Methods for Accommodating Nonproportional Hazards in Oncology: A Compromise Solution

Dominic Magirr (Novartis), Carl-Fredrik Burman (AstraZeneca)

5/7/2020

The common way to analyze two-arm randomized controlled trials with a time-to-event endpoint is using a log-rank test. This test is designed to work best when the new treatment has an effect that is approximately constant over time. Researchers working in immuno-oncology, however, frequently have the strong belief that the treatment effect in their studies will be delayed. They anticipate survival curves to be similar for a number of months, before diverging in favour of the new treatment. Alternative analysis methods more tailored to this situation have been widely studied, but not yet used for the primary analysis of confirmatory trials. Whether they ought to be is controversial, with conflicting views appearing in a recent discussion in this journal. Freidlin & Korn (2019) have stressed the need for caution in moving away from a tried and trusted method, while Uno et al. (2020) believe we should not be wedded to the log-rank test merely out of tradition. Huang et al. (2020) argue that it is the Kaplan-Meier estimate that is fundamental in assessing treatment effect size, not the hazard ratio. We see merits in the reasoning of all sets of authors, and believe there is a compromise way forward.

The arguments of Freidlin & Korn are based on two examples that we have approximately reproduced in Figure 1.

In example A, the log-rank test gives a one-sided p-value (in favour of the test drug) of p = 0.59, i.e. non-significant. If, instead, one had pre-specified a particular 'late-emphasis' test (Harrington and Fleming 1982), often denoted $G^{(0,1)}$, the p-value would have been less than 0.0001. Freidlin & Korn make the point that this apparent increase in power comes at a price. If one were to apply the $G^{(0,1)}$ test to the data in example B, one would find p < 0.001 *in favour of the test drug.* This behaviour is indeed troubling. One should be wary of using the $G^{(0,1)}$ test for regulatory decision making.

Nevertheless, one should also acknowledge that the $G^{(0,1)}$ test is somewhat extreme. It is a weighted log-rank test that gives a weight of exactly 0 to the first event, with weights increasing towards 1. This means that the *relative* weight of early events is close to 0%. A more cautious approach would give the first event a weight of 1, with weights increasing towards 2. Then the relative weight of early events would be no less than 50%. Such a "modestly-weighted logrank test" (MWLRT) was studied by Magirr & Burman (2019). It has high power when the treatment effect is delayed, as well as retaining good power when the treatment effect is truly constant over time. In example A, the MWLRT (with a maximum weight of 2) would give a one-sided p-value of 0.025; in example B, p = 0.94.

The MWLRT has the key property that if survival on the test drug is truly lower (or equal) to survival on control at all timepoints, then the probability of claiming a statistically significant improvement is less than 2.5% (assuming a conventional threshold is applied). This property is shared by the log-rank test, but not by the $G^{(0,1)}$ test. One could argue, therefore, that the log-rank test and MWLRT are both valid "gatekeeper" tests, a gatekeeper test being a minimum regulatory hurdle set by society as a blunt tool to prevent an abundance of false positive findings. Clearing such a hurdle is just the first step, with interpretation of treatment effect size coming from study of the Kaplan-Meier curves.

Now consider two groups of researchers, both developing an experimental drug for the same indication, where both drugs have a similar magnitude of benefit overall. Group A expects their drug to produce a constant effect over time. Group B expects their drug to improve long-term survival but it may not show any increase in survival for a number of months. For group A, the log-rank test hurdle is easier to clear than the MWLRT hurdle, and vice-versa for group B. By forcing both groups to use the log-rank test hurdle, one is disadvantaging group B, relative to group A. If designing a regulatory system from scratch, this is probably not something one would wish to do, especially considering long-term survival is often (though not always)

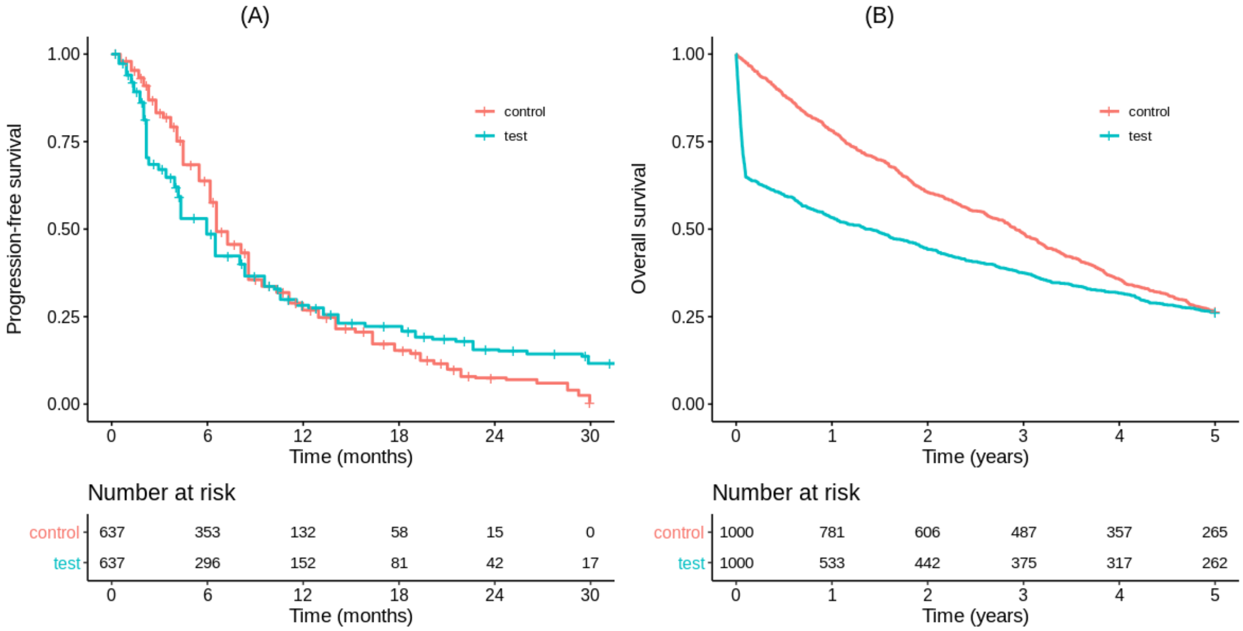

Figure 1: Figure 1. (A) Data reconstructed from Figure 1 of Freidlin & Korn (2019) using methods described in Guyot et al. (2012). (B) Data constructed according to the algorithm described in Figure A1 of Freidlin & Korn (2019)

more highly valued than short-term survival.

A final point to emphasize is that when a delayed effect is anticipated the difference in power between the log-rank test and MWLRT is not trivial. Freidlin & Korn discuss an increase in sample size of 10% coming from using the log-rank test instead of a weighted version. This inflation factor will depend on many parameters, but 20% is also realistic (see Magirr & Burman). Converted into time and cost savings, and multiplied by the number of studies, this would have major implications.

References

Freidlin, Boris, and Edward L Korn. 2019. "Methods for Accommodating Nonproportional Hazards in Clinical Trials: Ready for the Primary Analysis?" *Journal of Clinical Oncology* 37 (35). American Society of Clinical Oncology: 3455.

Guyot, Patricia, AE Ades, Mario JNM Ouwens, and Nicky J Welton. 2012. "Enhanced Secondary Analysis of Survival Data: Reconstructing the Data from Published Kaplan-Meier Survival Curves." *BMC Medical Research Methodology* 12 (1). Springer: 9.

Harrington, David P, and Thomas R Fleming. 1982. "A Class of Rank Test Procedures for Censored Survival Data." *Biometrika* 69 (3). Oxford University Press: 553–66.

Huang, Bo, Lee-Jen Wei, and Ethan B Ludmir. 2020. "Estimating Treatment Effect as the Primary Analysis in a Comparative Study: Moving Beyond P Value." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, JCO1903111.

Magirr, Dominic, and Carl-Fredrik Burman. 2019. "Modestly Weighted Logrank Tests." *Statistics in Medicine* 38 (20). Wiley Online Library: 3782–90.

Uno, Hajime, and Lu Tian. 2020. "Is the Log-Rank and Hazard Ratio Test/Estimation the Best Approach for Primary Analysis for All Trials?" *Journal of Clinical Oncology: Official Journal of the American Society* *of Clinical Oncology*, JCO1903097.