The hazards of period specific and weighted hazard ratios

How to adequately handle non-proportional hazards (NPH) in clinical trials is an important and timely question, particularly given recent advances in immuno-oncology treatments, in which survival curves may only separate after some delay or even cross each other. As Lin *et al* 2019 note, quantification of treatment effects is of crucial importance in addition to hypothesis testing. Under NPH, Lin et al suggest the use of piecewise hazard ratios (HRs), which they state can be used to describe the change in treatment effect over time, or alternatively a weighted HR corresponding to the 'winning' weighted log-rank test in the 'MaxCombo' test (Lin and León 2017).

We are concerned with the suggestion that changes in period specific HRs can be interpreted as changes in treatment effect. Our concern stems from the fact that period specific HRs are subject to selection bias, even in randomised trials (Aalen et al 2008, Hernán 2010, Aalen et al 2015, Martinussen et al 2018, Stensrud et al 2019)). This bias occurs due to the ubiquitous existence of unknown factors (so called frailty factors) which affect patients' time to event. Randomisation guarantees that the distribution of such factors in the trial at baseline is (on average) identical across treatment groups. Even so, at later times during follow-up, the survivors in the two treatment groups will generally have systematically different distributions of these factors. A period-specific HR contrasts the event rates in the next interval of time between these two groups of survivors, and hence reflects not only the effects of treatment but also the effects of the frailty factors, whose distribution may increasingly differ between the two groups of survivors.

As a simple example, consider testing a new treatment (A) vs standard-of-care (B) after surgery to remove cancer (adjuvant setting). Assume that surgery cures half of the patients and for the purposes of illustration the only factors which influence time to relapse are cure status and randomised treatment. Here cure status represents the unmeasured frailty factor. Suppose that, among non-cured patients, the HR (A vs B) for time to relapse is 0.5 at all times. Thus, A has a continuing benefit for the non-cured half of the population, but no effect in the cured half. Due to randomisation, each group of survivors initially contains 50% cured patients and the period-specific HR in an intention-to-treat analysis is expected to be 0.5 in the first part of follow-up. In subsequent periods of follow-up, since A reduces the hazard (relative to B) in non-cured patients, the proportion of cured patients in group B survivors becomes higher than the corresponding proportion in group A survivors, such that the period-specific HR may rise above 1, apparently (wrongly) suggesting that A is harmful relative to B in the later part of follow-up. This can happen even in situations where treatment A is beneficial for all (Aalen et al 2008, Aalen et al 2015, Martinussen et al., 2018). Likewise, a weighted HR may be above 1 even if survival on A is uniformly better than on B (Magirr and Burman 2019).

Importantly, these concerns about the lack of comparability between the two groups of survivors apply even when the (unconditional) HR comparing treatment groups is constant over time, raising doubts about the interpretability of the HR as a measure of treatment effect even when the proportional hazards assumption holds (Stensrud et al 2019). We believe the frailty factors also render weighted HRs (Lin and León 2017) difficult to interpret causally as effects solely attributable to treatment. These issues did not affect the simulations reported by Lin *et al* (2019) because their simulation studies generated datasets in which each patient's

hazard depended solely on treatment. Lastly we note that our critique is in line with the recent ICH E9 (2019) estimand addendum (A.3.2., principal stratum strategies), which explains that treatment effects defined by comparing outcomes in a subset defined by occurrence of a post-baseline event leads to confounding/selection bias. Period specific HRs are an example of this, since they consider only those patients who survive under their assigned treatment to the start of the period in question.

In line with the recent epidemiological and causal inference literature, we would discourage use of HRs for quantifying the effects of treatments or exposures (Hernán and Robins 2020). Instead, we recommend the use of better interpretable estimands, such as contrasts of survival probabilities at specified times, contrasts of specified quantiles (e.g. median survival), or differences/ratios of restricted mean survival time.

References

Aalen, O., Borgan, O. and Gjessing, H., 2008. "Chapter 6 Unobserved heterogeneity: The odd effects of frailty" in *Survival and event history analysis: a process point of view*, New York: Springer Science & Business Media, pp. 231- 270.

Aalen, O.O., Cook, R.J. and Røysland, K. (2015), "Does Cox analysis of a randomized survival study yield a causal treatment effect?," *Lifetime Data Analysis*, 21, 579-593.

Hernán, M.A. (2010), "The hazards of hazard ratios," *Epidemiology*, 21 13-15.

Hernán M.A., Robins J.M. (2020), "Causal Inference: What If," Boca Raton: Chapman & Hall/CRC

ICH E9 (R1) (2019), "ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials," available at https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf

Lin, R.S. and León, L.F. (2017), "Estimation of treatment effects in weighted log-rank tests," *Contemporary Clinical Trials Communications*, 8, 147-155.

Lin, R.S., Lin, J., Roychoudhury, S., Anderson, K.M., Hu, T., Huang, B., Leon, L.F., Liao, J.J., Liu, R., Luo, X. and Mukhopadhyay, P. (2019), "Alternative Analysis Methods for Time to Event Endpoints under Non-proportional Hazards: A Comparative Analysis," *Statistics in Biopharmaceutical Research*, epub

Magirr, D., Burman, C.F. (2019), "Modestly weighted logrank tests," *Statistics in Medicine* 38, 3782-3790.

Martinussen, T., Vansteelandt, S. and Andersen, P.K. (2018), "Subtleties in the interpretation of hazard ratios," *arXiv* 1810.09192.

Stensrud, M.J., Aalen, J.M., Aalen, O.O. and Valberg, M. (2019), "Limitations of hazard ratios in clinical trials. *European Heart Journal*," 40, 1378-1383.