Published in final edited form as:

Biometrics. 2009 December ; 65(4): 1194–1202. doi:10.1111/j.1541-0420.2008.01178.x.

Flexible Regression Model Selection for Survival Probabilities: With Application to AIDS

A. Gregory DiRienzo

Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.

dirienzo@biostat.harvard.edu

SUMMARY

Clinicians are often interested in the effect of covariates on survival probabilities at pre-specified study times. Since different factors can be associated with the risk of short-term and long-term failure, a flexible modeling strategy is pursued. Given a set of multiple candidate working models, an objective methodology is proposed that aims to construct consistent and asymptotically normal estimators of regression coefficients and average prediction error for each working model, that are free from the nuisance censoring variable. It requires the conditional distribution of censoring given covariates to be modeled. The model selection strategy uses stepup or stepdown multiple hypothesis testing procedures that control either the proportion of false positives or generalized familywise error rate when comparing models based on estimates of average prediction error. The context can actually be cast as a missing data problem, where augmented inverse probability weighted complete case (AIPWCC) estimators of regression coefficients and prediction error can be used (Tsiatis, 2006). A simulation study and an interesting analysis of a recent AIDS trial are provided.

Keywords

Average prediction error; Censored data; Doubly-robust estimator; Restricted moment model; Model misspecification; Nonparametric bootstrap; Simultaneous inference

1. Introduction

Researchers are often interested in evaluating the effects of a set of covariates on a survival time endpoint. In some settings, it is of importance to identify potential treatment by covariate interactions. For example, AIDS Clinical Trials Group (ACTG) study 398 (Hammer, Vaida, et al. 2002) was a randomized study of four salvage therapy regimens for HIV-1 infected subjects with loss of virologic suppression. In addition to a backbone regimen (Amprenavir, Abacavir, Efavirenz and Adefovir Dipivoxil), subjects received one of three protease inhibitors or placebo. Some other baseline covariates of interest were non-nucleoside reverse transcriptase inhibitor (NNRTI) experience (yes or no), $\log_{10}$ HIV-1 RNA copies/mL, and reverse-transcriptase codon 103 (mutant or not). The primary endpoint was defined as the time from randomization to going *off-track* for suppression of HIV-1 RNA copies/mL (Huang, DeGruttola, et al. 2001). Going off-track is an empirical definition of virologic failure that has been shown to generally agree with protocol-based definitions. Basically, at each follow-up visit, a subject is classified as either on-track for suppression of

Supplementary Materials

Web Appendices referenced in Sections 2, 3, 4 and 7 are available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>

HIV-1 RNA, or not (off-track). Going off-track generally refers to either a sharp, unexpected rebound in HIV-1 RNA, or failure to achieve a substantial initial drop in HIV-1 RNA. The censoring time in this study was the time from randomization to the first missed follow-up visit, or study closure when no follow-up visits were missed, which was observed in full for each subject. Since treatment efficacy may depend on subgroups of subjects as defined by baseline covariates, it is of interest to investigate potential treatment by covariate interactions on the risk of going off-track. Furthermore, it would be informative to know if factors associated with the risk of early failure (at or before week 8) differ from those associated with the risk of late failure (at or before week 24).

A common way to model survival data is thru the semiparametric approaches suggested by Cox (1972), Cheng, Wei, et al. (1995), and Kalbfleisch and Prentice (2002, Section 7). These methods are useful for a *global* assessment of the effect of covariates on survival. In fact, the Cox (1972) model can allow for dynamic covariate effects by including time dependent covariates, as nicely illustrated in Kalbfleisch and Prentice (2002, Section 6.4). To compare working models, a desirable criterion is average prediction error (Tian, Cai, et al., 2007 for uncensored outcomes; Uno, Cai, et al., 2007 for censored outcomes). However, with censored survival data, it is in general difficult to obtain estimates of average prediction error because the support of the censoring variable is typically much shorter than that of the survival time (Sinisi, Neugebauer, et al. 2006). Thus a *localized* approach to model selection may be preferred.

Uno, Cai, et al. (2007) considered a localized approach that focused on survival prediction at one or more pre-specified failure times. Their methodology permits flexible modeling of the effect of covariates on survival by allowing regression coefficients and link functions to arbitrarily vary with time in binary regression models. The inference procedures provided by Uno, Cai, et al. (2007) provide consistent estimation of and large-sample confidence *intervals* for regression coefficients and average absolute prediction error, that are free of the nuisance censoring variable. Although their results do not require that working survival models are properly specified, Uno, Cai, et al. (2007) do work under the assumption that the distribution of the censoring time is independent of both survival time and covariates.

Overall, the purpose of this paper is to develop an objective, statistically formal yet flexible localized survival model selection methodology under the assumption of conditional independence of censoring and survival times given covariates. Furthermore, the proposed methods provide *simultaneous* inference over all time points and model comparisons of interest, based on estimates of average prediction error, by using multiple testing procedures that control either the generalized familywise error rate or proportion of false positives.

More specifically, the proposed methodology acts locally by specifying a set of multiple, varying-sized candidate working models for survival prediction at each of several pre-specified failure times. The conditional distribution of censoring time given covariates needs to be modeled so that consistent and asymptotically normal estimators of regression parameters and average prediction error, that are free of the nuisance censoring variable, can be obtained. This paper assumes that the censoring variable is observed in full for each subject, which allows a flexible semiparametric modeling approach to be suggested. The Discussion addresses how to handle the case when the censoring variable is not observed in full for each subject. Consistent estimates of regression coefficients and average prediction error are obtained when the working model for censoring is properly specified. When the working model for censoring is misspecified, but the working model for survival given covariates is properly specified, estimated regression coefficients are consistent, but estimated average prediction error is in general not. However, simulations demonstrate that

fact may not preclude consistent selection (and estimation) of the most parsimonious candidate survival model.

At each pre-specified study time, t , estimation of regression coefficients and average prediction error of working survival models use so-called inverse probability weighted complete case estimators (IPWCC) (Uno, Cai, et al., 2007). The complete cases here are those subjects for which the value of $I(T \leq t)$ is able to be assessed, where T is the survival time. For those subjects censored at or before time t , the variable $I(T \leq t)$ is missing, otherwise it is observed. As illustrated in this paper, this framework allows use of augmented inverse probability weighted complete case estimators (AIPWCC) (Tsiatis, 2006), which attempt to gain efficiency by using information from subjects with missing $I(T \leq t)$, thereby making use of all available data.

The paper is organized as follows. Section 2 describes the background and notation. Estimation of regression parameters is detailed in Section 3 and of average prediction error in Section 4. Section 5 outlines the multiple testing procedures used in the context of the proposed model selection strategy that is detailed in Section 6. A simulation study is presented in Section 7 and an analysis of ACTG 398 in Section 8. Some final remarks are given in the Discussion.

2 Background and Notation

Let T denote the survival time and let C denote the potential censoring time. Denote the unique observed failure times by $t_1^*, \dots, t_{R^*}^*$, and suppose the time points of interest are $t_1, \dots, t_R$. For simplicity in the exposition below, consider only one time point $t \in (t_1, \dots, t_R)$, so-called t -year survival prediction. It should be noted that, in addition to $R > 1$ time points of interest, it is straight-forward to simultaneously accommodate several types of survival times. For example, in ACTG 398, one may be interested in the time to off-track and the time to treatment toxicity.

Let $X = \min(T, C)$, with $\delta = I(C > T)$, $I(\cdot)$ the indicator function, and let $\mathbf{Z}$ denote the column vector of baseline covariates. The methodology presented below requires that the conditional survival function for censoring given covariates be estimated. It is frequently the case in studies of HIV/AIDS that C is observed in full for each subject. This is assumed to be the case in the main body of the paper; the Discussion addresses the case when C is potentially censored. Suppose that the observed data consists of n independent realizations of $(X, \delta, C, \mathbf{Z})$ from the unknown probability distribution $P(T, C, \mathbf{Z})$, denoted $(X_i, \delta_i, C_i, \mathbf{Z}_i)$, $i = 1, \dots, n$.

Consider L pre-specified functions $\tilde{\mathbf{Z}}_\ell$ of $\mathbf{Z}$, $\ell = 1, \dots, L$. Each $\tilde{\mathbf{Z}}_\ell$ can be a subset of $\mathbf{Z}$ and can also include simple transformations of the covariates such as logarithms, squares and interaction terms; $\tilde{\mathbf{Z}}_\ell$ usually will include the constant 1 to correspond to an intercept term. For each $\tilde{\mathbf{Z}}_\ell$, denote the corresponding *working model* for $P(T \leq t | \tilde{\mathbf{Z}}_\ell)$ by $g_\ell(\beta_\ell^\top \tilde{\mathbf{Z}}_\ell)$, where $a^\top$ is the transpose of matrix a , and for fixed ℓ and t , $g_\ell(x)$ is a known, strictly increasing, differentiable function of $x \in (-\infty, \infty)$ and β_ℓ is an unknown column vector of parameters that is defined below, $\ell = 1, \dots, L$. For example, setting $g_\ell(\cdot) = 1 - \exp\{-\exp(\cdot)\}$ corresponds to a proportional hazards model; setting $g_\ell(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$ corresponds to a proportional odds model. Note that, for ease of exposition, t is suppressed in the notation $g_\ell(\cdot)$ and β_ℓ , however, these quantities are allowed to arbitrarily depend on time (and also survival type), as illustrated in Section 8.

To evaluate the fit of each of the L candidate survival models $g_\ell(\beta_\ell^\top \tilde{\mathbf{Z}}_\ell)$, consider average squared prediction error $d_\ell = E\{I(T \leq t) - g_\ell(\beta_\ell^\top \tilde{\mathbf{Z}}_\ell)\}^2$, $\ell = 1, \dots, L$, where expected value $E(\cdot)$ is with respect to $P(T, \mathbf{Z})$. Average squared prediction error is used here primarily for mathematical convenience when constructing AIPWCC estimators, as shown below. To evaluate the adequacy of the ℓ th working survival model, it is required that d_ℓ and β_ℓ , which do not depend on the nuisance censoring variable C , be consistently estimated.

This paper shows how to obtain consistent estimators of β_ℓ and d_ℓ that are asymptotically normal (and free of the nuisance variable C) when T and C are conditionally independent given $\mathbf{Z}$, written $T \perp C | \mathbf{Z}$, and the working model for $P(C > x | \mathbf{Z})$, $x \in \mathcal{T}_t^*$, is properly specified, where $\mathcal{T}_t^* = \{t_1^*, \dots, \min(t, t_{R^*}^*)\}$, regardless of whether $g_\ell(\beta_\ell^\top \tilde{\mathbf{Z}}_\ell)$ is properly specified.

The construction of AIPWCC estimators for β_ℓ and d_ℓ includes an augment term, which requires an estimate for $P(T \leq t | \mathbf{Z})$, say $\hat{P}^*(T \leq t | \mathbf{Z})$ based on the working model $P^*(T \leq t | \mathbf{Z})$, whereas IPWCC estimators do not. The IPWCC estimator of β_ℓ will be so-called *doubly robust* for those specific working survival models that are properly specified, that is, satisfy $P(T \leq t | \mathbf{Z}) = g_\ell(\beta_\ell^\top \tilde{\mathbf{Z}}_\ell)$; AIPWCC estimators of β_ℓ in addition require $P(T \leq t | \mathbf{Z}) = P^*(T \leq t | \mathbf{Z})$. Note that properly specified working models may be inefficient in the sense that they include regression terms whose coefficients are equal to 0. Doubly-robust here means that the estimator for β_ℓ corresponding to these specific working models will be consistent and asymptotically normal (and free of the nuisance variable C) when $T \perp C | \mathbf{Z}$, even if the model for $P(C > x | \mathbf{Z})$, $x \in \mathcal{T}_t^*$, is misspecified. These results are shown in Web Appendix A.

When $T \perp C | \mathbf{Z}$ and the model for $P(C > x | \mathbf{Z})$, $x \in \mathcal{T}_t^*$, is misspecified, it is not in general the case that corresponding estimators for d_ℓ are consistent and asymptotically normal when working survival models are properly specified. When estimating β_ℓ under these conditions, the corresponding estimating equation can be shown to be asymptotically equal to 0 at β_ℓ , its value in the absence of censoring. When estimating d_ℓ under these conditions, its asymptotic value equals some positive number, which is in general not equal to its value in the absence of censoring, d_ℓ . However, the value of d_ℓ is not of intrinsic interest itself; rather the working survival models corresponding to the smallest d_ℓ . Thus, if the ranking of the asymptotic values of average prediction error under these conditions is the same as those corresponding to d_ℓ (at least for the best fitting working models) then the most parsimonious working survival model can still be selected, and, if properly specified, have consistent estimators of β_ℓ . This appears to be the case in the simulations reported.

3 Estimation of β_ℓ

Using the results from Tsiatis (2006, section 10.2), define the AIPWCC estimating equation

$$\begin{aligned} \widehat{U}_\ell(\beta) &= \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i}{\widehat{P}^*(C > X_i \wedge t | \mathbf{Z}_i)} \tilde{\mathbf{Z}}_{i,t} \{I(X_i \leq t) - g_\ell(\beta^\top \tilde{\mathbf{Z}}_{i,t})\} \\ &\quad - \left\{ \frac{\Delta_i}{\widehat{P}^*(C > X_i \wedge t | \mathbf{Z}_i)} - 1 \right\} \tilde{\mathbf{Z}}_{i,t} \{\widehat{P}^*(T \leq t | \mathbf{Z}_i) - g_\ell(\beta^\top \tilde{\mathbf{Z}}_{i,t})\}, \end{aligned}$$

where $\Delta_i = \delta_i I(X_i \leq t) + I(X_i > t)$, $\tilde{\mathbf{Z}}_{i,\ell}$ is the value of $\tilde{\mathbf{Z}}_\ell$ for the i th subject, $\widehat{P}^*(C > x | \mathbf{Z}_i)$ is an estimate for $P(C > x | \mathbf{Z}_i)$ based on the working model $P^*(C > x | \mathbf{Z})$, $\widehat{P}^*(T < t | \mathbf{Z}_i)$ is an estimate for $P(T < t | \mathbf{Z}_i)$ based on the working model $P^*(T \leq t | \mathbf{Z})$ and fit using only those observations with $\Delta = 1$. Note that t is suppressed in the notation for Δ . The term in the second line of

$\hat{U}_\ell(\beta)$ is the augment term that attempts to gain efficiency by using information from subjects with $\Delta = 0$; setting this term to 0 for all subjects corresponds to an IPWCC estimating equation. Note that it is only necessary to evaluate $P^*(C > x | \mathbf{Z}_i)$ at the time points $\mathcal{T}_t^*$. Let $\hat{\beta}_\ell$ satisfy $\hat{U}_\ell(\hat{\beta}_\ell) = 0$. The estimating equation in Uno, Cai, et al. (2007) differs from that above in that no augment term is used and the censoring distribution is estimated marginally, not conditionally on $\mathbf{Z}$.

Note that when the covariate vector $\mathbf{Z}$ is categorical, a saturated model for $P(C > x | \mathbf{Z})$ can be constructed for each $x \in \mathcal{T}_t^*$, which is nonparametric and thus properly specified. When a saturated model for $P(C > x | \mathbf{Z})$ is not used for each $x \in \mathcal{T}_t^*$, for instance when $\mathbf{Z}$ consists of one or more continuous components, semiparametric approaches can be used. For example, a flexible way to estimate $P(C > x | \mathbf{Z})$ can be to specify a separate working model for each $x \in \mathcal{T}_t^*$, $P^*(C > x | \mathbf{Z}) = h_x(\gamma_x^\top \tilde{\mathbf{Z}}_C)$, where $h_x(\cdot)$ is a known, strictly increasing, differentiable function, γ_x is an unknown column vector of parameters, defined below, and $\tilde{\mathbf{Z}}_C$ is a function of $\mathbf{Z}$. Estimates of γ_x , denoted by $\hat{\gamma}_x$, are obtained by solving

$$0 = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{Z}}_{i,C} \{I(C_i > x) - h_x(\hat{\gamma}_x^\top \tilde{\mathbf{Z}}_{i,C})\},$$

where $\tilde{\mathbf{Z}}_{i,C}$ is the value of $\tilde{\mathbf{Z}}_C$ for the i th subject, $i = 1, \dots, n$. The estimator $\hat{\gamma}_x$ converges in probability to the limit γ_x , $x \in \mathcal{T}_t^*$.

The working model $P^*(T \leq t | \mathbf{Z})$ is estimated using only those observations with $\Delta = 1$, in which case $I(X \leq t) = I(T \leq t)$. As a result, one may consider $P^*(T \leq t | \mathbf{Z}) = h_T(\gamma_T^\top \tilde{\mathbf{Z}}_T)$, where $h_T(\cdot)$ is defined as $h_x(\cdot)$, $\tilde{\mathbf{Z}}_T$ is some function of $\mathbf{Z}$ and γ_T is estimated by $\hat{\gamma}_T$, with

$$0 = \frac{1}{n} \sum_{i=1}^n \Delta_i \tilde{\mathbf{Z}}_{i,T} \{I(X_i \leq t) - h_T(\hat{\gamma}_T^\top \tilde{\mathbf{Z}}_{i,T})\}$$

and $\tilde{\mathbf{Z}}_{i,T}$ is the value of $\tilde{\mathbf{Z}}_T$ for the i th subject, $i = 1, \dots, n$. The estimator $\hat{\gamma}_T$ converges in probability to the limit γ_T .

Now, to show consistency of the estimator $\hat{\beta}_\ell$, as given by Uno, Cai, et al. (2007), it suffices to show that $E\{\hat{U}_\ell(\beta)\} = E[\tilde{\mathbf{Z}}_\ell \{I(T \leq t) - g_\ell(\beta^\top \tilde{\mathbf{Z}}_\ell)\}] = \mu_\ell(\beta)$. Then, under the mild condition that there does not exist a β such that $P(\beta^\top \tilde{\mathbf{Z}}_{i,\ell} > \beta^\top \tilde{\mathbf{Z}}_{j,\ell} | T_i \leq t \leq T_j) = 1$, $i < j$, and using arguments similar to those in Web Appendix A of Tian, Cai, et al. (2007), it can be shown that $u_\ell(\beta) = 0$ has a unique solution, namely at $\beta = \beta_\ell$. Furthermore, if there does not exist β such that $\delta_i I(X_i \leq t \leq X_j) = 1$ implies $I(\beta^\top \tilde{\mathbf{Z}}_{i,\ell} \leq \beta^\top \tilde{\mathbf{Z}}_{j,\ell}) = 1$, $i \leq j$, $\hat{U}_\ell(\beta)$ has a unique solution, $\hat{\beta}_\ell$, for any finite n . Web Appendix A shows the consistency of $\hat{\beta}_\ell$ when $T \perp C | \mathbf{Z}$ and either $P^*(C > x | \mathbf{Z})$ is properly specified, $x \in \mathcal{T}_t^*$, or $g_\ell(\beta_\ell^\top \tilde{\mathbf{Z}}_\ell)$ (and $P^*(T \leq t | \mathbf{Z})$ when AIPWCC estimators are used) is properly specified.

4 Estimation of d_ℓ

Define the empirical average squared prediction error

$$\begin{aligned}
\widehat{D}_\ell &= \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i}{\widehat{P}^*(C > X_i \wedge t | \mathbf{Z}_i)} \left\{ I(X_i \leq t) - g_\ell(\widehat{\beta}_\ell^\top \tilde{\mathbf{Z}}_{i,\ell}) \right\}^2 \\
&\quad - \left\{ \frac{\Delta_i}{\widehat{P}^*(C > X_i \wedge t | \mathbf{Z}_i)} - 1 \right\} [\widehat{P}^*(T \leq t | \mathbf{Z}_i) \{1 - 2g_\ell(\widehat{\beta}_\ell^\top \tilde{\mathbf{Z}}_{i,\ell})\} + g_\ell^2(\widehat{\beta}_\ell^\top \tilde{\mathbf{Z}}_{i,\ell})] \\
&= \frac{1}{n} \sum_{i=1}^n A_{i,\ell}.
\end{aligned}$$

The statistic $\widehat{D}_\ell$ is an estimate of the so-called *apparent* error. The augment term that is subtracted in the second line of $\widehat{D}_\ell$ is included in an attempt to gain efficiency by using information from those subjects with missing $I(T \leq t)$ (Tsiatis, 2006, section 10.2). It is shown in Web Appendix A that $\widehat{D}_\ell$ converges uniformly in probability to d_ℓ when $P^*(C > x | \mathbf{Z})$ is properly specified, $x \in \mathcal{X}_t^*$.

5 Simultaneous Inference on (d_ℓ)

Suppose it is of interest to simultaneously test the sequence of null hypotheses $H_\ell : d_\ell = d_0$ versus the alternative $d_\ell < d_0$, where d_0 corresponds to some reference model. The estimator $\widehat{D}_\ell$ has an asymptotic normal distribution and, if $P^*(C > x | \mathbf{Z})$ is properly specified, $x \in \mathcal{X}_t^*$, is a consistent estimator for d_ℓ . Thus a test statistic on H_ℓ can be defined as $\hat{S}_\ell = (\widehat{D}_\ell - \widehat{D}_0)/\hat{\sigma}_\ell$, where $\hat{\sigma}_\ell = \{n^{-1} \text{var}(A_{i,\ell} - A_{i,0})\}^{1/2}$ and $\text{var}(\cdot)$ denotes sample variance.

When the working model for censoring is properly specified, the statistic $\hat{S}_\ell$ is asymptotically standard normal under H_ℓ for any $P(T, C, \mathbf{Z})$ so that $\hat{\pi}_\ell = \Phi(\hat{S}_\ell)$ asymptotically satisfies $P(\hat{\pi}_\ell \leq u) \leq u$, for $u \in (0, 1)$, $\Phi(\cdot)$ the standard normal cumulative distribution function. As a result, stepup or stepdown multiple hypothesis testing procedures (MTPs) can be used to provide asymptotic *strong* control of a given error rate, that is, for any $P(T, C, \mathbf{Z})$; this property is useful in this context since the $(\hat{\pi}_\ell)$ are likely to be correlated. The error rates considered are the generalized familywise error rate, gFWER(k), and proportion of false positives, PFP(θ). Romano and Shaikh (2006) provide stepup procedures that asymptotically bound above at target level α (i) the probability of more than k rejected true null hypotheses (gFWER(k)), and (ii) the probability that the proportion of true null hypotheses among the rejected hypotheses (PFP) is greater than θ , i.e., $P(\text{PFP} > \theta) < \alpha$ (PFP(θ)). Since the PFP(θ) retains the same interpretation regardless of the number of null hypotheses tested, it will be solely used for the remainder of the paper. Note that gFWER($k = 0$) and PFP($\theta = 0$) correspond to the familywise error rate (FWER). The false discovery rate, defined as $E(\text{PFP}) < \alpha$, is not used in this paper because it does not provide any control over the variability of the PFP.

The stepup algorithm given by Romano and Shaikh (2006) to control the PFP(θ) at level α is as follows. Suppose there are B null hypotheses to be tested. Put the observed p-values in descending order to obtain $\hat{\pi}_{(B)} \geq \dots \geq \hat{\pi}_{(1)}$ and let $H_{(B)}, \dots, H_{(1)}$ denote the corresponding null hypotheses. If $\pi_{(B)} \leq \alpha \alpha'_B / c(B, \theta)$ then reject all B null hypotheses. Otherwise, reject null hypotheses $H_{(b^*)}, \dots, H_{(1)}$, where b^* is the smallest index such that

$$\hat{\pi}_{(B)} > \frac{\alpha}{c(B, \theta)} \alpha'_B, \dots, \hat{\pi}_{(b^*+1)} > \frac{\alpha}{c(B, \theta)} \alpha'_{b^*+1},$$

and $\alpha'_b = \{\text{int}(\theta b) + 1\} / \{B - b + \text{int}(\theta b) + 1\}$, with $\text{int}(a)$ the integer part of a . If $b^* = 0$ then don't reject any hypotheses. The formula for the constant $c(B, \theta)$ is provided in Romano and

Shaikh (2006). Lehmann and Romano (2005) and Romano and Shaikh (2004) provide stepdown algorithms that provide strong control of the $gFWER(k)$ and $PFP(0)$.

6 Model Selection Strategy

The following is an approach that attempts to identify the best fitting working survival model with the smallest number of regression terms at each failure time of interest. The algorithm works in two general steps. The first is to identify the model comparisons of interest at each failure time under consideration; once these have been identified, an MTP from section 5 is then used to simultaneously test all corresponding null hypotheses across all failure times considered. To identify the model comparisons of interest, the algorithm first compares the fit of all working models to the null model, where only an intercept term is fit. If working models are identified that provide a significantly better fit to the data than the null model, then, the reference model is re-defined from the null model to the smallest working model among these with the best fit. Working models offering a better fit than this new reference are then compared to it. The algorithm iterates in this fashion, continuing to re-define the reference null model, until no working models are determined be significantly better than the reference. A step-by-step description is provided below, where for simplicity, t is consider fixed.

0. Set $i = 1$.
1. Define $H_\ell^{(i)}: d_\ell = d_0^{(i)}$ against the one-sided alternative $d_\ell < d_0^{(i)}$. Here $d_0^{(1)} = E\{I(T \leq t) - g_0^{(1)}(\beta_0^{(1)})\}^2$ corresponds to the null model with only an intercept. Test the sequence of null hypotheses $(H_\ell^{(i)})$ using an MTP as outlined in Section 5, denoting test statistics by $(\widehat{S}_\ell^{(i)})$ and p-values by $(\widehat{\pi}_\ell^{(i)})$.
2. Re-define the reference model to be the working model rejected from $(H_\ell^{(i)})$ with the smallest number of regressors with the smallest $\widehat{\pi}_\ell^{(i)}$; denote this model index by ℓ_i and its population average squared prediction error by $d_0^{(i+1)}$, so that $d_0^{(i+1)} = d_{\ell_i}$. If none of $(H_\ell^{(i)})$ are rejected then stop.
3. Define $H_\ell^{(i+1)}: d_\ell = d_0^{(i+1)}$, against $d_\ell < d_0^{(i+1)}$, where the index ℓ corresponds to those working models rejected from $(H_\ell^{(i)})$ with $\widehat{\pi}_\ell^{(i)} < \widehat{\pi}_{\ell_i}^{(i)}$; if no $\widehat{\pi}_\ell^{(i)} < \widehat{\pi}_{\ell_i}^{(i)}$ then stop. Otherwise test this sequence using an MTP as outlined in Section 5, where $\widehat{S}_\ell^{(i+1)}$ is now defined using the estimate of $d_0^{(i+1)}$ as a reference, with p-value $\widehat{\pi}_\ell^{(i+1)}$.
4. Repeat 2. and 3. above, incrementing $i = i + 1$ after each iteration.

The algorithm above is repeated for each $t \in (t_1, \dots, t_R)$; let NH_r denote the total number of null hypotheses tested for each t_r , $r = 1, \dots, R$. The total number of model comparisons of interest identified by this algorithm is $NH = NH_1 + \dots + NH_R$. In order to maintain the global false-positive rate, it is necessary to test all of the NH null hypotheses simultaneously using an MTP as outlined in Section 5. For each $t \in (t_1, \dots, t_R)$, the final model choice will correspond to the working model with the smallest number of regressors, with the smallest p-value, in the largest stage i that was rejected.

7 Simulation Study

A simulation study was performed to investigate the small sample properties of the proposed model selection methodology. For simplicity, only comparisons to the null model were considered (stage $i = 1$). Two probability distributions were initially used. First, for $P_1(T, C, \mathbf{Z})$, the vector $\mathbf{Z}$ consisted of 7 covariates from the probability distribution $P(\mathbf{Z})$ defined in Web Appendix B. The underlying survival time, say T^* , was generated from the

exponential distribution with hazard function $\lambda(t|\mathbf{Z}) = \frac{1}{4} \exp(0.5Z_7)$, $Z_7 \sim N(0, 1)$, the actual survival time was then $T = \text{round}(T^* + 1)$ so that the range for T was the positive integers. The censoring time C was defined as the minimum of an administrative censoring time, say C_1 , that was uniformly distributed on $(3, 10)$, and a loss to followup time, say C_2 , that was

exponential with hazard function $\lambda_{C_2}(t|\mathbf{Z}) = \frac{1}{8} \exp(0.25Z_7)$, the actual censoring time was then $C = \text{round}[\min\{C_1, (C_2 + 1)\}]$ so the range for C was the positive integers. In all cases the outcome of interest was $I(T \leq 3)$. Using a single random sample with $n = 200000$ observations, $P_1(T, C, \mathbf{Z})$ results in about 50% of survival times being censored and about 74% with $\Delta = 1$. In the absence of censoring, $P_1(T, \mathbf{Z})$ results in about 13%, 33% and 48% of subjects with failure at or before time 1, 2 and 3, respectively. Note this sample of $n = 200000$ was used only to approximate the parameters of $P_1(\cdot)$.

The second probability distribution, $P_2(\cdot)$, was defined as $P_1(\cdot)$ except for $\mathbf{Z}$. Here, $\mathbf{Z}$ had 12 covariates, the first 7 defined as Z_1 thru Z_7 from $P_1(\cdot)$, and the last 5, Z_8 thru Z_{12} , defined as Z_{12} thru Z_8 from $P_1(\cdot)$. This distribution was used to investigate the impact of increasing the number of null hypotheses on the small sample properties of the methodology. For both $P_1(\cdot)$ and $P_2(\cdot)$ the value of $\beta_{1,7}$ in the working model for $P(T \leq 3|Z_7)$ defined by $g((\beta_{0,7} + \beta_{1,7}Z_7))$ is 0.70, where $g(\cdot)$ is the anti-logit link function, $g(x) = e^x/(1 + e^x)$.

In all cases, the working model for $P(C > x|\mathbf{Z})$ was $h(\gamma_{0,x} + \gamma_{1,x}^\top \mathbf{Z})$, for $x = 1, 2, 3$, $h(\cdot)$ the anti-logit link function. Two settings were considered for when the working model for censoring was misspecified. First, a case when the functional form of a covariate was misspecified, the

conditional hazard for C_2 was $\lambda_{C_2}(t|\mathbf{Z}) = \frac{1}{8} \exp(0.25|Z_7|)$. Second, a case when an important covariate was omitted, $\lambda_{C_2}(t|\mathbf{Z}) = \frac{1}{8} \exp(0.25Z_7 + 0.5Z_7^2)$. The working model for $P(T \leq 3|\mathbf{Z})$, used in the augment terms, was $h(\gamma_{0,T} + \gamma_{1,T}^\top \mathbf{Z})$.

The distributions $P_1(\cdot)$ and $P_2(\cdot)$ have 7 and 12 covariates respectively, so that the total number of candidate working models for $I(T \leq 3)$, using only one-way terms, is $2^7 - 1 = 127$ and $2^{12} - 1 = 4095$. All working models for survival used the anti-logit link function. The best fitting, most parsimonious model for $I(T \leq 3)$ is the one that has Z_7 as the only predictor. For both $P_1(\cdot)$ and $P_2(\cdot)$, the population value of d_ℓ is 0.225 for models containing Z_7 and 0.250 for models not containing Z_7 .

The simulations were conducted as follows. Separately for each sample size and each distribution, a random sample $(\mathbf{Z}_i)_{i=1}^n$ was generated from $P_j(\mathbf{Z})$ a total of 1000 independent times. At each of the 1000 simulation iterations, $Z_{i,1}$ was randomly generated from $P_j(Z_1)$, then $Z_{i,2}$ was randomly generated from $P_j(Z_2|Z_{i,1})$, and so on. Then, at each simulation iteration, with $(\mathbf{Z}_i)_{i=1}^n$ considered fixed, the random samples $(T_i)_{i=1}^n$ and $(C_i)_{i=1}^n$ were independently generated from each respective distribution. The simulations, and the analysis of ACTG 398, were executed using Matlab version 7.4 (R2007a) and with 64-bit Intel Xeon 3.2 GHz processors.

The empirical power of several different MTPs corresponding to $H_7 : d_7 = d_0$, where $\ell = 7$ indexes the model with Z_7 as the only term, is shown in Table 1 for several sample sizes and for $P_1(\cdot)$ and $P_2(\cdot)$. The target level error was $\alpha = 0.1$ in all cases. The MTP with the largest power was the stepup procedure using the $\text{PFP}(\theta = 0.1)$. Figure 1 plots these powers for the stepup algorithm. Notice that, when increasing the number of null hypotheses tested, there is only a small reduction in power of the MTP using the $\text{PFP}(\theta = 0.1)$, however, the reduction when using the $\text{PFP}(\theta = 0)$ (FWER) is large. For $P_1(\cdot)$, the empirical FWER for all MTPs was 0.004 or less for $n = 100$, and was 0.002 or less for $n = 200$ thru 1000, with the vast majority equal to 0. For $P_2(\cdot)$, the empirical FWER of all MTPs was less than 0.024 for $n = 100$, less than 0.009 for $n = 200$, and 0.005 or less for $n = 300$ thru 1000, with the vast majority equal to 0. Note that with 1000 simulation iterations and a nominal level error of $\alpha = 0.1$, a normal-theory based 95% confidence interval for the population $\text{PFP}(\theta)$ extends ± 0.0186 from the empirical $\text{PFP}(\theta)$.

Define the probability distributions $P_3(\cdot)$ and $P_4(\cdot)$ as $P_1(\cdot)$ except with the first and second alternate definitions of C_2 above, respectively. The empirical FWER and powers to reject H_7 corresponding to $P_3(\cdot)$ and $P_4(\cdot)$ were all very similar to those presented for $P_1(\cdot)$, suggesting minor impact from misspecification $P(C > x|Z)$.

Table 2 presents the empirical mean and standard deviation of estimates for the coefficient $\beta_{1,7}$ and average squared prediction error d_7 for the candidate survival model $P(T \leq 3|Z_7) = g(\beta_{0,7} + \beta_{1,7}Z_7)$, for $P_1(\cdot)$, $P_3(\cdot)$ and $P_4(\cdot)$ and several sample sizes. Results for $P_2(\cdot)$ were very similar to those of $P_1(\cdot)$. For all distributions, empirical means of the estimates $\hat{\beta}_{1,7}$ and $\hat{D}_7$ are close to their population values of 0.70 and 0.225, respectively, with empirical standard deviations decreasing with increasing sample size. Define the working model index $\ell = 1$ by $P(T \leq 3|Z_1) = g(\beta_{0,1} + \beta_{1,1}Z_1)$; note that $H_1 : d_1 = d_0$ is true. As shown in Table 2, the empirical mean and standard deviation of $\hat{S}_1$, corresponding to H_1 , differ somewhat from the target values 0 and 1, with variance overestimation present. The empirical mean of $\hat{S}_7$, corresponding to H_7 , increases with sample size, and, although variance overestimation is present, its empirical standard deviation is closer to 1 compared to $\hat{S}_1$. Since the statistic $\hat{D}_\ell$ is asymptotically linear, the nonparametric bootstrap can be used to estimate the variance of $\hat{D}_\ell - \hat{D}_0$ (Dudoit, van der Laan and Pollard, 2004). The numbers in parentheses under $\hat{S}_7$ correspond to using the nonparametric bootstrap for standard error estimation instead of $\hat{\sigma}_7$; here, at each simulation iteration, the estimates $\hat{P}^*(C > x|Z)$ and $\hat{P}^*(T \leq 3|Z)$ were calculated once on the observed data and fixed thru the 100 nonparametric bootstrap iterations. Although there were slight improvements to the means and standard deviations from using the bootstrap, the empirical powers and sizes were very similar to tests that used $\hat{\sigma}_7$ for standard error estimation.

8 Analysis of ACTG 398

Consider model selection for the risk of going off-track at or before week 8 (early failure) and at or before week 24 (late failure). The time scale used in this analysis was weeks, as subjects were scheduled for follow up visits at week 2, 4, 8, 16 and every 8 weeks thereafter until week 48. Note that actual visit week was used in the analysis and not scheduled visit week. The censoring time was the time from randomization to going off-study, or study closure, whichever occurred first, and thus was observed in full for each subject. Full baseline data was available for $n = 450$ subjects, of those, a total of 331 subjects were observed to go off-track; 163 at or before week 8 and 284 at or before week 24. There were observed off-track events at weeks 1 thru 24; the shortest censoring time was at week 1. For early and late failure, the percent of subjects with $\Delta = 1$ was 98.4% and 96.4%, respectively.

The following four baseline covariates were considered: treatment group (TRT), coded as 1=additional PI, 0=placebo; NNRTI experience (NN), coded as 1=naive, 0=experienced; $\log_{10}$ HIV-1 RNA (RNA); and reverse-transcriptase codon 103, coded as 1=wildtype, 0=mutant. Besides these four main effect terms, it was of interest to identify potential treatment by covariate interactions, denoted by TRT*NN, TRT*RNA and TRT* 103. By imposing the restriction that working models that include an interaction term also include the corresponding main effect terms, at a given study time of interest, $t \in (8, 24)$, the total number of candidate survival models of one term, two terms, up thru seven terms, is 34.

The working model for the censoring distribution, $P^*(C > x|\mathbf{Z})$, included all seven regression terms, and was separately estimated at each observed off-track time less than or equal to 24 weeks. The working model $P^*(T \leq t|\mathbf{Z})$, for use in the augment terms, $t = 8, 24$, also included all seven regression terms. The link function used in all cases was the anti-logit. The nonparametric bootstrap with 500 realizations was used for variance estimation in test statistics, for which the estimated working models $\hat{P}^*(C > x|\mathbf{Z})$ and $\hat{P}^*(T \leq t|\mathbf{Z})$ were calculated once on the observed dataset then fixed for use in each bootstrap iteration.

In all cases, the MTP used was the stepup procedure with the PFP($\theta = 0.1$) and $\alpha = 0.1$. To identify the model comparisons of interest, first, for week 8, 34 hypotheses were tested with the null model as the reference model (stage $i = 1$). The results are displayed in Table 3. A total of 30 of 34 hypotheses were rejected. The smallest working model rejected with

smallest p-value had RNA as the only term ($\hat{\pi}_{\ell_1}^{(1)} = 0.001$). Using this model as the reference model, the second stage ($i = 2$) tested 21 hypotheses, corresponding to those models in Table 3 with $\hat{\pi}_{\ell}^{(1)} < \hat{\pi}_{\ell_1}^{(1)}$. Of these 21 hypotheses, the smallest model rejected with smallest p-value had 2 terms, NN and RNA ($\hat{\pi}_{\ell_2}^{(2)} = 0.0034$). Using this model as the reference model, stage $i = 3$ tested 13 hypotheses, none were rejected. Thus the total number of model comparisons identified for week 8 was $NH_1 = 34 + 21 + 13 = 68$.

For week 24, the first stage ($i = 1$) tested 34 hypotheses using the null model as the reference model. The smallest model rejected with smallest p-value had NN as the only term

($\hat{\pi}_{\ell_1}^{(1)} = 0.001$). Using this model as the reference model, stage $i = 2$ tested 21 hypotheses, corresponding to those models with $\hat{\pi}_{\ell}^{(1)} < \hat{\pi}_{\ell_1}^{(1)}$. Of these 21 hypotheses, the smallest model rejected with smallest p-value had 3 terms, NN, TRT and RNA ($\hat{\pi}_{\ell_2}^{(2)} = 0.0088$). Using this model as the reference model, stage $i = 3$ tested 11 hypotheses, none were rejected. For week 24, the total number of model comparisons identified was $NH_2 = 34 + 21 + 11 = 66$.

The final step was to simultaneously test all of the $NH = 68 + 66 = 134$ hypotheses corresponding to all model comparisons of interest. For week 8, the smallest model with smallest p-value (0.0034) from latest stage rejected ($i = 2$) had 2 terms, NN and RNA. For a fixed value of RNA, being naive to NNRTIs at baseline was estimated to reduce the odds of early failure by 70% compared to being experienced; for a fixed value of NN, a $\log_{10}$ increase in baseline viral load was estimated to increase the odds of early failure by 2.2 fold. For week 24, the smallest model with smallest p-value (0.0054) from latest stage rejected ($i = 2$) had 4 terms, NN, TRT, RNA and TRT*NN. Thus a treatment effect was identified on the risk of late failure but not early failure. For a fixed value of RNA: the additional PI was estimated to reduce the odds of late failure by 29% for NNRTI naive subjects, and 70% for NNRTI experienced subjects compared to placebo; compared to having NNRTI experience, being naive to NNRTIs was estimated to reduce the odds of late failure by 75% for subjects receiving an additional PI, and 89% for subjects receiving placebo. For a fixed value of NN

and TRT, a $\log_{10}$ increase in baseline viral load was estimated to increase the odds of late failure by 1.9 fold.

In conclusion, NNRTI experience and HIV-1 RNA level at baseline were found to be significantly associated with the risk of early and late failure. There was a treatment effect detected only on the risk of late failure, where, after adjusting for baseline NNRTI experience and HIV-1 RNA level, the additional PI was more beneficial for NNRTI experienced subjects than for naive subjects.

9 Discussion

This paper provides a statistically formal yet flexible approach to objective model selection for censored survival data, which can be especially useful when covariate effects depend on time. Consistent estimators of regression parameters and average prediction error for candidate survival models, that are free of the nuisance censoring variable, are obtained when the conditional censoring distribution given covariates is correctly modeled. When the censoring variable is completely observed for each subject in the dataset, this article shows how flexible working models for the conditional censoring distribution can be formed. When the model for censoring is misspecified, consistent estimators of regression parameters require proper specification of the candidate survival model given covariates and $P^*(T \leq t|\mathbf{Z})$ when AIPWCC estimators are used; however estimators of average prediction error are not necessarily consistent in this case. Nevertheless, one may still be able to identify the most parsimonious working survival model and consistently estimate its regression parameters when the censoring distribution is mismodeled, as was evidenced in the simulation study.

The potential censoring time C may not be completely observed for each subject, for example when T is the time to death. In such cases, one can reverse the roles of T and C and use the semiparametric modeling approaches mentioned in the Introduction or those presented in this paper to form working models for $P(C > x|\mathbf{Z})$. However, having to select two censored-data regression models simultaneously can be burdensome.

Apparent error was used instead of cross validated error (Shao, 1993; 1996) to estimate average squared prediction error because of the savings in computation. The proposed model selection strategy is designed to potentially accommodate a large number of candidate working survival models. The small-sample correction that cross validation provides to estimates of average prediction error may be negligible in the test statistics $\hat{S}_\ell$, which are based on the difference between two prediction error estimates. However, one may choose to implement cross-validated prediction error estimates, perhaps as a descriptive statistic for a selected survival model.

Although only the PFP(θ) was used in the simulations and analysis of ACTG 398, one could have used the gFWER(k). However, unlike the PFP(θ), interpretation of the gFWER(k) depends on the total number of null hypotheses tested.

When relatively sparse covariates are considered, for example, treatment by categorical covariate interactions, the derivative of the estimating equation corresponding parameters of the working models for censoring or survival may not be of full rank. In these settings, one may use the trust-region dogleg method (Moré, Garbow, et al., 1980; Powell, 1970) instead of the Newton-Raphson algorithm for root finding.

Acknowledgments

This work was supported by grant R01-AI058836 from the U.S. National Institutes of Health. The author is grateful to an Associate Editor and Referee whose comments improved the clarity and presentation of the paper.

References

- Cheng SC, Wei LJ, Ying Z. Analysis of Transformation Models with Censored Data. *Biometrika*. 1995; 82:835–845.
- Cox DR. Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society, Series B*. 1972; 34:187–220.
- Dudoit S, van der Laan M, Pollard K. Multiple Testing. Part I. Single-Step Procedures for Control of General Type I Error Rates. *Statistical Applications in Genetics and Molecular Biology*, 3, Article 13. 2004 <http://www.bepress.com/sagmb/vol3/iss1/art13>.
- Hammer S, Vaida F, Bennett K, et al. Dual vs Single Protease Inhibitor Therapy Following Antiretroviral Treatment Failure: A Randomized Trial. *Journal of the American Medical Association*. 2002; 288(2):169–180. [PubMed: 12095381]
- Huang W, DeGruttola V, Fischl M, et al. Patterns of Plasma Human Immunodeficiency Virus Type-1 RNA Response to Antiretroviral Therapy. *Journal of Infectious Diseases*. 2001; 183(10):1455–1465. [PubMed: 11319681]
- Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. 2nd ed.. New York: John Wiley & Sons, Inc.; 2002.
- Lehmann EL, Romano JP. Generalizations of the Familywise Error Rate. *Annals of Statistics*. 2005; 33:1138–1154.
- Moré, J.; Garbow, B.; Hillstom, K. *User Guide for MINPACK 1*. Argonne National Laboratory; 1980. Rept. ANL-80-74.
- Powell, MJD. A Fortran Subroutine for Solving Systems of Nonlinear Algebraic Equations. In: Rabinowitz, P., editor. *Numerical Methods for Nonlinear Algebraic Equations*. 1970.
- Romano, J.; Shaikh, A. Technical Report No. 2004–31. Dept. of Statistics, Stanford Univ.; 2004. On Control of the False Discovery Proportion.
- Romano J, Shaikh A. Stepup Procedures for Control of Generalizations of the Familywise Error Rate. *Annals of Statistics*. 2006; 34:1850–1873.
- Shao J. Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*. 1993; 88:486–494.
- Shao J. Bootstrap Model Selection. *Journal of the American Statistical Association*. 1996; 91:655–665.
- Sinisi SE, Neugebauer R, van der Laan MJ. Cross-validated Bagged Prediction of Survival. *Statistical Applications in Genetics and Molecular Biology*. 2006; 5(1) Article 12.
- Tian L, Cai T, Goetghebeur E, Wei LJ. Model Evaluation Based on the Distribution of Estimated Absolute Prediction Error. *Biometrika*. 2007; 94:297–311.
- Tsiatis, AA. *Semiparametric Theory and Missing Data*. New York: Springer; 2006.
- Uno H, Cai T, Tian L, Wei LJ. Evaluating Prediction Rules for t -Year Survivors with Censored Regression Models. *Journal of the American Statistical Association*. 2007; 102:527–537.

Figure 1.

Empirical power to reject H_7 for stepup procedure with $\text{PFP}(\theta = 0)$ ('o') and $\text{PFP}(\theta = 0.1)$ ('*') for $P_1(\cdot)$ (dashed lines) and $P_2(\cdot)$ (solid lines).

Table 1

Simulation results: empirical power of PFP(θ) at target level $\alpha = 0.1$

MTP		Sample size (n)										
		100	200	300	400	500	600	700	800	900	1000	
$P_1(\cdot)$:												
stepup,	$\theta = 0$	0.015	0.055	0.188	0.383	0.600	0.766	0.898	0.964	0.988	0.995	
	$\theta = 0.1$	0.037	0.158	0.388	0.624	0.827	0.922	0.976	0.996	0.997	1	
	stepdown,	$\theta = 0$	0.021	0.095	0.286	0.501	0.717	0.847	0.941	0.983	0.996	0.999
		$\theta = 0.1$	0.032	0.102	0.269	0.441	0.664	0.801	0.912	0.967	0.985	0.997
$P_2(\cdot)$:												
stepup,	$\theta = 0$	0.002	0.006	0.016	0.060	0.131	0.253	0.390	0.582	0.710	0.840	
	$\theta = 0.1$	0.033	0.099	0.268	0.512	0.727	0.876	0.959	0.984	0.994	0.999	
	stepdown,	$\theta = 0$	0.003	0.012	0.031	0.089	0.194	0.365	0.503	0.685	0.813	0.894
		$\theta = 0.1$	0.025	0.038	0.069	0.148	0.215	0.359	0.505	0.659	0.770	0.861

Table 2

Simulation results: empirical mean and standard deviation

Sample Size	$\hat{\beta}_{1,7}$	$\hat{D}_7$	$\hat{S}_1$	$\hat{S}_7$
$P_1(\cdot)$:				
$n = 200$	0.7116	0.2223	0.3972	2.1165 (2.1972)
	0.2007	0.0115	0.3106	0.6252 (0.6865)
$n = 400$	0.7038	0.2235	0.3814	2.9850 (3.0011)
	0.1364	0.0080	0.2955	0.6072 (0.6594)
$n = 800$	0.6994	0.2242	0.3959	4.2219 (4.2542)
	0.0966	0.0057	0.3088	0.6120 (0.6958)
$P_3(\cdot)$:				
$n = 200$	0.7145	0.2219	0.4012	2.1347 (2.1474)
	0.2098	0.0121	0.3104	0.6431 (0.6936)
$n = 400$	0.7047	0.2233	0.3880	3.0062 (3.0541)
	0.1422	0.0084	0.2993	0.6177 (0.7126)
$n = 800$	0.6981	0.2242	0.3936	4.2415 (4.3108)
	0.1010	0.0060	0.3015	0.6256 (0.7165)
$P_4(\cdot)$:				
$n = 200$	0.7416	0.2196	0.4012	2.3034 (2.3268)
	0.2624	0.0151	0.3044	0.7684 (0.8164)
$n = 400$	0.7388	0.2208	.3781	3.3012 (3.3251)
	0.1744	0.0103	0.2873	0.7221 (0.8066)
$n = 800$	0.7314	0.2218	0.3776	4.6680 (4.7224)
	0.1241	0.0074	0.2949	0.7360 (0.8074)

Note: For each sample size, first row is mean, second row is standard deviation.

Table 3

Week 8 results from ACTG 398

t	$\widehat{\pi}_\ell^{(1)}$	Regression terms
1	0.0017	103, NN
2	0.0014	NN
3	0.0013	103, NN, TRT, TRT*103
4	0.0013	103, NN, TRT
5	0.0013	103, NN, TRT, TRT*NN
6	0.0012	103, NN, TRT, TRT*103, TRT*NN
7	0.0011	NN, TRT
8	0.0011	NN, TRT, TRT*NN
9	0.0010	RNA
10	0.0010	TRT, RNA, TRT*RNA
11	0.0008	TRT, RNA
12	0.0003	103, TRT, RNA, TRT*103, TRT*RNA
13	0.0003	103, TRT, RNA, TRT*RNA
14	0.0002	103, RNA
15	0.0002	103, TRT, RNA, TRT*103
16	0.0002	103, TRT, RNA
17	0.0000	103, NN, RNA
18	0.0000	NN, RNA
19	0.0000	103, NN, TRT, RNA, TRT*103, TRT*RNA
20	0.0000	103, NN, TRT, RNA, TRT*NN, TRT*RNA
21	0.0000	103, NN, TRT, RNA, TRT*RNA
22	0.0000	103, NN, TRT, RNA, TRT*103, TRT*NN, TRT*RNA
23	0.0000	103, NN, TRT, RNA, TRT*103
24	0.0000	103, NN, TRT, RNA, TRT*NN
25	0.0000	NN, TRT, RNA, TRT*NN, TRT*RNA
26	0.0000	NN, TRT, RNA, TRT*RNA
27	0.0000	103, NN, TRT, RNA
28	0.0000	103, NN, TRT, RNA, TRT*103, TRT*NN
29	0.0000	NN, TRT, RNA, TRT*NN
30	0.0000	NN, TRT, RNA

Note: Those values of $\widehat{\pi}_\ell^{(1)}$ given by 0.0000 are < 0.0001 .