Biometrics 60, 335–343
June 2004

Analyzing Time-to-Event Data in a Clinical Trial When an Unknown Proportion of Subjects Has Experienced the Event at Entry

Raji Balasubramanian* **and Stephen W. Lagakos**

Department of Biostatistics, Harvard University School of Public Health, Huntington Avenue, HIV+
Boston, Massachusetts 02115, U.S.A.
**email*: rbalasub@hsph.harvard.edu

Summary. In some clinical trials, where the outcome is the time until development of a silent event, an unknown proportion of subjects who have already experienced the event will be unknowingly enrolled due to the imperfect nature of the diagnostic tests used to screen potential subjects. For example, commonly used diagnostic tests for evaluating HIV infection status in infants, such as DNA PCR and HIV Culture, have low sensitivity when given soon after infection. This can lead to the inclusion of an unknown proportion of HIV-infected infants into clinical trials aimed at the prevention of transmission from HIV-positive mothers to their infants through breastfeeding. The infection status of infants at the end of the trial, when they are more than a year of age, can be determined with certainty. For those infants found to be infected with HIV at the end of the trial, it cannot be determined whether this occurred during the study or whether they were already infected when they were enrolled. In these settings, estimates of the cumulative risk of the event by the end of the study will overestimate the true probability of event during the study period and hypothesis tests comparing two or more intervention strategies can also be biased. We present inference methods for the distribution of time until the event of interest in these settings, and investigate issues in the design of such trials when there is a choice of using both imperfect and perfect diagnostic tests.

Key words: Clinical trials; Imperfect diagnostic tests; Time-to-event methods.

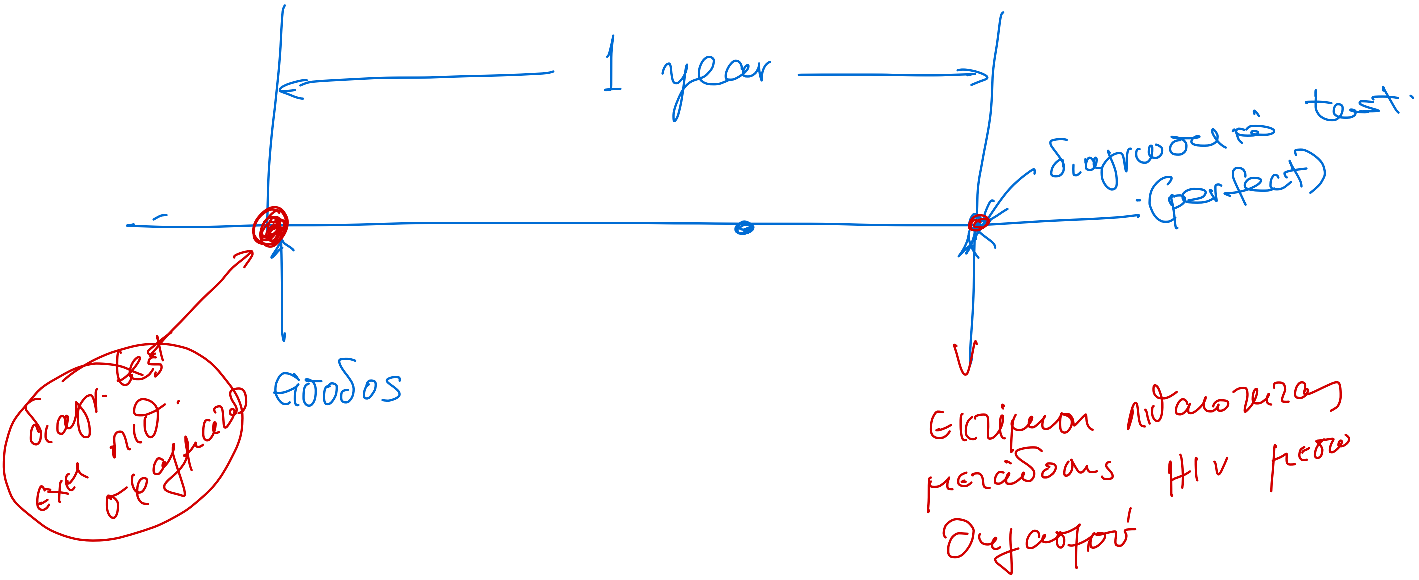
1. Introduction

Often clinical trials are designed to evaluate the time to some silent event among subjects who have not experienced the event of interest at the time of enrollment into the trial. However, in some cases, the test used to screen patients for entry is imperfect, so that some patients who have already had the event will unintentionally be enrolled. If a perfect diagnostic test were available at the end of the trial, then a patient found not to have had the event by study's end could not have had it at study entry. However, if at the end of the trial a patient is found to have had the event, then we cannot determine whether it occurred before entry or during the trial. Thus, the proportion of all subjects who have had the event at the end of the study overestimates the probability of developing the event during the study.

Our interest in this problem arose from our participation in a clinical trial currently underway in Botswana that includes a comparison of two strategies for preventing HIV infection in infants during their first 18 months of life. Newborns of HIV-infected mothers are screened for HIV using standard diagnostic tests (DNA PCR) and those infants who test negative for HIV are randomly assigned to one of the prevention strategies and then followed. The sensitivity of these diagnostic tests is known to be low for several days following infection (Dunn et al., 1995, 2000), and thus infants who became infected in uterus or during birth can be unknowingly enrolled into the trial. Suppose that all infants are tested for HIV infection at 18 months of age, using a definitive diagnostic test such as ELISA/Western Blot. Then the proportion of infants found to be infected by that time would overestimate the probability of HIV transmission during the study period. Tests comparing intervention strategies can also give biased results.

Another example of the setting considered in this article arises in some studies of Hepatitis B and Leishmaniasis, which cause liver and spleen disorders, respectively (Sundar et al., 2002; Hadziyannis et al., 2003). Such studies might require that, at entry, all subjects are in a particular disease state, such as no fibrosis or presenting evidence of Leishmaniasis; however, a definitive test to assess this, such as liver biopsy or splenic aspirate, carries some risk and inconvenience to the patients. Thus, some patients might be screened for eligibility using a less invasive yet imperfect diagnostic test (Harith et al., 1986; Martin and Friedman, 1998).

The purpose of this article is to develop statistical methods for clinical trials where the endpoint of interest is the time until some silent event whose occurrence is assessed only periodically and with possibly imperfect diagnostic tests, so that some unrecognized patients may have already had the event at the time they are enrolled. Because, as in the second example, both an imperfect and a perfect diagnostic test might sometimes be available, we also examine the relative efficiency of study designs that use different proportions of perfect versus imperfect diagnostic tests, as the increased risk


1 year
διαγνωστικό test: (perfect)
εισοδος
διαγν. test εχει πιθ. σφαλματο
Εκτίμηση πιθανοτητας μετάδοσης HIV μεσω θηλασμού

or inconvenience of the former might outweigh their better diagnostic properties in some settings. In Section 2, we define the key underlying variables and establish notation. In Section 3, we consider estimation of the identifiable aspects of the distribution of time until the event of interest (e.g., HIV infection or liver fibrosis) during the study period. In Sections 4 and 5, we consider hypothesis testing and design, respectively, and in Section 6, we discuss some related issues and extensions of the proposed methods.

2. Notation and Probability Elements

Let T denote the time of occurrence of an event that can take values in $(-\infty, \infty)$, with 0 denoting the time of enrollment of an individual into the trial. We denote the probability density function (p.d.f.) and cumulative distribution function (c.d.f.) of T by $f(\cdot)$ and $F(\cdot)$, respectively. The goal is to make an inference about the distribution function for developing the event during the study; i.e., either

$$F_0(t) \overset{\text{def}}{=} F(t) - F(0) = \int_0^t f(u)\,du$$

or

$$F_c(t) \overset{\text{def}}{=} \Pr(T \leq t \mid T > 0) = F_0(t)/[1 - F(0)],$$

for $t > 0$. Note that $F_0(\cdot)$ is a subdistribution function since $F_0(\infty) < 1$ when $F(0) > 0$. Without loss of generality, we assume that the eligibility criterion to enter the study is a negative diagnostic test result at screening. Let E denote the result of the diagnostic test used to determine whether the subject is enrolled into the trial, with $E = 1$ indicating that the diagnostic test is negative, in which case the individual is enrolled, and $E = 0$ indicating that the diagnostic test is positive, in which case the subject is not enrolled. The joint distribution of (T, E) is characterized by the marginal distribution of T, given by $f(\cdot)$, and the conditional distribution of E given T, denoted by

$$g(t) \overset{\text{def}}{=} \Pr[E = 1 \mid T = t] \quad \text{for } t \in (-\infty, \infty).$$

It follows that the marginal distribution of E is given by

$$\Pr[E = 1] = \int_{-\infty}^{\infty} g(t) f(t)\,dt.$$

The conditional distribution of T, given enrollment into the study, is

$$\begin{aligned} F_g(\tau) &= \Pr[T \leq \tau \mid E = 1] \\ &= \frac{\displaystyle\int_{-\infty}^{\tau} g(t) f(t)\,dt}{P[E = 1]} \\ &= \frac{\displaystyle\int_{-\infty}^{\tau} g(t) f(t)\,dt}{\displaystyle\int_{-\infty}^{\infty} g(t) f(t)\,dt}. \end{aligned}$$

In the first example used to motivate this problem, we considered a single diagnostic test used to determine enrollment into the trial and a perfect diagnostic test administered at the end of trial to determine whether the event had occurred by then. In other settings, perfect or imperfect diagnostic tests may be administered at various time points during the trial, including to subjects who are not enrolled into the trial.

3. Estimation

We begin this section by examining the common pretest–posttest situation in which each subject is given a diagnostic test to determine enrollment into a trial and is then given another diagnostic test at the end of trial to determine whether the event of interest has occurred. We allow perfect and imperfect diagnostic tests, and for the latter initially assume that the test sensitivity and specificity are time independent. We then consider more general settings where multiple diagnostic tests can be given during a trial, and where diagnostic tests can have a time-dependent sensitivity.

3.1 *Pretest–Posttest Trials*

Suppose that the imperfect test has a time-independent sensitivity p_1 and specificity p_0. That is, if t denotes the time of occurrence of the event and τ denotes the time the imperfect test is administered, then the probability that the test is negative for the occurrence of the event, in which case the subject is enrolled into the trial, is

$$\Pr[E = 1 \mid T = t, \tau] = \begin{cases} 1 - p_1 & \text{for } t \leq \tau, \\ p_0 & \text{for } t > \tau. \end{cases}$$

We assume that $1 - p_1 < p_0$; that is, that the probability of a negative test result is greater if the event of interest has not yet occurred. Define π_j to be the probability that test j is positive when given at time 0, and define π_{jk} to be the conditional probability that test k is positive when given at time $\tau > 0$, given that test j was negative when given at time 0, where the subscripts j and k equal 1 for the perfect test and 2 for the imperfect test. Then it is easily shown that $\pi_1 = F(0)$, $\pi_2 = p_1 F(0) + (1 - p_0)[1 - F(0)]$,

$$\pi_{11} = \frac{F(\tau) - F(0)}{[1 - F(0)]},$$

$$\pi_{12} = \frac{p_1[F(\tau) - F(0)] + (1 - p_0)[1 - F(\tau)]}{1 - F(0)},$$

$$\pi_{21} = \frac{(1 - p_1)F(0) + p_0[F(\tau) - F(0)]}{(1 - p_1)F(0) + p_0[1 - F(0)]},$$

and

$$\pi_{22} = \frac{p_1(1 - p_1)F(0) + p_0 p_1[F(\tau) - F(0)] + p_0(1 - p_0)[1 - F(\tau)]}{(1 - p_1)F(0) + p_0[1 - F(0)]},$$

where in the last expression we have assumed that the two imperfect test results are conditionally independent, given the time of the event.

Suppose that $K_1 \geq 0$ subjects are assessed for eligibility at $t = 0$ using a perfect diagnostic test and that $K_2 > 0$ are assessed using an imperfect diagnostic test. Let N_1 and N_2 denote the number of these subjects that test negative $(E = 1)$ and are enrolled into the trial, and assume that all $N_1 + N_2$ are subsequently tested at time τ, denoting the end of the trial, with either the perfect or imperfect diagnostic test. Let N_{ij} denote the number of the N_i subjects that are evaluated at the end of the trial using the perfect $(j = 1)$ and imperfect

($j = 2$) diagnostic tests, and let r_{ij} denote the corresponding number that test positive for the occurrence of the event. We allow the N_{ij} to be determined adaptively by any known deterministic or probabilistic function of (K_1, K_2, N_1, N_2). Then it can be shown (Appendix A) that the likelihood function is proportional to

$$L = \prod_{j=1}^{2} \pi_j^{K_j - N_j} \prod_{k=1}^{2} [(1-\pi_j)\pi_{jk}]^{r_{jk}} [(1-\pi_j)(1-\pi_{jk})]^{N_{jk} - r_{jk}}. \tag{1}$$

In Section 5, we consider the special case where only the perfect test is used at the end of the study (i.e., $N_{12} = N_{22} = 0$). Here, the likelihood function simplifies to

$$L = \prod_{j=1}^{2} \pi_j^{K_j - N_j} [(1-\pi_j)\pi_{j1}]^{r_{j1}} [(1-\pi_j)(1-\pi_{j1})]^{N_{j1} - r_{j1}}. \tag{2}$$

In Section 4, we consider another special case where all screening tests are imperfect (i.e., $K_1 = N_1 = 0$) and only the perfect test is used at the end of the study (i.e., $N_{22} = 0$). In this setting, the likelihood further simplifies to

$$L = \pi_2^{K_2 - N_2} [(1-\pi_2)\pi_{21}]^{r_{21}} [(1-\pi_2)(1-\pi_{21})]^{N_{21} - r_{21}}. \tag{3}$$

Yet another special case is when a gold standard (perfect diagnostic test) is not available. In this case, the likelihood function is equal to (1), but where $K_1 = N_1 = N_{21} = 0$.

It follows from the above that, at best, the only identifiable aspects of $F(\cdot)$ are $F(0)$ and $F(\tau)$. When p_0, p_1 are known, then $F(0)$ and $F(\tau)$ are in general identifiable without making any additional assumptions. When either p_0 or p_1 are unknown, nonidentifiability can result unless one reduces the dimensionality of the unknown parameter vector by making some additional assumptions (see, for details, Balasubramanian and Lagakos, 2003). Maximum likelihood estimates for these can be obtained by numerical maximization of the log likelihood subject to the constraint $F(\tau) \geq F(0)$; however, the inverse of the expected information, used to estimate their covariance matrix, is obtainable in closed form (see Appendix B).

Under mild conditions, the estimators of $F(0)$ and $F(\tau)$ can be shown to be consistent and asymptotically normal as $K_1 \to \infty$ and $K_2 \to \infty$.

3.2 *More General Experimental Conditions*

The setting described in Section 3.1 can be generalized to allow each subject to receive several types of perfect or imperfect diagnostic tests at multiple times during the trial. For example, an imperfect diagnostic test might be given at time 0 (to determine whether a patient is enrolled) and then monthly for the duration of the trial, at which time a perfect diagnostic test is given. Furthermore, the imperfect diagnostic tests might have sensitivities/specificities that are time dependent and, additionally, subjects that test positive at $t = 0$ might also be followed in some settings, as we see below. In general, as long as some subjects are screened for enrollment using an imperfect diagnostic test, the phenomenon of interest in this article—that some subjects will have already had the event of interest upon enrollment—is present.

A general approach for estimating the identifiable aspects of $F(\cdot)$ is given in Balasubramanian and Lagakos (2003) for settings where there can be multiple types of diagnostic tests, multiple test times, and where the sensitivity of an imperfect diagnostic test can be a function of the elapsed time between the event and the time the diagnostic test is given. In these more general settings, the identifiable aspects of $F(\cdot)$ will depend on the times that diagnostic tests are administered as well as the form of the time-dependent sensitivities of the imperfect diagnostic tests. If $F(t)$ is estimable at $t = 0$, then the desired parameters $F_0(\tau)$ and $F_c(\tau)$ will be estimable for those $\tau > 0$ for which $F(\tau)$ is estimable.

To illustrate these points, we consider a randomized trial of newborns of HIV-infected mothers, where infants that test negative for HIV infection, using an imperfect diagnostic test (such as DNA PCR), are randomized to one of several feeding strategies. Infants who are enrolled are then evaluated at age τ using a perfect diagnostic test (such as ELISA/Western Blot) to determine whether they have become infected. Because the Botswana trial that motivated our interest is ongoing, we use the results from two other clinical trials aimed at preventing mother-to-child transmission of HIV to obtain an estimate of $F(\cdot)$ that might be reflective of the distribution of time until HIV infection for infants in this setting.

The first is a trial recently conducted in Tanzania (Fawzi et al., 1998). To see the impact of the imperfect diagnostic test used for screening, suppose that newborns are screened at day 30 using DNA PCR, and evaluated at age $\tau = 2$ years of age for infection. Using the estimate of $F(\cdot)$ for this trial obtained by Balasubramanian and Lagakos (2003), the probability of already being infected by the time of randomization is 0.23, the unconditional probability that an infant will be enrolled is $\Pr(E = 1) = 0.79$, and the probability of being infected by 2 years of age of 0.38. These calculations assume a specificity of 98%, that the sensitivity of PCR within 2 weeks following infection is 70%, and that the sensitivity more than 2 weeks following infection is 93% (see, for details, Balasubramanian and Lagakos, 2003). Figure 1 gives the resulting estimators of $F_g(\tau)$, the conditional distribution of being infected by age τ, given enrollment into the trial, and of $F_c(\tau)$, the conditional distribution of becoming infected during the trial, given that the infant was truly uninfected at enrollment. The bias in the naive estimator $F_g(\cdot)$ is evident. In the same setting, but with a randomization on day 7 after birth, a smaller bias resulted due to a smaller probability of HIV transmission during the 2 weeks preceding randomization.

As a second example, we consider protocol 076 of the AIDS Clinical Trials Group (Connor et al., 1994), which was aimed at preventing HIV transmission during pregnancy and at birth (Balasubramanian and Lagakos, 2001). Suppose that the infants are screened at birth using DNA PCR and evaluated at age $\tau = 2$ years for infection. Using the sensitivity, specificity, and estimate of $F(\cdot)$ corresponding to times prior to birth for this study from Balasubramanian and Lagakos (2001), the probability of already being infected by the time of randomization is 0.22 and the unconditional probability that an infant will be enrolled is $\Pr(E = 1) = 0.91$. Then if the postpartum probability of infection for an infant breast-fed up to 2 years of age were 0.18, then the resulting estimates of $F_g(\tau)$ and $F_c(\tau)$ are 0.34 and 0.23, respectively. In this setting, the bias of the naive estimator is heightened due to a high probability of HIV transmission in the 2 weeks prior to randomization

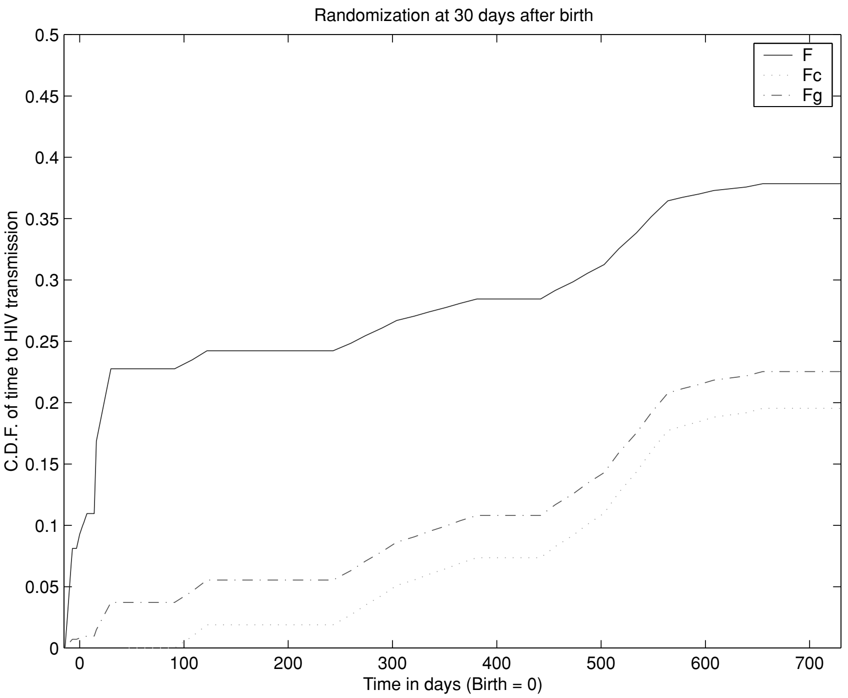


Figure 1. Estimates of $F(\cdot)$, $F_{\mathrm{g}}(\tau)$, and $F_{\mathrm{c}}(\tau)$ for randomization at 30 days after birth.

and the low sensitivity of DNA PCR when given shortly after infection. These examples illustrate that the bias of the naive estimator $F_{\mathrm{g}}(\cdot)$ can be substantial in this setting of an HIV prevention trial because of the risk of infection at or shortly before birth and the lower sensitivity of the DNA PCR assay when given shortly after the time of infection.

4. Hypothesis Testing

A primary goal in a randomized clinical trial is to compare treatment groups. Suppose patients are randomized to one of the two treatment groups, say A and B, and let $F^x(t)$ and $f^x(t)$ represent the c.d.f. and p.d.f. of T for treatment group x, respectively, where $x = \mathrm{A}, \mathrm{B}$. Define $F_0^x(\cdot)$ and $F_{\mathrm{c}}^x(\cdot)$, similarly. We consider the null hypothesis

$$\mathrm{H}_0 : F_{\mathrm{c}}^{\mathrm{A}}(\tau) = F_{\mathrm{c}}^{\mathrm{B}}(\tau), \quad \tau > 0$$

that the treatment groups have the same distribution of time until the event of interest, conditional on the event not having occurred prior to randomization. This null hypothesis is equivalent to testing the equality of the unconditional subdistribution functions F_0^{A} and F_0^{B}. Two natural statistics for testing H_0 are given by

$$Z_0 = \frac{\hat{F}_0^{\mathrm{A}}(\tau) - \hat{F}_0^{\mathrm{B}}(\tau)}{\sqrt{\mathrm{Var}\big(\hat{F}_0^{\mathrm{A}}(\tau) - \hat{F}_0^{\mathrm{B}}(\tau)\big)}} = \frac{\hat{F}^{\mathrm{A}}(\tau) - \hat{F}^{\mathrm{B}}(\tau)}{\sqrt{\mathrm{Var}\big(\hat{F}^{\mathrm{A}}(\tau) - \hat{F}^{\mathrm{B}}(\tau)\big)}},$$

$$Z_{\mathrm{c}} = \frac{\hat{F}_{\mathrm{c}}^{\mathrm{A}} - \hat{F}_{\mathrm{c}}^{\mathrm{B}}}{\sqrt{\mathrm{Var}\big(\hat{F}_{\mathrm{c}}^{\mathrm{A}} - \hat{F}_{\mathrm{c}}^{\mathrm{B}}\big)}},$$

where for Z_0 we have used the fact that $F^{\mathrm{A}}(0) = F^{\mathrm{B}}(0)$ due to the randomization and where the estimates for F_0 and F_{c} are the maximum likelihood estimators described in Section 3. The variance estimates can be obtained from the observed information and using the Delta method. Both test statistics are asymptotically $N(0, 1)$ under H_0 and thus can be used to assess H_0.

An alternative approach for testing treatment equality is to simply compare outcomes among all subjects enrolled into the trial, even though an unknown proportion have had the event at entry. To illustrate the possible problems with this approach, suppose that all patients are screened for enrollment using an imperfect diagnostic test and that all are evaluated at the end of the trial using a perfect test, that is, $K_1 = 0$ and $N_{22} = 0$. Let $\pi_{21}^x(\tau)$ denote the value of $\pi_{21}(\tau)$ for subjects in treatment group x, for $x = \mathrm{A}$ and B, respectively, and consider the test statistic

$$Z_n = \frac{\hat{\pi}_{21}^{\mathrm{A}}(\tau) - \hat{\pi}_{21}^{\mathrm{B}}(\tau)}{\sqrt{\mathrm{Var}\big(\hat{\pi}_{21}^{\mathrm{A}}(\tau) - \hat{\pi}_{21}^{\mathrm{B}}(\tau)\big)}},$$

where the estimates of $\pi_{21}(\tau)$ are the observed proportion of subjects who test positive for event at τ among all those enrolled in the study (i.e., those who satisfied $E = 1$ at entry), and

$$\mathrm{Var}\big(\hat{\pi}_{21}^{\mathrm{A}}(\tau) - \hat{\pi}_{21}^{\mathrm{B}}(\tau)\big) = \pi_{21}^{\mathrm{A}}(\tau)\big[1 - \pi_{21}^{\mathrm{A}}(\tau)\big] E\left(\frac{1}{N_2^{\mathrm{A}}}\right) + \pi_{21}^{\mathrm{B}}(\tau)\big[1 - \pi_{21}^{\mathrm{B}}(\tau)\big] E\left(\frac{1}{N_2^{\mathrm{B}}}\right),$$

where N_2^{A} and N_2^{B} are the number of enrolled subjects in treatment groups A and B, respectively. Note that Z_n is

Table 1
Observed (OR^*) *versus true* (OR) *odds ratios for hypothetical values of* ρ, $F_c^A(\tau)$, *and* $F_c^B(\tau)$

ρ	F_c^A	F_c^B	$\pi_{21}^A(\tau)$	$\pi_{21}^B(\tau)$	OR	OR^*
0.020	0.10	0.05	0.118	0.069	2.111	1.805
0.050	0.10	0.05	0.145	0.098	2.111	1.570
0.100	0.10	0.05	0.190	0.145	2.111	1.383
0.020	0.15	0.05	0.167	0.069	3.353	2.705
0.050	0.15	0.05	0.193	0.098	3.353	2.207
0.100	0.15	0.05	0.235	0.145	3.353	1.811

asymptotically equivalent to a Fisher's exact test based on those subjects enrolled in the study.

We refer to Z_n as a naive test because it does not take explicit account of the fact that some subjects may have already experienced the event prior to enrollment. Note that Z_n actually tests the null hypothesis

$$\mathrm{H}_n : \pi_{21}^{\mathrm{A}}(\tau) = \pi_{21}^{\mathrm{B}}(\tau),$$

which in general is not equivalent to H_0. To see their connection, note that

$$\begin{aligned} &\Pr(T \le \tau \mid T > 0, E = 1) \\ &\quad = \frac{\Pr(T \le \tau, E = 1 \mid T > 0)}{\Pr(E = 1 \mid T > 0)} \\ &\quad = \frac{\displaystyle\int_0^\tau \Pr(E = 1 \mid t, T > 0) f(t \mid T > 0)\, dt}{\Pr(E = 1 \mid T > 0)}. \end{aligned}$$

If the specificity of the imperfect diagnostic test does not depend on the future time of occurrence of the event, then the right-hand side of this equation simplifies to $F_c(\tau)$. Thus, if we define $\rho = \Pr(T \le 0 \mid E = 1)$, it follows that

$$\frac{\pi_{21}(\tau) - \rho}{1 - \rho} = F_c(\tau).$$

Since ρ does not depend on the treatment group, it follows that when the specificity of the imperfect diagnostic test is time independent, then H_n is equivalent to H_0 and thus a naive test, such as Z_n, will also be a valid test of H_0.

When H_0 does not hold, the naive approach will, in general, give biased estimates of treatment differences. To illustrate this, Table 1 gives theoretical values of $\pi_{21}^{\mathrm{A}}(\tau)$ and $\pi_{21}^{\mathrm{B}}(\tau)$ for different values of ρ, $F_c^{\mathrm{A}}(\tau)$, and $F_c^{\mathrm{B}}(\tau)$, again for the case where $K_1 = 0$ and $N_{22} = 0$. The inclusion of false negatives in the odds ratio comparing the two treatments, A and B, attenuates the true underlying differences between the two treatments. As ρ increases, the difference between the true odds ratio (based on $F_c^{\mathrm{A}}(\tau)$ and $F_c^{\mathrm{B}}(\tau)$) and the observed odds ratio (based on $\pi_{21}^{\mathrm{A}}(\tau)$ and $\pi_{21}^{\mathrm{B}}(\tau)$) also increases.

It does not necessarily follow that the corresponding naive statistical tests, such as Z_n, will have less power than bias-adjusted tests, such as Z_0 and Z_c. To explore this, we first consider the pretest–posttest setting described in Section 3.1 with $K_1 = 0$ and $N_{22} = 0$; that is, where all subjects are screened with the imperfect diagnostic test and assessed at the end of the trial with the perfect test, and then compute the asymptotic relative efficiency (ARE) of the estimators of $F_0(\tau)$ to that of $\pi_{21}(\tau)$. Details are given in Appendix B.

Suppose that subjects are screened at entry by an imperfect diagnostic test (i.e., $K_1 = 0$), and that all those who test negative at entry are evaluated at some later time τ using a perfect test. Figure 2 presents plots of ARE of the bias-adjusted to the naive test as a function of the sensitivity of the screening test. Figure 2a–2d represent values of $\{F(0), F_0(\tau)\}$ equal to $\{(0.50,0.25), (0.25, 0.50), (0.20, 0.05), (0.05, 0.20)\}$, respectively. For each of the four cases of $\{F(0), F_0(\tau)\}$, we consider values of specificity equal to 0.75 and 0.90, denoted by dotted and solid lines, respectively. Note that these results do not depend on the total number of subjects screened for entry into the study.

For all four choices of $\{F(0), F_0(\tau)\}$, the asymptotic variance of the estimate of $F_0(\tau)$ is higher than that of the naive estimator, $\hat{\pi}_{21}(\tau)$, when sensitivity/specificity of the screening test is low. For high values of sensitivity/specificity, the adjusted estimator is at least as good as, and in some cases, more precise than the naive estimator. In panels b and d, where a higher proportion of events occur during the study, estimates of $\pi_{21}(\tau)$ perform significantly better than the adjusted estimator for all values of sensitivity. When a higher proportion of events occur prior to the study (panels a and c), the adjusted estimator actually performs somewhat better than the naive estimator at values of sensitivity greater than 0.80.

These results suggest that in cases where the diagnostic test has relatively low error rates, estimates based on the proposed methods may be preferable, especially in settings where the event rate prior to study entry is high. When the diagnostic test has poor diagnostic properties, hypothesis tests constructed based on the naive estimator may be preferable.

Similar results were obtained when the sensitivity of the test was assumed to be time dependent, based on models for the behavior of DNA PCR and HIV culture assays for detecting HIV infection in infants (Balasubramanian and Lagakos, 2001). Such situations arise specifically in HIV vertical transmission studies, where diagnostic tests to detect HIV in infants are highly specific only after a few weeks following infection (Dunn et al., 2000). The results for the simulation study imply that tests based on the adjusted estimator may be preferable, especially when the specificity and maximum sensitivity of the diagnostic tests are relatively high (details available upon request). However, when diagnostic tests are administered at multiple times during the study, tests based on proposed methods may have decreased power as a result of $F(\cdot)$ being estimated from a larger number of parameters. Note that in settings where there is a single pretest and posttest, the only identifiable aspects of $F(\cdot)$ are $F(0)$ and $F(\tau)$. This implies that the null hypothesis H_0 can be satisfied even in cases where there may be an initial treatment difference, that is no longer present by the end of the study. In these situations, additional data from tests administered during the study could help identify additional aspects of $F(\cdot)$ and hence aid in testing hypotheses of treatment differences during the study.

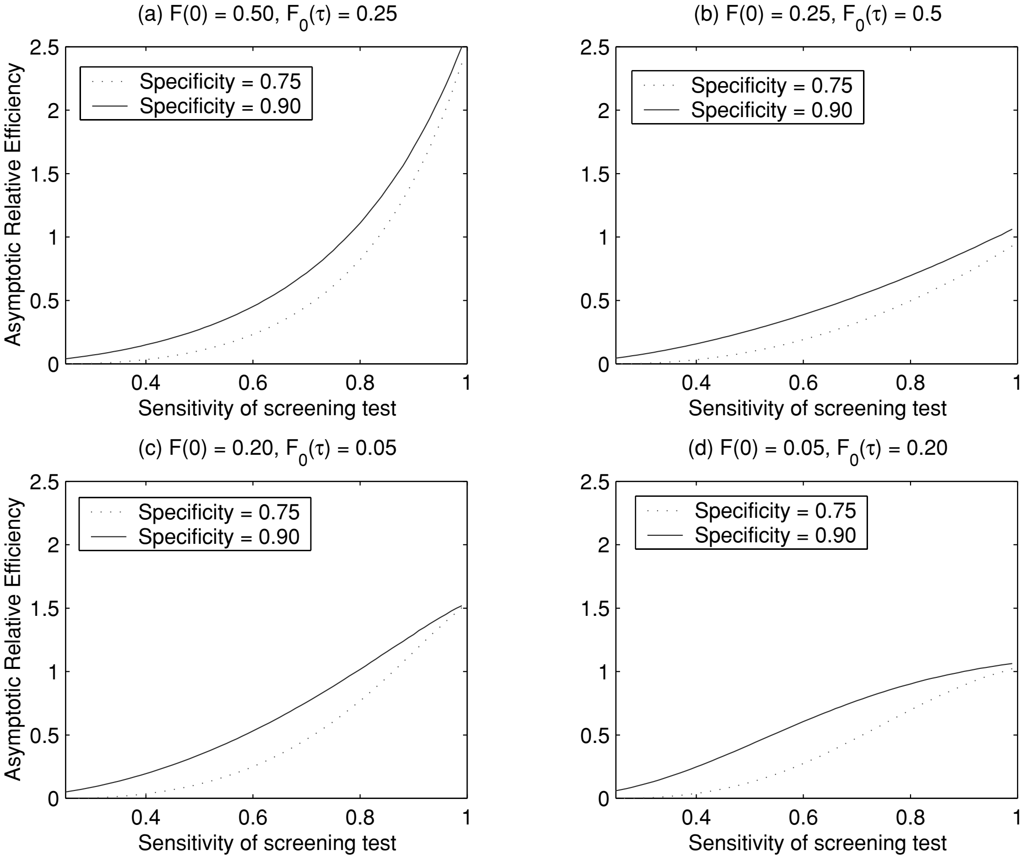


Figure 2. ARE of the estimate of $F_0(\tau)$ to that of $\pi_{21}(\tau)$ (i.e., $\frac{\mathrm{Var}[\hat{\pi}_{21}(\tau)]}{\mathrm{Var}[\hat{F}_0(\tau)]}$).

5. Study Design

To discuss issues related to study design, we consider a clinical trial in which K_1 and K_2 subjects are screened for eligibility using a perfect and imperfect diagnostic test, respectively, and where all are evaluated at the end of the trial using the perfect diagnostic test. It is obvious that use of the perfect diagnostic test for all subjects is the most efficient from a statistical perspective. However, in many settings there will be tradeoffs between the cost and/or invasiveness of the former and the possibility of errors with the latter. To illustrate the tradeoff between the sample sizes of the groups given the perfect versus imperfect screening tests for this setting, we evaluate the relative sample sizes needed among different designs to achieve the same power for a variety of choices for the sensitivity (p_1) and specificity (p_0) of the imperfect diagnostic test, the probability, $F(0)$, of being infected by the time of randomization, and the probability, $F_0(\tau)$, of being infected between randomization and the end of the trial (τ).

Suppose that $\{F(0),\ F_0(\tau)\}$ can take the set of values $\{(0.50, 0.25), (0.25, 0.50), (0.20, 0.05), (0.05, 0.20)\}$ and that $\{p_0,\ p_1\}$ can take the set of values $\{(0.90, 0.90), (0.90, 0.75), (0.75, 0.90), (0.75, 0.75)\}$. Figure 3 presents the percentage increase that is required in the number of subjects screened by imperfect tests in order to compensate for a corresponding percentage decrease in the number of subjects screened by perfect tests, where the objective is to obtain the same accuracy in the estimate of $F_0(\tau)$ as is attained when there are equal number of subjects screened by perfect and imperfect tests. The relative sample sizes are based on the asymptotic variances of the resulting estimators of $F_0(\tau)$, obtained from the inverse of the expected information matrix (see analytic expressions in Appendix B). The expression for the percentage increase in the sample size of the group screened by the imperfect test depends on K_1 and K_2 only through the percentage decrease in the sample size of the group screened by the perfect test (details available upon request).

We see that for a given set of values of $(F(0),\ F_0(\tau))$, the percent increase in the size of the group screened by imperfect tests is lower for tests with higher sensitivity and specificity. Moreover, the tradeoff between the number of subjects tested by perfect and imperfect screening tests is dependent on the values of the underlying distribution, i.e., $(F(0),\ F_0(\tau))$, with the rate of change of the percentage increase in the number of imperfect tests being the slowest for the case corresponding to $(F(0),\ F_0(\tau)) = (0.05,\ 0.20)$ and fastest for the case corresponding to $(F(0),\ F_0(\tau)) = (0.20,\ 0.05)$ (see Figure 3c and 3d). The increase in sample size is not extreme in cases where the imperfect test has relatively high sensitivity and specificity. In addition, the tradeoff is relatively less severe in situations where the event rate during the study (i.e., $F_0(\tau)$) is expected to be relatively higher than that prior to entry (i.e., $F(0)$).

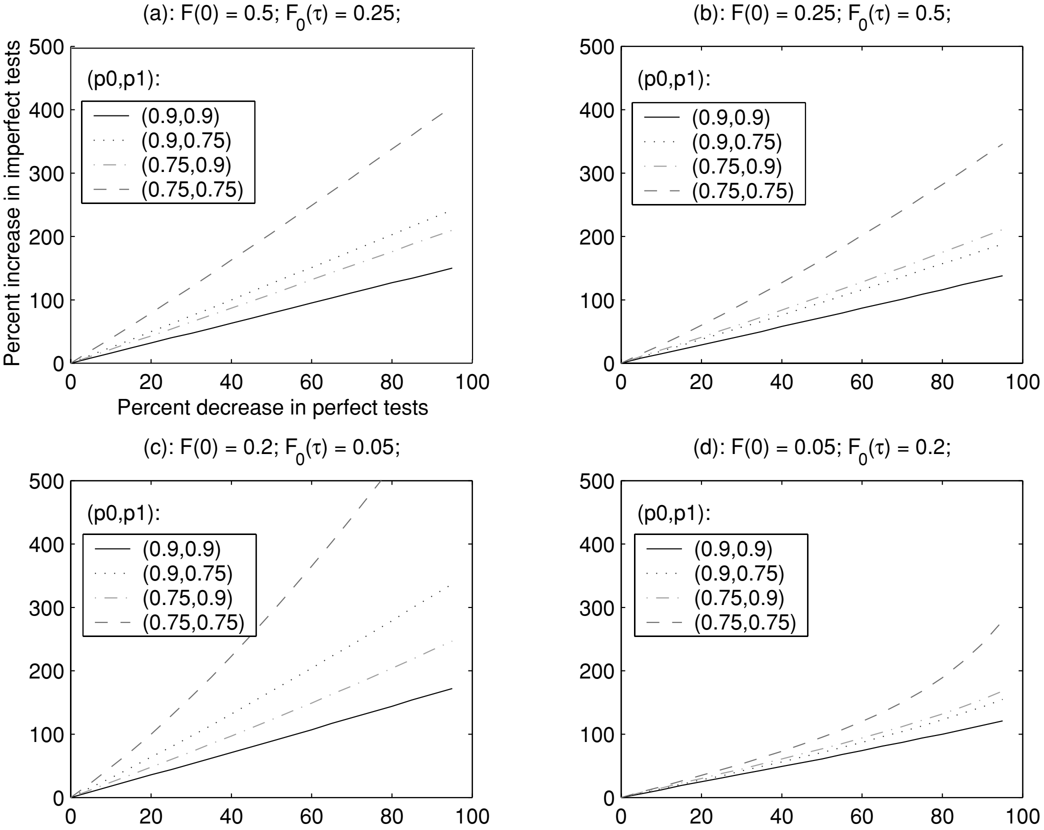


Figure 3. Percentage increase in the number of imperfect screening tests required for a given percentage decrease in the number of perfect screening tests in order to obtain the same level of accuracy in the MLE of $F_0(\tau)$ as obtained with equal numbers of perfect and imperfect screening tests.

These results imply that studies can be designed to maintain high power to detect differences between treatment groups even when the number of subjects screened by the perfect test is lowered. As an example, consider the randomized study of Oral Miltefosine in treating Indian visceral Leishmaniasis by Sundar et al. (2002). The study enrolled subjects who were diagnosed to have Leishmaniasis by the presence of *Leishmania* in splenic aspirates, through an invasive procedure. The endpoint of interest was the absence of Leishmaniasis, also determined at the study end via splenic aspirates. In such settings, the benefit of an imperfect test for screening may outweigh the cost/inconvenience of a perfect test. Here it may be useful to consider study designs that involve only a few subjects screened by the perfect test, but compensate by increasing the number of subjects screened by the imperfect test.

6. Discussion

In this article, we propose methods to estimate and test hypotheses in clinical trials, where the diagnostic test used to allow enrollment into the study may be subject to error. It was seen in Table 1 that standard estimates of event rates can be biased. The bias is small when the sensitivity and specificity of the diagnostic tests are high, but can be substantial for poorer diagnostic tests. For testing hypotheses, the naive tests are valid if the specificity of the diagnostic test is not time dependent. The bias-adjusted methods are somewhat preferable when the screening test is almost perfect, yet the naive estimator is preferable when the diagnostic test has low sensitivity and specificity. For most intermediate situations, the power of both methods is similar.

In the second motivating example used in this article, both imperfect and perfect (or near perfect) diagnostic tests are available to determine eligibility for enrollment into a trial and to evaluate whether an event has occurred during the trial. We showed in Section 5 that in certain settings, use of an imperfect diagnostic test may require only a small increase in sample size to achieve the same power as the perfect diagnostic test. Thus, when the perfect test carries risk or discomfort to the patient, or when it is more expensive, use of the imperfect test may be preferable because the excess costs in number of patients could be more than offset by the reduced risk/inconvenience to patients or costs of the diagnostic test. A variation of the design setting considered in Section 5 is where no diagnostic test is used to screen for enrollment. The naive statistical test can still be used and, as illustrated in Section 4, will often have good statistical properties as compared to a corrected test. This suggests that in some settings, such as trials to compare treatments for prevention of prostate cancer in elderly men where the definitive diagnostic test is a prostatic biopsy, an initial diagnostic test might be skipped altogether and thereby reduce risk/discomfort to participants.

Throughout this article, we have assumed that testing negative on the diagnostic test used at screening would lead to enrollment into the trial. In other settings, however, testing positive on a diagnostic test would be the condition for enrollment. Here the same methods apply, but with the roles of p_0 and p_1 reversed.

The methods developed in this article can be extended in several ways. In some settings, the "perfect" diagnostic test for evaluating the occurrence of event at the end of the study may in fact be subject to error. For example, in studies involving liver disorders, a liver biopsy although considered to be the gold standard, may still be error prone. In these settings, naive estimators of event rates will no longer be appropriate but the methods proposed in this article can be extended for such applications. The proposed methods could also be easily extended to accommodate situations where all subjects in the study do not have the same follow-up time.

Finally, in this article, we consider the setting where a fixed number of subjects are screened at entry and those who test negative for the event are enrolled into the study. That is, we assume that the number of subjects enrolled in the study is random and the number of subjects screened for potential entry into the study is fixed. Other study designs could involve screening as many patients as needed to enroll a fixed number of subjects into the study. It would be useful to extend the methods developed in this article to accommodate other study designs.

Acknowledgements

This work was supported by National Institutes of Health grant AI24643 and a predoctoral fellowship from the Howard Hughes Medical Institute. We are grateful to the editors and referees for their helpful comments.

Résumé

Dans les essais cliniques où le critère de jugement est le délai jusqu'à l'apparition d'un événement présentant une phase de latence avant d'être détectable, une proportion inconnue de sujets qui auront déjà l'événement sera incluse dans l'étude sans qu'on le sache, du fait de l'imperfection des tests diagnostiques utilisés pour trier les sujets éligibles. Par exemple, les tests diagnostiques communément utilisés pour évaluer le statut VIH chez les nourrissons, tel que la PCR DNA et la culture du virus, ont une faible sensibilité quand ils sont faits tôt après l'infection. Ceci peut mener à l'inclusion d'une proportion inconnue de nourrisson infectés par le VIH dans des essais cliniques visant à prévenir la transmission du virus de la mère VIH positive vers leur nourrisson au travers de l'allaitement. Le statut infectieux des nourrissons à la fin de l'essai, quand ils sont âgés de plus de 12 mois, peut être déterminé avec certitude. Pour ces enfants diagnostiqués infectés à la fin de l'étude, il est impossible de déterminer si la séroconversion a eu lieu pendant l'étude ou si l'enfant était déjà infecté au moment de son inclusion. Dans ces cas, les estimations du risque cumulé d'événement à la fin de l'essai surestimeront la vraie probabilité d'événement pendant la période d'étude et les hypothèses des tests comparant deux stratégies d'intervention ou plus, peuvent aussi être biaisées. Nous présentons des méthodes d'inférence à utiliser dans de tels cas, pour obtenir la distribution du délai jusqu'à la survenue de l'événement d'intérêt, et nous recherchons des méthodes à proposer dans le protocole de tels essais quand il existe à la fois des tests diagnostiques imparfaits et parfaits.

References

Balasubramanian, R. and Lagakos, S. W. (2001). Estimation of the timing of perinatal transmission of HIV. *Biometrics* **57,** 1048–1058.

Balasubramanian, R. and Lagakos, S. W. (2003). Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika* **90,** 171–182.

Connor, E., Sperling, R., Gelber, R., et al. (1994). Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. *New England Journal of Medicine* **331,** 1173–1180.

Dunn, D. T., Simonds, R. J., Bultery, M., Kalish, L. A., Moye, J., deMaria, A., Kind, C., Rudin, C., Denamur, E., Krivine, A., Loveday, C., and Newell, M. L. (2000). Interventions to prevent vertical transmission of HIV-1: effect on viral detection rate in early infant samples. *AIDS* **14,** 1421–1428.

Dunn, D. T., Brandt, C., Krivine, A., et al. (1995). The sensitivity of HIV-1 DNA polymerase chain reaction in the neonatal period and the relative contributions of intra-uterine and intra-partum transmission. *AIDS* **9,** F7–F11.

Fawzi, W., Msamanga, G., Spiegelman, D., Urassa, E., McGrath, N., Mwakagile, D., Antelman, G., Mbise, R., Herrera, G., Kapiga, S., Willet, W., and Hunter, D. (1998). Randomized trial of effects of vitamin supplements on pregnancy outcomes and T cell counts in HIV-1 infected women in Tanzania. *Lancet* **351,** 1477–1482.

Hadziyannis, S., Tassopoulos, N., Heathcote, E., Chang, T., Kitis, G., Rizzetto, M., Marcellin, P., Lim, S., Goodman, Z., Wulfson, M., Xiong, S., Fry, J., and Brosgart, C. (2003). Adefovir dipivoxil for the treatment of hepatitis B e antigen-negative chronic hepatitis B. *New England Journal of Medicine* **348,** 800–807.

Harith, E., Kolk, A. H., Kager, P., Leeuwenburg, J., Muigai, R., and Kiugu, S. (1986). A simple and economical direct agglutination test for serodiagnosis and sero-epidemiological studies of visceral leishmaniasis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **80,** 583–587.

Martin, P. and Friedman, L. (1998). Assessment of liver function and diagnostic studies. In *Handbook of Liver Disease*, L. S. Friedman and E. B. Keeffe (eds), 1–14. Philadelphia, Pennsylvania: Churchill Livingstone.

Sundar, S., Jha, T., Thakur, C., Engel, J., Sindermann, H., Fischer, C., Junge, K., Bryceson, A., and Berman, J. (2002). Oral miltefosine for Indian visceral leishmaniasis. *New England Journal of Medicine* **347,** 1739–1746.

Received May 2003. *Revised September* 2003.
Accepted November 2003.

Appendix A

Derivation of Likelihood

Let $D_i = (\tau_i, v_i, r_i)$ refer to the observed data for ith subject, for $i = 1, \ldots, K_1 + K_2$, where $\tau_i = (0, \tau)$ refers to the times of the diagnostic tests, $v_i = (v_{i1}, v_{i2})$ refers to the types of the corresponding tests, and $r_i = (r_{i1}, r_{i2})$ is the vector of corresponding test results. We take v_{ij} equal to 1 or 2 for a perfect or imperfect diagnostic test, respectively, and r_{ij} equal to 1 for a positive test for the occurrence of event of interest and 0 otherwise. Note that for those subjects who test positive for event at entry, no further tests will be administered. Then, the general form of the likelihood can be expressed as

$$\begin{aligned}
L &= \prod_{i=1}^{K_1+K_2} g(D_i)\\
&= \prod_{i=1}^{K_1+K_2} \int_{-\infty}^{\infty} g(\tau_{i1}, v_{i1}, r_{i1}, \tau_{i2}, v_{i2}, r_{i2} \mid t) f(t)\, dt\\
&= \prod_{i=1}^{K_1+K_2} \int_{-\infty}^{\infty} g(\tau_{i1}, v_{i1}, r_{i1} \mid t)\\
&\qquad \times g(\tau_{i2}, v_{i2}, r_{i2} \mid \tau_{i1}, v_{i1}, r_{i1}, t) f(t)\, dt\\
&\propto \prod_{i=1}^{K_1+K_2} \int_{-\infty}^{\infty} g(r_{i1} \mid \tau_{i1}, v_{i1}, t) g(r_{i2} \mid \tau_{i2}, v_{i2}, t) f(t)\, dt.
\end{aligned}$$

We have assumed that (a) the results of the diagnostic tests are independent conditional on the time of event, t (i.e., $g(r_{i2} \,|\, \tau_{i2},\, v_{i2},\, r_{i1},\, v_{i1},\, t_{i1},\, t) = g(r_{i2} \,|\, \tau_{i2},\, v_{i2},\, t)$) and (b) the times and types of diagnostic tests are noninformative with respect to T (i.e., $g(\tau_{i1},\ v_{i1} \,|\, t) = g(\tau_{i1},\ v_{i1})$ and that $g(\tau_{i2},\, v_{i2} \,|\, r_{i1},\, \tau_{i1},\, v_{i1},\, t) = g(\tau_{i2},\, v_{i2} \,|\, r_{i1},\, \tau_{i1},\, v_{i1})$). In other words, the rules for determining whether to administer the second diagnostic test and type of test given could be probabilistic or deterministic functions of the results of the first test, but does not depend on knowledge of the time of the event of interest. Under these assumptions, the likelihood can be expressed as

$$L \propto \prod_{j=1}^{2} \pi_j^{K_j - N_j} \prod_{k=1}^{2} [(1-\pi_j)\pi_{jk}]^{r_{jk}} [(1-\pi_j)(1-\pi_{jk})]^{N_{jk}-r_{jk}}.$$

Appendix B

Analytic Expressions for Expected Information Matrix

To derive the expected information matrix, we assume that $N_{j1} = \alpha_j N_j$ for $j = 1, 2$, where α_j are known constants and satisfy $\alpha_j < 1$. Let the expected information matrix be denoted by $I = (I_{ij})$.

The components of I for the general form of the likelihood in equation (1) are

$$\begin{aligned}
I_{11} &= \frac{K_1}{F(0)} + \frac{(p_0+p_1-1)^2 K_2}{p_1 F(0) + (1-p_0)[1-F(0)]} + \frac{\alpha_1 K_1}{1-F(\tau)}\\
&\quad + \frac{(1-\alpha_1)(p_0-1)^2 K_1}{(p_1[F(\tau)-F(0)] + (1-p_0)[1-F(\tau)]}\\
&\quad + \frac{(1-\alpha_1)p_0^2 K_1}{(1-p_1)[F(\tau)-F(0)] + p_0[1-F(\tau)]}\\
&\quad + \frac{\alpha_2(1-p_1)^2 K_2}{(1-p_1)F(0) + p_0[F(\tau)-F(0)]} + \frac{\alpha_2 p_0 K_2}{1-F(\tau)}\\
&\quad + \frac{(1-\alpha_2)[p_1(1-p_1) - p_0(1-p_0)]^2 K_2}{p_1(1-p_1)F(0) + p_0 p_1[F(\tau)-F(0)] + p_0(1-p_0)[1-F(\tau)]}\\
&\quad + \frac{(1-\alpha_2)\left[(1-p_1)^2 - p_0^2\right]^2 K_2}{(1-p_1)^2 F(0) + p_0(1-p_1)[F(\tau)-F(0)] + p_0^2[1-F(\tau)]},
\end{aligned}$$

$$\begin{aligned}
I_{21} &= I_{12}\\
&= \frac{\alpha_1 K_1}{1-F(\tau)} + \frac{(1-\alpha_1)(p_0-1)(p_0+p_1-1)K_1}{p_1[F(\tau)-F(0)] + (1-p_0)[1-F(\tau)]}\\
&\quad + \frac{(1-\alpha_1)p_0(p_0+p_1-1)K_1}{(1-p_1)[F(\tau)-F(0)] + p_0[1-F(\tau)]}\\
&\quad + \frac{\alpha_2(1-p_1)p_0 K_2}{(1-p_1)F(0) + p_0[F(\tau)-F(0)]} + \frac{\alpha_2 p_0 K_2}{1-F(\tau)}\\
&\quad + \frac{(1-\alpha_2)[p_1(1-p_1) - p_0(1-p_0)][p_0 p_1 - p_0(1-p_0)]K_2}{p_1(1-p_1)F(0) + p_0 p_1[F(\tau)-F(0)] + p_0(1-p_0)[1-F(\tau)]}\\
&\quad + \frac{(1-\alpha_2)\left[(1-p_1)^2 - p_0^2\right]\left[p_0(1-p_1) - p_0^2\right] K_2}{(1-p_1)^2 F(0) + p_0(1-p_1)[F(\tau)-F(0)] + p_0^2[1-F(\tau)]},
\end{aligned}$$

$$\begin{aligned}
I_{22} &= \frac{\alpha_1 K_1}{F(\tau)-F(0)} + \frac{\alpha_1 K_1}{1-F(\tau)}\\
&\quad + \frac{(1-\alpha_1)(p_0+p_1-1)^2 K_1}{p_1[F(\tau)-F(0)] + (1-p_0)[1-F(\tau)]}\\
&\quad + \frac{(1-\alpha_1)(1-p_1-p_0)^2 K_1}{(1-p_1)[F(\tau)-F(0)] + p_0[1-F(\tau)]}\\
&\quad + \frac{\alpha_2 p_0^2 K_2}{(1-p_1)F(0) + p_0[F(\tau)-F(0)]} + \frac{\alpha_2 p_0 K_2}{1-F(\tau)}\\
&\quad + \frac{(1-\alpha_2)[p_0 p_1 - p_0(1-p_0)]^2 K_2}{p_1(1-p_1)F(0) + p_0 p_1[F(\tau)-F(0)] + p_0(1-p_0)[1-F(\tau)]}\\
&\quad + \frac{(1-\alpha_2)\left[p_0(1-p_1) - p_0^2\right]^2 K_2}{(1-p_1)^2 F(0) + p_0(1-p_1)[F(\tau)-F(0)] + p_0^2[1-F(\tau)]}.
\end{aligned}$$

For the special case when all tests at study end are perfect (see equation [2]), the above expressions simplify as follows:

$$\begin{aligned}
I_{11} &= \frac{K_1}{F(0)} + \frac{(p_0+p_1-1)^2 K_2}{p_1 F(0) + (1-p_0)[1-F(0)]} + \frac{K_1}{1-F(\tau)}\\
&\quad + \frac{(1-p_1)^2 K_2}{(1-p_1)F(0) + p_0[F(\tau)-F(0)]} + \frac{p_0 K_2}{1-F(\tau)},\\
I_{21} &= I_{12}\\
&= \frac{K_1}{1-F(\tau)} + \frac{p_0(1-p_1)K_2}{(1-p_1)F(0) + p_0[F(\tau)-F(0)]} + \frac{p_0 K_2}{1-F(\tau)},\\
I_{22} &= \frac{K_1}{F(\tau)-F(0)} + \frac{K_1}{1-F(\tau)}\\
&\quad + \frac{p_0^2 K_2}{(1-p_1)F(0) + p_0[F(\tau)-F(0)]} + \frac{p_0 K_2}{1-F(\tau)}.
\end{aligned}$$

When we further assume that all screening tests are imperfect (see equation [3]), we obtain

$$\begin{aligned}
I_{11} &= \frac{(p_0+p_1-1)^2 K_2}{p_1 F(0) + (1-p_0)[1-F(0)]}\\
&\quad + \frac{(1-p_1)^2 K_2}{(1-p_1)F(0) + p_0[F(\tau)-F(0)]} + \frac{p_0 K_2}{1-F(\tau)},\\
I_{21} &= I_{12}\\
&= \frac{p_0(1-p_1)K_2}{(1-p_1)F(0) + p_0[F(\tau)-F(0)]} + \frac{p_0 K_2}{1-F(\tau)},\\
I_{22} &= \frac{p_0^2 K_2}{(1-p_1)F(0) + p_0[F(\tau)-F(0)]} + \frac{p_0 K_2}{1-F(\tau)}.
\end{aligned}$$