Risk-stratified imputation in survival analysis

Richard E Kennedy, Kofi P Adragni, Hemant K Tiwari, Jenifer H Voeks, Thomas G Brott and George Howard

Clin Trials published online 1 July 2013
DOI: 10.1177/1740774513493150

CLINICAL TRIALS ARTICLE Clinical Trials 2013; 0: 1–10

Risk-stratified imputation in survival analysis

Richard E Kennedy[a], Kofi P Adragni[b], Hemant K Tiwari[a], Jenifer H Voeks[c], Thomas G Brott[d] and George Howard[a]

Background Censoring that is dependent on covariates associated with survival can arise in randomized trials due to changes in recruitment and eligibility criteria to minimize withdrawals, potentially leading to biased treatment effect estimates. Imputation approaches have been proposed to address censoring in survival analysis; while these approaches may provide unbiased estimates of treatment effects, imputation of a large number of outcomes may over- or underestimate the associated variance based on the imputation pool selected.

Purpose We propose an improved method, *risk-stratified imputation*, as an alternative to address withdrawal related to the risk of events in the context of time-to-event analyses.

Methods Our algorithm performs imputation from a pool of replacement subjects with similar values of both treatment and covariate(s) of interest, that is, from a risk-stratified sample. This stratification prior to imputation addresses the requirement of time-to-event analysis that censored observations are representative of all other observations in the risk group with similar exposure variables. We compared our risk-stratified imputation to case deletion and bootstrap imputation in a simulated dataset in which the covariate of interest (study withdrawal) was related to treatment. A motivating example from a recent clinical trial is also presented to demonstrate the utility of our method.

Results In our simulations, risk-stratified imputation gives estimates of treatment effect comparable to bootstrap and auxiliary variable imputation while avoiding inaccuracies of the latter two in estimating the associated variance. Similar results were obtained in analysis of clinical trial data.

Limitations Risk-stratified imputation has little advantage over other imputation methods when covariates of interest are not related to treatment. Risk-stratified imputation is intended for categorical covariates and may be sensitive to the width of the matching window if continuous covariates are used.

Conclusions The use of the risk-stratified imputation should facilitate the analysis of many clinical trials, in which one group has a higher withdrawal rate that is related to treatment. *Clinical Trials* 2013; **0:** 1–10. http://ctj.sagepub.com

Introduction

A central assumption of survival analysis is that censoring is independent of exposure variables associated with survival, implying that the exposure variable for individuals who are censored is

[a]Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA, [b]Department of Mathematics and Statistics, University of Maryland, Baltimore, MD, USA, [c]Department of Neurosciences, Medical University of South Carolina, Charleston, SC, USA, [d]Department of Neurology, Mayo Clinic, Jacksonville, FL, USA

Author for correspondence: George Howard, Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, 1665 University Boulevard, Birmingham, AL 35294-0022, USA.
Email: ghoward@uab.edu

10.1177/1740774513493150

representative of all other individuals in the risk group [1]. This assumption is not always met in the analysis of time-to-event randomized clinical trials, in which censoring is dependent on covariates that are also associated with survival, such as symptomatic status.

Censored data in survival analysis are comparable to missing (or coarsened) data in other types of analyses. Multiple imputation is a powerful method for dealing with missing or coarsened data that are applicable in a variety of contexts [2]. Taylor *et al*. [3] developed a modification of multiple imputation specific to survival analysis. Under their method, the follow-up time and outcome (event/non-event at that follow-up time) for a censored individual is randomly selected from other individuals in the risk set at the time of censoring, a form of hot-deck imputation [4]. Imputation proceeds from the participant with the smallest censoring time to the largest censoring time, so that an observation that is imputed as censored is not subsequently imputed again (regardless of whether the selected observation in the hot-deck process was censored at a later time), and imputed values are not included in the risk set for imputation of other censored individuals. Multiple imputation is achieved by repeatedly generating bootstrap samples from the original dataset, with each bootstrap sample undergoing the imputation procedure. The multiple datasets are then combined using standard formulae. Hsu *et al*. [5] modified this approach to include auxiliary variables in the imputation process. Their method focuses on continuous auxiliary variables, defining two risk scores based on the relationship of the auxiliary variables to the failure time and the censoring time. These risk scores are then used to define a neighborhood around the individual to be imputed, from which the imputation values are randomly selected using the method of Taylor *et al*. [3]

Multiple imputation also depends critically on the data coarsening mechanism, which is usually assumed to be coarsened completely at random (CCAR) or coarsened at random (CAR) using the terminology of Heitjan and Rubin [6] and Tsiatis [7]. Violations of these assumptions would presumably lead to biased estimates under multiple imputation, although it is often difficult to identify such violations in practice [8,9]. In this article, we describe a motivating example from the Carotid Revascularization Endarterectomy versus Stenting Trial (CREST) [10], where the withdrawal mechanism was related to measured exposure variables, corresponding to the CAR mechanism. In contrast to the approach of Taylor *et al*. [3], we distinguish between censored observations, in which the event does not occur during the study period (type 1 censoring), and withdrawals, in which the subject terminates participation prior to the end of the study period (type 2 and random censoring). We then consider two factors as potentially influencing the likelihood of censoring: (1) the factor of interest (i.e., the 'treatment') that is the focus of the randomized trial, and (2) a 'nuisance' covariate that is not the focus of the randomized trial, but which may be associated with the risk of outcome events.

We envision five potential scenarios:

1. Neither differences in outcome nor the risk of withdrawal was related to treatment or the covariate.
2. Treatment, but not the covariate, is related to outcome events, but neither the treatment nor covariate affects the likelihood of withdrawal.
3. Both treatment and the covariate are related to outcome events, but the likelihood of withdrawal is not related to either treatment or the covariate.
4. Both treatment and the covariate are related to the outcome, and withdrawal was related to treatment but not the covariate.
5. Both the treatment and the covariate are related to both the outcome and withdrawal.

The first three scenarios correspond to a CCAR mechanism, while the last two correspond to a CAR mechanism, so that sensitivity of the methods to the analytical assumptions can be evaluated. Standard analytical approaches for survival analysis in clinical trials assume one of the first three cases, specifically that censoring is not related to the treatment. The imputation methods such as those of Taylor *et al*. [3] are readily applied to the first four cases. However, the last case was not addressed by the approach of Taylor, and his proposed bootstrap imputation may not be suitable. In general, the use of hot-deck imputation for a large number of individuals can underestimate the variance, as the range of sample variability is restricted by the variability in the replacement set [11]. This underestimation would be accentuated in a hot-deck procedure that does not properly account for the effects of the covariate on the outcome and censoring. Alternatively, the imputation of a large number of outcomes may overestimate the variance if the pool from which the imputed values are drawn is not sufficiently similar to the individual whose values are being imputed.

Herein, we describe the properties of a multiple imputation approach, *risk-stratified imputation*, that we demonstrate to have the potential to reduce or remove the bias arising from withdrawals related to measured exposure variables, as well as the over- or underestimation of the variance in other imputation approaches. This is a specific implementation, focusing on categorical covariates, of the multiple imputation procedures described in the recent National

Research Council report on missing data [12] applied to withdrawals in clinical trials. In addition, because randomization implies equal exposure to the changing risk of events introduced through changes in eligibility, the approach can remove the bias introduced by both measured and unmeasured characteristics of eligibility (i.e., it is not required to identify symptomatic status as the factor that potentially changes the risk of events over the course of the trial). Hence, as shown in later sections, the use of our approach can potentially correct for biases introduced by desired changes to reduce withdrawals, and changes in eligibility criteria when the association with the risk of events is not fully described, and thereby reduce or avoid a systematic bias frequently ignored in randomized trials. We illustrate the properties of the risk-stratified imputation through a series of simulation studies. These methods were also applied to the CREST data and demonstrate that the magnitude of this particular bias in this particular trial was minimal (perhaps because of a lack of differences in treatment efficacy).

Methods and simulation studies

Survival model

Let $T_1, T_2, \ldots, T_n$ denote the event times for the n subjects in the study. In the survival model, an observation will be considered *censored* if the event does not occur during the study period, that is, the subject completes the study without an event. An observation will be considered *withdrawn* if the subject terminates participation prior to the end of the study period and prior to the occurrence of an event. (This is in contrast to many survival analyses, where withdrawals are counted as part of the censoring process.) An observation will be considered a *failure* if an event occurs prior to the censoring and withdrawal times. Let $C_1, C_2, \ldots, C_n$ and $W_1, W_2, \ldots, W_n$ denote the corresponding potential censoring and withdrawal times, respectively. The observed data for subject i are $Y_i = \min(T_i, C_i, W_i)$. Thus, the classification of an individual as failure, withdrawn, or censored is as in Table 1.

The survival data were generated using a model with a separate event hazard $h(t)$ and withdrawal hazard $w(t)$ as a function of time t. The hazard functions incorporated both a treatment x_1 and a covariate x_2 given by the functions ψ and ϕ as

$$\psi(x_1, x_2) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$$

$$\phi(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where α_0, α_1, and α_2 are the coefficients associated with events, and β_0, β_1, and β_2 are the coefficients associated with withdrawals. The treatment and covariate were categorical variables with two levels (0, 1) and three levels (−1, 0, 1), respectively. The event times and withdrawal times were generated using exponential hazard functions

$$h(t) = h_0(t)\exp[\tau_1 \psi(x)]$$

$$w(t) = U[0, \tau_2] \cdot \exp[\phi(t)]$$

The tuning parameters τ_1 and τ_2 were chosen so that about 10%–15% of the subjects would have failure, about 10% would withdraw, and the remaining 75%–80% would be censored by the end of the study, where these parameter values were chosen to provide proportions that are similar to the proportions in our motivating example.

The five scenarios discussed above can be implemented by appropriate choices of the α and β parameters, specifically:

1. Both the event hazard and withdrawal hazard were independent of the treatment and of the covariate, so that $\alpha_1 = 0$, $\alpha_2 = 0$, $\beta_1 = 0$, and $\beta_2 = 0$.
2. The event hazard was dependent on the treatment, while the withdrawal hazard was independent of the treatment and the covariate, so that $\alpha_1 = 1$, $\alpha_2 = 0$, $\beta_1 = 0$, and $\beta_2 = 0$.
3. The event hazard was dependent on the treatment and the covariate, while the withdrawal hazard was independent of the treatment and the covariate, so that $\alpha_1 = 1$, $\alpha_2 = 1$, $\beta_1 = 0$, and $\beta_2 = 0$.
4. The event hazard was dependent on the treatment and the covariate, while the withdrawal

Table 1. Classification of individuals within the survival model

	$T_i \le C_i$		$T_i \ge C_i$		
	$W_i > T_i$	$W_i \le T_i$	$W_i \le T_i$		$W_i > T_i$
			$W_i \le C_i$	$W_i > C_i$	
Event type	Failure	Withdraw	Withdraw	Censored	Censored

hazard was dependent on the treatment, so that $\alpha_1 = 1$, $\alpha_2 = 1$, $\beta_1 = 1$, and $\beta_2 = 0$.
5. Both the event hazard and the withdrawal hazard were dependent on the treatment and the covariate, so that $\alpha_1 = 1$, $\alpha_2 = 1$, $\beta_1 = 1$, and $\beta_2 = 1$.

A series of $r = 10,000$ replicates were generated for each scenario. There were 200 observations (subjects) for each level of the treatment and covariate for a total sample size of 1200 for each replication.

Imputation

In addition to straightforward analyses of the resulting data, three different imputation strategies were used. The first was the bootstrap imputation described by Taylor *et al.* [3], where a bootstrap sample (of the same size as the original data) was created by selecting with replacement from observations in the original simulated data. For each censored subject (either from withdrawal or lack of an event during the study period), a pool of observations was created consisting of all individuals experiencing failure or censoring with time greater than the censoring time for the subject to be imputed. A single subject was randomly selected from the pool, and the values for the subject to be imputed were then set to the outcome (observed event versus censored outcome) and follow-up time (at the time of the outcome) for the selected subject. In the bootstrap imputation, bootstrapping was used to introduce uncertainty to the observed data being employed for imputation. This differs from the typical application of bootstrapping to approximate a distribution nonparametrically and requires only a small number of bootstrap samples. For this analysis, the bootstrapping was performed $m = 10$ times. The second strategy was the auxiliary variable imputation described by Hsu *et al.* [5]. Using a proportional hazards (PH) model, two risk scores were defined based on the relationship of the auxiliary variables to the failure time and the censoring time, respectively. These risk scores are weighted and then used to define a nearest neighborhood (NN = 10 based on the original description of the algorithm) around the individual to be imputed, from which the imputation values are randomly selected using the method of Taylor *et al.* [3]. The third strategy was the risk-stratified imputation, our proposed new approach. For this approach, a pool of observations was created consisting of all individuals with the same treatment and covariate values as the subject being imputed and experiencing failure or censoring with an event or censoring time greater than the withdrawal time for the subject being imputed to condition on survival up to the time of withdrawal. A single subject was randomly selected from the pool, and the values for the subject to be imputed were then set to the values for the selected subject. The risk-stratified imputation was performed $m = 10$ times. Hence, one difference between our approach and that of Taylor is that rather than selecting a random person from the treatment group, we select a random person from the treatment group with a similar value of the covariate associated with both outcome and censoring (in the motivating example, person with the same treatment and symptomatic status). The bootstrap, auxiliary variable, and risk-stratified imputation are depicted in Figure 1(a) to (c) to illustrate these differences.

Analysis

Data were analyzed using a Cox PH model [13] with the treatment and covariate. Model fitting was performed using the *survival* package version 2.35 [14] in the R programming environment version 2.10 [15]. For the three imputation methods, each imputed dataset was analyzed separately. The m imputed datasets were then combined using Rubin's formulae [16]. For comparison, a fourth analysis was conducted on the original simulated data by removing subjects who withdrew (case-wise deletion).

Performance comparison

Performance of the four analytic approaches was evaluated using the root mean square error (RMSE), percent coverage for the treatment effect, and the average length of the 95% confidence interval. The RMSE is a standardized metric for assessing the performance of an estimate relative to a known value, computed as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{r} (\hat{\alpha}_i - \alpha_1)^2}{r}}$$

where α_1 is the true value of the coefficient for treatment effect, and $\hat{\alpha}_i$ is the estimated coefficient on the ith of r replications. Smaller values of the RMSE indicate statistics with reduced bias compared to larger values. A second measure of performance was the coverage. For each of the r replicates, the mean and standard error for the estimated $\hat{\alpha}_i$ in the Cox model were used to construct a 95% confidence interval across the m imputations using Rubin's formulae [16]. The percent coverage, which assesses the accuracy of the variance estimate, was the percentage of confidence intervals containing the true value of α_1, computed as

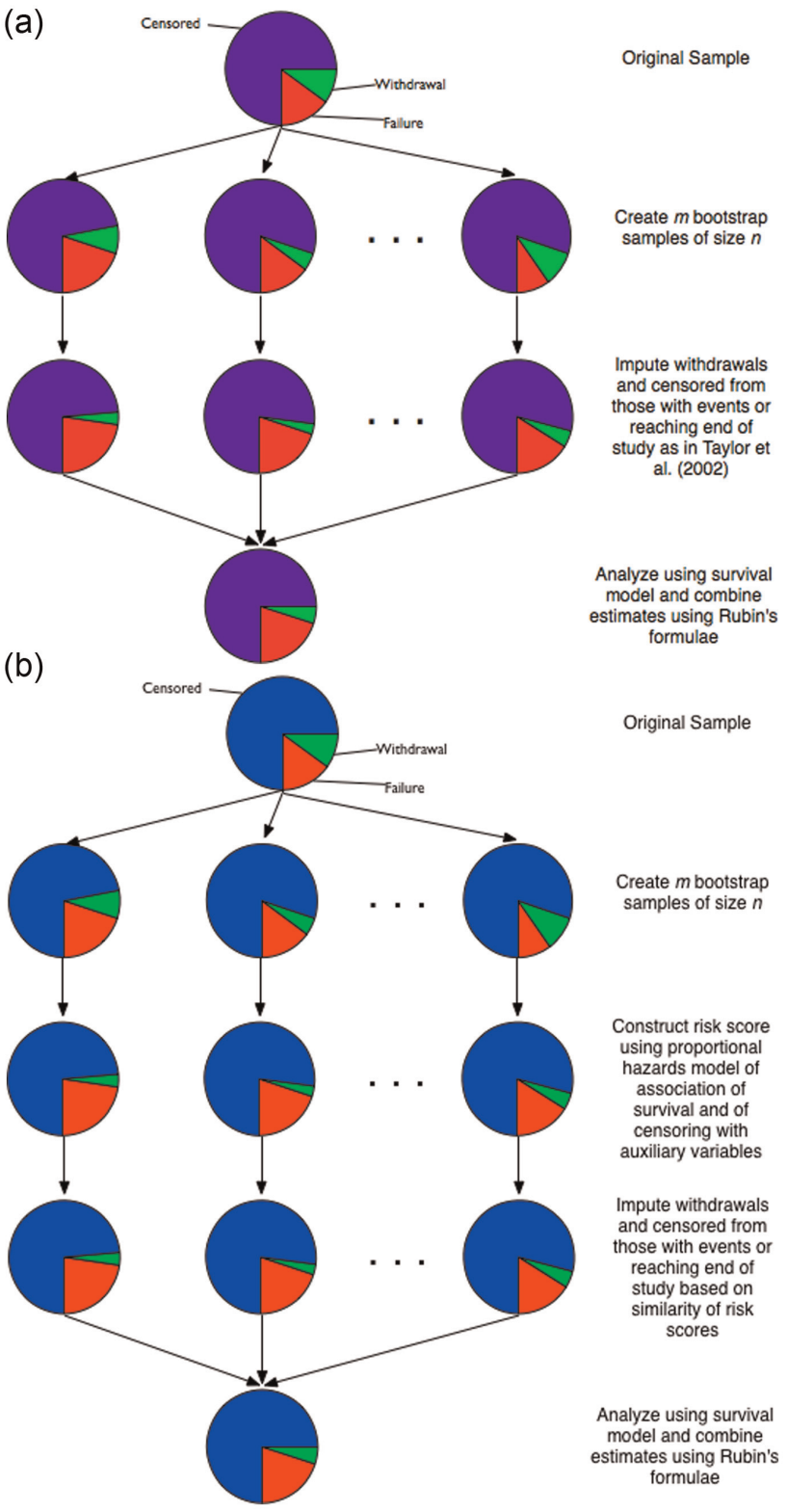

Figure 1. (Continued)

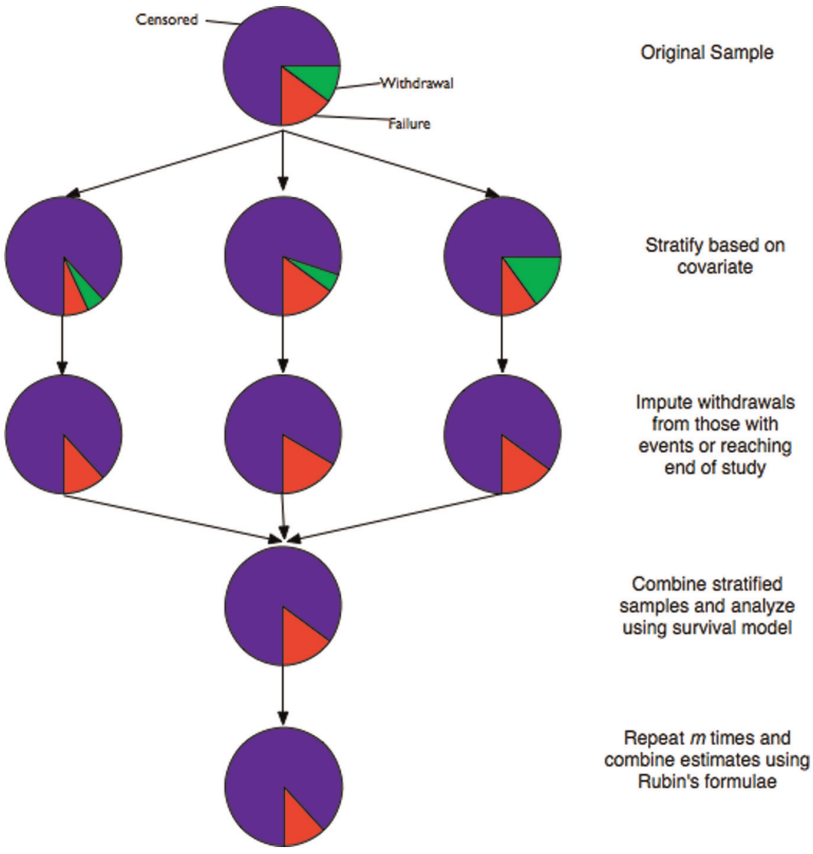

Figure 1. (a) Graphical flowchart of the bootstrap imputation, illustrating the imputation of all censored observations (including withdrawals) and lack of stratification. (b) Graphical flowchart of the auxiliary variable imputation, illustrating the imputation of all censored observations (including withdrawals) and lack of stratification. (c) Graphical flowchart of the risk-stratified imputation, illustrating the imputation of only withdrawals after stratification based on covariate(s) of interest.

$$Coverage = \frac{\sum_{i=1}^{r} \#\{C_{L,i} \leq \alpha_1 \leq C_{U,i}\}}{r}$$

where # is the count of members in the set, and $C_{L,i}$ and $C_{U,i}$ are the lower and upper boundaries of the 95% confidence interval for α_1 respectively on the *i*th replicate, so that the coverage should be 95% for our analyses. Lower coverage would indicate falsely reduced variance estimates, resulting in an excessive (inappropriate) number of positive test results, while higher coverage would indicate overestimation of the variance, resulting in a loss of power to detect significant treatment differences. The average length of the confidence interval was calculated as

$$\text{Mean CI Interval} = \frac{\sum_{i=1}^{r} |C_{U,i} - C_{L,i}|}{r}$$

with smaller values for the average interval indicating tighter confidence bounds and increased statistical power to detect significant treatment differences.

Results

Results for each of the four analytic strategies (case-wise deletion, Taylor's bootstrap imputation, Hsu's auxiliary variable imputation, and our proposed risk-stratified imputation) for each of the five simulation scenarios are given in Table 2. For the first three scenarios (where the assumptions of survival analysis are uniformly met), the rate of withdrawal did not depend on the treatment or the covariate, corresponding to a CCAR pattern of withdrawals. As expected, all four analytic strategies gave accurate estimates of the treatment effect. When the rate of withdrawal depended on the treatment or on the treatment and the covariate, analysis using only the complete data (after removing withdrawals) led to bias in estimation of the treatment effect, which would also be expected. All three imputation strategies led to accurate estimates of the treatment effect, although the RMSE statistic indicates that the risk-stratified imputation slightly outperformed the bootstrap and auxiliary variable imputation in accuracy. Of concern, the bootstrap imputation consistently

Table 2. Estimated values for the treatment effect (α_1) and performance measures under each of the five simulation scenarios

			Complete data	Risk-stratified imputation	Bootstrap imputation (Taylor *et al.* [3])	Auxiliary variable imputation (Hsu *et al.* [5])
Scenario 1:	Failure: 19%;	Mean	0.00042	0.00049	0.00057	0.00074
$\alpha_0 = 1$, $\alpha_1 = 0$, $\alpha_2 = 0$;	withdraw: 7%;	(SE)	(0.0949)	(0.0950)	(0.1007)	(0.0921)
$\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 0$	censored: 75%	RMSE	0.0949	0.0950	0.1006	0.0921
		Coverage	95.0	94.9	99.2	98.8
		Mean CI length	0.3692	0.3696	0.5746	0.4954
Scenario 2:	Failure: 13%;	Mean	1.0039	1.0058	1.0098	0.9534
$\alpha_0 = 0$, $\alpha_1 = 1$, $\alpha_2 = 0$;	withdraw: 7%;	(SE)	(0.0013)	(0.0013)	(0.0013)	(0.0013)
$\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 0$	censored: 80%	RMSE	0.1261	0.1265	0.1336	0.1357
		Coverage	95.0	94.9	99.2	97.6
		Mean CI length	0.4928	0.4934	0.7565	0.6671
Scenario 3:	Failure: 17%;	Mean	0.9983	1.0019	1.0056	0.9081
$\alpha_0 = 0$, $\alpha_1 = 1$, $\alpha_2 = 1$;	withdraw: 7%;	(SE)	(0.0011)	(0.0011)	(0.0012)	(0.0011)
$\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 0$	censored: 76%	RMSE	0.1102	0.1103	0.1180	0.1416
		Coverage	95.1	95.2	99.3	94.7
		Mean CI length	0.4344	0.4349	0.6780	0.5811
Scenario 4:	Failure: 16%;	Mean	0.9230	0.9968	0.9335	0.8532
$\alpha_0 = 0$, $\alpha_1 = 1$, $\alpha_2 = 1$;	withdraw: 13%;	(SE)	(0.00114)	(0.00114)	(0.00122)	(0.00113)
$\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 0$	censored: 71%	RMSE	0.1373	0.1143	0.1386	0.1850
		Coverage	88.8	94.5	98.2	88.2
		Mean CI length	0.4451	0.4449	0.6916	0.6050
Scenario 5:	Failure: 16%;	Mean	0.9383	0.9912	0.9522	0.8757
$\alpha_0 = 0$, $\alpha_1 = 1$, $\alpha_2 = 1$;	withdraw: 18%;	(SE)	(0.00114)	(0.00115)	(0.00122)	(0.00114)
$\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$	censored: 66%	RMSE	0.1293	0.1150	0.1306	0.1683
		Coverage	90.8	94.9	98.5	91.5
		Mean CI length	0.4419	0.4412	0.6849	0.6065

RMSE: root mean square error; SE: standard error; CI: confidence interval.
α_0, α_1, and α_2 represent the intercept (base rate), treatment effect, and nuisance covariate effect on the rate of events, respectively, while β_0, β_1, and β_2 represent the corresponding effects on the rate of withdrawals.

had larger than expected coverage and average confidence interval length across all five scenarios, implying that the Taylor's approach appears to be uniformly liberal in testing. This appears to be due to a combination of greater error in estimating the mean and overestimation of the variance using this method. The performance of auxiliary variable imputation was more complex, with larger than expected coverage in the first two scenarios, but smaller than expected coverage for the last two scenarios where withdrawal was related to treatment and the covariate. This appears to be due to greater error in estimating the mean, together with overestimation of the variance in the first two scenarios and excessive reduction in variance in the last two scenarios.

Application of method to true CREST data

We applied the multiple imputation approach developed herein to the CREST, which compared carotid artery stenting (CAS) to carotid endarterectomy (CEA) to reduce cardiovascular outcomes in 2502 patients with high-grade carotid stenosis [10]. In this study, the proportion of patients withdrawing was related to treatment (p = 0.0019), with 68 (5.4%) of the 1262 patients randomized to receive stenting (CAS) withdrawing consent or being lost to follow-up, compared to 106 (8.6%) of the 1240 assigned to CEA doing so [17]. Enrollment was initially limited to symptomatic patients; after approximately 30% of the patients had been recruited, eligibility was broadened to enrollment of asymptomatic patients to reduce withdrawals. This implied that asymptomatic patients were recruited disproportionately during the period in which withdrawals from the study had been minimized, creating a covariate (symptomatic vs asymptomatic status) that was associated with a lower (p = 0.0043) proportion of withdrawal for the asymptomatic patients (5.4%) than for the symptomatic patients (8.3%). The combined effect of a higher withdrawal rate in the CEA treatment group and a higher withdrawal of

symptomatic patients implies that a higher proportion of high-risk patients have been removed from the CEA group relative to the CAS group, thereby systematically biasing the study to the benefit of the CEA group. Under the alternative hypothesis that treatment affected outcome, this would be similar to the fifth scenario in our simulations. In the original analysis of the CREST dataset, the Kaplan–Meier estimate of the hazard ratio was not significantly different between the two groups (hazard ratio for stenting versus endarterectomy = 1.11; 95% confidence interval = 0.81–1.51; p = 0.51). Reanalysis of the CREST data using risk-stratified imputation with symptomatic status as the covariate for stratification gave an estimated hazard ratio associated with treatment of 1.10 (95% confidence limits about this estimated hazard were 0.54–2.23), while bootstrap imputation gave a slightly higher hazard ratio of 1.16 (and wider 95% confidence limits of 0.49–2.77). As such, while withdrawals related to symptomatic status could potentially introduce bias, these findings support that this is not the case for the CREST study. In this motivating example, the differences in treatment efficacy were not apparent, and this may contribute to the lack of differences in results from the alternative approaches. However, this lack of difference provided support for the claim made in the primary study paper that the treatments did not differ.

Discussion

In clinical trials (and other longitudinal studies), a substantial number of subjects may withdraw prior to the conclusion of the study. Such subjects must be properly dealt with in the analytic phase to prevent biased estimates of the treatment effect. In addition, issue could be raised that not including withdrawals from trials in the analysis is a violation of the intention to treat principle (where all randomized patients should be included in the analysis in their assigned treatment group). We have raised the issue of changing enrollment criteria specifically in the CREST, but would suggest most trials have policies that could naturally lead to changing withdrawal rates over time. It is common for trials to strive to improve methods to reduce withdrawals; as such, as a study progresses, a goal is to constantly work to have the withdrawal (censoring) rates change to lower levels. The alternative, not constantly striving to reduce the proportion of patients withdrawing from a study, is simply not consistent with good trial conduct. However, maintaining a constant high withdrawal rate across the duration of the study would avoid this issue addressed herein. In addition, it is common (as in the CREST) that eligibility criteria will change, reflecting experience in recruitment as the trial progresses. In the CREST, the broadened eligibility criteria were important to ensure the generalizability of the results (approximately 80% of revascularizations performed in the United States are done in asymptomatic patients, who were initially excluded from our trial) and to assist in meeting recruitment goals (recruitment was lagging substantially when eligibility was broadened to include asymptomatic patients). To the extent that the eligibility criteria are associated with event rates, the combination of the successful changes during the conduct of the trial to reduce withdrawals and changes also during the conduct of the trial in eligibility criteria holds the possibility to bias many randomized trials.

In survival analysis, withdrawals are often viewed as a form of censoring. If the withdrawal mechanism is not related to survival, the usual methods may be applied to the data, but withdrawal that is dependent on covariates related to survival requires adjustments to reduce bias due to differential censoring among groups. Robins and Finkelstein [18] proposed a method for censoring that depends on covariates related to survival, the inverse probability of censoring weighted (IPCW) method, in which each observation is weighted by the inverse of the individual being censored at time t. Weights are constructed using a PH model of the relationship between censoring and auxiliary variables and appear more sensitive to misspecification of the censoring model than other methods. An alternative approach is that of multiple imputation, which is commonly used in fields outside of survival analysis. Taylor *et al.* [3] proposed a bootstrap imputation procedure to replace censored and withdrawn observations in survival analysis. For this and other imputation methods, it should not be assumed that subjects drop out for reasons unrelated to the treatment; we have provided the example of the CREST study to demonstrate how this assumption may be violated. We also describe a modified imputation procedure, the risk-stratified imputation, to address the problem that arises in studies where withdrawal depends on the treatment.

Our imputation procedure extends the approach of Taylor *et al.* [3] in two critical ways:

1. The bootstrap imputation imputes values for all censored patients, corresponding to both censored and withdrawal groups in our classification scheme. In contrast, the risk-stratified imputation imputes values only for the withdrawal group; there is no imputation for those in the censored group. The former imputes a substantial portion of the individuals, requiring a large pool from which imputed values may be drawn. An insufficiently large pool could falsely reduce the variance estimates and give an

inflated amount of power. The latter will impute only values for subjects who withdraw but not for subjects who are censored. This greatly increases the pool of potential subjects from which the imputed value may be drawn and, as seen in our simulations, would potentially avoid the false reduction in variance that can occur with the bootstrap imputation. This modification also makes the bootstrapping step of Taylor *et al.* [3] unnecessary.

2. The bootstrap imputation imputes by randomly selecting a subject from the study, who is still active at the time of censoring. Our procedure stratifies the risk of the subjects as a function of the covariate and imputes by randomly selecting a subject with a similar value of the covariate. If the covariate is related to the risk of events, this procedure imputes data for the withdrawn person using a person who is at similar risk, so that we are using a risk-stratified random person. In addition, by stratifying based on treatment, the risk-stratified imputation only selects from a pool of individuals within the same treatment group to estimate the survival time if a subject had not withdrawn but without consideration of whether the individual remained on his or her assigned treatment. In contrast, the bootstrap imputation selects from a pool of individuals at risk at the time a subject withdrew, which may include individuals from different treatment groups. Although the former pool would naturally seem to be more representative of the withdrawn subject than the latter pool, there is no empirical data to justify one over the other.

Hsu *et al.* [5] proposed modifications to the approach of Taylor *et al.* [3] to include auxiliary variables in the imputation process. Although categorical variables can be used, their method primarily focuses on continuous auxiliary variables. This differs from our risk-stratified imputation focused on categorical auxiliary variables, in which the pool of potential values consists of all individuals within the stratum. Furthermore, Hsu *et al.* [5] perform imputation for all censored patients, while the risk-stratified imputation only imputes values for the withdrawal group. Wang *et al.* [19] propose an alternative approach for hot-deck imputation using predictive mean matching (PMM) to include covariates in the imputation process, albeit in the context of imputing recurrent unobserved events rather than a single terminal event. As with Hsu *et al.* [5], this method focuses primarily on continuous covariates and performs imputation for all missing events.

As seen in Table 2, all three methods for dealing with coarsened data perform well when the withdrawal is unrelated to the treatment. In such circumstances, it is well known that analysis of complete cases is sufficient for obtaining unbiased estimates of the treatment effect [18]. Analysis of complete cases is also well known to be inappropriate when withdrawal is related to treatment or to a related covariate [20]. However, the appropriateness of various imputation methods is unclear. In our simulations, all three imputation methods lead to unbiased estimates of the treatment effect when the withdrawal depends on treatment. The RMSE statistic favors estimates from the risk-stratified imputation over the bootstrap imputation and auxiliary variable imputation, although the difference between the first two is not large. However, the bootstrap imputation and auxiliary variable imputation also lead to problems in estimating the variance, reflected by excessively large or excessively small coverage. In the former case, the imputation of a large number of individuals (particularly without stratification) may lead to pools for imputation containing individuals sufficiently dissimilar to the one being imputed to affect results. In the latter case, the imputation of a large number of individuals in these procedures may falsely reduce the variance of the sample, and the bootstrapping step may not be sufficient to address this problem. Although the latter was only observed with auxiliary variable imputation in our simulations, it has been reported with bootstrap imputation with a censoring rate lower than we used [3]. In contrast, our risk-stratified imputation imputes only the much smaller number of individuals who withdrew, so that the variance and coverage remain appropriate. The use of risk stratification in selecting the imputed values addresses the requirement for survival analysis that the individuals who withdrew are representative of all other individuals in the risk group with similar exposure variables. However, as with other uses of stratification, our risk-stratified imputation is appropriate for a small number of categorical covariates; use of high-dimensional covariates would naturally lead to excessively small pools for imputation and corresponding increase in the variance. Conversely, use of a small number of covariates could potentially lead to biased estimates due to large pools for imputation that contain individuals dissimilar to the subject whose values are to be imputed, but the low values of the RMSE and appropriate coverage in our simulations suggest this effect is minimal. Reanalysis of the CREST data shows that both the bootstrap imputation and the risk-stratified imputation give similar hazard ratio estimates as the original analysis. It is likely that the imputation techniques do not show a strong advantage over the original analytic method due to the lack of differences in efficacy between the two treatments. This is similar to Hsu *et al.* [5] who found no difference between treatment

and placebo in AIDS data from the AIDS Clinical Trial Group–019 (ACTG-019) clinical trial using multiple imputation. However, as with our simulations, the bootstrap imputation leads to wider confidence limits than the risk-stratified imputation in the CREST data. This provides evidence favoring the risk-stratified imputation in a real dataset and further strengthens the conclusion drawn from our simulation studies.

Conclusion

We propose a risk-stratified imputation procedure for addressing the problem of nonrandom withdrawals in the context of survival analysis. This approach eliminates the bias present using analyses that assume random withdrawal, while avoiding the over- or underestimation of the variance that can occur with bootstrap and auxiliary variable imputation. The risk-stratified imputation will facilitate the analysis of many clinical trials involving time-to-event data, in which one group experiences a higher withdrawal rate and the withdrawal rate is related to treatment.

Funding

This work was partially supported by the National Institute of Neurological Disorders and Stroke (NINDS; grant number RO1 NS 038384), by supplemental funding from Abbott Vascular Solutions (formally Guidant) including donations of Accunet and Acculink Systems equivalent to approximately 15% of the total study costs, and by the National Heart, Lung, and Blood Institute (NHLBI; grant number T32 HL 072757).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. **Collett D.** *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC, Boca Raton, FL, 2003.
2. **Rubin D.** Multiple imputation after 18+ years. *J Am Stat Assoc* 1996; **91**: 473–89.
3. **Taylor JMG, Murray S, Hsu C-H.** Survival estimation and testing via multiple imputation. *Stat Probabil Lett* 2002; **58**: 221–32.
4. **Little RJA, Rubin DB.** *Statistical Analysis with Missing Data*. Wiley, New York, 1987.
5. **Hsu C-H, Taylor JM, Murray S, Commenges D.** Survival analysis using auxiliary variables via non-parametric multiple imputation. *Stat Med* 2006; **25**: 3503–17.
6. **Heitjan D, Rubin D.** Ignorability and coarse data. *Ann Stat* 1991; **19**: 2444–53.
7. **Tsiatis AA.** *Semiparametric Theory and Missing Data*, ch. 7. Springer, New York, 2006.
8. **Little R.** A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc* 1988; **83**: 1198–202.
9. **Potthoff RF, Tudor GE, Pieper KS, Hasselblad V.** Can one assess whether missing data are missing at random in medical studies? *Stat Methods Med Res* 2006; **15**: 213–34.
10. **Sheffet AJ, Roubin G, Howard G, *et al*.** Design of the Carotid Revascularization Endarterectomy vs. Stenting Trial (CREST). *Int J Stroke* 2010; **5**: 40–46.
11. **Pérez A, Dennis RJ, Gil JFA, Rondón MA, López A.** Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. *Stat Med* 2002; **21**: 3885–96.
12. **National Research Council.** *The Prevention and Treatment of Missing Data in Clinical Trials*. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC, 2010.
13. **Cox DR.** Regression models and life-tables. *J Roy Stat Soc B* 1972; **34**: 187–220.
14. **Therneau TM, Grambsch PM.** *Modeling Survival Data*. Springer-Verlag, New York, 2000.
15. **R Development Core Team.** *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, 2009.
16. **Rubin DB.** *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.
17. **Brott TG, Hobson RW, Howard G, *et al*.** Stenting versus endarterectomy for treatment of carotid-artery stenosis. *N Engl J Med* 2010; **363**: 11–23.
18. **Robins JM, Finkelstein DM.** Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; **56**: 779–88.
19. **Wang C-N, Little R, Nan B, Harlow SD.** A hot-deck multiple imputation procedure for gaps in longitudinal recurrent event histories. *Biometrics* 2011; **67**: 1573–82.
20. **Enders CK.** A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosom Med* 2006; **68**: 427–36.