Multivariate Permutation Test to Compare Survival Curves for Matched Data

Stefania Galimberti

Abstract In the absence of randomization, the comparison of an experimental treatment with respect to the standard may be done based on a matched design. When there is a limited set of cases receiving the experimental treatment, matching of a proper set of controls in a non fixed proportion is convenient. In order to deal with the highly stratified survival data generated by multiple matching, we have extended to this setting the multivariate permutation testing approach, since standard non parametric methods for the comparison of survival curves cannot be applied. We have demonstrated the validity of the proposed method with simulations, and we will illustrate its application to data from an observational study for the comparison of bone marrow transplantation and chemotherapy in the treatment of paediatric leukaemia.

1 Introduction

In clinical trials with right-censored failure time outcome, inference often focuses on the comparison of survival curves. When data from observational studies are used to explore the role of different treatments, the main problem is to limit the biases due to the lack of randomization. Matching on relevant baseline features can be used in order to increase the comparability between subjects treated with an experimental therapy (cases) and those receiving standard treatment (controls). In settings where matching is done with a variable number of controls (multiple matching), highly stratified data are produced, with strata containing a few, possibly censored, observations. For this reason, the statistical comparison of survival in the two groups cannot be directly addressed by means of the usual non parametric procedures, such as the log-rank test. The stratified version of these tests, which should account for matching, is inefficient when the number of strata increases and the stratum size is small [7]. Furthermore,

[1] Stefania Galimberti, Center of Biostatistics for Clinical Epidemiology, Department of Clinical Medicine and Prevention, University of Milano-Bicocca; stefania.galimberti@unimib.it

these methods are less sensitive when proportional hazard is not satisfied and this might often be the case, especially in the clinical setting that motivated this work, i.e. the comparison of bone marrow transplantation and chemotherapy in the treatment of leukemia [1, 4].

The comparison of the survival curves for highly stratified data due to non-paired matching was addressed by Galimberti et al. [1], who proposed a weighted Kaplan-Meier estimator for the curve of controls and a non parametric permutation test on the survival difference at one pre-fixed time point. The aim of this work is to extend the comparison of survival at multiple time points, using the multivariate permutation approach originally introduced by Pesarin [5]. In the proposed procedure, differently from Pesarin and Salmaso [6], there is no need to resort to the missing data framework in order to deal with censoring.

2 Methods

The multidimensional permutation approach introduced by Pesarin is here adapted to the context of survival analysis with matched data for comparing the marginal survival of cases (S_1) and controls (S_2). A non parametric combination of a finite number of dependent permutation tests performed at different time points is proposed here.

The standard hypotheses on the survival distributions are rephrased into a certain number q of sub-hypotheses H_{0i}, H_{1i}, (i=1,...,q):

$$H_0: \; S_1(t) = S_2(t) \text{ becomes } H_0: \bigcap_{i=1}^{q} H_{0i} \text{ with } H_{0i}: \; S_1(t_i) = S_2(t_i)$$

$$H_1: \; S_1(t) \neq S_2(t) \text{ becomes } H_1: \bigcup_{i=1}^{q} H_{1i} \text{ with } H_{1i}: \; S_1(t_i) \neq S_2(t_i)$$

We considered two versions of the partial univariate test statistics T_i'. They are based on the distance of the two survival curves or on their complementary log-log transformation [2] estimated at q different times points $(t_1,..,t_i,..,t_q)$:

$$_{S}T_i' = \hat{S}_1(t_i) - \hat{S}_2^{w}(t_i)$$

$$_{\log S}T_i' = \log\{-\log[\hat{S}_1(t_i)]\} - \log\{-\log[\hat{S}_2^{w}(t_i)]\}$$

where $\hat{S}_1(.)$ is the usual Kaplan-Meier estimate of the survival distribution for cases, while $\hat{S}_2^{w}(.)$ is a weighted version [1] for the matched controls.

If J is the number of cases (experimental group), then the control group size is of $\sum_{j=1}^{J} m_j$, where m_j is the number of matched controls for case j. In this group, the weighted Kaplan-Meier estimator is used, with stratum specific weights to account for the variable number of subjects in each stratum:

$$\hat{S}_2^{w}(t) = \prod_{u:u\le t}\left(1 - \frac{\sum_{j=1}^{J} w_j d_j(u)}{\sum_{j=1}^{J} w_j r_j(u)}\right)$$

where $d_j(u)$ and $r_j(u)$ are the number of events and the number of subjects at risk at time u in stratum j, respectively, while $w_j=1/m_j$ is the weight applied to the m_j controls in stratum j.

The q partial tests are performed by considering the permutation distribution of the estimated distances ${}_S T_i^{'}$ and ${}_{logS} T_i^{'}$. (i=1,...,q). These distributions are approximated following a strategy that considers B random samples from the population of all possible permutations. Each permutation sample is obtained by assigning, within each observed stratum of (m_j+1) subjects, one subject to the experimental treatment and the remaining m_j to the standard treatment.

Once the q p-values $\hat{\lambda}_i$ are obtained for each first order test, they are combined in a unidimensional second-order test statistic $T^{''}$, in order to summarize the whole information obtained in the partial tests. Two alternatives are considered:

$$T_F^{''} = -2\sum_{i=1}^{q} \log(\hat{\lambda}_i) \qquad \text{Fisher transform}$$

$$T_T^{''} = \max_{1\le i\le q} (1-\hat{\lambda}_i) \qquad \text{Tippett transform}$$

Different ways of defining the time points (t_1,.., t_i,.., t_q) involved in the comparison of the two survival functions are explored: i) fixed time points, ii) time points identified by percentiles of the overall event distribution and iii) all the observed event times included between the 10^{th} and the 90^{th} percentiles of the overall event distribution.

3 Simulations protocol and results

The performance of these tests are evaluated through simulations. The size and the power are calculated for different alternatives under proportional hazards and for three different scenarios of non-proportionality (e.g. survival curves showing both an early and a late difference or crossing each other). The behavior of the proposed tests is also explored in the presence of different number of strata (k=30, 50, 100) and of different degree of strata heterogeneity (no or low strata effect). Censoring was uniform over 3-6 in order to have approximately 28-38 percent of censoring, on average.

For all the configurations studied, the results are based on 1000 samples and each of them uses B=2000 Monte Carlo permutations in order to define the empirical permutation distributions. Definitions of the q time points involved in the comparison are as follows: i) 8 or 4 fixed equi-spaced times from 0.5-1 to 4, ii) 9 time points identified by the 10^{th}, 20^{th},.., 90^{th} percentiles of the overall event distribution, and iii) all the observed failure times between the 10^{th} and the 90^{th} percentiles of the overall event distribution.

For comparative purposes, the ordinary stratified log-rank test, the modified log-rank test for highly stratified data of Schoenfeld and Tsiatis [7] and the Cox model with a sandwich robust standard error for the treatment effect [3], were applied to the simulated data.

The extended simulation study shows good performances of the proposed test in terms of alpha coverage. Under the alternative, as expected, power increases with increasing number of strata under every scenario and the test has a good behavior also in the situation of crossing hazards. No major gain in power is achieved by increasing

the number of fixed time points from 4 to 8 and similar results are obtained when points are fixed at percentiles of the event distribution or take 80% of all observed event times. The fixed time point strategy seems generally more satisfactory for the proportional hazards and the late difference scenarios. A clear advantage in the use of the complementary log-log transform is seen only in the early difference scenario; conversely the survival difference seems to offer an advantage in the crossing hazards setting. The performance under the Fisher and Tippet functions is very similar, except for an advantage of Tippett in the crossing hazards scenario.

The ordinary stratified log-rank test has inflated size, while the modified log-rank test by Schoenfeld and Tsiatis and the robust test from the Cox model including only the covariate for treatment has size close to the nominal level. The multivariate permutation tests perform better than the two latter tests in terms of power in the crossing hazards and early difference scenarios, while they have a similar behavior in the presence of proportional hazards.

4 Conclusions

A valid global test for the comparison of survival curves in the context of highly stratified data produced by matching on a non-fixed proportion is provided, by extending the multivariate permutation approach proposed by Pesarin. The proposed procedure is proven to be at least equal or superior to the modified log-rank test for highly stratified data and to the robust Cox model. Preference to this test should be definitely given when the proportional hazards assumption does not hold, even in non matched data.

References

1. Galimberti, S., Sasieni, P., Valsecchi, M.G.: A weighted Kaplan-Meier estimator for matched data with application to the comparison of chemotherapy and bone-marrow transplant in leukemia. Statistics in Medicine 21:3847-3864 (2002) doi: 10.1002/sim.1357
2. Klein, J.P., Logan, B., Harhoff, M., Andersen, P.K.: Analyzing survival curves at a fixed point in time. *Statistics in Medicine* 26:4505-4519 (2007). doi: 10.1002/sim.2864
3. Lee, E.W., Wei, L.J., Amato, D.A.: Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In: Survival analysis: state of the art. Dordrecht: Kluwer Academic pp 237-247 (1992)
4. Logan, B.R., Klein, J.P., Zhang, M.J.: Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. Biometrics 64:733-740 (2008) doi: 10.1111/j.1541-0420.2007.00975.x
5. Pesarin, F.: Multivariate Permutation Tests: With Applications in Biostatistics. Wiley: Chichester pp 133-179 (2001)
6. Pesarin, F., Salmaso, L.: Permutation Tests for Complex Data: Theory, Applications and Software. Wiley: Chichester pp 295-300 (2010)
7. Schoenfeld, D.A., Tsiatis, A.A.: A modified log-rank test for highly stratified data. *Biometrika*, 74:167-175 (1987)