Published in final edited form as:

Biometrics. 2013 March ; 69(1): 52–61. doi:10.1111/j.1541-0420.2012.01818.x.

Latent Subgroup Analysis of a Randomized Clinical Trial Through a Semiparametric Accelerated Failure Time Mixture Model

L. Altstein* and

Novartis Institutes for Biomedical Research, Cambridge, MA 02139, USA

G. Li*

Department of Biostatistics, School of Public Health, University of California at Los Angeles, Los Angeles, CA 90095

Summary

This paper studies a semiparametric accelerated failure time mixture model for estimation of a biological treatment effect on a latent subgroup of interest with a time-to-event outcome in randomized clinical trials. Latency is induced because membership is observable in one arm of the trial and unidentified in the other. This method is useful in randomized clinical trials with all-or-none noncompliance when patients in the control arm have no access to active treatment and in, for example, oncology trials when a biopsy used to identify the latent subgroup is performed only on subjects randomized to active treatment. We derive a computational method to estimate model parameters by iterating between an expectation step and a weighted Buckley-James optimization step. The bootstrap method is used for variance estimation, and the performance of our method is corroborated in simulation. We illustrate our method through an analysis of a multicenter selective lymphadenectomy trial for melanoma.

Keywords

All-or-none non-compliance; Buckley-James estimator; Clinical trials; Competing risks; EM algorithm; Non-proportional hazards model; Treatment efficacy

1. Introduction

Subgroup analyses aim to estimate treatment effects within groups of the patient population distinguished by baseline characteristics and are typically auxiliary to the primary *intention-to-treat* analysis comparing groups defined by the randomization assignment (Lagakos, 2006; Wang et al., 2007). Our interest lies in scenarios where the auxiliary treatment effect pertains only to a specific subgroup that is partially obscured by randomization. This problem occurs in oncology trials of solid tumor cancers such as breast, thyroid and melanoma, when patients with metastatic lymph nodes may stand to benefit from complete lymph node dissection, whereas patients with absent metastases do not and are potentially harmed by the surgery (Gipponi et al., 2004). Clinical trials designed to evaluate the usefulness of sentinel node biopsy in tumor staging may randomize one arm to receive the

*lily.altstein@novartis.com. *vli@ucla.edu.

Supplementary Materials

Web Appendices and Tables referenced in Sections 3.1 and 4.2 are available with this paper at the *Biometrics* website on Wiley Online Library.

biopsy, with subsequent lymph node dissection for patients whose biopsies are positive for metastasis (“node-positive”), and the other arm to the current (nonsurgical) standard-of-care. Alternatively, the protocol may specify dissection for all patients randomized to treatment, regardless of nodal status. Both of these designs induce latent subgroups in the control arm because membership to the node-positive and -negative groups is revealed through a biomarker available only in the treatment arm.

When the interpretation or definition of a treatment effect differs among levels of a baseline covariate, the intention-to-treat test assesses the *effectiveness* of a new treatment strategy that applies to the patient population in aggregate. In the preceding oncology examples, a natural corollary is a test of whether lymph node dissection improves survival in the node-positive subgroup. We refer to this estimand as *biological efficacy* because it pertains to the subgroup of patients who stand to benefit medically from the treatment. These ideas were discussed by Sommer and Zeger (1991) in the context of noncompliance in the treatment arm with a binary outcome and formalized in the extension of instrumental variables methods for causal inference to general types of all-or-none compliance behavior and outcome distributions (Angrist et al., 1996; Imbens and Rubin, 1997). Statistical models in this framework postulate a mixture structure to the population and aim to estimate the causal or biological effect of treatment in the (partially) unidentified class of compliers who would always accept the randomization assignment.

The majority of the broad literature on treatment noncompliance focuses on continuous and discrete outcomes (Goetghebeur and Molenberghs, 1996; Cuzick et al., 1997; Little and Yau, 1998), and missing data (Frangakis and Rubin, 1999; Baker, 2000; Mealli et al., 2004; Peng et al., 2004; O’Malley and Normand, 2004). Relatively few methods pertain to latent subgroup analysis of right-censored survival endpoints, owing in part to the difficulty in nonparametric estimation of an unobserved distribution (Imbens and Rubin, 1997b; Abadie, 2002). Frangakis and Rubin (1999) and Elashoff et al. (2012) examined nonparametric estimation of the survival distribution of compliers in the control arm under different censorship assumptions, but their estimates are not monotone decreasing nor do the methods accommodate covariates. Several semiparametric models have been proposed, all of which require some form of proportional hazards assumption (Follmann, 2000; Cuzick et al., 2007; Loeys and Goetghebeur, 2003). Although the proportional hazards approach is by far the most popular method for regression analysis of right censored survival data, it does have limitations since the proportional hazards assumption is stringent and does not always hold in practice. For the standard regression problem with right censored data, a number of alternative models such as accelerated failure time models have been studied to complement the proportional hazards approach (Klein and Moeschberger, 2003). However, it is nontrivial to extend them for latent subgroup survival analysis (Robins and Tsiatis, 1991; Altstein et al., 2011). Robins and Tsiatis (1991) studied a class of rank-preserving structural failure time models (RPSFTM) for within-subject survival differences that compare a subject’s observed survival time (T_i) with a latent survival time (U_i) if he/she were never to receive treatment. In the case of all-or-none noncompliance, the RPSFTM assumes that T_i and U_i differ by a factor and that the factor is the same for all individuals. This implies that if two subjects have identically observed failure times and observed treatment histories, then they would have had identical latent failure times if treatment had always been withheld, which might be biologically implausible. It is also not trivial for RPSFTM to incorporate covariates. Altstein et al. (2011) developed a method for parametric subgroup analysis through an accelerated failure time (AFT) mixture model. This model does not refer to a latent baseline failure time and can naturally incorporate covariates. Maximum likelihood estimates of biological efficacy and other covariate effects were achieved through the EM algorithm (Dempster et al., 1997). Flexibility was introduced by allowing the residual error distributions and covariate effects to differ in the subgroups. The method offers easy

implementation in standard statistical packages such as SAS or R and a straightforward framework for maximum likelihood inference, but these advantages come at the expense of parametric assumptions. In practice, it may not be easy to identify a proper parametric model. Moreover, we illustrate in Sections 4 and 5 that the parametric approach can lead to substantial bias if the parametric model is mis-specified.

The main purpose of this paper is to introduce a semiparametric version of the AFT mixture model for latent subgroup analysis with a completely unknown error distribution in each subgroup. Although the semiparametric AFT model has been extensively studied in the literature (Buckley and James, 1979; Louis, 1981; Wei, 1992; Tsiatis, 1990; Jin et al., 2006; Zeng and Lin, 2007), its adoption to latent subgroup analysis is nontrivial due to unobserved latent subgroup membership in the control arm. As discussed in the beginning of Section 3, direct maximization of the likelihood for the semiparametric AFT model does not yield consistent regression parameter estimates. We develop an iterative algorithm to estimate the model parameters including the unknown error distribution in each subgroup. The iterative algorithm involves an expectation step to estimate the conditional probability of subgroup membership given the observed data for each subject in the control arm, which requires nonparametric hazard estimation. The estimated probabilities are then used in a weighted Buckley and James (1979) step to update all parameter estimates. In keeping with our parametric method, the Buckley-James approach is intuitive – it is often referred to as a semiparametric EM algorithm – and the computation is relatively facile. The biological efficacy in our model is quantified by a parametric shift in the unknown survival distribution of the subgroup of interest. Compared to the parametric approach of Altstein et al. (2011), the semiparametric AFT mixture model allows us to avoid the difficulties in specifying a parametric form for the error distributions and thus is more flexible and robust. The model incorporates covariates that can interact with subgroup status. It also provides a framework to examine the distributional assumption of a parametric AFT mixture model.

The paper is organized as follows. Section 2 describes the model and assumptions. Section 3 studies estimation and inference. Section 4 presents simulation studies to evaluate the performance of our method. Section 5 illustrates our method using data from a melanoma trial. Further discussions are given in Section 6.

2. The Semiparametric AFT Mixture Model

2.1 Notation

We will denote the observed data on the possibly right-censored outcome as (Y_i, δ_i) , $i = 1, \dots, n$, where Y_i is the minimum of the log-transformed failure time T_i and a censoring time C_i , $\delta_i = I(T_i \leq C_i)$. These variables may depend on an observed time-fixed covariate vector $\mathbf{Z}_i = (Z_{1i}, \dots, Z_{qi})'$ such that $T_i \propto C[\mathbf{Z}_i, R]$, where R denotes the randomization assignment. We will assume that only two subgroups exist in the population, $G_i = g$, $g = 0, 1$. The observed data, denoted by $\mathbf{X}_{\text{obs}}$, consist of $(y_i, \delta_i, \mathbf{z}_i, g_i, i : R_i = 1)$ in the treatment arm and $(y_i, \delta_i, \mathbf{z}_i, i : R_i = 0)$ in control.

2.2 Model and Assumptions

Subgroup membership in this framework is an inherent characteristic of the subject, biological or otherwise, that is unaffected by the randomization assignment. This advocates a mixture model conditional on subgroup:

$$T_i | (G_i = g) = \psi_g R_i + \mathbf{Z}_i' \beta_g + \varepsilon_{gi}, \quad (1)$$

where T_i is the log-transformed failure time, $\epsilon_g \sim S_g$, $g = 0, 1$, and S_g denotes the survivor function $S_g(e) = P(\epsilon_g > e)$. It is a mixture model because the subgroup indicator G_i is not observable in the control arm. The biological efficacy of treatment is defined in the *treatable* subgroup of patients ($g = 1$) but not in the untreatable subgroup, usually because the latter does not stand to benefit from the treatment, which is the case with node-negative patients, or because treatment receipt in this subgroup is not determined by the randomization assignment, which is the case in a latent class noncompliance setting. In terms of model parameters, this translates to $\psi_0 \equiv 0$. This assumption is known as the *exclusion restriction* in the noncompliance literature (Angrist et al., 1996).

The semiparametric AFT mixture model (1) allows the survival distributions of the error terms, S_1 and S_0 , to remain unspecified. We will focus on the most general case, where S_1 and S_0 are both unspecified with unknown means $\alpha_g = E(\epsilon_g)$ and variances $\sigma_g^2 = \text{Var}(\epsilon_g)$ ($g = 0, 1$). We assume that $\mathbf{Z}_i$ does not include the standard intercept term to ensure the identifiability of α_g . This model requires the relatively weak assumptions of log-linearity in the covariates, and homoscedasticity of error variances within subgroup. The estimation method requires smoothness and differentiability of S_1 and S_0 . The parametric portion of this model equates biological efficacy with a location shift in the survival distribution of the treatable subgroup.

3. Estimation and Inference

The likelihood for parameter vector $\Theta = (p, \psi_1, \beta_1, \beta_0, S_1, S_0)$ based on the observed data is

$$L(\Theta|\mathbf{X}_{\text{obs}}) = \prod_{i:R_i=1} \{p dF_1(y_i - \psi_1 - \mathbf{z}'_i \beta_1)^{\delta_i} S_1(y_i - \psi_1 - \mathbf{z}'_i \beta_1)^{1-\delta_i} \}^{g_i} \{(1-p) dF_0(y_i - \mathbf{z}'_i \beta_0)^{\delta_i} S_0(y_i - \mathbf{z}'_i \beta_0)^{1-\delta_i} \}^{1-g_i}$$

$$\prod_{i:R_i=0} \{p dF_1(y_i - \mathbf{z}'_i \beta_1) + (1-p) dF_0(y_i - \mathbf{z}'_i \beta_0)\}^{\delta_i} \{p S_1(y_i - \mathbf{z}'_i \beta_1) + (1-p) S_0(y_i - \mathbf{z}'_i \beta_0)\}^{1-\delta_i}, \quad (2)$$

where $F_g(e) = 1 - S_g(e)$ and $dF(x) = F(x) - F(x-)$ for any distribution function F . This likelihood is difficult to work with because the unobserved subgroup indicators in control preclude its factorization into separate components for the mixing proportion and each subgroup. The “complete” data scenario of known subgroup status on all subjects yields a much simpler likelihood:

$$L(\Theta|\mathbf{X}_{\text{com}}) = \prod_{i=1}^n p^{g_i} (1-p)^{1-g_i} \{dF_1(y_i - \psi_1 r_i - \mathbf{z}'_i \beta_1)^{\delta_i} S_1(y_i - \psi_1 r_i - \mathbf{z}'_i \beta_1)^{1-\delta_i}\}^{g_i} \{dF_0(y_i - \mathbf{z}'_i \beta_0)^{\delta_i} S_0(y_i - \mathbf{z}'_i \beta_0)^{1-\delta_i}\}^{1-g_i} \quad (3)$$

$$= L(p|\mathbf{X}_{\text{com}}) L(\psi_1, \beta_1, F_1|\mathbf{X}_{\text{com}}) L(\beta_0, F_0|\mathbf{X}_{\text{com}}), \quad (4)$$

where $\mathbf{X}_{\text{com}} = (\mathbf{X}_{\text{obs}}, g_i; i: R_i = 0)$. In the fully parametric case, the EM algorithm was a natural choice because it exploited the tractability of (3). However, nontrivial modifications are required for the current model since the error distributions F_0 and F_1 are completely unspecified. In particular, direct maximization of the complete data likelihood does not lead to plausible estimates of the regression parameters since the profile likelihood after profiling out the nonparametric error distributions in (3) is a step function of the regression parameters. Below we develop an estimation method that can be regarded as an extension of the EM algorithm that iterates between an expectation (E-) step and a Buckley and James (1979) type (BJ-) optimization step. The BJ-optimization step is an extension of the

weighted least squares method for right censored data, which involves a further iteration algorithm that updates the regression parameter estimates and the error distribution estimates alternatively. The derivation is provided in Web Appendix A.

3.1 E-BJ Algorithm

3.1.1 E-Step—The log of the complete data likelihood (3) is linear in the unobserved subgroup indicators in the control arm, so the E-step amounts to taking the expectation of $(G_i; i: R_i = 0)$ conditional on the observed data, $\mathbf{X}_{\text{obs}}$, and the current estimate of the parameter vector $\Theta^{(m)} = (p^{(m)}, \psi_1^{(m)}, \beta_1^{(m)}, \beta_0^{(m)}, S_1^{(m)}, S_0^{(m)})$. It yields the following imputed variable for the subgroup indicator g_i in the control arm ($R_i = 0$):

$$g_i^{(m)} = \begin{cases} \frac{p^{(m)}h_1^{(m)}(e_{1i}^{(m)})S_1^{(m)}(e_{1i}^{(m)})}{p^{(m)}h_1^{(m)}(e_{1i}^{(m)})S_1^{(m)}(e_{1i}^{(m)}) + (1-p^{(m)})h_0^{(m)}(e_{0i}^{(m)})S_0^{(m)}(e_{0i}^{(m)})} & \text{if } \delta_i = 1, \\ \frac{p^{(m)}S_1^{(m)}(e_{1i}^{(m)})}{p^{(m)}S_1^{(m)}(e_{1i}^{(m)}) + (1-p^{(m)})S_0^{(m)}(e_{0i}^{(m)})} & \text{if } \delta_i = 0, \end{cases} \quad (5)$$

where $e_{1i}^{(m)} = y_i - \psi_1^{(m)}r_i - \mathbf{z}_i'\beta_1^{(m)}$, $e_{0i}^{(m)} = y_i - \mathbf{z}_i'\beta_0^{(m)}$, $i = 1, \dots, n$ are the current estimates of the residuals, and

$$h_g^{(m)}(e) = \int \frac{1}{b} K\left(\frac{e-u}{b}\right) d\Lambda_g^{(m)}(u), \quad (6)$$

is a kernel estimate of the hazard function of S_g . In (6), b is the smoothing bandwidth, K is a nonnegative kernel function satisfying certain properties, and

$$\Lambda_g^{(m)}(e) = \int_0^e \frac{dS_g^{(m)}(u)}{S_g^{(m)}(u-)} \quad (7)$$

is the cumulative hazard function of $S_g^{(m)}(e)$. We discuss the specification of the smoothing procedure and sensitivity of the estimation method to it in Section 3.3.

In the treatment arm ($R_i = 1$), subgroup status is fully observable so $g_i^{(m)} = g_i$.

3.1.2 BJ-Step—According to the complete data factorization of (4), there are three components to the estimation of $\Theta^{(m+1)}$ conditional on $g_i^{(m)}$. The new estimate of the treatable proportion is simply given by maximum likelihood:

$$p^{(m+1)} = \underset{p}{\operatorname{argmax}} \left\{ \sum_{i=1}^n g_i^{(m)} \log(p) + (1 - g_i^{(m)}) \log(1 - p) \right\} = \frac{1}{n} \sum_{i=1}^n g_i^{(m)}.$$

Each of the remaining portions of the likelihood corresponds to one of the subgroups and requires two steps to update: the first estimates $S_g^{(m+1)}$ conditional on the finite dimensional parameters, and the second produces the new estimates $(\psi_1^{(m+1)}, \beta_1^{(m+1)}, \beta_0^{(m+1)})$ conditional on $S_g^{(m+1)}$, $g = 0, 1$. We illustrate this procedure on the treatable subgroup ($g = 1$).

Let $e_{11} < e_{12} < \dots < e_{1D}$ be the distinct uncensored values of the current estimates of the residuals $e_{1i}^{(m)} = y_i - \psi_1^{(m)}r_i - \mathbf{z}_i'\beta_1^{(m)}$, $i = 1, \dots, n$. Since the residuals are on the log scale, $e_{10} = -\infty$ and $e_{1,D+1} = \infty$. For simplicity, we omit superscripts because this order does not depend on $\Theta^{(m)}$. We define the following weighted counts of failures at e_{1j} , and individuals at risk just prior to e_{1j} , $j = 1, \dots, D$:

$$d_{1j}^{(m)} = \sum_{i=1}^n g_i^{(m)} I(e_{1i}^{(m)} = e_{1j}, \delta_i = 1),$$

$$n_{1j}^{(m)} = \sum_{i=1}^n g_i^{(m)} I(e_{1i}^{(m)} \geq e_{1j}).$$

The new estimate of S_1 is a weighted Kaplan-Meier estimate

$$S_1^{(m+1)}(e) = \begin{cases} 1 & \text{if } e < e_{11} \\ \prod_{e_{1j} < e} \left(1 - \frac{d_{1j}^{(m)}}{n_{1j}^{(m)}}\right) & \text{if } e \geq e_{11}. \end{cases}$$

on the triplets $(e_{1i}^{(m)}, \delta_i, g_i^{(m)})$, $i = 1, \dots, n$.

The updates $(\psi_1^{(m+1)}, \beta_1^{(m+1)})$ are the solutions to the weighted least squares normal equations:

$$\begin{aligned} \sum_{i=1}^n g_i^{(m)} r_i (y_{1i}^{(m)} - \psi_1 r_i - \mathbf{z}_i \beta_1) &= 0 \\ \sum_{i=1}^n g_i^{(m)} z_{1i} (y_{1i}^{(m)} - \psi_1 r_i - \mathbf{z}_i \beta_1) &= 0 \\ &\vdots \\ \sum_{i=1}^n g_i^{(m)} z_{qi} (y_{1i}^{(m)} - \psi_1 r_i - \mathbf{z}_i \beta_1) &= 0, \end{aligned}$$

where

$$y_{1i}^{(m)} = \begin{cases} y_i & \text{if } \delta_i = 1 \\ \psi_1 r_i + \mathbf{z}_i' \beta_1 + \frac{\sum_{e_{1k}^{(m)} > e_{1i}^{(m)}} w_{1k}^{(m+1)} e_{1k}^{(m)}}{S_1^{(m+1)}(e_{1i}^{(m)})} & \text{if } \delta_i = 0, \end{cases}$$

$i = 1, \dots, n$, and the weights $w_{1k}^{(m+1)}$ are the steps of $S_1^{(m+1)}$.

The derivation of $(\beta_0^{(m+1)}, S_0^{(m+1)})$ is analogous and uses weights $1 - g_i^{(m)}$.

The E- and BJ-steps are iterated until convergence to yield the final estimates $(\hat{\rho}, \hat{\psi}_1, \hat{\beta}_1, \hat{\beta}_0, \hat{S}_1, \hat{S}_0)$.

3.2 Variance Estimation

The asymptotic distribution and variance estimation of the original Buckley-James estimator have been studied extensively in the literature. Ritov (1990), Tsiatis (1990) and Lai and Ying (1991) proved the asymptotic normality of the estimator, but their variance estimators involve nonparametric estimation of the density function and its derivative, which are not appealing in practice. The variance estimator originally proposed by Buckley and James

(1979) may be inconsistent (Lai and Ying, 1991). Several authors have suggested resampling methods, including Ritov (1990), Stare et al. (2001) (through implementation of the `bj()` function in R) and Jin et al. (2006). Derivation of the variance matrix is even more complicated in the presence of subgroups; hence, we use the bootstrap to approximate the standard errors of $(\hat{\psi}_1, \hat{\beta}_1, \hat{\beta}_0)$ and have corroborated its performance in simulation.

3.3 Sensitivity to Hazard Smoother

Our estimation procedure requires specifications of a kernel function and a bandwidth in a hazard smoother defined by (6). In practice, one may use an Epanechnikov kernel defined as

$K(x) = 0.75(1 - x^2), |x| \leq 1$, with the global bandwidth of $b_{g0} = c \cdot \max\{e_{gi}\} / (8n_{gu}^{\frac{1}{5}})$, where $\max\{e_{gi}\}$ is the largest residual in the g^{th} subgroup, n_{gu} is the weighted count of uncensored observations in the g^{th} subgroup, $g = 0, 1, i = 1, \dots, n$, and c is a scaling factor. The bandwidth is updated at each step of the estimation algorithm. Mueller and Wang (1990)

suggested $b_0 = \max\{e_i\} / (8n_u^{\frac{1}{5}})$ as a starting point for their algorithm to optimize the global bandwidth based on minimization of the MSE. In a comprehensive simulation study of their methods, Hess et al. (1999) found that this initial estimate does not much differ from the optimal value. Our extension of b_0 to b_{g0} performed well in our limited empirical studies, and we specified $c = 0.95$ except when the weighted risk set in the g^{th} subgroup shrank below ten, at which point we set $c = 2$. We found empirically that boundary corrections and adaptive bandwidths did not improve the estimation sufficiently to warrant the gross increase in computation time necessary for their implementation. Furthermore, the extension of these methods to the subgroup problem is not straightforward; in particular we found the standard methods of specifying the boundary correction to be unsatisfactory in estimation of subgroup-specific hazard functions. The performance of our smoother, which we will refer to as the “reference”, did not prove to be particularly sensitive to modest modifications of c , including modifications to the tail correction, as illustrated in the end of Section 5.

3.4 Model Assessment

A simple method to assess lack of fit of the semiparametric AFT mixture model is to compare the estimated quantities from the full model using all data with the respective estimates based on each observed subgroup in the treatment arm. For example, in the two-sample case with no covariates, one can evaluate the agreement between the observed and estimated survival curves for the subgroups. The observed survival curves are the Kaplan-Meier survival curves of the treatable subgroup ($g = 1$) and the untreatable subgroup ($g = 0$) formed from data on individuals randomized into the treatment arm. The estimated survival curves from the semiparametric model are the weighted Kaplan-Meier estimates based on the final set of estimated residuals and parameters. A divergence between the observed and estimated survival curves would suggest lack of fit of the model.

Furthermore, the semiparametric AFT mixture model can be used to evaluate the distributional assumption of the parametric AFT mixture models of Altstein et al. (2011) by examining the agreement between the estimated error survival curves from the semiparametric model and their counterparts from the parametric model.

4. Simulation Studies

4.1 Parametric versus Semiparametric

We first present a simulation study to compare the performance of the semiparametric AFT mixture model with the fully parametric Weibull AFT mixture model (Altstein, Li, and Elashoff (2011)). Table 1 reports the bias and mean squared error of the efficacy parameter estimates from the two methods under six model configurations with sample size $n = 500$,

treatable mixing proportion $p = 0.8$, censoring rate $CR = 50\%$, and true value of the biological efficacy $\psi_1 = 0.6$. The six scenarios include symmetric error distributions (normal, logistic), extreme-value error distributions (skewed to the left), Gumbel and Gamma error distributions (skewed to the right) and bimodal error distributions (normal mixture). The Weibull AFT mixture model is correct when the error distributions are extreme-value distributions, and mis-specified in the other five scenarios. Each entry is based on 3,000 Monte Carlo samples.

All simulations use the same method of hazard smoothing: the nonnegative and symmetric Epanechnikov kernel, defined as $K(x) = 0.75(1-x^2)$, $|x| \leq 1$; satisfying $\int_{-\infty}^{\infty} K(x)dx = 1$. The bandwidth is proportional to $n_{gu}^{-\frac{1}{5}}$, where n_{gu} is the weighted count of uncensored observations in the g^{th} subgroup, $g = 0, 1$, and it is augmented when the risk set in the respective subgroup is less than ten. This choice of hazard smoother was discussed in Section 3.3.

It is seen from Table 1 that when the fully parametric Weibull AFT mixture model is correctly specified with extreme-value errors, both methods have small bias. In this case, the parametric method is more efficient with a smaller MSE as expected. In the other five scenarios for which the Weibull model is mis-specified, the semiparametric method still has small bias whereas the parametric method is seriously biased, which leads to a much larger MSE for the parametric method. Therefore the parametric method is not robust against mis-specification of the error distributions.

4.2 Estimation of Biological Efficacy Under Various Sample Conditions

There are many potential factors influencing the performance of the method for the semiparametric AFT mixture model, including treatable proportion (p), censoring rate (CR), magnitudes of biological efficacy and other covariate effects, residual error distributions F_1 and F_0 , and hazard smoother. Below we present two simulation studies in response to these issues. The first looks at the behavior of the estimator under various characteristics in models with no other covariates and the second investigates its performance when covariates are added to the model; all are based on 400 repetitions with 200 bootstrap samples each unless otherwise noted. Sensitivity to the hazard smoother are discussed in Section 5.

The first simulation study evaluates the bias, bootstrap standard error estimate, and coverage of the semiparametric method under various combinations of treatable proportion, censoring rate, biological effect of treatment, and error distribution. The error distributions were specified as normal and extreme-value distributions, leading to log-normal and Weibull survival models. These distributions are specified by their mean and variance in the case of the normal and by their location and scale otherwise.

Table 2 contains partial results from this simulation study with sample size $n = 500$. Each line of the table corresponds to one specification and reports the empirical mean of the estimated biological efficacy ($\hat{\psi}_1$), the empirical standard error of the efficacy ($\widehat{SE}(\hat{\psi}_1)$), which is defined as the standard deviation of the estimates in simulation, the average bootstrapped standard error estimate ($\widehat{SE}_b(\hat{\psi}_1)$), and the proportion of 95% normal confidence intervals that cover the true efficacy (CP). Except for the last row, the bootstrap seems to perform well under all scenarios and the bias tends to grow larger when the treatable proportion is small and the censoring rate is high. The last row represents a scenario with a small treatment proportion and a high censoring rate for which a larger sample size would be needed for our method to work well.

Results from the second simulation for covariate models are reported in Web Table 1 of Web Appendix B. It is observed that the estimation algorithm and bootstrap procedure seem to perform well in covariate models even when the censoring rate is high and the treatable proportion is low.

5. An Example

Our interest in this problem issues from the Multicenter Selective Lymphadenectomy Trial-I (MSLT-I), a clinical trial to evaluate the effectiveness of sentinel node biopsy (SNB) in guiding treatment of melanoma (Morton, Thompson, Essner, Elashoff, Stern, Nieweg, Roses, Karakousis, Mozzillo, Reingten, Wang, Glass, Cochran, and the Multicenter Selective Lymphadenectomy Trial Group (1999)). At the time the trial began in 1994, the standard-of-care consisted of wide excision of the primary melanoma followed by observation with lymphadenectomy only upon clinical evidence of recurrence. However, patients with nodal metastases have a poor prognosis and may stand to benefit from immediate lymphadenectomy. MSLT-I randomized patients to receive an experimental course of therapy (treatment arm) or the standard-of-care (control arm). All patients in the treatment arm received SNB, and those with biopsies positive for metastasis underwent an immediate lymphadenectomy. Node-negative patients received the same treatment as controls.

The intention-to-treat analysis of MSLT-I determines whether the strategy of SNB plus immediate lymphadenectomy improves survival in the diseased population in aggregate. It does not estimate the biological efficacy of the surgery on the subgroup of patients with nodal metastases. The third interim analysis of the trial Morton et al. (2006) attempted to address this estimand by comparing node-positive patients in treatment to control patients who subsequently developed recurrence, but this does not produce an unbiased estimate of biological efficacy because there is not a one-to-one correspondence with positive SNB at the time of randomization and the eventual development of a clinically evident recurrence.

Previously, we applied a fully parametric Weibull mixture model to reanalyze two endpoints from the 2006 analysis: overall disease free survival (DFS) and distant disease free survival (DDFS). The latter is defined as time until the development of recurrence in a distant node or death without distant recurrence and is considered more relevant than overall DFS. The objective was to see whether the trial provides sufficient evidence that immediate lymphadenectomy improves DFS and/or DDFS relative to the control treatment in patients with sentinel-node metastases. We revisit this analysis by fitting an analogous set of semiparametric AFT mixture models. As an illustration, we present a two-sample analysis to estimate biological efficacy in a model with no covariates. The parameter estimates, 95% normal confidence intervals and p-values for testing a treatment difference based on bootstrapped standard errors are presented in Tables 3. The bootstrapped standard errors were calculated using 200 bootstrap samples. It is seen that both the fully parametric and semiparametric two-sample tests conclude that immediate lymphadenectomy significantly prolongs both overall and distant disease free survival with the fully parametric tests having smaller p-values. However, the fully parametric analysis yields substantially larger estimates for the efficacy parameter which is measured on the log-scale. The inflated treatment efficacy by the fully parametric analysis could be due in part to possible misspecification of the parametric model as discussed below.

Figure 1 plots the observed and predicted survival curves from the semiparametric AFT mixture model for the two-sample analysis with respect to DFS (left panel) and DDFS (right panel) using the method described in Section 3.4. They are nearly indistinguishable, indicating no lack of fit of the semiparametric AFT mixture model. We have also

superimposed the estimated survival functions resulting from the parametric model (grey curves), which reveals a poor fit of the parametric model, especially in the right tail of the node positive subgroup in the control arm. The poor fit of the parametric model can lead to a biased estimate of the efficacy parameter and explains why the parametric model produces inflated efficacy estimates than the semiparametric analysis in Table 3.

Because the semiparametric AFT mixture models seem to fit the data well, we argue that the resulting estimated subgroup survival functions can be used to check if the hazard of node+ in treatment is proportional to that in control. Figure 2 depicts the plots of $\log\{-\log(\text{Survival of node + in treatment})\} - \log\{-\log(\text{Survival of node + in control})\}$ using estimates from the semiparametric AFT mixture models for DFS and DDFS. The plot is expected to be a constant over time under the proportional hazards assumption. This is clearly not the case for both DFS and DDFS, which indicates that the hazards between the two subgroups (node + in treatment versus node+ in control) are unlikely to be proportional for DFS and DDFS.

We also performed similar analysis adjusting for possible confounding factors including Breslow's score and the presence of ulceration. The general conclusions are similar and thus are not reported here.

Finally, we examined the role of the hazard smoother (6) in our estimation procedure by running sensitivity analyses on the choice of kernel, bandwidth and tail correction. We also considered boundary kernels, which are intended to mitigate the extremes of the data when relatively few events occur and the standard bandwidth window extends beyond the range of the data, and locally optimal bandwidth methods, which optimize the bandwidth at each target estimation time according to the local characteristics of the data. These methods were proposed by Gasser and Mueller (1979), Mueller and Wang (1990) and studied thoroughly by Hess et al. (1999). The performance of our reference smoother did not prove to be particularly sensitive to modest modifications of c , including modifications to the tail correction. We illustrate this in Figure 3 by plotting four versions of the final estimate of the smoothed hazard for disease-free survival in the treatable subgroup of MSLT-I. Each line corresponds to a different smoother: the reference, the reference without augmentation of c in the tail, one based on a larger bandwidth ($c = 1.2$ and 2.5 in the tail), and one based on a smaller bandwidth ($c = .85$ and 1.5 in the tail). The smoothed curves are nearly identical until the right extreme of the data, at which point the curve corresponding to the reference without the tail correction becomes more sensitive to fluctuations in the hazard. The deviations in the efficacy parameter estimate $\hat{\psi}_1$, its estimated standard error (in the legend of the figure), and the shape of smoothed curves are consistent with our expectations – larger bandwidths lead to smoother estimates with less variability – and are unlikely to have practical impact.

6. Discussion

Unlike Altstein et al. (2011), which benefitted from a fully parametric likelihood, development of the current semiparametric model and Buckley-James technique for estimation is much more complex. In order to consistently estimate the conditional probability that a subject is treatable in the E-step of our algorithm, we smoothed the estimated hazard function in each subgroup. This adds a layer of complexity to the estimation algorithm and complicates theoretical investigation of the properties of the resulting estimates of biological efficacy and other covariate effects. Further research to develop large sample properties of our estimator is warranted.

The semiparametric AFT mixture model provides a useful alternative to the proportional hazards approach (Cuzick et al., 2007; Loeys and Goetghebeur, 2003) for latent subgroup

analysis with missing membership. We point out that the Buckley-James (1979) approach requires homoscedastic errors in the semiparametric AFT model. It can lead to biased and inconsistent parameter estimates in the presence of heteroscedastic errors. The AFT model and the proportional hazards model make different assumptions and should be regarded as complementary, as opposed to competing, models (Louis, 1981; Wei, 1992). A further question is how to determine which model is more appropriate in practice. Section 3.4 provides a simple graphical method to assess lack of fit of the semiparametric AFT mixture model. However, to the best of our knowledge, no general method has been discussed in the literature on how to check the proportional hazards assumption within the context of latent subgroup analysis with missing membership, even though there is a rich literature on diagnostic methods for the simple Cox regression setting with no missing membership. Further development of diagnostic tools for latent subgroup models is desired.

In the simple case with no missing membership, there are a number of other models such as additive risks models for regression analysis with right censored data. It would be useful to examine these models and make them available for the latent subgroup analysis problem. For example, one could try to relax the proportional hazards assumption by including time-dependent covariates. Such a model could also provide a framework to evaluate the proportional hazards assumption. Another possible approach is to consider the AFT model with smoothed error distributions (Komarek et al., 2005) which may allow direct estimation of regression parameters for latent subgroup analysis using the maximum penalized likelihood approach. Finally, our method can be extended to more complex situations with multiple subgroups as in Angrist et al. (1996) and Cuzick et al. (2007).

R code for the E-BJ algorithm is available from the authors upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Thomas A. Louis, the editor, an Associate Editor, and a referee for their insightful and constructive comments that led to significant improvements of the paper. We also thank Drs. R. Elashoff and D.L. Morton for providing the MLST-I data and X.Y. Lu for his programming assistance. Gang Li's research was supported in part by NIH grant CA016042 and NIH grant P01AT003960.

References

- Abadie A. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*. 2002; 97(457):284–292.
- Altstein L, Li G, Elashoff R. A method to estimate treatment efficacy among latent subgroups of a randomized clinical trial. *Statistics in Medicine*. 2011; 30(7):709–717. [PubMed: 21394747]
- Angrist J, Imbens G, Rubin D. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*. 1996; 91(434):444–455.
- Baker S. Analyzing a randomized cancer prevention trial with a missing binary outcome, an auxiliary variable, and ali-or-none compliance. *Journal of the American Statistical Association*. 2000; 95(449):43–50.
- Buckley J, James I. Linear regression with censored data. *Biometrika*. 1979; 66(3):429–436.
- Cuzick J, Edwards R, Segnan N. Adjusting for non-compliance and contamination in randomized clinical trials. *Statistics in Medicine*. 1997; 16(9):1017–1029. [PubMed: 9160496]
- Cuzick J, Sasieni P, Myles J, Tyrer J. Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2007; 69(4):565–588.

- Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1997; 39(1):1–38.
- Elashoff R, Li G, Zhou Y. Nonparametric inference for assessing treatment efficacy in randomized clinical trials with a time-to-event outcome and all-or-none compliance. *Biometrika*. 2012; 99(2): 393–404.
- Follmann D. On the effect of treatment among would-be treatment compliers: an analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association*. 2000; 95(452):1101–1109.
- Frangakis C, Rubin D. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing out comes. *Biometrika*. 1999; 86(2):365–379.
- Gasser, T.; Mueller, H. Kernel estimation of regression functions. In: Gasser; Rosenblatt, editors. *Smoothing techniques for curve estimation*. Vol. 43. Heidelberg: Springer Verlag; 1979. p. 23–67.
- Gipponi M, Solari N, Di Somma F, Bertoglio S, Cafiero F. New fields of application of the sentinel lymph node biopsy in the pathologic staging of solid neoplasms: review of literature and surgical perspectives. *Journal of surgical oncology*. 2004; 85(3):171–179. [PubMed: 14991890]
- Goetghebeur E, Molenberghs G. Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association*. 1996; 91(435): 928–934.
- Hess K, Serachitopol D, Brown B. Hazard function estimators: a simulation study. *Statistics in medicine*. 1999; 18(22):3075–3088. [PubMed: 10544307]
- Imbens G, Rubin D. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*. 1997; 25(1):305–327.
- Imbens G, Rubin D. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*. 1997b; 64(4):555–574.
- Jin Z, Lin D, Ying Z. On least-squares regression with censored data. *Biometrika*. 2006; 93(1):147–161.
- Klein, J.; Moeschberger, M. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed.. Springer-Verlag; 2003.
- Komarek A, Lesaffre E, Hilton J. Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*. 2005; 14(3):726–745.
- Lagakos S. The challenge of subgroup analyses-reporting without distorting. *The New England Journal of Medicine*. 2006; 354(16) 1667–xxxx.
- Lai T, Ying Z. Large sample theory of a modified buckley-james estimator for regression analysis with censored data. *The Annals of Statistics*. 1991; 19:1370–1402.
- Little R, Yau L. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using rubin's causal model. *Psychological Methods*. 1998; 3:147–159.
- Loeys T, Goetghebeur E. A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics*. 2003; 59(1):100–105. [PubMed: 12762446]
- Louis T. Nonparametric analysis of an accelerated failure time model. *Biometrika*. 1981; 68(2):381–390.
- Mealli F, Imbens G, Ferro S, Biggeri A. Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics*. 2004; 5(2):207–222. [PubMed: 15054026]
- Morton D, Thompson J, Cochran A, Mozzillo N, Elashoff R, Essner R, Nieweg O, Roses D, Hoekstra H, Karakousis C, Reintgen D, Coventry B, Glass E, Wang H. the Multi center Selective Lymphadenectomy Trial Group. Sentinel-node biopsy or nodal observation in melanoma. *New England Journal of Medicine*. 2006; 355(13):1307–1317. [PubMed: 17005948]
- Morton D, Thompson J, Essner R, Elashoff R, Stern S, Nieweg O, Roses D, Karakousis C, Mozzillo N, Reintgen D, Wang H, Glass E, Cochran A. the Multicenter Selective Lymphadenectomy Trial Group. Validation of the accuracy of intraoperative lymphatic mapping and sentinel lymphadenectomy for early-stage melanoma: a multicenter trial. *Annals of Surgery*. 1999; 230(4): 453–463. [PubMed: 10522715]

- Mueller H, Wang JL. Locally adaptive hazard smoothing. *Probability Theory and Related Fields*. 1990; 85(4):523–538.
- O'Malley A, Normand S. Likelihood methods for treatment noncompliance and subsequent nonresponse in randomized trials. *Biometrics*. 2004; 61(2):325–334. [PubMed: 16011678]
- Peng Y, Little RA, Raghunathan T. An extended general location model for causal inferences from data subject to noncompliance and missing values. *Biometrics*. 2004; 60(3):598–607. [PubMed: 15339281]
- Ritov Y. Estimation in a linear regression model with censored data. *The Annals of Statistics*. 1990; 18:303–328.
- Robins J, Tsiatis A. Correcting for noncompliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics A*. 1991; 20:2609–2631.
- Sommer A, Zeger S. On estimating efficacy from clinical trials. *Statistics in Medicine*. 1991; 10(1): 45–52. [PubMed: 2006355]
- Stare J, Harrell F, Heinzl H. an s-plus program to fit linear regression models to censored data using the buckley-james method. *Computer Methods and Programs :in Biomedicine*. 2001; 64(1):45–52. [PubMed: 11084232]
- Tsiatis A. Estimating regression parameters using linear rank tests for censored data. *Annals of Statistics*. 1990; 18:354–372.
- Wang R, Lagakos S, Ware J, Hunter D, Drazen J. Statistics in medicine-reporting of subgroup analyses in clinical trials. *The New England Journal of Medicine*. 2007; 357(21):2189–2194. [PubMed: 18032770]
- Wei L. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*. 1992; 11:1871–1879. [PubMed: 1480879]
- Zeng D, Lin D. Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*. 2007; 102(480):1387–1396.

Figure 1.
Estimated and Observed Survival Curves in the Two-Sample Analysis of MSLT-I

Figure 2. Plot of $\log(-\log(\text{Survival of node + in treatment})) - \log(-\log(\text{Survival of node + in control}))$ for DFS (Left Panel) and DDFS (Right Panel) in MSLT-I

Figure 3.
Sensitivity of the Hazard Smoother on DFS in Node-Positive Subgroup

Simulated bias and mean squared error (MSE) of the efficacy parameter estimates $\hat{\psi}_1$ and $\hat{\psi}_{1p}$ from the semiparametric AFT mixture model and a parametric Weibull AFT mixture model under six scenarios with extreme-value (EV), normal (N), logistic (Log), Gumbel (GB), Gamma (GM), and normal mixture error distributions (F_0 and F_1). Here $n = 500$, treatable mixing proportion $p = 0.8$, censoring rate $= 50\%$, and biological efficacy $\psi_1 = 0.6$. The Weibull model is correct under the first scenario and mis-specified under other scenarios. Each entry is based on 3,000 Monte Carlo samples

Table 1

F_1	F_0	$\text{Bias}(\hat{\psi}_1) \times 10^3$	$\text{Bias}(\hat{\psi}_{1p}) \times 10^3$	$\text{MSE}(\hat{\psi}_1) \times 10^3$	$\text{MSE}(\hat{\psi}_{1p}) \times 10^3$	$\frac{\text{MSE}(\hat{\psi}_{1p})}{\text{MSE}(\hat{\psi}_1)}$
EV(2, .4)	EV(2, .3)	-1.6	-1.7	6.8	4.4	0.65
N(1.5, .16)	N(2, .16)	1.4	-28.7	3.0	4.3	1.45
log(2.5, .4)	log(2, .3)	-4.6	70.8	10.5	14.5	1.39
GBL(2, .4)	GBL(2, .3)	2.0	78.5	3.3	22.2	6.67
GM(8, .25)	GM(10, .25)	1.6	-96.3	6.4	19.6	3.03
$\frac{1}{2}N(1, .1) + \frac{1}{2}N(2, .1)$	$\frac{1}{2}N(1.5, .1) + \frac{1}{2}N(2.5, .1)$	1.3	-27.7	6.4	11.8	1.85

Table 2

Performance of the semiparametric model without covariates under various treatable mixing proportions (p), censoring rates (CR), and normal (N) and extreme-value (EV) error distributions (F_1 and F_0). Here $n = 500$ and true biological efficacy ψ_1 is 0.6. Entries for $\hat{\psi}_1$, $\widehat{SE}(\hat{\psi}_1)$, $\widehat{SE}_b(\hat{\psi}_1)$, and CP are the empirical mean of the estimated efficacy, empirical standard error of ψ_1 , empirical mean of the bootstrapped standard error estimate, and the proportion of 95% confidence intervals that cover the true efficacy, respectively, based on 400 Monte Carlo samples.

F_1	F_0	p	CR	$\hat{\psi}_1$	$\widehat{SE}(\hat{\psi}_1)$	$\widehat{SE}_b(\hat{\psi}_1)$	CP
N(1.5, .16)	N(2, .16)	0.8	20%	0.60	0.050	0.050	0.930
			50%	0.60	0.056	0.055	0.955
			80%	0.60	0.077	0.078	0.950
			0.5	0.59	0.070	0.073	0.968
	N(2, .16)	0.5	20%	0.59	0.072	0.078	0.960
			50%	0.59	0.091	0.098	0.980
			80%	0.60	0.111	0.115	0.958
			0.3	0.57	0.106	0.119	0.969
	EV(2, .3)	0.8	20%	0.57	0.144	0.143	0.938
			50%	0.59	0.053	0.056	0.967
			80%	0.60	0.056	0.058	0.950
			0.5	0.61	0.069	0.074	0.959
EV(2, .4)	EV(2, .3)	0.8	20%	0.59	0.081	0.087	0.964
			50%	0.59	0.085	0.088	0.962
			80%	0.57	0.105	0.104	0.949
			0.3	0.57	0.116	0.133	0.967
	EV(2, .3)	0.5	20%	0.59	0.115	0.130	0.962
			50%	0.59	0.157	0.237	0.950
			80%	0.50			
	EV(2, .3)	0.3	20%	0.57	0.116	0.133	0.967
			50%	0.59	0.115	0.130	0.962
			80%	0.50			

Table 3

Two-sample analysis of MSLT-I data on disease free survival (DFS) and distant disease free survival (DDFS).

a. Summary Statistics on Survival Endpoints (N=1264)			
	Treatment, Node + (N= 122)	Treatment, Node - (N= 642)	Control (N= 500)
Recurrence at any site - %	44.3	16.0	26.8
Distant recurrence - %	36.9	12.3	17.8
DFS (months) - median	37.9	59.1	54.5
DDFS (months) - median	43.8	59.8	59.9

b. Parametric Analysis (N=1264)			
	Efficacy ($\hat{\psi}_1$)	95% Interval	p-value
Disease Free Survival	1.593	(1.167, 2.019)	2.4E-13
Distant Disease Free Survival	0.937	(0.119, 1.755)	0.025

c. Semiparametric Analysis (N=1264)			
	Efficacy ($\hat{\psi}_1$)	95% Interval	p-value
Disease Free Survival	1.132	(0.592, 1.672)	4.0E-5
Distant Disease Free Survival	0.760	(0.046, 1.474)	0.037