Published in final edited form as:

Stat Med. 2008 November 20; 27(26): 5456–5470. doi:10.1002/sim.3365.

Incorporating Validation Subsets into Discrete Proportional Hazards Models for Mismeasured Outcomes

Amalia S. Magaret

Department of Laboratory Medicine, University of Washington, Box 359928, Seattle, WA 98195

Abstract

Standard proportional hazards methods are inappropriate for mismeasured outcomes. Previous work has shown that outcome mismeasurement can bias estimation of hazard ratios for covariates. We previously developed an Adjusted Proportional Hazards method that can produce accurate hazard ratio estimates when outcome measurement is either non-sensitive or non-specific. That method requires that mismeasurement rates (the sensitivity and specificity of the diagnostic test) are known. Here, we develop an approach to handle unknown mismeasurement rates. We consider the case where the true failure status is known for a subset of subjects (the validation set) until the time of observed failure or censoring. Five methods of handling these mismeasured outcomes are described and compared. The first uses only subjects on whom complete data are available (validation subset) while the second uses only mismeasured outcomes (naïve method). Three other methods include available data from both validated and non-validated subjects. Through simulation, we show that inclusion of the non-validated subjects can improve efficiency relative to use of the complete case data only; and that inclusion of some true outcomes (the validation subset) can reduce bias relative to use of mismeasured outcomes only. We also compare the performance of the validation methods proposed using an example dataset.

Keywords

failure; survival; validation subset; proportional hazards; measurement error

1 Introduction

Outcomes may be mismeasured for several reasons; and use of these mismeasured outcomes may be preferable to more accurate alternatives in some circumstances. Gold standard tests are sometimes prohibitively expensive, inappropriately invasive for the research context, or logistically infeasible to perform (or evaluate) in the research environment. Often alternative measures with some acknowledged mismeasurement are more affordable or more easily assessable. However, mismeasurement in outcomes is known to bias estimates of outcome summary measures and of exposure influence in many contexts including relative risk estimation [1,2,3], the general linear model [4,5], binary regression [6,7], and survival analysis.

Some specific contexts on mismeasurement in time-to-event outcomes have been addressed. These include improved estimation of true status for outcomes requiring adjudication through use of “auxiliary variables” [8], estimation of cumulative risk of false positives given repeated screening for an outcome with a non-specific diagnostic tool [9], estimation

of the perinatal transmission of HIV-1, including risk factors, when the sensitivity of the outcome measurement tool (polymerase chain reaction or PCR) changes with time since infection [10], and estimation of vaccine efficacy on disease incidence when the outcome lacks specificity [11,12,13]. Others have addressed mismeasurement in proportional hazards models, but have considered covariate rather than outcome mismeasurement. Validation sampling and auxiliary covariates have been applied in proportional hazards models to improve efficiency when covariates are measured with error [14,15,16].

For the general setting of time-to-event outcomes measured in discrete time, Richardson and Hughes [17] derived an EM algorithm for the product limit estimate of the survivor function when the binary outcome measure was subject to error. Meier et al. [18] extended this work to include estimation of covariate effects, developing an Adjusted Proportional Hazards model (APH) for mismeasured outcomes. Meier et al. showed that the APH model provided accurate hazard ratio estimates when outcome mismeasurement occurs and mismeasurement rates are known. However, when the mismeasurement rates were inaccurately known, bias could be substantial. See table 1 where their results are reproduced; a 14.5% bias in the regression covariate results from a 2% inaccuracy in assumed specificity. Thus, when sensitivity and specificity are not known with reasonable precision and accuracy, inference using APH may be no more reliable than that using standard proportional hazards.

Imperfect diagnostic measures may be used even when their accuracy is not well quantified. For example, a diagnostic test may perform in unexpected ways when used to detect a new strain or clade of an antigen, or when it is used in a new population. Diagnostic tests may also have imprecise or unknown mismeasurement rates when false results have been detected but insufficient testing has been performed to accurately define sensitivity and/or specificity. Therefore, a new method is desired to estimate hazard ratios for the context where outcome mismeasurement occurs but sensitivity and specificity are unknown.

Unquantified outcome mismeasurement rates have sometimes been addressed by performing a gold standard diagnostic test (having 100% sensitivity and specificity) on some portion of the sample: a validation subset. Pepe [19,20] developed validation subset techniques for the general regression model where a surrogate or mismeasured outcome is available on all subjects and the outcome measure is available on a subset. These methods have not previously been applied to mismeasured time-to-event outcomes.

We adapt the validation techniques from Pepe's work to the discrete-time Adjusted Proportional Hazards (APH) model for mismeasured outcomes. Use of the validation subset obviates the need for advanced knowledge of the mismeasurement rates of the surrogate outcome (sensitivity and specificity) and removes the potential to induce bias in assuming incorrect mismeasurement rates. We estimate cumulative survival and covariate effects when outcomes are measured at pre-determined discrete time points, the surrogate or screening outcome is subject to mismeasurement, and the true outcome status is available at all time points on a subset of participants.

2 Context

We apply the use of imperfectly measured outcomes to discrete survival data, where hazard ratio estimation is of interest. Auxiliary variables are not required and missed visits are not permitted. Subjects are assumed to enroll prior to failure and to be followed until the first positive screening test outcome or censoring. For example, our context includes clinical trials in which subjects are evaluated for an outcome of interest at common, pre-determined time points. We define the imperfectly measured outcome as the “screen”, and assume the gold standard outcome (GS) is perfectly sensitive and specific. For each subject, we observe the screen until either first observed failure or censoring, which must be uninformative. For

a subset, the validation set, we also know the history of the GS until the observed event time. In this work we assume that GS confirmation of false positives in real time is not possible, so that the validation subset is selected and evaluated after study completion. Therefore, GS failure status following positive screening results is not available. Selection of validation subjects depends on the method applied, but for most methods it must be a completely random subset of the study sample. Stated differently, the true outcome must be missing completely at random (MCAR). For the mean score method, selection of the validation subset can depend on both observed covariates and the observed value of the screen and therefore true failure status need only be missing at random (MAR) [21].

3 Notation

The following general notation was described by Pepe [19,20] and is used to introduce the models. For outcomes measured at a single time point, let Y refer to the true outcome, S the screen, and V the validation subset. Both Y and S are available on n_V subjects in the validation set, while Y is not measured on the remaining $n_{\bar{V}}$ subjects, and $n = n_V + n_{\bar{V}}$. $\mathbf{X}$ is the covariate vector of interest, and β parameterizes the distribution of $f_\beta(Y|X)$ and ψ that of $f_\psi(S|Y, X)$.

For discrete-time, proportional hazards models, the outcome data (Y in the general notation) are expressed as follows. We define the time of event T which may be a failure or censoring. As will be explained in §4.3, we define failure status at each time point rather than at T only. Therefore A_k is the failure status at $k \in \{1, 2, \dots, T\}$, where $A_k = 1$ if failure occurred at k and $A_k = 0$ if failure had not yet occurred at k . The failure history up to time $l \leq T$ expressed as the set of failure statuses: $H_l = \{A_k\}$, $k \in \{1, 2, \dots, l\}$. The true failure data for subject i are the time of event t_i and history $h_{it_i} = \{a_{i1}, a_{i2}, \dots, a_{it_i}\}$. For example, when subject i failed at $t_i = 3$, $h_{it_i} = \{0, 0, 1\}$. Failure status A_k is defined for subjects in the validation set from time points 1 to T , unless a false positive screen occurs as described below. Failure status A_k is not known at any time for subjects in the non-validation set. The validation subset (n_V) describes the set of participants on whom the gold standard is measured at every observed time point, while it is never measured on those in $n_{\bar{V}}$.

Due to the imperfect sensitivity and/or specificity of the surrogate (screen), the observed outcome (S in the general notation) may differ from the truth; therefore, we also define the observed time of event (either failure or censoring) as T^o and the observed failure history

$$H_{T^o}^o = \{A_k^o\}, \quad k \in \{1, 2, \dots, T^o\}, \text{ with individual values } t_i^o \text{ and } h_{it_i^o}^o = \{a_{i1}^o, a_{i2}^o, \dots, a_{it_i^o}^o\}.$$

Expressions H_T (failure history up to the time of event) can be derived from A_T (failure status at the time of event), as can $H_{T^o}^o$ from $A_{T^o}^o$, since all statuses prior to the first failure or censoring are 0 by construction.

Since the gold standard is not performed at time points occurring after observed failures, subjects who have a false positive outcome are censored at T^o and will not be observed to T . If censoring occurs at time j , neither the true nor observed outcomes will contain information occurring later than j .

For discrete-time, proportional hazards models, we also use the following parameter and covariate notation. Let the baseline hazard at each time point be represented by $\lambda_0 = (\lambda_{01}, \lambda_{02}, \dots, \lambda_{0\tau})$, where τ is the last evaluated time point (maximum follow-up length). The covariate vector of interest is $\mathbf{X} = (X_1, X_2, \dots, X_p)$ and the corresponding vector of hazard ratio coefficients is $\beta = (\beta_1, \beta_2, \dots, \beta_p)$. Only external (fixed or time-dependent) covariates are considered [22], meaning that $\mathbf{X}_i$ may change over time, so long as its value is not generated by the individual under study. Let $\mathbf{X}^S$ be the subset of $\mathbf{X}$ on which the screen may

additionally depend after conditioning on Y , such that $f(S | Y, \mathbf{X}^S) \neq f(S | Y)$. This subset of covariates must be discrete while others within $\mathbf{X}$ need not be. The parameter vector of

interest is $\Omega = \begin{pmatrix} \lambda_0 \\ \beta \end{pmatrix}$. Let sensitivity and specificity be represented by $\theta_{\mathbf{x}}$ and $\phi_{\mathbf{x}}$, respectively, where their values may depend on covariates.

4 Methods

Four discrete-time proportional hazards methods for mismeasured outcomes are introduced. The first method is standard proportional hazards (PH) involving a single outcome. We will apply it two ways: complete case analysis (§4.1.1) will be performed on the validation subset, ignoring subjects on whom the gold standard result is not available, and naive analysis (§4.1.2) will be performed using the screening outcome only on the complete cohort. Three other methods, the parametric (§4.2), empirical (§4.3), and mean-score (§4.4) methods include all available outcome measures on validated and non-validated subjects. Both the parametric and empirical methods are based on Pepe [19]. The mean score method was described by Pepe and Reilly et al. [20]. The parametric method relies on correct specification of $f_{\psi}(S | Y, X)$, while the others allow this distribution to be unspecified.

4.1 Proportional hazards

The discrete proportional hazards model for discrete survival data [22] is applied in two ways as shown below.

4.1.1 Using only complete case data—We use only the gold standard test outcomes. Subjects in the non-validation set are not used in estimation Ω of and therefore this method is expected to be least efficient, though simplest to compute. It provides consistent estimation of hazard ratios when the gold standard test result is MAR. The likelihood for discrete proportional hazards applied to the validation set only is $L^C = \prod_{i \in V} f_{\Omega}(a_{it_i} | \mathbf{X}_i)$ where:

$$f_{\Omega}(a_{it_i} | \mathbf{X}_i) = \prod_{j=1}^{t_i-1} \left\{ (1 - \lambda_{0j}) e^{(\mathbf{x}'_j \beta)} \right\} \left\{ 1 - (1 - \lambda_{0t_i}) e^{(\mathbf{x}'_{t_i} \beta)} \right\}^{a_{it_i}} \left\{ (1 - \lambda_{0t_i}) e^{(\mathbf{x}'_{t_i} \beta)} \right\}^{(1-a_{it_i})}. \quad (4.1)$$

4.1.2 Using only the screening outcome (naive)—We apply the same computation to the mismeasured outcome, available on all subjects. The likelihood when applied to the screening outcome is $\mathcal{L}^S = \prod_{i=1}^n f_{\Omega}(a_{it_i}^o | \mathbf{X}_i)$ where:

$$f_{\Omega}(a_{it_i}^o | \mathbf{X}_i) = \prod_{j=1}^{t_i^o-1} \left\{ (1 - \lambda_{0j}) e^{(\mathbf{x}'_j \beta)} \right\} \left\{ 1 - (1 - \lambda_{0t_i^o}) e^{(\mathbf{x}'_{t_i^o} \beta)} \right\}^{a_{it_i^o}^o} \left\{ (1 - \lambda_{0t_i^o}) e^{(\mathbf{x}'_{t_i^o} \beta)} \right\}^{(1-a_{it_i^o}^o)}. \quad (4.2)$$

4.2 Validation subset using parametrically designated distribution

This method parameterizes the distribution of observed outcomes conditioned on true outcomes: $f(S | Y, X)$. We assume that the sensitivity and specificity of the screen may depend on known covariates, but not on time since enrollment or on time before or since true failure. We also assume that the accuracy of each test is independent of any other test result; for example, that a subject with a previous false negative result is not inclined to

repeat this result (though subject-specific accuracy rates may be characterized by known covariates). The likelihood to be maximized is:

$$\mathcal{L}_{\beta, \psi}^P = \prod_{i \in V} f_{\beta} (y_i | \mathbf{x}_i) f_{\psi} (s_i | y_i, \mathbf{x}_i) \times \prod_{j \in \bar{V}} f_{\beta, \psi} (s_j | \mathbf{x}_j)$$

where

$$f_{\beta, \psi} (s_j | \mathbf{x}_j) = \sum_y \left[f_{\beta} (y | \mathbf{x}_j) f_{\psi} (s_j | y, \mathbf{x}_j) \right]. \quad (4.4)$$

These formulae can be adapted for discrete survival analysis to read:

$$\mathcal{L}_{\Omega, \theta, \phi}^P = \prod_{i \in V} f_{\Omega} (a_{it_i} | \mathbf{x}_i) f_{\theta, \phi} (a_{it_i}^o | a_{it_i}, \mathbf{x}_i) \times \prod_{j \in \bar{V}} f_{\Omega, \theta, \phi} (a_{jt_j}^o | \mathbf{x}_j) \quad (4.5)$$

where $f_{\Omega} (a_{it_i} | \mathbf{x}_i)$ is computed as in equation (4.1). Subjects in the non-validation set contribute the marginal probability of their observed screening results given covariates. For the j th subject, $f_{\Omega, \theta, \phi} (a_{jt_j}^o | \mathbf{x}_j)$ is computed by summing over all possible true failure statuses up to the last observed screen as follows: they either failed at $k \in \{1, 2, \dots, t_j^o\}$ or were censored at t_j^o .

$$f_{\Omega, \theta, \phi} (a_{jt_j}^o | \mathbf{x}_j) = \sum_{k=1}^{t_j^o} \left[f_{\Omega} (a_{jk}=1 | \mathbf{x}_j) \times f_{\theta, \phi} (a_{jt_j}^o | a_{jk}=1, \mathbf{x}_j) \right] + \left[f_{\Omega} (a_{jt_j}^o=0 | \mathbf{x}_j) \times f_{\theta, \phi} (a_{jt_j}^o | a_{jt_j}^o=0, \mathbf{x}_j) \right]. \quad (4.6)$$

This method makes use of the following parametric distribution of observed outcomes conditioned on true failure status.

$$\begin{aligned} f_{\theta, \phi} (a_{jt_j}^o | a_{jk}=1, k \leq t_j^o, \mathbf{x}_j) &= \phi_{\mathbf{x}_j}^{k-1} (1 - \theta_{\mathbf{x}_j})^{t_j^o - k} (1 - \theta_{\mathbf{x}_j})^{(1 - a_{jt_j}^o)} \theta_{\mathbf{x}_j}^{(a_{jt_j}^o)} \quad \text{and} \\ f_{\theta, \phi} (a_{jt_j}^o | a_{jt_j}^o=0, \mathbf{x}_j) &= \phi_{\mathbf{x}_j}^{t_j^o - 1} \phi_{\mathbf{x}_j}^{(1 - a_{jt_j}^o)} (1 - \phi_{\mathbf{x}_j})^{(a_{jt_j}^o)}. \end{aligned} \quad (4.7)$$

Covariate-specific estimates of sensitivity $\theta_{\mathbf{x}_j}$ and specificity $\phi_{\mathbf{x}_j}$ may be needed such as in a multi-center study where mismeasurement rates vary by site. However θ and ϕ could be substituted into (4.7) when mismeasurement rates do not vary with $\mathbf{X}$. Estimates of sensitivity and specificity are produced by the parametric method along with β and λ_0 .

4.3 Validation subset using empirically computed distribution

We use validation data to construct the distribution of observed data conditioned on true outcomes. The likelihood is computed using:

$$\mathcal{L}_\beta^E = \prod_{i \in V} f_\beta(y_i | \mathbf{x}_i) \prod_{j \in V} \widehat{f}_\beta(s_j | \mathbf{x}_j), \quad (4.8)$$

where $\widehat{f}_\beta(S | \mathbf{X})$ is computed using the same computation as (4.4) with $f_\beta(S | Y, \mathbf{X})$ replaced by the empirical estimate $\widehat{f}(S | Y, \mathbf{X}^S)$ and where $\mathbf{X}^S$ is the subset of $\mathbf{X}$ on which $f(S | Y)$ is assumed to depend. We count occurrences of outcomes to estimate the conditional distribution of observed outcomes in the validation set.

$$\widehat{f}(s_j | y, \mathbf{x}_j^S) = \frac{\sum_{i \in V} I(s_i = s_j, y_i = y, \mathbf{x}_i^S = \mathbf{x}_j^S)}{\sum_{i \in V} I(y_i = y, \mathbf{x}_i^S = \mathbf{x}_j^S)}. \quad (4.9)$$

We adapt the empirical method for survival analysis and rewrite the log likelihood as:

$$\mathcal{L}_\Omega^E = \prod_{i \in V} f_\Omega(a_{it_i} | \mathbf{x}_i) \times \prod_{j \in V} \widehat{f}_\Omega(a_{jt_j}^o | \mathbf{x}_j) \quad (4.10)$$

where:

$$\begin{aligned} \widehat{f}_\Omega(a_{jt_j}^o | \mathbf{x}_j) = & \sum_{k=1}^{r_j^o} \left[f_\Omega(a_{jk}=1 | \mathbf{x}_j) \times \widehat{f}(a_{jt_j}^o | a_{jk}=1, \mathbf{x}_j^S) \right] \\ & + \left[f_\Omega(a_{jt_j}^o=0 | \mathbf{x}_j) \times \widehat{f}(a_{jt_j}^o | a_{jk}=0, \mathbf{x}_j^S) \right]. \end{aligned} \quad (4.11)$$

and $\widehat{f}(a_{jt_j}^o | a_{jk}, \mathbf{x}_j^S)$ is estimated from the validation subset as follows:

$$\widehat{f}(a_{jt_j}^o | a_{jk}, \mathbf{x}_j^S) = \frac{\sum_{i \in V} I(a_{it_i}^o = a_{jt_j}^o, h_{it_i} \supset a_{jk}, \mathbf{x}_i^S = \mathbf{x}_j^S)}{\sum_{i \in V} I(h_{it_i} \supset a_{jk}, \mathbf{x}_i^S = \mathbf{x}_j^S)}. \quad (4.12)$$

The expression $h_{it_i} \supset a_{jk}$ indicates that the failure history for subject i includes the status of the j th subject at the k th time point. Subjects in the validation set contribute to this computation for each true failure status, not only at the time of event.

In the empirical method it is assumed that test accuracy may depend on known covariates, but not on time from or since failure (ie sensitivity does not increase prior to failure). However, since the probability of each test outcome is determined separately for those still under observation at each time point, test accuracy may depend on time since enrollment. Test results may also be dependent, since for example, the probability of consecutive false negatives within a subject need have no relation to the probability of a single false negative test. Therefore the empirical method may be more generally applied than the parametric method (§4.2) as it does not assume (or estimate) constant sensitivity or specificity. However, in order that observations in the non-validation set can contribute to the likelihood of this method, most combinations of X^S and T must be observed in the validation set. For this reason, X^S should be discrete.

4.4 Validation subset using Mean Score Method

The fourth technique weights each complete case data observation to incorporate the information provided by the non-validated subset. Recall that we have n_V participants in the validation subset and $n_{\bar{V}}$ with outcome assessment by screening test only. Then $n_V(s_i, \mathbf{x}_i)$ and $n_{\bar{V}}(s_i, \mathbf{x}_i)$ are the numbers of participants with screening outcome and covariates equal to $(s_i, \mathbf{x}_i)$. A solution is achieved by minimizing the mean score of the weighted likelihood. Explicit estimates of sensitivity and specificity are not produced.

When the covariates of interest in the regression model are discrete, the score function of the weighted log likelihood to be minimized is:

$$\sum_{i \in V} \frac{\partial}{\partial \beta} \left[\log f_{\beta}(y_i | \mathbf{x}_i) \right] \left\{ 1 + \frac{n_{\bar{V}}(s_i, \mathbf{x}_i)}{n_V(s_i, \mathbf{x}_i)} \right\} \quad (4.13)$$

or, in our notation,

$$\sum_{i \in V} \frac{\partial}{\partial \Omega} \left[\log f_{\Omega}(a_{it_i} | \mathbf{x}_i) \right] \left\{ 1 + \frac{n_{\bar{V}}(a_{it_i}^o, \mathbf{x}_i)}{n_V(a_{it_i}^o, \mathbf{s}_i)} \right\} \quad (4.14)$$

A Newton-Raphson algorithm is used to solve, first computing the naive Ω using the validation dataset as the starting estimate, and updating after each j th iteration until

convergence: $\widehat{\Omega}^{j+1} = \widehat{\Omega}^j + \widehat{I}_n^{-1}(\widehat{\Omega}^j) \widehat{S}(\widehat{\Omega}^j)$ where $S(\Omega) = \frac{\partial}{\partial \Omega} [\log f_{\Omega}(A_T | \mathbf{X})]$ and

$$\widehat{I}_n(\Omega) = - \frac{\partial}{\partial \Omega} (\widehat{S}(\Omega)).$$

5 Simulations

In order to compare the methods on estimation of hazard ratios, datasets were simulated. Discrete survival outcomes were created with a single binary covariate ($X = 0$ or 1) and with 5 time points of observation. Sample sizes were computed in order to achieve 50% power to detect a difference using only the validation subset [23]. This low initial power allows us to detect potentially large increases in power when the non-validated subjects are included. With a baseline hazard rate of $\lambda_0 = .10$ when $X = 0$ and $\beta = 0.7$ ($\lambda_{|X=1} = .19$), we need 80 subjects in the validation set (40 placebo, 40 intervention) to achieve 50% power. With $\lambda_0 = .20$ and $\beta = -0.4$ ($\lambda_{|X=1} = .14$), we need 210 in the validation set.

One-thousand repetitions were performed at each set of conditions, in order to achieve an approximate $\pm 1\%$ confidence interval for coverage and power. Sensitivity and specificity were imposed on the simulated events to create the observed outcomes T^o and $H_{T^o}^o$; and a validation subset was selected for which true outcomes were also retained. Imposed sensitivity, specificity and the choice and size of the validation subset varied according to the specific hypothesis tested, as described below.

Several questions were of interest: 1a) does inclusion of the non-validated subjects increase efficiency relative to the model using only the complete case data, 1b) does inclusion of the validated subjects increase accuracy relative to the naive model using only mismeasured data, 2) does the model that empirically estimates $f(S | Y)$ outperform the parametric

estimate when the parametric distribution is incorrectly specified, and 3) does the mean score method outperform all others when the validation subset is not chosen at random?

Three groups of simulations were performed to answer these questions. In the first group, we added non-validated subjects such that the validation subset of the data n_V was created to

be either $\frac{1}{2}$ or $\frac{1}{3}$ of the total number of subjects n . The validation subset was selected randomly. All four methods were compared when imposed sensitivity was 80% and imposed specificity was 95% for the screen relative to the gold standard. In the second group, we violated one of the assumptions of the parametric analysis, and imposed a time-varying specificity. Specificity decreased linearly with time since enrollment, for all subjects, from 95% at $t = 1$ to 55% at $t = 5$. The validation subset was again chosen randomly. In the third group, we examined violation of the assumption that the validation subset is representative of the study population. We performed the gold standard diagnostic tool on (or we included

in the validation subset) all participants observed to fail under the screening test and $\frac{1}{2}$ or $\frac{1}{3}$ of the remaining. Here we considered the impact of verification bias [24].

5.1 Results

Tables 2-4 correspond to questions 1-3 described above (§5). Convergence rates of the algorithms were $\geq 98\%$ for all methods over all conditions except for the mean score method in the first two conditions of table 3, when the convergence rate was 85-88%. The high false positive rate for those conditions contributed to having relatively few subjects at later time points and therefore somewhat unstable ratios used in estimation.

Table 2 shows results when simulated sensitivity and specificity were 80% and 95%, respectively, and the validation subjects were selected at random. The naive method, though relatively efficient and powerful, has up to 29.6% bias and coverage below 87%. All validation methods demonstrated an overall reduction in the mean square error relative to the complete case data method. Bias was under 5% under all conditions under validation subset methods and mostly insignificant with coverage ranging from 94-96%. A substantial increase in power (up to 37% higher) and in efficiency (up to 2.5-fold higher) relative to the complete case method was gained when using either the parametric or the empirical means of incorporating the validation subset. The increase in efficiency was not quite as large for the mean score method.

Table 3 shows relative performance when specificity depends on time since enrollment; specifically, specificity was simulated to decrease linearly with time. The naive method demonstrated even greater bias (above 55%) and poor coverage (below 45%) due to the larger number of false positive outcomes. All validation methods, however, provided appropriate coverage (94-96%). Both the parametric method (which assumed specificity was constant) and the empirical method showed an improvement in efficiency over the complete case data method; and, to a lesser degree, so did the mean score method. There is significant bias in β under all conditions for the parametric method (2.9-11.5% bias) and to a lesser degree in a few other methods (bold). The empirical method had highest power: 5-8% higher than the parametric method and 7-15% higher than the mean score; however, all methods demonstrate a reduction in the mean square error relative to the complete case method.

The last sets of simulations were performed by validating all observed failures ($A_{ro}^o=1$) plus some proportion of subjects not observed to fail ($A_{ro}^o=0$). As in the first set of conditions, specificity and sensitivity were constant at 95% and 80%, respectively, over all time points. Table 4 shows that the complete case data technique has 11.4-22.7% bias with only 87-94%

coverage, though power is inflated above the anticipated 50% since there are more events observed. The complete case method is anticipated to be biased since the validation set is mainly comprised of those detected positive by screen. The naive method, which assumes all screen positives are true, has even higher bias (above 23.4%) and lower coverage (below 87%). Some significant bias was also demonstrated for the validation methods though it was small (between 2.2 and 4.8%, shown in bold). All validation methods are less biased though similarly efficient relative to the complete case data method; and all show comparable improvements in power relative to the complete case method (10-17% greater) and improvements in coverage (1-8% greater).

The parametric method provides estimates of sensitivity and specificity of the screening test relative to the gold standard, while the other validation methods implicitly estimate accuracy of the screen. For the parametric method, average sensitivity over 1000 simulations was estimated between 79% and 81% in every case presented in tables 2-4 (standard deviation between 2% and 6%), close to the true value of 80%. Similarly, when imposed specificity was 95% at all time points (tables 2 and 4), average specificity was estimated between 94.9% and 95.0% for each conditions (standard error 0.5%-1.4%). When imposed specificity varied over time (tables 3), the parametric method average estimate of specificity was also near the anticipated average value of 82.8% (ranged between 82.6% and 83.7%, standard deviation 1.2% - 1.9%). This anticipated average value results from specificity varying between 95% at time point 1 and 55% at time point 5, with fewer subjects remaining at each later time point.

6 Application: HIV acquisition

We demonstrate the relative performance of these methods on clinical data. In an observational, prospective, cohort study of commercial sex workers in Mombasa Kenya, subjects were tested regularly for HIV-1 and other sexually transmitted infections (STIs). Of 6413 screened subjects, 2010 subjects were HIV seronegative at screening and had at least one additional monthly visit. Of these, 1738 had enrollment behavioral characteristics available, resulting in a total of 1122 person-years of follow-up and 128 HIV-1 seroconversions observed. Of interest were the relationships between demographic factors, hormonal contraceptive use, incident STIs and HIV-1 seroconversion. Complete results, study procedures and participant characteristics from an earlier cohort are described in detail elsewhere [25].

As part of study protocol, HIV testing was performed monthly using a single ELISA with confirmation of positives by a second ELISA, and if necessary by tests on additional blood specimens in the case of contradictory results. While it was anticipated that single ELISA can provide false positive results, no false negative results were expected and therefore negative ELISAs were not confirmed. The reported sensitivities and specificities of the two antibody tests used were 100.0% and 99.0% [26] for the first-line ELISA (detect-HIV by Biochem Immunosystems) and 100.0% and 99.9% [27] for the confirmatory ELISA (Recombigen by Cambridge Biotech). To demonstrate the performance of the proposed methods, we considered the first ELISA result to be the *screening test* and the HIV status confirmed via an additional ELISA to be the *gold standard*. We selected a random proportion of participants to define the validation subset and considered the confirmatory ELISAs (performed only at time points when first-line ELISA positive) on these participants only. This approach satisfies the assumptions of all three validation methods since the first ELISA is known to be 100% sensitive. Therefore no false negative results can occur and this approach is equivalent to applying a gold standard to all time points of all subjects selected as part of the validation subset. Based on screened and confirmed HIV statuses, we found the first ELISA to be 99.6% specific.

We applied Cox discrete proportional hazards (PH) to confirmed HIV status on all subjects in order to assess risk factors under the gold standard, and we considered the resulting hazard ratios to be the most accurate available in this population. To assess relative efficiency and accuracy of methods that do not rely on complete outcome ascertainment, we then compared the obtained hazard ratio estimates to those estimated from six additional proportional hazards models: 1) the PH method applied to the first ELISA result, 2),3) the Adjusted Proportional Hazards method (APH) imposing 2 values of specificity, and 4)-6) all three validation subset methods including confirmed HIV status on a 33% validation subset.

We compared the methods in estimating risk factors for HIV acquisition (Figure 1). We found that place of employment was a significant risk factor for HIV acquisition, with women working in bars having a 2.8-fold increased hazard of acquisition relative to those working in nightclubs ($\hat{\beta} = 1.04, p < .001$). The hazard ratio estimate for employment in a bar using proportional hazards applied to the first ELISA result was lower than that based on confirmed HIV status (HR= 2.1; $\hat{\beta} = 0.74, p < .001$). Adjusted Proportional Hazards (APH) methods estimates were wide and highly subject to assumed specificity: HR= 3.3 ($\hat{\beta} = 1.19, p = .004$) when the assumed specificity was 99.6% and HR= 4.1 ($\hat{\beta} = 1.42, p = .011$) when the assumed specificity was 99.5%. Empirical, parametric and mean score validation methods estimates of $\hat{\beta}$ more closely matched that based on confirmed HIV status: HR= 2.8 ($\hat{\beta} = 1.06, p < .001$), HR= 2.6 ($\hat{\beta} = 0.97, p = .001$) and HR= 3.2 ($\hat{\beta} = 1.17, p < .001$), respectively. All the validation subset methods described herein have higher efficiency than the APH method and better accuracy than PH applied to the screening results only; and their confidence intervals followed closely that of the estimate based on confirmed HIV status for the complete cohort (shaded region).

Since the sensitivity and specificity of these tests were higher than those used in the simulations, 1000 additional simulations were performed under similar conditions to the data above: sensitivity = 100%, specificity = 99.5%, $\lambda_0 = .10, \beta = 1, 1200$ -1800 total

observations, and $\frac{1}{2}$ or $\frac{1}{3}$ of participants in the validation set. These simulations confirmed that even with at high specificity all three validation methods provided excellent coverage (94-96%), power (100%), non-significant bias (< 2%), and increased precision (relative efficiency 1.9-3.3) over the complete case data methods.

7 Discussion

We presented three techniques to incorporate validation subsets into discrete proportional hazards models. We compared their performance in terms of bias, coverage, power and mean square error when various scenarios were assumed.

With regard to the first question posed in (§5), we consider the findings of table 2. When mismeasurement was MCAR and therefore the assumptions of all models held, all validation subset methods demonstrated increased efficiency and power relative to the complete case data, and reduced bias and improved coverage relative to the naive method. The mean score method provided the lowest gains for these conditions. This may be because, unlike other validation methods, the mean score method links true outcomes between subjects in the validation and non-validation subsets on the observed time of event T^0 and not at other time points.

To address question 2 we compared the empirical and parametric estimation methods when the parametric distribution was incorrectly specified. Under greatly varying specificity, the empirical method outperformed the parametric method in terms of power. However, we found that the methods performed similarly when smaller variations in specificity were

imposed (data not shown), and concluded that the empirical estimation method may be relatively robust to departures from the assumption of time-constant mismeasurement.

Lastly, we tested whether the mean score method performs best when the assumption of random sampling of the validation set was not met (table 4). The mean score method performed nearly identically to the parametric and empirical methods in terms of power, bias and coverage, all demonstrating an improvement over the complete case method. Any of these methods might be a good choice under these specific conditions.

The data example confirms both that the APH is highly sensitive to mismeasurement accuracy rates and that the validation subset methods provide improved accuracy and precision. The example also confirms the apparent preferability of the empirical and parametric methods, as their estimates include the entire confidence interval obtained using true HIV status, while that of the mean score method does not.

We recommend that sample size for the validation set be chosen so that the anticipated number of subjects at risk in each arm at the final time point is no less than 10; since convergence problems were absent with 12-15 expected subjects at time point 5 but present when the expected number was 7-10. This can be achieved during the study design phase and by grouping visits if necessary.

Limitations of our findings include the inability to remove bias completely in some simulations even when all assumptions were met. However, this bias was small (under 5%) and was much less than that found using methods for which assumptions were violated. We found that bias was reduced to negligible amounts in all simulated cases when data is simulated with perfect sensitivity (data not shown), indicating that when there is less uncertainty in the timing of events, performance of the validation methods improve. Another limitation is the examination of a relatively small set of conditions: hazard of failure between .1 and .2, hazard ratio between .7 and 2. It is difficult to derive general theoretical conclusions since estimation occurs via numerical maximization and the variance is not straightforward to compute. These findings may not apply under other conditions; however, time-to-event studies are unlikely to be designed for events with a higher than 20% hazard rate.

Other methods that were not compared include multiple imputation techniques for the true outcome, where imputed values could be derived based on observed relationships between true and surrogate outcomes in the validation subset. This was developed by Chen [28] for mismeasured outcomes in the general linear model but might be adapted for time-to-event data.

Under all conditions evaluated, we found that validation subset methods can be beneficial in improving power and reducing bias when computing hazard ratios with mismeasured outcomes. Similar benefits in estimation of cumulative survival were achieved but are not shown for reasons of space. The empirical estimate appeared to perform at least as well as the others under all conditions considered, was robust to departures from the assumption of random sampling for the validation subset, and might be a good general choice to use when assumptions cannot be tested.

8 Acknowledgments

This work was partly funded by the Fred Hutchinson Cancer Research Center and later by NIH grant AI-30731-13 under Dr. Anna Wald. Colleagues James P. Hughes and Barbra A. Richardson provided valuable critique and guidance.

References

1. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology*. 1977; 105:488–495. [PubMed: 871121]
2. Green MS. Use of predictive value to adjust relative risk estimates biased by misclassification of outcome status. *American Journal of Epidemiology*. 1983; 117:98–105. [PubMed: 6823958]
3. Irwig LM, Groeneweld HT, Simpson JM. Correcting for measurement error in an exposure-response relationship based on dichotomising a continuous dependent variable. *The Australian Journal of Statistics*. 1990; 32:261–269.
4. Buonaccorsi JP. Correcting for nonlinear measurement errors in the dependent variable in the general linear model. *Communications in Statistics - Theory and Methods*. 1993; 22(10):2687–2702.
5. Buonaccorsi JP. Measurement error in the response in the general linear model. *Journal of the American Statistical Association*. 1996; 91:633–642.
6. Magder L, Hughes JP. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*. 1997; 146(2):195–203. [PubMed: 9230782]
7. Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*. 1999; 86:843–855.
8. Snapinn SM. Survival analysis with uncertain endpoints. *Biometrics*. 1998; 54:209–218. [PubMed: 9574966]
9. Gelfand AE, Wang F. Modelling the cumulative risk for a false-positive under repeated screening events. *Statistics in Medicine*. 2000; 19(14):1865–1879. [PubMed: 10867676]
10. Balasubramanian R, Lagakos SW. Estimation of the timing of perinatal transmission of HIV. *Biometrics*. 2001; 57:1048–1058. [PubMed: 11764243]
11. Halloran ME, Longini IM. Using validation sets for outcomes and exposure to infection in vaccine field studies. *American Journal of Epidemiology*. 2001; 155(5):391–398. [PubMed: 11532779]
12. Chu H, Halloran ME. Estimating vaccine efficacy using auxiliary outcome data and a small validation sample. *Statistics in Medicine*. 2004; 23(17):2697–2711. [PubMed: 15316953]
13. Scharfstein DO, Halloran ME, Chu H, Daniels MJ. On estimation of vaccine efficacy using validation samples with selection bias. *Biostatistics*. 2006; 7(4):615–629. [PubMed: 16556610]
14. Zhou H, Pepe MS. Auxiliary covariate data in failure time regression. *Biometrika*. 1995; 82(1): 139–149.
15. Wang CY. Augmented inverse probability weighted estimator for cox missing covariate regression. *Biometrics*. 2001; 57(2):414–419. [PubMed: 11414564]
16. Cheng HY. Cox Regression in cohort studies with validation sampling. *Journal of the Royal Statistical Society Series B*. 2002; 64(1):51–62.
17. Richardson BA, Hughes J. Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics*. 2000; 1(3):341–54. [PubMed: 12933514]
18. Meier AS, Richardson BA, Hughes JP. Discrete proportional hazards models for mismeasured outcomes. *Biometrics*. 2003; 59(4):947–954. [PubMed: 14969473]
19. Pepe MS. Inference using surrogate outcome data and a validation sample. *Biometrika*. 1992; 79:355–365.
20. Pepe MS, Reilly M, Fleming TR. Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*. 1994; 42:137–160.
21. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd edn. Wiley; New Jersey: 2002.
22. Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. 2nd edn. Wiley; New York: 2002.
23. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry losses to follow-up noncompliance and stratification (Corr: V42 p1009). *Biometrics*. 1986; 42:507–519. [PubMed: 3567285]

24. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983; 39:207–215. [PubMed: 6871349]
25. Martin HL, Nyange PM, Richardson BA, Lavreys L, Mandaliya K, Jackson JD, Ndinya-Achola JO, Kreiss J. Hormonal contraception sexually transmitted diseases and risk of heterosexual transmissions of human immunodeficiency virus type 1. *Journal of Infectious Diseases*. 1998; 178:1053–1059. [PubMed: 9806034]
26. Galli RA, Castriciano S, Fearon M, Major C, Choi KW, Mahony J, Chernesky M. Performance characteristics of recombinant enzyme immunoassay to detect antibodies to human immunodeficiency virus type 1 HIV-1 and HIV-2 and to measure early antibody responses in seroconverting patients. *Journal of Clinical Microbiology*. 1996; 34(4):999–1002. [PubMed: 8815130]
27. Kuun E, Brashaw M, Heyns ADP. Sensitivity and specificity of standard and rapid HIV-antibody tests evaluated by seroconversion and non-seroconversion low-titre panels. *Vox Sanguinis*. 1997; 72(1):11–15. [PubMed: 9031494]
28. Cheng HY. A robust imputation method for surrogate outcome data. *Biometrika*. 2000; 87(3):711–716.

Figure 1.

Estimation of the hazard ratio of HIV acquisition for employment in a bar versus nightclub. Comparing estimates based on confirmed HIV status with three techniques using screening test only (assuming 100% sensitivity and varying assumptions regarding specificity) and three techniques using validation subsets: parametric (Para), empirical (Emp) and the mean score method (MS). The shaded area encompasses the confidence interval based on confirmed HIV status.

Table 1

Effects of mismeasurement when specificity is 90%: estimation of $\beta = 1.3$ from 1000 simulations of discrete proportional hazards models (PH and APH) performed on the mismeasured outcome with varying assumed values of specificity.

Method	Specificity Simulated	Specificity Assumed	% Bias in $\hat{\beta}$
PH	90%	100%	-34.7% ($\pm 1\%$)
APH	90%	90%	0.4% ($\pm 1\%$)
APH	90%	88%	14.5% ($\pm 1\%$)

Table 2

Comparing estimation of β , the coefficient for the binary covariate, when sensitivity = 80%, specificity = 95% using 5 methods: proportional hazards using either the mismeasured outcome (Naive) or complete case only (Comp) and validation subset methods using either parametric distribution of $F(S|Y)$ (Para), empirical distribution of $F(S|Y)$ (Emp) or the mean score method (MS).

Obs.	p_V	λ_0	β	Method	%Bias	Cov.	Power	Var.	RelEff	MSE
160	1/2	0.1	0.7	PH: Naive	-28.6%	83%	71%	0.040	3.2	0.081
				PH: Comp	3.4%	95%	50%	0.127	(ref)	0.127
				Para	1.3%	95%	77%	0.071	1.8	0.071
				Emp	-1.7%	95%	76%	0.068	1.9	0.068
				MS	-0.3%	95%	67%	0.084	1.5	0.084
240	1/3	0.1	0.7	PH: Naive	-27.7%	79%	86%	0.029	4.4	0.066
				PH: Comp	1.7%	94%	49%	0.127	(ref)	0.128
				Para	3.0%	94%	87%	0.056	2.3	0.056
				Emp	-0.8%	95%	85%	0.054	2.4	0.054
				MS	0.4%	95%	72%	0.076	1.7	0.076
420	1/2	0.2	-0.4	PH: Naive	-24%	87%	70%	0.015	2.7	0.024
				PH: Comp	0.9%	96%	51%	0.040	(ref)	0.040
				Para	0.5%	96%	77%	0.022	1.8	0.022
				Emp	-0.3%	96%	76%	0.022	1.8	0.022
				MS	-3.9%	95%	73%	0.022	1.8	0.022
630	1/3	0.2	-0.4	PH: Naive	-23.4%	85%	85%	0.010	3.9	0.019
				PH: Comp	-0.1%	96%	52%	0.039	(ref)	0.039
				Para	-0.7%	95%	89%	0.015	2.5	0.015
				Emp	-1.8%	95%	89%	0.015	2.5	0.015
				MS	-4.4%	95%	76%	0.020	1.9	0.020

Bias significantly different from zero is shown in bold. Abbreviated column headings are as follows: "Obs." = number of observations, " p_V " = probability of inclusion in the validation set, " λ_0 " = simulated baseline hazard, " β " = simulated covariate coefficient, "Cov." = coverage, "Var" = variance, "RelEff" = efficiency relative to the complete case method, and "MSE" = mean square error.

Table 3

Comparing estimation of β , the coefficient for the binary covariate, when sensitivity = 80%, and specificity decreases from 95% at $t = 1$ to 55% at $t = 5$ using 5 methods: proportional hazards using either the mismeasured outcome (Naive) or complete case only (Comp) and validation subset methods using either parametric distribution of $F(S|Y)$ (Para), empirical distribution of $F(S|Y)$ (Emp) or the mean score method (MS).

Obs.	p_V	λ_0	β	Method	%Bias	Cov.	Power	Var.	RelEff	MSE
160	1/2	0.1	0.7	PH: Naive	-60.4%	33%	35%	0.031	4.6	0.21
				PH: Comp	3.1%	95%	47%	0.143	(ref)	0.143
				Para	-5.1%	95%	54%	0.098	1.5	0.099
				Emp	-2.9%	95%	59%	0.097	1.5	0.098
				MS	-0.8%	96%	50%	0.120	1.2	0.121
240	1/3	0.1	0.7	PH: Naive	-60.9%	13%	51%	0.019	7.9	0.201
				PH: Comp	1.9%	95%	43%	0.151	(ref)	0.152
				Para	-11.5%	94%	54%	0.085	1.8	0.092
				Emp	-2.7%	96%	62%	0.085	1.8	0.086
				MS	-1.1%	95%	54%	0.112	1.4	0.112
420	1/2	0.2	-0.4	PH: Naive	-55.2%	45%	38%	0.011	4.2	0.06
				PH: Comp	2.5%	96%	48%	0.046	(ref)	0.046
				Para	-2.9%	96%	59%	0.031	1.5	0.031
				Emp	1.4%	95%	64%	0.031	1.5	0.031
				MS	-2.2%	95%	57%	0.033	1.4	0.033
630	1/3	0.2	-0.4	PH: Naive	-55.1%	27%	55%	0.007	6.3	0.056
				PH: Comp	0.3%	96%	48%	0.044	(ref)	0.044
				Para	-10.7%	94%	68%	0.022	2.0	0.024
				Emp	-2.0%	95%	75%	0.023	1.9	0.023
				MS	-5.0%	96%	60%	0.029	1.5	0.030

Bias significantly different from zero is shown in bold. Abbreviated column headings are as follows: "Obs." = number of observations, " p_V " = probability of inclusion in the validation set, " λ_0 " = simulated baseline hazard, " β " = simulated covariate coefficient, "Cov." = coverage, "Var" = variance, "RelEff" = efficiency relative to the complete case method, and "MSE" = mean square error.

Comparing estimation of β , the coefficient for the binary covariate, when sensitivity = 80%, specificity = 95%, and all observed failures and a random subset of censorings are validated using 5 methods: proportional hazards using either the mismeasured outcome (Naive) or complete case only (Comp) and validation subset methods using either parametric distribution of $F(S|Y)$ (Para), empirical distribution of $F(S|Y)$ (Emp) or the mean score method (MS).

Table 4

Obs.	$p_V \mid A_{T^0}^O = 0$	λ_0	β	Method	%Bias	Cov.	Power	Var.	RelEff	MSE
160	1/2	0.1	0.7	PH: Naive	-28.6%	83%	71%	0.040	1.4	0.081
				PH: Comp	-11.4%	94%	74%	0.056	(ref)	0.062
				Para	2.6%	95%	85%	0.057	1.0	0.057
				Emp	2.2%	95%	84%	0.057	1.0	0.057
				MS	-1.0%	95%	84%	0.054	1.0	0.054
240	1/3	0.1	0.7	PH: Naive	-27.7%	79%	86%	0.029	1.4	0.066
				PH: Comp	-18.6%	90%	81%	0.040	(ref)	0.057
				Para	1.6%	94%	94%	0.040	1.0	0.040
				Emp	1.1%	95%	94%	0.040	1.0	0.040
				MS	-1.7%	94%	93%	0.039	1.0	0.039
420	1/2	0.2	-0.4	PH: Naive	-24.0%	87%	70%	0.015	1.2	0.024
				PH: Comp	-15.6%	93%	71%	0.018	(ref)	0.022
				Para	-0.7%	96%	83%	0.018	1.0	0.018
				Emp	-1.0%	96%	83%	0.018	1.0	0.018
				MS	-4.8%	95%	82%	0.017	1.1	0.017
630	1/3	0.2	-0.4	PH: Naive	-23.4%	85%	85%	0.010	1.2	0.019
				PH: Comp	-22.7%	87%	79%	0.012	(ref)	0.020
				Para	-0.7%	95%	96%	0.012	1.0	0.012
				Emp	-1.0%	95%	96%	0.011	1.1	0.011
				MS	-4.5%	95%	95%	0.012	1.0	0.012

Bias significantly different from zero is shown in bold. Abbreviated column headings are as follows: "Obs." = number of observations, " p_V " = probability of inclusion in the validation set, " λ_0 " = simulated baseline hazard, " β " = simulated covariate coefficient, "Cov." = coverage, "Var" = variance, "RelEff" = efficiency relative to the complete case method, and "MSE" = mean square error.