Estimating Covariate-Adjusted Log Hazard Ratios for Multiple Time Intervals in Clinical Trials using Nonparametric Randomization Based ANCOVA

Patricia F. MOODIE, Benjamin R. SAVILLE, Gary G. KOCH, Catherine M. TANGEN

The hazard ratio is a useful tool in randomized clinical trials for comparing time-to-event outcomes for two groups. Although better power is often achieved for assessments of the hazard ratio via model-based methods that adjust for baseline covariates, such methods make relatively strong assumptions, which can be problematic in regulatory settings that require prespecified analysis plans. This article introduces a nonparametric method for producing covariate-adjusted estimates of the log weighted average hazard ratio for nonoverlapping time intervals under minimal assumptions. The proposed methodology initially captures the means of baseline covariables for each group and the means of indicators for risk and survival for each interval and group. These quantities are used to produce estimates of interval-specific log weighted average hazard ratios and the difference in means for baseline covariables between two groups, with a corresponding covariance matrix. Randomization-based analysis of covariance is applied to produce covariate-adjusted estimates for the interval-specific log hazard ratios through forcing the difference in means for baseline covariables to zero, and there is variance reduction for these adjusted estimates when the time to event has strong correlations with the covariates. The method is illustrated on data from a clinical trial of a noncurable neurologic disorder.

Key Words: Analysis of covariance; Linear models equating covariate means; Survival; Time-to-event; Weighted least squares.

1. Introduction

Randomized clinical trials often incorporate time-to-event measures as primary outcomes. Under the assumption of proportional hazards between the two groups, in which the hazard of a given group is a measure of the instantaneous event rate at a given time, the Cox proportional hazards model (Cox 1972) facilitates the comparison of instantaneous event rates for two groups via the log hazard ratio. Under less stringent assumptions, traditional nonparametric methods for comparing time-to-event data in two groups include the Wilcoxon and logrank tests (Gehan 1965; Peto and Peto 1972), although proportional hazards with zero as the log hazard ratio naturally applies when there are no differences between the two groups. Better power for the comparison between the two groups via the Cox model or nonparametric methods can be achieved by adjusting for covariates; see Jiang et al. (2008) who discussed a supporting simulation study for this consideration. Such covariance adjustment is straightforward and commonly implemented for the Cox proportional hazards model, but is not widely implemented for the logrank and Wilcoxon tests, even
though randomization-based methods with such capability do exist in the literature (Tangen and Koch 1999b).

For randomized clinical trials in regulatory settings, one must specify the primary outcomes and primary analyses a priori (LaVange et al. 2005). This is problematic when implementing the Cox proportional hazards model in an analysis plan, because one cannot evaluate the proportional hazards assumption in advance of data collection. For example, consider a clinical trial of 722 patients with an incurable neurologic disorder. Patients were randomized to test treatment or control in a 2:1 ratio, resulting in 480 and 242 patients in the test treatment and control groups, respectively. The primary aim was to compare time to disease progression within 18 months for test treatment versus control in a setting for which the test treatment was expected to delay the disease progression but not prevent it. To achieve better power, the investigators also identified 22 covariates and 3 geographical strata for adjustment in the analysis. For the comparison of the two treatments, the analysis included a logrank test without covariate adjustment \((p\text{-value} = 0.048)\) and a Cox proportional hazards model with baseline covariate adjustment \((p\text{-value} = 0.002)\). The Cox model with baseline covariates assumes proportional hazards not only for the treatment parameter, but also for each covariate in the model, adjusting for all other predictors in the model. Upon reviewing the results, a regulatory agency was concerned about the validity of the proportional hazards assumption and the difference in magnitude of the observed treatment effect between the unadjusted and the adjusted Cox analysis. There is a known tendency for the Cox adjusted analysis to have a point estimate for the log hazard ratio that is further from the null than the unadjusted estimate, as well as a somewhat larger standard error (Tangen and Koch 1999b, 2000). However, model-checking and fine-tuning the analysis post hoc in order to address these concerns can be viewed as exploratory and can produce misleading results (Lewis et al. 1995).

The investigators could have avoided these concerns by specifying a nonparametric analysis that accommodates covariance adjustment a priori. Tangen and Koch (1999b) proposed using randomization-based nonparametric analysis of covariance (RBANCOVA) for comparing differences in logrank or Wilcoxon scores with adjustment for relevant baseline covariates. A limitation of their method is that it does not enable the estimation of a hazard ratio with a corresponding confidence interval, which has great appeal in interpretation compared to the difference in logrank or Wilcoxon scores. Tangen and Koch (2000) used nonparametric randomization-based ANCOVA to estimate covariance-adjusted hazard ratios through estimating incidence density ratios for intervals from a partition of time. Their method has the advantage of providing an interpretable hazard ratio, but its limitation is that incidence density ratios are only applicable as estimators of hazard ratios when the survival times follow a piecewise exponential distribution for the intervals from the partition of time. In this article, these methods are improved upon via a nonparametric randomization based method for estimating covariate-adjusted hazard ratios for nonoverlapping time intervals from a partition of time under minimal assumptions. Estimates for interval-specific hazard ratios are produced, from which one can assess homogeneity of the estimates across time intervals, and ultimately produce a common hazard ratio across intervals with covariate adjustment. The manuscript is organized as follows. In Section 2, the general strategy of the method is outlined, with details provided in the Appendices; in Section 3 the method is applied to the illustrative example; and a discussion is provided in Section 4.

2. Methods

For test and control treatments as \(i = 1, 2\) respectively, let \(S_i(t)\) denote the survivorship function for the probability of surviving (or avoiding) some event for at least time \(t\) where \(t > 0\) is inherently continuous. The hazard function, \(h_i(t) = -S_i'(t)/S_i(t)\), where \(S_i'(t) = dS_i(t)/dt\) is the first derivative function of \(S_i(t)\) with respect to \(t\), is the instantaneous event rate of group \(i\) at time \(t\). A hazard ratio \(\theta(t) = h_1(t)/h_2(t)\) is then a useful measure for comparing the hazard of group 1 to group 2.

By partitioning time into \(J\) nonoverlapping consecutive time intervals (see Appendix A.1 for complete details), one can decompose the survivorship function for group 1 in the \(j\)th interval, \(S_1(t_j)\), where \(t_j\) is the time upper bound of interval \(j\), into the product of the survivorship function for the previous interval and the conditional survivorship function for interval \(j\) given survival of interval \((j - 1)\), written as \(S_1(t_{j-1}) \exp\{-\int_{t_{j-1}}^{t_j} h_1(x)dx\}\). Noting that \(h_1(x) = \theta(x)h_2(x)\), one can use the Second Mean Value Theorem for Integrals, as discussed by Moodie et al. (2004), to formulate a log weighted average hazard ratio for group 1 relative to group 2 in interval \(j\), assuming that the probability of at least one event is greater than zero for each group during this interval (so that there is at least one observed event through sufficient sample size). This log weighted average hazard ratio is based on a log(-log) transformation of the conditional probabilities for survival of the \(j\)th interval given survival of the \((j - 1)\)th interval, or \(\pi_{ij} = S_1(t_j)/S_1(t_{j-1})\). Application of the log(-log) transformation to the estimator of \(\pi_{ij}\) requires that at least one event has occurred in each group in every interval under consideration. As noted by Moodie et al. (2004), the weights for the log average hazard ratio for interval \(j\) are the hazards for the
control group in the \( j \)th interval. Such weights arguably have clinical relevance as they place emphasis on those times when the hazard function in the control group is large, which can be particularly helpful when evaluating the potentially beneficial effect of a test treatment. The proposed estimator does not assume proportional hazards within or across intervals; however, if the hazards are proportional in a given interval \( j \), the log weighted average hazard ratio for interval \( j \) is equal to the proportionality constant of interval \( j \) on a log scale.

For estimation of the log weighted hazard ratio, one first estimates the extent of risk \( r_{ijk} \) and survival \( s_{ijk} \) for patient \( k \) for interval \( j \) in group \( i \) for each of the \( J \) time intervals. As shown in Appendix A.2, if a patient is censored, \( r_{ijk} \) equals 1 for all of the intervals prior to censoring, equals 0 for all of the intervals after the interval with censoring, and equals 0.5 or the proportion of the interval that was at risk prior to censoring for the interval with censoring. Also, \( s_{ijk} \) equals 1 for all of the intervals prior to censoring, equals 0 for all of the intervals after the interval with censoring, and equals 0.5 or the proportion of the interval that was at risk prior to censoring for the interval with censoring. If a patient has an event, \( r_{ijk} \) equals 1 for all of the intervals prior to and including the interval with the event, and equals 0 for all intervals after the interval with the event. Also, \( s_{ijk} \) equals 1 for all of the intervals prior to the event, equals 0 for the interval with the event, and equals 0 for all intervals after the event.

A vector \( f_{ik} \) is created that contains the risk estimates \( r_{ijk} \), survivorship estimates \( s_{ijk} \), and \( M \) relevant baseline covariates \( x_{jk} \). Means are produced for each of these respective components, and a corresponding covariance matrix is estimated. By using the ratio of the mean survivorship over the mean risk, one can construct estimates of the conditional probability for survival of the \( j \)th interval given survival of the \((j-1)\)th interval, and use a log(-log) transformation of these values to estimate the log weighted average hazard ratio for interval \( j \). A corresponding covariance matrix is produced via linear Taylor series methods.

One then constructs a vector \( d \) that contains the estimated log weighted average hazard ratio for each interval, and the differences in means of the baseline covariates, along with a consistent estimator of the covariance matrix \( V_d \). Weighted least squares regression, as discussed by Stokes et al. (2000, Chapter 13), is then used to produce covariate adjusted estimates \( \hat{b} \) for the log hazard ratios through forcing the difference in means for covariates to 0. More specifically, as discussed by Koch et al. (1998), Tangen and Koch (1999b), and LaVange et al. (2005), nonparametric randomization based ANCOVA has invocation for \( d \) by fitting the linear model \( X = [I_J, 0_{JM}]' \) to \( d \) by weighted least squares, in which \( 0_{JM} \) is the \((J \times M)\) matrix of 0’s, and \( I_J \) is the \((J \times J)\) identity matrix, with \( J \) as the number of intervals and \( M \) as the number of baseline covariables. The resulting covariance adjusted estimator \( \hat{b} \) for the log hazard ratios is given by

\[
\hat{b} = (X'V_d^{-1}X)^{-1}X'V_d^{-1}d. \tag{1}
\]

A consistent estimator for the covariance matrix of \( \hat{b} \) is \( V_b \) in (2).

\[
V_b = (X'V_d^{-1}X)^{-1} \tag{2}
\]

The estimators \( \hat{b} \) have an approximately multivariate normal distribution when the sample sizes for each group are sufficiently large for \( d \) to have an approximately multivariate normal distribution (e.g., each interval of each group has at least 10 patients with the event and 10 patients who complete the interval without the event and each group has an initial sample size of at least 80 patients). In this regard, simulation studies (Moodie et al. 2004) suggest that sample sizes larger than 100 may be required if the values of the survivorship functions involved in computing the log hazard ratio are close to one or zero.

The rationale for randomization-based covariance adjustment is the expected absence of differences between test treatment and control groups for the means \( \bar{x} \) of the covariables. A related criterion for evaluating the extent of random imbalances between the test and control groups for the \( \bar{x} \) is \( Q_0 \) in (3).

\[
Q_0 = (d - Xb)'V_d^{-1}(d - Xb). \tag{3}
\]

This criterion approximately has the chi-squared distribution with \( M \) degrees of freedom.

The homogeneity of the adjusted log hazard ratios in \( b \) across the \( J \) time intervals can have assessment with the criterion \( Q_{\text{homog.b}} \) in (4)

\[
Q_{\text{homog.b}} = b'C'(CV_dC')^{-1}Cb, \tag{4}
\]

where \( C = [I_{(J-1)}, -1_{(J-1)}] \). This criterion approximately has the chi-squared distribution with \((J - 1)\) degrees of freedom. When homogeneity of the adjusted log hazard ratios in \( b \) does not have contradiction by \( Q_{\text{homog.b}} \), then the common adjusted log hazard ratio \( b_{\text{homog}} \) can have estimation by weighted least squares as shown in (5). A consistent estimator for its variance is \( \hat{v}_{b_{\text{homog}}} = (1'V_b^{-1}1_J)^{-1} \).

\[
b_{\text{homog}} = (1'V_b^{-1}b)/(1'V_b^{-1}1_J). \tag{5}
\]

The estimator \( b_{\text{homog}} \) approximately has a normal distribution with \( \hat{v}_{b_{\text{homog}}} \) as the essentially known variance. Accordingly, a two-sided 0.95 confidence interval for the common hazard ratio for the comparison between groups 1 and 2 with randomization based covariance adjustment
is \( \exp \{ b_{\text{homog}} \pm 1.96 \sqrt{\text{var}(b_{\text{homog}})} \} \). When homogeneity for \( b \) has contradiction by the result for (4), an average log hazard ratio \( \bar{b} = (1/J) \hat{b} \) may still be of interest when the signs of the elements of \( b \) are predominantly the same. A consistent estimator for its variance is \( (1/J) \hat{V}_{\hat{b}1} (1/J) = \hat{v}_{\hat{b}} \), and the two-sided 0.95 confidence interval for the corresponding average hazard ratio is \( \exp \{ \bar{b} \pm 1.96 \sqrt{\hat{v}_{\bar{b}}} \} \).

The proposed method should have reasonably good properties in terms of Type I error of statistical tests and coverage of confidence intervals when the values of the survivorship functions involved are not too close to one or zero relative to the initial sample sizes for each group, as well as relative to the number of patients with the event and the number of patients who complete the interval without the event in each time interval in each group (e.g., the initial sample size for each group is greater than 80 and each time interval for each group has at least 10 patients with the event and at least 10 patients who complete the interval without the event). This structure enables \( d \) to have an approximately multivariate normal distribution with \( V_d \) as its essentially known covariance matrix; see Koch et al. (1995, 1977, 1985, 1998) and Koch and Bhapkar (1982). Beyond these general guidelines, how many intervals to use and how to specify them so as to have better power are questions beyond the scope of this paper. In most situations, the number of intervals \( J \) would be between 3 and 10 so as to provide as comprehensive information as feasible for the survivorship functions \( S_i(t) \) with respect to the expected number of events throughout the duration of follow-up \( T \), but without making the dimension of \( d \) so large that its estimated covariance matrix \( V_d \) loses stability. Also, the respective intervals would usually have equal lengths \( (T/J) \) when this specification is reasonably compatible with roughly similar numbers of events per interval, but information concerning the \( S_i(t) \) from previous studies could support an alternative specification for roughly similar numbers of events per interval. Additional future research with simulation studies can shed light on whether other specifications for \( J \) and the ranges of time for the intervals can provide better power than the previously stated guidelines.

### 3. Application

The proposed method is illustrated for the previously described clinical trial of 722 patients with an incurable neurologic disorder. The original protocol for this clinical trial identified 22 covariables a priori and 3 geographical strata to be incorporated into the analysis, including demographics, vital sign covariates, and disease onset and severity covariates. For the purposes of this illustration, covariate-adjustment will focus on all 22 covariables as well as 2 dummy variables for geographical region, resulting in 24 covariates total. Collinearity diagnostics of these 22 covariables (variance inflation factors) did not indicate any problems with collinearity.

A graph of the Kaplan-Meier estimates for the probability of no progression is provided in Figure 1. The benefit of treatment versus control appears to be the largest around 12 months, but lessens somewhat by 18 months.

---

Figure 1. Kaplan Meier survival estimates.
The logrank test statistic for comparing test treatment versus control yields a \( p \)-value of 0.048, indicating a borderline significant difference at the \( \alpha = 0.05 \) level. An unadjusted Cox proportional hazards model results in an estimated hazard ratio of 0.797 with 0.95 confidence interval as (0.636, 0.998) and \( p \)-value = 0.048, showing a significant benefit for test treatment versus control. Adjusting for the 24 baseline covariates, an adjusted Cox proportional hazards model yields an estimated hazard ratio of 0.642 with 0.95 confidence interval as (0.508, 0.813) and \( p \)-value = 0.002, showing a substantially stronger benefit for test treatment versus control compared to the unadjusted counterpart. The parameter estimate for the adjusted log hazard ratio (−0.443) corresponds to a much larger effect size than the unadjusted parameter estimate (−0.227), but also has a slightly larger standard error (0.120 compared to 0.115, respectively), as shown in Table 1.

In general, one typically adjusts for covariates for the purpose of achieving variance reduction through smaller estimated standard errors. For the Cox proportional hazards model, such a trend was not observed, primarily because the parameter estimate of the adjusted Cox model applies to a specific subpopulation corresponding to patients with specific baseline covariate values, whereas the unadjusted parameter estimate is an unrestricted population average log hazard ratio; see Koch et al. (1998) and Tangen and Koch (2000). Despite the lack of variance reduction in the adjusted model, the corresponding parameter estimate shows a substantial increase in effect size, and hence results in a smaller \( p \)-value compared to the unadjusted counterpart, but this pattern of results can create dilemmas for regulatory reviewers with concerns for exaggerated estimates of effect sizes.

For the application of nonparametric randomization based analysis of covariance (RBANCOVA) methods that do not require the proportional hazards assumption, the 18-month study was divided into six time intervals, each 3 months long (with this specification for the intervals being in harmony with the guidelines at the end of Section 2, particularly with the recognition that 24 covariables and 6 intervals leads to the dimension of \( d \) being 30 which is near what the sample size of 722 might reasonably support for stability of the estimated covariance matrix \( V_d \) of \( d \)). Interval-specific covariate adjusted estimates of the hazard ratios were obtained using the proposed nonparametric RBANCOVA methods (see Table 2). With the exception of the 12–15 month interval, all intervals show a benefit for treatment versus control.

### Table 1. Estimated common log hazard ratios across intervals for time to disease progression

<table>
<thead>
<tr>
<th>Method</th>
<th>Covariate Adj.</th>
<th>Estimate</th>
<th>SE</th>
<th>( p )-value</th>
<th>HR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBANCOVA</td>
<td>No</td>
<td>−0.236</td>
<td>0.116</td>
<td>0.042</td>
<td>0.790 (0.629, 0.991)</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>−0.231</td>
<td>0.090</td>
<td>0.011</td>
<td>0.794 (0.665, 0.948)</td>
</tr>
<tr>
<td>Cox</td>
<td>No</td>
<td>−0.227</td>
<td>0.115</td>
<td>0.048</td>
<td>0.797 (0.636, 0.998)</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>−0.443</td>
<td>0.120</td>
<td>0.002</td>
<td>0.642 (0.508, 0.813)</td>
</tr>
<tr>
<td>Tangen and Koch (2000)</td>
<td>Yes</td>
<td>−0.263</td>
<td>0.089</td>
<td>0.003</td>
<td>0.77 (0.64, 0.91)</td>
</tr>
</tbody>
</table>

Table 2. RBANCOVA: Interval-specific estimated log hazard ratios for time to disease progression

<table>
<thead>
<tr>
<th>Adjusted</th>
<th>Interval (months)</th>
<th>Estimate</th>
<th>SE</th>
<th>HR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>0–3</td>
<td>−0.128</td>
<td>0.363</td>
<td>0.880 (0.433, 1.791)</td>
</tr>
<tr>
<td></td>
<td>3–6</td>
<td>−0.370</td>
<td>0.262</td>
<td>0.691 (0.413, 1.156)</td>
</tr>
<tr>
<td></td>
<td>6–9</td>
<td>−0.399</td>
<td>0.267</td>
<td>0.671 (0.398, 1.133)</td>
</tr>
<tr>
<td></td>
<td>9–12</td>
<td>−0.506</td>
<td>0.246</td>
<td>0.663 (0.372, 0.976)</td>
</tr>
<tr>
<td></td>
<td>12–15</td>
<td>0.255</td>
<td>0.322</td>
<td>1.291 (0.687, 2.426)</td>
</tr>
<tr>
<td></td>
<td>15–18</td>
<td>0.025</td>
<td>0.290</td>
<td>1.025 (0.580, 1.810)</td>
</tr>
<tr>
<td>Yes</td>
<td>0–3</td>
<td>−0.024</td>
<td>0.343</td>
<td>0.976 (0.498, 1.911)</td>
</tr>
<tr>
<td></td>
<td>3–6</td>
<td>−0.228</td>
<td>0.243</td>
<td>0.796 (0.494, 1.281)</td>
</tr>
<tr>
<td></td>
<td>6–9</td>
<td>−0.447</td>
<td>0.254</td>
<td>0.639 (0.389, 1.051)</td>
</tr>
<tr>
<td></td>
<td>9–12</td>
<td>−0.464</td>
<td>0.231</td>
<td>0.629 (0.400, 0.988)</td>
</tr>
<tr>
<td></td>
<td>12–15</td>
<td>0.120</td>
<td>0.304</td>
<td>1.127 (0.622, 2.044)</td>
</tr>
<tr>
<td></td>
<td>15–18</td>
<td>−0.083</td>
<td>0.275</td>
<td>0.920 (0.536, 1.579)</td>
</tr>
</tbody>
</table>

*Test of homogeneity across intervals, unadjusted \( p = 0.408 \), adjusted \( p = 0.605 \)
Despite the fact that the 12–15 month interval shows an increased risk in disease progression for test treatment versus control, a test for homogeneity for the treatment parameter estimates across the 6 intervals yields a p-value of 0.605, which does not contradict the assumption of a common log hazard ratio. The covariate adjusted hazard ratio is 0.794 with 0.95 confidence interval as (0.665, 0.948) and p-value = 0.011, indicating a significant benefit for delaying disease progression for test treatment versus control. Note that the nonparametric RBANCOVA estimate of the hazard ratio is similar to the unadjusted Cox estimate, but the confidence interval is narrower for the nonparametric RBANCOVA estimate. An assessment for random imbalances in the baseline covariates between the test and control groups yields a p-value of 0.120, which does not suggest noteworthy baseline imbalances between the two groups. The number of events by treatment group and also by interval are provided in Table 3, with these being sufficient for invocation of nonparametric RBANCOVA methods.

For comparison, one can also apply weighted least squares methods without covariate adjustment to produce unadjusted log hazard ratios for each of the six intervals. The unadjusted common hazard ratio is 0.790 with 0.95 confidence interval as (0.629, 0.991) and p-value = 0.042, with test of homogeneity p-value = 0.408. Hence, the adjusted hazard ratio has a comparable effect size to the unadjusted counterpart and a narrower 0.95 confidence interval, which is not the case in the Cox model. This is because the nonparametric RBANCOVA adjusted hazard ratio provides a population average hazard ratio estimate with differences in baseline covariates forced to zero, while the adjusted Cox hazard ratio provides a hazard ratio estimate for specific subpopulations corresponding to covariate profiles; see Koch et al. (1998) and Tangen and Koch (2000). One can also note that the standard error of the adjusted nonparametric RBANCOVA log hazard ratio (0.090) is smaller than the standard error of the unadjusted nonparametric RBANCOVA log hazard ratio (0.116), meaning that the nonparametric RBANCOVA covariate adjustment produced a variance reduction in the parameter estimate for the comparison between the two treatments. Similar patterns are observed for comparing the interval specific covariate-adjusted hazard ratios to the interval specific unadjusted hazard ratios (Table 2).

4. Discussion

Tangen and Koch (2000) used these same data to illustrate a nonparametric randomization based method for estimating hazard ratios via log incidence density ratios. For data from an exponential distribution, the incidence density is an efficient estimator for the hazard ratio. The authors estimated incidence density ratios (IDR) for 3 intervals, each 6 months long, and produced a common estimate across all intervals. Their nonparametric procedure adjusted for the same 24 baseline covariates. As observed in the current illustration, the authors noted a treatment benefit during the first 12 months (IDR 0.64 and 0.66 for 0–6 months and 6–12 months, respectively), and an increased risk in months 12–18 (IDR 1.22). The common IDR was 0.77 (see Table 1), compared to a common hazard ratio of 0.794 in the current illustration. The standard error of the covariate adjusted common log IDR (0.089) observed by Tangen and Koch (2000) was nearly identical to the standard error of the common hazard ratio (0.090) from the methods in this article for this example. Their approach does not require the proportional hazards assumption (i.e., a constant hazard ratio) throughout the entire follow-up period; but its scope can be limited by the requirement of a piecewise exponential distribution in order to use the incidence density ratio as an estimator of the hazard ratio. However, sufficiently large sample size and sufficiently many events for the two groups can often enable the use of sufficiently many intervals so that the assumption of a piecewise exponential distribution can be realistic.

Tangen and Koch (1999b) also used these data to illustrate a nonparametric method for analyzing time to event outcomes with logrank and Wilcoxon scores. Their method produced covariate-adjusted differences in means of logrank and Wilcoxon scores (−0.088 and −0.109, respectively) and corresponding p-values (0.034 and 0.011, respectively) for comparing time to disease progression in test treatment versus control via nonparametric randomization based ANCOVA. Although their method is an effective tool for evaluating treatment dif-

---

Table 3. Summary of number of events per interval by group

<table>
<thead>
<tr>
<th>Group</th>
<th>0–3</th>
<th>3–6</th>
<th>6–9</th>
<th>9–12</th>
<th>12–15</th>
<th>15–18</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>12</td>
<td>25</td>
<td>24</td>
<td>29</td>
<td>13</td>
<td>17</td>
<td>120</td>
</tr>
<tr>
<td>Test</td>
<td>21</td>
<td>35</td>
<td>34</td>
<td>39</td>
<td>38</td>
<td>38</td>
<td>205</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>60</td>
<td>58</td>
<td>68</td>
<td>51</td>
<td>55</td>
<td>325</td>
</tr>
</tbody>
</table>

---

Statistics in Biopharmaceutical Research: Vol. 0, No. 0
ferences in time-to-event outcomes under minimal assumptions, the estimate of differences in logrank or Wilcoxon scores lacks appeal in interpretation, and does not provide an estimate of the hazard ratio with corresponding confidence interval.

As noted by Tangen and Koch (2000), the regulatory agency that reviewed this study was concerned with the large difference in \( p \)-values between the adjusted and unadjusted Cox models for evaluating treatment differences, and whether the adjusted Cox model overstated the treatment effect size. In addition, there was concern whether the proportional hazards assumption was violated in these data, that is, whether proportional hazards holds for each predictor in the model, adjusting for all other predictors in the model. The proposed method has great appeal in such regulatory settings. Like the Cox model, it provides an adjusted estimate of the log hazard ratio, but does so without the assumption of proportional hazards for each covariate within or across intervals. Essentially the only assumptions of the proposed method are valid randomization to the respective treatment groups and noninformative censoring of the event time. Moreover, the adjusted log hazard ratio produced by the method is similar in magnitude to the unadjusted counterpart, but has better precision in terms of a smaller standard error of the log hazard ratio.

To accommodate situations where the proportional hazards assumption is not realistic for Cox models with baseline (i.e., not time-dependent) covariate predictors, Cox (1972) proposed more complicated models with time-dependent covariates and suggested the product of a covariate and time as an example of one of many possible functions that could parametrically represent the interrelationships between the outcome and a covariate over time. Fisher and Lin (1999) discussed the opportunities as well as the complications associated with using time-dependent covariates in Cox models and advised caution in interpreting these models. Additionally, time-dependent covariates can be particularly difficult to interpret in randomized clinical trials because they can have confounding with the treatment effect.

Although ideal for randomized clinical trials requiring prespecified analysis plans, the nonparametric randomization-based methods in this article do not address all aspects of interest for analysis; see Koch et al. (1998) and Tangen and Koch (1999a,b, 2000). For example, because the proposed method provides a population-averaged test treatment effect, the method does not provide an assessment of homogeneity of the test treatment effect across specific patient subpopulations, nor does it provide an assessment for relevant covariates. In contrast, model-based methods such as the Cox proportional hazards model can address baseline and time-dependent covariable effects, treatment by covariables interactions, and assessment of treatment effects for specific subpopulations, but do so under more stringent assumptions. A useful approach is to use both nonparametric randomization-based methods and model-based methods as complementary approaches for addressing the research questions of interest. For example, in regulatory settings requiring prespecified analysis plans, it may be advantageous to prespecify the nonparametric randomization-based method as the primary analysis under minimal assumptions, with Cox proportional hazards models implemented as supportive analyses for further understanding of the data.

The proposed method could also be extended to accommodate multivariate time-to-event settings. For example, consider two events of interest: time to disease progression or death (whichever comes first) and time to death. One could produce a hazard ratio for each event type for several intervals, and then create an average hazard ratio across event type and interval with nonparametric randomization based covariate adjustment. Related multivariate approaches have been highlighted for the Cox model and logrank test (Wei et al. 1989; Saville et al. 2010), but these methods lack the capability to produce an estimate of the average hazard ratio under minimal assumptions. The methods in this article can additionally have extension to randomized studies with 3 to 5 treatment groups with sufficient sample size through straightforward adaptations of the structure discussed by Tangen and Koch (2001).

Appendix

A.1 Specifications for Average Hazard Ratios within Time Intervals

For test and control treatments as \( i = 1,2 \) respectively, let \( S_i(t) \) denote the survivorship function for the probability of surviving (or avoiding) some event for at least time \( t \) where \( t > 0 \) is inherently continuous. Given that \( S_i(t) \) has continuous derivatives through order 2 for all \( t \leq T \) with \( T \) being the duration of follow-up for the event and with \( S_1(t) > 0 \), the corresponding hazard function
\[
h_i(t) = -S_i(t)/S_i(t) \quad \text{where} \quad S_i'(t) = dS_i(t)/dt \quad \text{is the first derivative for} \quad S_i(t) \quad \text{with respect to} \quad t.
\]
Accordingly, it follows that \( S_i(t) = \exp \left( - \int_0^t h_i(x) \, dx \right) \). With \( \theta(x) = h_1(x)/h_2(x) \) as the hazard ratio function for group 1 relative to group 2, \( S_1(t) \) can have expression as in (A.1)
\[
S_1(t) = \exp \left( - \int_0^t \theta(x) h_2(x) \, dx \right). \quad (A.1)
\]
For the partition of time into \( J \) consecutive and adjacent intervals \( (t_0 = 0, t_1], (t_1, t_2], \ldots, (t_{J-1}, t_J = T] \), (A.1) im-

Estimating Covariate-Adjusted Log Hazard Ratios for Multiple Time Intervals in Clinical Trials
plies the structure in (A.2).

\[ S_1(t_j) = \exp \left\{ -\sum_{j'=1}^{j} \left( \int_{t_{j'-1}}^{t_{j'}} \theta(x) h_2(x)dx \right) \right\} \]

\[ = S_1(t_{j-1}) \exp \left( -\int_{t_{j-1}}^{t_j} \theta(x) h_2(x)dx \right), \quad \text{(A.2)} \]

Because \( f_{t_{j-1}} h_1(x)dx > 0 \), under the assumption that the probability of at least one event is greater than zero for each interval in each group, it follows from the Second Mean Value Theorem for Integrals (Moodie et al. 2004) that (A.2) can have alternative expression as (A.3)

\[ S_1(t_j) = S_1(t_{j-1}) \exp \left( -\tilde{\theta}_j \int_{t_{j-1}}^{t_j} h_2(x)dx \right). \quad \text{(A.3)} \]

For (A.3), \( \tilde{\theta}_j \) represents a weighted average hazard ratio for group 1 relative to group 2 in the \( j \)th interval in the sense shown in (A.4)

\[ \tilde{\theta}_j = \left\{ \frac{\int_{t_{j-1}}^{t_j} \theta(x) h_2(x)dx}{\int_{t_{j-1}}^{t_j} h_2(x)dx} \right\} / \left\{ \int_{t_{j-1}}^{t_j} h_2(x)dx \right\}. \quad \text{(A.4)} \]

Because \( \exp \left( -\int_{t_{j-1}}^{t_j} h_2(x)dx \right) = S_2(t_j)/S_2(t_{j-1}) \), it follows that (A.3) can have expression as (A.5)

\[ \{ S_1(t_j)/S_1(t_{j-1}) \} = \{ S_2(t_j)/S_2(t_{j-1}) \} \tilde{\theta}_j. \quad \text{(A.5)} \]

The quantities \( \pi_{j} = S_1(t_j)/S_1(t_{j-1}) \) represent conditional probabilities for survival of the \( j \)th interval given survival of the \( (j-1) \)th interval. From (A.5), it follows that \( \log \pi_{j} = \log \tilde{\theta}_j \log \pi_{j-1} \), and so the log average hazard ratio \( \log \tilde{\theta}_j \) for the \( j \)th interval can be shown as (A.6)

\[ \log \tilde{\theta}_j = \left\{ \log e (-\log e \pi_{j}) - \log e (-\log e \pi_{j-1}) \right\}. \quad \text{(A.6)} \]

Through (A.6), estimation of the \( \log \tilde{\theta}_j \) is possible from time to event data for group 1 and group 2 through the methods outlined in Appendix A.2. These methods also address the comparison between group 1 and group 2 in terms of the \( \{ \tilde{\theta}_j \} \).

As noted previously, \( \tilde{\theta}_j \) represents a weighted hazard ratio for the comparison between group 1 and group 2 in the \( j \)th interval with the weights being the hazards \( h_2(x) \) for the control group in the \( j \)th interval. Because such weights are larger where the hazards \( h_2(x) \) for the control group are higher, they arguably have clinical relevance for the comparison between the test and control groups for the extent to which they survive the event of interest across the respective time intervals; see Moodie et al. (2004).

### A.2 Construction of Hazard Ratio Estimators With Randomization Based Covariance Adjustment

Let \( i = 1, 2 \) index test and control treatments respectively. Let \( j = 1, 2, \ldots, J \) index a prespecified set of consecutive and adjacent time intervals \( (0, t_1], (t_1, t_2], \ldots, (t_{(j-1)}, t_j] \) within which the first occurrence of an event of interest could occur or not. Let \( k = 1, 2, \ldots, n_i \) index the patients in group \( i \). Let \( y_{ik} \) equal the duration of follow-up for patient \( k \) in group \( i \) and let \( u_{ik} = 0, 1 \) according to whether patient \( k \) in group \( i \) respectively survives or has the event of interest during the time for their follow-up. Let \( x_{ik} = (x_{ik1}, \ldots, x_{ikM})' \) denote the vector of \( M \) numeric baseline covariables for patient \( k \) in group \( i \), where any categorical covariable with \( L \) categories has expression as a set of \( (L - 1) \) indicator variables.

Patient \( k \) in group \( i \) is at risk for the occurrence of the event of interest in interval \( j \) if \( y_{ik} > t_{(j-1)} \) and their extent of risk in interval \( j \) corresponds to \( r_{ijk} \) in (A.7)

\[ r_{ijk} = \begin{cases} 1, & \text{if } (y_{ik} - t_{(j-1)})/(t_j - t_{(j-1)}) > 0 \text{ and } t_{(j-1)} < y_{ik} < t_j \text{ and } u_{ik} = 0 \\ 0, & \text{if } y_{ik} \leq t_{(j-1)} \end{cases} \quad \text{(A.7)} \]

If patient \( k \) in group \( i \) is at risk in interval \( j \), their extent of survival for interval \( j \) corresponds to \( s_{ijk} \) in (A.8)

\[ s_{ijk} = \begin{cases} r_{ijk}, & \text{if } u_{ik} = 0 \text{ or } y_{ik} > t_j \\ 0, & \text{if } u_{ik} = 1 \text{ and } y_{ik} \leq t_j \end{cases} \quad \text{(A.8)} \]

For both \( r_{ijk} \) and \( s_{ijk} \), there can be replacement of \( (y_{ik} - t_{(j-1)})/(t_j - t_{(j-1)}) \) by 0.5 when the specific value of \( y_{ik} \) in \( t_{(j-1)} < y_{ik} < t_j \) is unknown; the corresponding assumption is that such \( y_{ik} \) randomly have a uniform distribution within the \( j \)th interval.

Let \( s_{ik} = (s_{i1k}, \ldots, s_{iJk})' \), \( r_{ik} = (r_{i1k}, \ldots, r_{iJk})' \), and let \( f_{ik} = (s_{i1k}, r_{i1k}, x_{ik})' \). Under the assumption that the \( n_i \) patients in group \( i \) represent a very large population in a sense comparable to a simple random sample, then an unbiased estimator for the covariance matrix of the vector \( f_i = (\Sigma_{k=1}^{n_i} f_{ik}/n_i) \) of means of the \( f_{ik} \) for the \( i \)th group is \( V_{f_i} \) in (A.9)

\[ V_{f_i} = \sum_{k=1}^{n_i} \{ f_{ik} - \bar{f}_i \} (f_{ik} - \bar{f}_i)' / \{ n_i (n_i - 1) \}. \quad \text{(A.9)} \]

Let \( g_{ij} = \log e (-\log e (\tilde{s}_{ij}/\tilde{r}_{ij})) \) where \( \tilde{s}_{ij} = \{ \Sigma_{k=1}^{n_i} s_{ijk}/n_i \} \) and \( \tilde{r}_{ij} = \{ \Sigma_{k=1}^{n_i} r_{ijk}/n_i \} \) and let \( g_i = (g_{i1}, \ldots, g_{iJ})' \). It then follows that \( g_i \) has the structure in (A.10)

\[ g_i = \log e (\{ f_{ij} - L \} \log e (\tilde{s}_{ij}/\tilde{r}_{ij})). \quad \text{(A.10)} \]

where \( \tilde{s}_i = (\tilde{s}_{i1}, \ldots, \tilde{s}_{iJ})' \) and \( \tilde{r}_i = (\tilde{r}_{i1}, \ldots, \tilde{r}_{iJ})' \). Thus, on the basis of linear Taylor series methods as discussed by Stokes et al. (2000); Koch et al. (1972); and Koch et al.
(1977), a consistent estimator for the covariance matrix of \( F_I = (g'_1, \bar{x}(/^{1})) \) is \( V_F \) in (A.11).

\[
V_F = \begin{bmatrix}
H_I & 0_{J,M} \\
0_{M,J} & I_M
\end{bmatrix} V_{F1} \begin{bmatrix}
H_I & 0_{J,M} \\
0_{M,J} & I_M
\end{bmatrix}^{T},
\]

(A.11)

where \( H_I = D^{-1}_{\alpha}(g'_1I_I - I_D)D_{\alpha}(g'_1I_I)^T \), with \( D_{\alpha} \) denoting the diagonal matrix with \( \alpha \) as diagonal elements.

Let \( d = (F_1 - F_2) = (g'_1 - g'_2, \bar{x}(/^{1}) - \bar{x}(/^{2})) \) represent the vector of differences between groups 1 and 2 with respect to \( \{F_i\} \). A consistent estimator for the covariance matrix of \( d \) is \( V_d = (V_{F1} + V_{F2}) \). The first \( J \) elements of \( d \) are estimators of the log hazard ratios for the comparisons between groups 1 and 2 for the \( J \) time intervals, and the last \( M \) elements are estimators for the differences between groups 1 and 2 for means of the respective covariables.

When the initial sample sizes \( n_i \) and the number of patients with the event and the number of patients who complete the interval without the event in each interval in each group are sufficiently large (e.g., \( n_i > 80 \) and each interval of each group has at least 10 patients with the event and 10 patients who complete the interval without the event), then \( d \) approximately has a multivariate normal distribution with \( V_d \) as its essentially known covariance matrix; see Koch et al. (1995, 1977, 1985, 1998); and Koch and Bhapkar (1982).

As discussed by Koch et al. (1998), Tangen and Koch (1999b), and LaVange et al. (2005), randomization based covariance adjustment has invocation for \( d \) by fitting the linear model \( X = [I_J, 0_{M}] \) to \( d \) by weighted least squares. The resulting covariance adjusted estimator \( b \) for the interval-specific log hazard ratios is given by

\[
b = (X'V_d^{-1}X)^{-1}X'V_d^{-1}d
\]

(A.12)

where \( V_{d,gg} \) is the submatrix of \( V_d \) that corresponds to the covariances between \( (g'_1 - g'_2) \) and \( (\bar{x}(/^{1}) - \bar{x}(/^{2})) \) and \( V_{d,xx} \) is the submatrix of \( V_d \) that corresponds to the covariance matrix of \( (\bar{x}(/^{1}) - \bar{x}(/^{2})) \); that is, \( V_d \) has the structure shown in (A.13).

\[
V_d = \begin{bmatrix}
V_{d,gg} & V_{d,gx} \\
V_{d,gx} & V_{d,xx}
\end{bmatrix}
\]

(A.13)

A consistent estimator for the covariance matrix of \( b \) is \( V_b \) in (A.14).

\[
V_b = (X'V_d^{-1}X)^{-1} = (V_{d,gg} - V_{d,gx}V_{d,xx}^{-1}V_{d,gx})
\]

(A.14)

The rationale for randomization based covariance adjustment is the expected absence of differences between test and control groups for the means \( \bar{x} \) of the covariables. A related criterion for evaluating the extent of random imbalances between test and control groups for the \( \bar{x} \) is \( Q_0 \) in (A.15).

\[
Q_0 = (d - Xb)V_d^{-1}(d - Xb)
\]

(A.15)

\[
= (\bar{x}(/^{1}) - \bar{x}(/^{2}))V_{d,xx}^{-1}(\bar{x}(/^{1}) - \bar{x}(/^{2}))
\]

This criterion approximately has the chi-squared distribution with \( M \) degrees of freedom.

Acknowledgments

We would like to acknowledge the reviewers for their insightful comments and suggestions for improving the manuscript.

[Received February 2010. Revised April 2010.]

REFERENCES


9


About the Authors

Patricia F. Moodie is Research Scholar, Department of Mathematics and Statistics, University of Winnipeg, 515 Portage Avenue, Winnipeg MB R3B 2E9, Canada. Benjamin R. Saville is Assistant Professor of Biostatistics, Department of Biostatistics, Vanderbilt University School of Medicine, S-2323 Medical Center North, 1161 21st Avenue South, Nashville, TN 37232-2158. Gary G. Koch is Professor of Biostatistics, Department of Biostatistics, University of North Carolina at Chapel Hill, CB #7420, Chapel Hill, NC 27599. Catherine M. Tangen is Member, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, M3-C102, Seattle, WA 98109-1024. E-mail for correspondence: b.saville@vanderbilt.edu.