagnostic test.

In the case of a lifelong infectious disease (e.g. HIV) and an imperfect diagnostic screening test, trial designs often call for positive test results to be followed at the next time point by a confirmatory test. If the confirmatory test is positive, then the subject is assumed to have the disease and follow-up testing ceases. If the confirmatory test is negative, the subject is assumed not to have the disease, and continues on in the study. Let \( b_i \) \( (i = 1, 2, \ldots, N) \) indicate a positive confirmatory test result for subject \( i \) at time \( t_i^{obs} + 1 \) (0 = confirmatory test not done, 1 = confirmatory test was positive for disease/condition). Then, \( B_j = \sum_{i=1}^{N} I(t_i^{obs} = j \land b_i = 1) \) represents the number of confirmed positive tests at time \( j \), and \( D_j - B_j \) represents the number of unconfirmed positive tests at time \( j \). Note that if a confirmatory test is negative, positive results at prior time points are ignored, and follow-up testing continues.

3. TREATABLE SELF-LIMITING INFECTIOUS DISEASE

We approach the problem of less than 100% specificity and/or sensitivity of the outcome measure as one of incomplete data, meaning that the observed outcomes are only a portion of the true information since they could be false positive or false negative test results. Using available information about the specificity and sensitivity of the diagnostic test for the outcome, we outline a self-consistent algorithm to obtain unbiased estimates of the survival distribution in the presence of this incomplete data (Dempster et al., 1977). In this case of a treatable self-limiting infectious disease, subjects are not tested after their first positive test since, after a positive test result, they are treated to resolve the disease. Individuals with false positive tests are effectively censored. Note that we assume here that the probability that a treatable self-limiting infectious disease resolves without treatment is zero. Details of the derivation of this algorithm are presented in the Appendix.

3.1. EM algorithm

Step 1. Obtain initial estimates \( \hat{p}_j^0 \) of the \( p_j \) \( (j = 1, 2, \ldots, T) \). One could use the usual Kaplan Meier estimates assuming the sensitivity and specificity of the test for the outcome are 100% plus an adjustment factor so that all starting estimates are greater than zero: \( \hat{p}_j^0 = (D_j/N_j) + \epsilon_1 \) where \( \epsilon_1 \) is a very small positive number (e.g. \( \epsilon_1 = 10^{-5} \)).
Step 2 (E-step). At the kth iteration compute

\[
\hat{u}_j^k = \frac{\hat{a}_j^k}{\hat{a}_j^k + (1 - \gamma) \sum_{l=1}^{j} \hat{g}_j^k} \quad (j = 1, 2, \ldots, T)
\]

\[
\hat{w}_{jm}^k = \frac{1 - \gamma \hat{g}_{jm}^k}{\hat{a}_j^k + (1 - \gamma) \sum_{l=1}^{j} \hat{g}_j^k} (j = 1, 2, \ldots, T) \quad (m = 1, 2, \ldots, j)
\]

\[
\hat{v}_{jm}^k = \frac{\hat{g}_{jm}^k}{\sum_{l=1}^{j} \hat{g}_j^k + \frac{1 - \phi}{\phi} \hat{a}_j^k} (j = 1, 2, \ldots, T) \quad (m = 1, 2, \ldots, j)
\]

and

\[
\hat{z}_j^k = \frac{1 - \phi \hat{a}_j^k}{\frac{1 - \phi}{\phi} \hat{a}_j^k + \sum_{l=1}^{j} \hat{g}_j^k} (j = 1, 2, \ldots, T)
\]

where

\[
\hat{a}_j^k = \phi^j \prod_{l=1}^{j} (1 - \hat{p}_l^k)
\]

\[
\hat{g}_{jm}^k = \gamma (1 - \gamma)^{j-1} \hat{p}_m^k \quad m = 1
\]

\[
= \gamma \phi^{m-1} (1 - \gamma)^{j-m} \hat{p}_m^k \prod_{k=1}^{m-1} (1 - \hat{p}_k^k) \quad 1 < m \leq j
\]

\[
= 0 \quad m > j.
\]

Here \(\hat{u}_j^k\) is an estimate of the conditional probability of being truly uninfected at time \(j\) given that a negative test result was observed at time \(j\), and \(\hat{w}_{jm}^k\) is an estimate of the conditional probability of being truly infected at time \(m\) given that a negative test result was observed at time \(j \geq m\) (and all previous test results were negative). Similarly, \(\hat{v}_{jm}^k\) is an estimate of the conditional probability of being infected at time \(m\) given that a positive result was first observed at time \(j \geq m\) (and all previous test results were negative), and \(\hat{z}_j^k\) is an estimate of the probability of being truly uninfected at time \(j\) given that a positive test result was observed at time \(j\).

Step 3 (M-step). At the kth iteration compute:

\[
\hat{p}_{j+1}^k = \frac{\sum_{t=1}^{T} D_t \hat{u}_{jt} + M_t \hat{w}_{jt}^{k+1}}{\sum_{t=1}^{T} D_t (\hat{z}_t^k + \sum_{l=j}^{T} \hat{v}_{lt}) + M_t (\hat{a}_t^k + \sum_{l=j}^{T} \hat{w}_{lt}^{k+1})}
\]

Step 4. Return to Step 2 and compute the \(\hat{u}_{j+1}^k\) and \(\hat{z}_{j+1}^k\) for \(j = 1, 2, \ldots, T\) and \(\hat{v}_{jm}^k\) and \(\hat{w}_{jm}^{k+1}\) for \(j = 1, 2, \ldots, T\) and \(m \leq j\) by replacing the \(\hat{p}_j^k\) with the \(\hat{p}_{j+1}^k\) computed in Step 3. Then repeat Step 3 and continue to iterate until \(\hat{p}_j^k - \hat{p}_{j+1}^k < \epsilon\) for \(\epsilon\) very small (e.g. \(\epsilon = 10^{-7}\)). Take \(\hat{p}_j = \hat{p}_{j+1}^k\).

Step 5. Compute the estimate of the cumulative distribution function:

\[
\hat{F}(j) = 1 - \prod_{m=1}^{j} (1 - \hat{p}_m).
\]
4. Lifelong Infectious Disease

Recall that for this case, individuals with positive tests are re-tested at the next time point to confirm their positive result, unless they are censored in between the two time points. It is assumed that the confirmatory test is 100% specific and 100% sensitive. In general, most confirmatory tests for infectious diseases are highly sensitive and specific to detect all false positive screening results. The main reasons that the confirmatory test is not used for initial screening are generally administrative (e.g. higher cost and labor, or the test is more difficult to implement in a particular setting). Hence, in the area of infectious disease research, the assumption that the confirmatory test for a lifelong infectious disease is 100% sensitive and 100% specific is reasonable. Details of the derivation of this algorithm are presented in the Appendix.

4.1. EM algorithm

Step 1. Obtain initial estimates \( \hat{p}_j^0 \) of the \( p_j \) (\( j = 1, 2, \ldots, T \)) as described in Section 3.1, Step 1.

Step 2 (E-step). At the \( k \)th iteration compute \( \hat{\phi}_j^k, \hat{\psi}_jm^k, \hat{\psi}_jm^k, \) and \( \hat{\psi}_j^k \) as explained in Section 3.1, and also compute

\[
\hat{\psi}_jm^k = \frac{1 - \varphi}{\varphi} \hat{\phi}_j^k \hat{p}_j^{k+1} + \sum_{l=1}^{J} \hat{\psi}_jl^k
\]

and

\[
\hat{\psi}_j^k = \frac{1 - \varphi}{\varphi} \hat{\phi}_j^k \hat{p}_j^{k+1} + \sum_{l=1}^{J} \hat{\psi}_jl^k
\]

for \( j = 1, 2, \ldots, T \) and \( m \leq j \). Here, \( \hat{\phi}_j^k, \hat{\psi}_jm^k, \) and \( \hat{\psi}_j^k \) have the interpretation given in Section 3.1, while \( \hat{\psi}_jm^k \) is an estimate of the conditional probability of being truly infected at time \( m \) given that a positive test result was observed at time \( j \geq m \) and no confirmatory testing was done. \( \hat{\psi}_j^k \) is an estimate of the conditional probability of being truly infected at time \( m \) given that a positive test result was observed at time \( j \geq m \) and a positive confirmatory test was observed at time \( j \) at time \( j + 1 \). \( \hat{\psi}_j^k \) is an estimate of the conditional probability of being truly negative at time \( j \) given that a positive test result was observed at time \( j \), and a positive confirmatory test was observed at time \( j + 1 \).

Step 3 (M-step). At the \( k \)th iteration compute

\[
\hat{p}_1^{k+1} = \frac{\sum_{l=1}^{J} ((D_l - B_l) (\hat{\psi}_lm^k + \hat{\psi}_lm^j + \hat{\psi}_lm^k + \hat{\psi}_lm^k))}{\sum_{l=1}^{J} ((D_l - B_l) (\hat{\psi}_lm^k + \hat{\psi}_lm^j + \hat{\psi}_lm^k)) + B_l (\hat{\psi}_lm^k + \sum_{l=1}^{J} \hat{\psi}_lm^k)}
\]

\[
\hat{p}_j^{k+1} = \frac{\sum_{l=1}^{J} ((D_l - B_l) (\hat{\psi}_lm^k + \hat{\psi}_lm^j + \hat{\psi}_lm^k + \hat{\psi}_lm^k))}{\sum_{l=1}^{J} ((D_l - B_l) (\hat{\psi}_lm^k + \hat{\psi}_lm^j + \sum_{l=1}^{J} \hat{\psi}_lm^k)) + B_l (\hat{\psi}_lm^k + \sum_{l=1}^{J} \hat{\psi}_lm^k + \hat{\psi}_lm^k)}
\]

for \( j = 2, 3, \ldots, T \).

Step 4. Return to Step 2 and compute the \( \hat{\psi}_jm^{k+1}, \hat{\psi}_jm^{k+1}, \hat{\psi}_jm^{k+1}, \hat{\psi}_jm^{k+1}, \) and \( \hat{\psi}_j^{k+1} \) (\( j = 1, 2, \ldots, T \) and \( m \leq j \)) by replacing the \( \hat{p}_j^k \) with the \( \hat{p}_j^{k+1} \) computed in Step 3. Then repeat Step 3 and continue to iterate until \( |\hat{p}_j^k - \hat{p}_j^{k+1}| < \varepsilon \) for \( \varepsilon \) small (e.g. \( \varepsilon = 10^{-7} \)). Take \( \hat{p}_j = \hat{p}_j^{k+1} \).
Step 5. Compute the estimate of the cumulative distribution function:

$$\hat{F}(j) = 1 - \prod_{m|j} (1 - \hat{p}_m).$$

5. Estimation of standard errors

We use the supplemented EM (SEM) algorithm described by Meng and Rubin (1991) to obtain estimates of the variance–covariance matrices of the hazard estimates in our two EM algorithms. We then use these estimates to obtain variance estimates for the cumulative distribution function estimates. Throughout, let \(D^*_j\) represent the number of subjects truly infected at time \(j\), \(N^*_j\) represent the number of subjects truly at risk of infection at time \(j\), and \(p^*_j\) represent the estimate of \(p^*_j\) obtained at convergence of the EM algorithm.

5.1. Treatable self-limiting infectious disease

Let \(I_c = \sum_{j=1}^T [N^*_j + \ln(1 - p^*_j) + D^*_j \ln p^*_j + M^*_j \ln(1 - p^*_j)]\) (the complete data log-likelihood). Then, \(I_c(p) = \frac{\hat{L}_c}{\hat{p} \hat{p}}\), and

$$I_{oc}(p) = E[-I_c(p) \mid D, N, M, p, \phi, \gamma]$$

evaluated at \(p = p^*\). \(I_{oc}(p)\) is a \(T\) by \(T\) diagonal matrix with diagonal elements:

$$[I_{oc}(p)]_{jj} = \frac{E(D^*_j \mid D, N, M, p, \phi, \gamma)(1 - 2p^*_j)}{p^*_j(1 - p^*_j)^2} + \frac{E(N^*_j \mid D, N, M, p, \phi, \gamma)}{(1 - p^*_j)^2}$$

$$\sum_{i=j}^T (D_i \hat{v}^*_i + M_i \hat{w}^*_i)(1 - 2p^*_j) \frac{1}{p^*_j(1 - p^*_j)^2} + \frac{\hat{M}_i \hat{w}^*_i}{(1 - p^*_j)^2}$$

for \(j = 1, 2, \ldots, T\), and \(\hat{v}^*_i\) and \(\hat{w}^*_i\) are defined in Section 3.1 and estimated at \(p^*_j\). The variance–covariance matrix of the hazard ratios \(p\) is then \(\text{Var}(p) = I_{oc}^{-1} + I_{oc}^{-1} \hat{R} (I - \hat{R})^{-1}\) where \(\hat{R}\) is estimated using the SEM algorithm as described in Meng and Rubin (1991) and in the Appendix. Given \(V\), \(\text{Var}(\hat{F}(j)) = \hat{F}^2(j) \sum_{|t|<j}(1 - \hat{p}_t)\text{Var}(\hat{p}_j)\).

5.2. Lifelong infectious disease

In this case, \(I_{oc}(p)\) is a \(T\) by \(T\) diagonal matrix with diagonal elements:

$$[I_{oc}(p)]_{jj} = \frac{E(D^*_j \mid D, B, N, M, p, \phi, \gamma)(1 - 2p^*_j)}{p^*_j(1 - p^*_j)^2} + \frac{E(N^*_j \mid D, B, N, M, p, \phi, \gamma)}{(1 - p^*_j)^2}$$

$$\sum_{i=j}^T (D_i \hat{v}^*_i + B_i \hat{\tilde{v}}^*_i + M_i \hat{z}^*_i)(1 - 2p^*_j) \frac{1}{p^*_j(1 - p^*_j)^2} + \frac{\hat{M}_i \hat{z}^*_i}{(1 - p^*_j)^2}$$

for \(j = 1, 2, \ldots, T\), where \(E(D^*_j \mid D, \ldots) = \sum_{i=1}^T [(D_i - B_i) \hat{v}^*_i + B_i \hat{\tilde{v}}^*_i + M_i \hat{w}^*_i], E(D^*_j \mid D, \ldots) = B_j \hat{\tilde{v}}^*_j - 1 + \sum_{i=1}^T (D_i - B_i) \hat{v}^*_i + B_i \hat{\tilde{v}}^*_i + M_i \hat{w}^*_i \) \((j = 2, 3, \ldots, T), E(N^*_j \mid D, \ldots) = E(D^*_j \mid D, \ldots) / p^*_j \) \((j = 1, 2, \ldots, T)\), and \(\hat{v}^*_i, \hat{\tilde{v}}^*_i, \hat{\tilde{v}}^*_i, \hat{w}^*_i, \) and \(\hat{z}^*_i\) are as defined in Section 4.1 and estimated at \(p^*_j\). The variance–covariance matrix of the hazard ratios \(p\) is then \(V = I_{oc}^{-1} + I_{oc}^{-1} \hat{R} (I - \hat{R})^{-1}\) where \(\hat{R}\)
Table 1. Unadjusted and adjusted estimates of the CDF of time to detection of HIV-1 infection in infants in the Nairobi breast feeding trial

<table>
<thead>
<tr>
<th></th>
<th>Breast feeders</th>
<th>Breast feeders adjusted</th>
<th>Formula feeders</th>
<th>Formula feeders</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unadjusted estimate (95% CI)</td>
<td>Estimate (95% CI)</td>
<td>Unadjusted estimate (95% CI)</td>
<td>Adjusted estimate (95% CI)</td>
</tr>
<tr>
<td>Birth</td>
<td>0.088 (0.033, 0.143)</td>
<td>0.078 (0.021, 0.136)</td>
<td>0.017 (0.000, 0.040)</td>
<td>0.014 (0.000, 0.035)</td>
</tr>
<tr>
<td>6 Weeks</td>
<td>0.216 (0.136, 0.296)</td>
<td>0.213 (0.124, 0.301)</td>
<td>0.079 (0.030, 0.129)</td>
<td>0.080 (0.027, 0.132)</td>
</tr>
</tbody>
</table>

is estimated using the EM algorithm as described in Meng and Rubin (1991) and in the Appendix. Given $V$,

$$\text{Var}(\hat{F}(j)) = \hat{F}(j)^2 \sum_{i|t<j} (1 - \hat{p}_t) \text{Var}(\hat{p}_j).$$

### 6. EXAMPLES

We present two examples. The first is a lifelong infectious disease (HIV) with a nonspecific diagnostic test, and the second is a treatable self-limiting infectious disease (gonorrhea) with a nonsensitive diagnostic test.

#### 6.1. Lifelong infectious disease example

These data are from a randomized clinical trial of breast feeding versus formula feeding in HIV-1 infected pregnant women in Nairobi, Kenya. The specific aim of the study was to determine the frequency of HIV-1 transmission through breastmilk. Four hundred and twenty five HIV-1 seropositive women were enrolled at approximately 32 weeks gestation and randomized to breast feed or formula feed their infants. Infants were tested by PCR for HIV-1 DNA at birth (cord blood), 6 weeks, 14 weeks, 6 months, and then every 3 months up to 24 months. Further details of the study are available in Nduati et al. (2000).

To illustrate our proposed methodology and avoid the complications associated with longer follow-up periods (see Hughes and Richardson (1999)), this analysis of time to detection of HIV-1 includes only those infants who had PCR testing at birth, and only includes infections detected at birth or at 6 weeks (thus, the results of this subset analysis differ slightly from the results reported in Nduati et al. (2000)).

To obtain less biased estimates of the cdf of time to detection of HIV-1 we set the specificity ($\phi$) and sensitivity ($\gamma$) of the PCR test at birth on cord blood samples equal to 95% and 100%, respectively, and assumed the specificity and sensitivity of the PCR test at 6 weeks (confirmatory test) was 100% (Simonon et al., 1994). In addition, we include a variable in the analysis that indicates whether or not a positive test was confirmed at the next time point. Table 1 shows the adjusted and unadjusted estimates for the breast feeding and formula feeding groups. As expected, due to some false positive PCR results on cord blood samples at birth, the adjusted estimates indicate that fewer infections occurred at birth compared to the unadjusted estimates.

#### 6.2. Treatable self-limiting infectious disease example

Data are from a prospective cohort study of HIV-1 seronegative prostitutes in Mombasa, Kenya. This cohort was formed in 1993 as part of the Preparation for AIDS Vaccine Evaluations (PAVE) initiative
Fig. 1. Unadjusted and adjusted estimates of cumulative probability of first GC infection (error bars = 1 standard error).

and continued as part of the HIV Network for Prevention Trials (HIVNET) program of the United States National Institutes of Health. Study population characteristics and details of study procedures have been reported previously (Martin et al., 1998). Briefly, women seronegative for HIV-1 were followed monthly and tested for HIV-1 and other sexually transmitted diseases to determine the HIV-1 incidence rate and correlates of seroconversion. At each visit, women underwent pelvic examinations during which endocervical swab specimens were collected for culture of Neisseria gonorrhoeae (GC) infection, which is often asymptomatic in women. Data are from the period between February 1993 and March 1999.

This analysis of time to first GC infection includes all women in the cohort until they were lost to follow-up or their last study visit (i.e. were censored), or until they tested positive for GC. Buimer et al. estimate that the sensitivity of culture for gonorrhea performed on cervical swab specimens is 50.0% versus a gold standard of a positive ligase chain reaction (LCR) or positive culture, while Carroll et al. estimate the sensitivity to be 58.3% when compared to a gold standard of at least two positive results for LCR on urine, LCR on swab, and culture on swab (Buimer et al., 1996; Carroll et al., 1998). To obtain less biased estimates of the cdf of time to first GC infection in this population of prostitutes, we set the sensitivity ($\gamma$) equal to 55% and specificity ($\phi$) equal to 100% in the EM algorithm. Figure 1 shows the unadjusted product limit estimates, assuming sensitivity of endocervical swab cultures for GC is 100%,
Table 2. Hypothetical study: inflation of sample size estimates based on unadjusted estimates of time to first GC infection in the Mombasa prostitute cohort

<table>
<thead>
<tr>
<th>Months of follow-up per woman</th>
<th>Percent inflation of sample size using unadjusted estimates (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
</tr>
</tbody>
</table>

and the adjusted estimates assuming a sensitivity of 55%. As expected, the adjusted estimates indicate more infections occurred at earlier time points than were actually observed, due to the low sensitivity of the test. Note that even the adjusted estimates may be biased if some of the GC infections self-resolved and were never detected.

To illustrate the effects of low sensitivity of culture for GC, we propose a hypothetical example of a randomized intervention study in this population that we hypothesize will decrease the incidence of GC, and where LCR will be used for testing. Assuming a 1 : 1 randomization scheme and a two-sided test, the sample size for such a study with follow-up of length \( j \) can be summarized by the total number of events, \( n \), necessary to detect a particular difference between two groups given the \( \beta \)-level and \( \alpha \)-level of the test (George and Desu, 1974). Given \( n \), along with \( \hat{F}^{\text{unadj}}(j) \) and \( \hat{F}^{\text{adj}}(j) \), the unadjusted and adjusted cumulative incidence rate at time \( j \), respectively, the inflation factor for the number of person years needed to see \( n \) events if one were to use the unadjusted rates to estimate the sample size is given by \( \hat{F}^{\text{adj}}(j)/\hat{F}^{\text{unadj}}(j) \). Table 2 gives the percent inflation of the sample size for such an intervention study for various follow-up periods \( (j) \). From this example, it is evident that if such a study were designed based on the unadjusted estimates of the time to first GC infection shown in Figure 1, the sample size for this hypothetical study would be inflated anywhere from 2% to 80%, depending on the amount of time each woman was followed.

To illustrate the effects of <100% specificity of the diagnostic test for GC in this example, we estimated the cdf assuming the sensitivity of the GC culture was 55% and varied the specificity in the algorithm from 100% to 96%. In Figure 2 we see that in this setting, where the sensitivity of the diagnostic test is low,
Fig. 2. Adjusted estimates of cumulative probability of first GC infection with sensitivity = 55%, and varying specificity.

and where the observed hazard of acquiring GC at each testing time is low (range: 0.01–0.06), that even a 1% change in the specificity of the test can have a large effect on the estimates of the cdf.

7. DISCUSSION

We have presented new methods for obtaining product limit estimates when the diagnostic test for an infectious disease is less than 100% sensitive and/or less than 100% specific. The outlined methods are useful for obtaining less biased estimates of the cdf of the time to first event when the outcome is measured with uncertainty. For example, these methods could be used in sample size calculations when one is planning to implement a study using a more sensitive diagnostic test using preliminary data based upon a less sensitive test.

Magder and Hughes (1997) used a similar approach (EM algorithm) to estimate parameters in a logistic regression when the outcome is measured with uncertainty. In the context of infectious diseases, the analysis methods developed by Magder and Hughes (1997) are applicable when investigating the prevalence of a disease, for instance in a cross-sectional study. In contrast, the methods developed in this paper are applicable to longitudinal studies investigating the incidence of a disease.
While the methods proposed here are useful, there are several limitations. First, our methods do not allow for recurring events, but instead, only allow for estimation of the cdf of the time to first event. Recurring event data is common in infectious disease research. In our example of a treatable self-limiting infectious disease, we only analyzed women up to their first occurrence of GC. It could be of equal interest to analyze these data including all occurrences of GC for all women, which would require development of methods for recurring events. However, when one looks at recurrent events, appropriate adjustments for misclassification will, in general, depend on assumptions regarding the natural history of disease. Second, we do not address the case of a non-treatable self-limiting infectious disease. This is again a more complex situation, and an area for future research. Finally, these methods could be extended to continuous time, although such data would generally be interval censored, as is the case for the example data from the Nairobi Breastfeeding Trial, if we were estimating time of infection rather than time to detection. Extension of our methods to interval censored data situations would be useful, but is not trivial.

With the constant development of more sensitive and specific diagnostic tests for diseases, the statistical methods developed here are of increasing use to data analysts and clinicians. However, much work remains to develop methods that can be implemented in more complex data situations.

ACKNOWLEDGEMENTS

This work was supported in part by University of Washington Royalty Research Grant #1779, by NIAID grant N01-AI45200, and by NIH grants R01-AI39932 and 2-R01-AI29168-10. We would like to thank Dr Joan Kreiss for the use of the data from the Mombasa cohort, and Drs Joan Kreiss and Ruth Nduati for the data from the Nairobi Breastfeeding Trial for the examples. Finally, we would like to thank Maureen Reiner for assistance with programming and analyses.

APPENDIX: DERIVATION OF EM ALGORITHMS

Let \( t_i \) \((i = 1, 2, \ldots, N)\) denote the true time of outcome (event or censoring) for subject \( i \), and let \( d_i \) \((i = 1, 2, \ldots, N)\) indicate the true outcome for subject \( i \): 0 = truly censored, 1 = truly positive for disease/condition being tested. Then using notation outlined in Section 2 above, the complete data log likelihood for subject \( i \) is:

\[
\ell_i = \sum_{j=1}^{t_i-1} \ln(1 - p_j) + d_i \ln(p_i) + (1 - d_i) \ln(1 - p_i)
\]

and the total complete data log likelihood is then \( l = \sum_{i=1}^{N} \ell_i \). Let \( \hat{p}^0 = (\hat{p}_1^0, \hat{p}_2^0, \ldots, \hat{p}_T^0) \) denote the starting estimates of the \( p_j \).

A.1. Derivation of the EM algorithm for treatable self-limiting infectious disease

**E-Step** Using the notation above and in Sections 2 and 3.1, the expected value of \( l \) given the observed data is:

\[
\sum_{j=1}^{T} \left[ D_j \left( \sum_{m=1}^{j} \ln(p_m) + \sum_{l=1}^{m-1} \ln(1 - p_l) \right) \hat{v}_{jm}^0 \right] + D_j \left( \sum_{m=1}^{j} \ln(1 - p_m) \hat{z}_{jm}^0 \right)
\]

\[
+ M_j \left( \sum_{m=1}^{j} \ln(p_m) + \sum_{l=1}^{m-1} \ln(1 - p_l) \right) \hat{w}_{jm}^0 + M_j \left( \sum_{m=1}^{j} \ln(1 - p_m) \hat{u}_{jm}^0 \right)
\]

\[
\sum_{j=1}^{T} \left[ D_j \left( \sum_{m=1}^{j} \ln(p_m) + \sum_{l=1}^{m-1} \ln(1 - p_l) \right) \hat{v}_{jm}^0 \right] + D_j \left( \sum_{m=1}^{j} \ln(1 - p_m) \hat{z}_{jm}^0 \right)
\]

\[
+ M_j \left( \sum_{m=1}^{j} \ln(p_m) + \sum_{l=1}^{m-1} \ln(1 - p_l) \right) \hat{w}_{jm}^0 + M_j \left( \sum_{m=1}^{j} \ln(1 - p_m) \hat{u}_{jm}^0 \right)
\]
Finally, setting these derivatives equal to zero, and solving for the \( p_j \) results in the estimates:

\[
\hat{p}_j = \frac{\sum_{i=1}^{T} (D_{ij}v^0_i + M_j\hat{w}^0_j)}{\sum_{i=1}^{T} (D_{ij}v^0_i + M_j\hat{w}^0_j)}
\]

### A.2. Derivation of the EM algorithm for lifelong infectious disease

#### E-Step

Using the notation above, and in Sections 2 and 4.1, the expected value of \( l \) given the observed data is:

\[
\sum_{j=1}^{T} \left[ (D_j - B_j) \left( \sum_{m=1}^{j} \left( \ln(p_m) + \sum_{i=1}^{m-1} \ln(1 - p_i) \right) v^0_{jm} \right) + (D_j - B_j) \left( \sum_{m=1}^{j} \ln(1 - p_m) \hat{z}^0_j \right) \right]
\]

+ \[
B_j \left( \sum_{m=1}^{j} \left( \ln(p_m) + \sum_{i=1}^{m-1} \ln(1 - p_i) \right) E^0_{jm} \right) + B_j \left( \ln(p_{j+1}) + \sum_{m=1}^{j} \ln(1 - p_m) s^0_j \right)
\]

+ \[
M_j \left( \sum_{m=1}^{j} \left( \ln(p_m) + \sum_{i=1}^{m-1} \ln(1 - p_i) \right) \hat{w}^0_{jm} \right) + M_j \left( \sum_{m=1}^{j} \ln(1 - p_m) \hat{u}^0_j \right)
\]

#### M-Step

Taking the derivatives of (1) with respect to the \( p_j \) for \( j = 1, 2, \ldots, T \) we get:

\[
\delta E(l \mid d^{\text{obs}}, t^{\text{obs}}, \phi, \gamma, \hat{p}^0) \over \delta p_j
\]

\[
= \sum_{i=j}^{T} \frac{D_{ij}v^0_i}{p_j} - \sum_{i=j+1}^{T} \frac{D_{ij}v^0_i}{1 - p_j} - \sum_{i=j}^{T} \frac{D_{ij}z^0_i}{1 - p_j}
\]

+ \[
\sum_{i=j}^{T} \frac{B_{ij}\hat{v}^0_{ij}}{p_j} - \sum_{i=j+1}^{T} \frac{B_{ij}\hat{z}^0_i}{1 - p_j} - \sum_{i=j}^{T} \frac{B_{ij}\hat{s}^0_i}{1 - p_j} + \frac{B_{ij}\hat{s}^0_{i-1}}{p_j}
\]

+ \[
\sum_{i=j}^{T} \frac{M_j\hat{w}^0_{ij}}{p_j} - \sum_{i=j+1}^{T} \frac{M_j\hat{v}^0_{ij}}{1 - p_j} - \sum_{i=j}^{T} \frac{M_j\hat{u}^0_i}{1 - p_j}
\]

Finally, setting these derivatives equal to zero, and solving for the \( p_j \), results in the estimates:
Finally, setting these derivatives equal to zero, and solving for the $p_j$, results in the estimates:

$$ \hat{p}_{1}^{k+1} = \frac{\sum_{i=1}^{T} [(D_i - B_i) \hat{v}_{1i}^k + B_i \hat{b}_{1i}^k + M_i \hat{u}_{1i}^k]}{\sum_{i=1}^{T} [(D_i - B_i) \hat{v}_{1i}^k + B_i (\hat{v}_{1i}^k + \sum_{l=2}^{T} \hat{v}_{li}^k) + M_i (\hat{u}_{1i}^k + \sum_{l=2}^{T} \hat{u}_{li}^k)]} $$

$$ \hat{p}_{j}^{k+1} = \frac{B_{j-1} \hat{v}_{ji}^{k+1} + \sum_{l=j}^{T} [(D_i - B_i) \hat{v}_{li}^k + B_i (\hat{v}_{li}^k + \sum_{l=j+1}^{T} \hat{v}_{li}^k) + M_i (\hat{u}_{li}^k + \sum_{l=j+1}^{T} \hat{u}_{li}^k)]}{B_{j-1} \hat{v}_{ji}^{k+1} + \sum_{l=j}^{T} [(D_i - B_i) \hat{v}_{li}^k + B_i (\hat{v}_{li}^k + \sum_{l=j+1}^{T} \hat{v}_{li}^k) + M_i (\hat{u}_{li}^k + \sum_{l=j+1}^{T} \hat{u}_{li}^k)]} $$

for $j = 2, 3, \ldots, T$.

### A.3. Derivation of estimate of $R$ for variance estimation

For each EM algorithm an estimate of the matrix $R$ for estimation of the variance estimates is obtained by first running the EM algorithm to get $p^*$ (the vector of estimates of $p$ at convergence of the EM algorithm), and saving the estimates of $p$ at each iteration of the algorithm (i.e. $\hat{p}_0, \hat{p}_1, \ldots$). Then, the $k$th iteration of the SEM algorithm is performed by following these steps for each $i = 1, 2, \ldots, T$:

1. Calculate $\hat{p}^{sk}(i) = (p^{*}_1, p^{*}_2, \ldots, p^{*}_{i-1}, \hat{p}_i^k, p^{*}_{i+1}, \ldots, p^{*}_T)$ (i.e. all elements of this vector are set equal to the final EM estimates, except the $i$th element which is set at $\hat{p}_i^k$, the estimates of $p_i$ saved at the $k$th iteration of the EM algorithm). Treating $\hat{p}^{sk}(i)$ as the current estimate of $p$, run one iteration of the EM algorithm to obtain $\hat{p}^k(i)$.

2. For $j = 1, 2, \ldots, T$, Calculate the ratios:

$$ r_{ij}^k = \frac{\hat{p}_j^k(i) - p^{*}_j}{\hat{p}_i^k - p^{*}_i} $$

3. To obtain the final estimate of $r_{ij}$, repeat steps 1 and 2 until $|r_{ij}^{t+1} - r_{ij}^t| < \sqrt{\varepsilon}$ for some $t^*$, where $\varepsilon$ is the convergence value used in the EM algorithm.

$\hat{R}$ is the $T$ by $T$ matrix comprised of the final estimates of the $r_{ij}$ for $i = 1, 2, \ldots, T$ and $j = 1, 2, \ldots, T$.

### References


[Received July 22, 1999; first revision January 21, 2000; second revision February 28, 2000; accepted for publication March 22, 2000]