"We are a community dedicated to leading and promoting the use of statistics within the healthcare industry for the benefit of patients."

Transcript of Questions from the recent Journal Club Webinar: Survival Analysis, 24/03/21

1. **What happens in the simulations if we choose different parameters in the Fleming-Harrington test, for instance 0.5 and 0.5 or 0.33 and 0.33 or 0.33 and 0.67?**
   The F-H weight functions are of the form S(t)^rho * (1 - S(t)) ^ gamma. For any gamma > 0, this means that early events will get weight very close to zero, and this opens the door to the issue I highlighted: it would be possible to construct examples where the experimental treatment is uniformly worse than control, but the expected value of the test statistic goes in the wrong direction.

2. **Are you implicitly proposing separating testing from estimation and confidence interval? Isn't that challenging for those interpreting the results?**
   Thank you, this is a very good question. My answer is yes, I do think in this specific context it is a good idea to separate testing and estimation. In my opinion, the estimand that is of true scientific interest is the whole survival curves on the two arms (even this may be too simple in a heterogeneous population). There is no single-number summary measure (that can be reliably estimated) that can adequately capture this highly multi-dimensional information (I think all authors on this topic agree about this). Of course, in every study report and scientific publication the K-M curves (or other estimates of survival curves) have a very prominent role. When it comes to designing a study and (at least a starting point for) decision making, we do need something more than this. Personally, I am happy for that something to be a test. If there were a really good summary measure available, I would immediately use it (with confidence interval) and de-prioritise the test. But this isn't the case. As it is, I prefer to have a somewhat messy and holistic approach that is capable of capturing the treatment effect, rather than a neat and tidy package that only reflects one slice of things. I would also say that this separation of testing and estimation is what already happens at the moment. Even though a number emerges from the Cox model, under highly non-proportional hazards it no longer has its interpretation as S_1(t) = S_0(t) ^ HR.

3. **Thanks - great presentation; I was wondering how detailed the weights are design in your experience. Are you playing around with parametric models and then optimize the weights given a prior or is it more heuristic?**
   Thank you. In our Stats in Medicine paper (Magirr & Burman, 2019) we do suggest it's possible to optimize the choice of x^* based on power under a set of assumptions. However, I also think one attraction of the modestly-weighted test is its robustness to model misspecification (not just the event distribution but also recruitment), so personally I quite like the heurisitic approach, thinking about it as similar to an average landmark analysis from x^* to the end of follow-up.

4. **In terms of having an interpretable summary, could one represent the risk-set and # events over time (K-M type) and use the weight to interpret the difference in the groups?**
   Thank you, that's a good point. I didn't have so much time to discuss alternative approaches, but comparing weighted K-M estimates has been proposed, and I think it's an attractive option. I would say the modestly-weighted test is somewhat similar, as I alluded to with the "average landmark" analogy, but perhaps has a slight advantage in that it is very easy to compute and one doesn't need to pre-specify the end of follow-up (or worry about stability in the tail of the distribution).

5. **Really great talk and the correspondence between the weighted tests and the score tests are really interesting. I just wondered why the FH(1,1) test and not the FH(0,1) test was chosen and if this has the same disadvantages wrt. the scores?**
   Thank you. I chose the FH(1,1) for this talk to relate it to the pembrolizumab example where they actually used this approach, but the same things happen for the FH(0,1) test, which is what I discussed in the paper.

6. **My question to Dominic, thank you for the nice presentation. As we all know the most popular way of analysing is the Cox PH and hazard ratio, and p-value from logrank is reported. There was i time i used post proportional hazard methods because the assumption of the PH was not met. However, all the reviewers were expected to see HR reported. Do you think the perception of the reviewers need change and be open to other methods like yours? What is your experience on this?**
   Thank you. Yes, I think you're right, and I would like to see this change. I sense a certain unwillingness to look beyond standard methods even from statisticians in companies that are developing immunotherapy drugs. I think this is understandable to some extent. It is a difficult topic, and while the standard analysis is inefficient when we can anticipate the delay, it will rarely be truly dreadful. But this is a low bar. If we do hundreds of trials that are 10% larger (say) than they need to be, then this adds up to a lot of waste.

7. **What's the regulatory perspective of these weighted log-rank tests that basically downweigh the contribution of events from participants who had the worst outcomes in the trial?**
   Thank you for this very important question. I think this probably requires a long discussion, but there are a few points that I'd like to make. (1) A specific point about the modestly-weighted test. As I explained in my presentation, this does not downweight early time periods any more than a landmark analysis, and therefore for any regulator who is willing to accept a test based on a landmark analysis, it would be inconistent to rule out a modestly-weighted test on this basis. (2) A more general point about the word "weighted". Regulators are smart people who weigh things up all the time. That's their job. I know that might sound flippant, but it's important to

point out that this shouldn't be a dirty word. (3) Regarding the standard log-rank test, it's important to emphasize that this does not assign equal weight to each patient; it assigns equal weight to each event. This means that every potential censoring distribution corresponds to a different weighting of the time periods. It's not that a standard logrank test is unweighted, rather, the weighting is done for you without you noticing. Maybe this weighting will be reasonable, maybe it won't. (4) More generally with right-censored survival data: every analysis is a weighted analysis. I've seen senior regulators write something along the lines of they are only interested in differences in mean survival. That's fair enough, but you can't have it. It's not possible to estimate reliably. If you take the restricted mean survival time instead, then this is a re-weighting of what you are truly interested in. We don't tend to think about it this way, and it's rare to hear a complaint about this similar to a complaint about a weighted logrank test. (5) Finally, even if hypothetically we could measure everybody's survival time and therefore the arithmetic mean on each arm: should we in that case never consider a weighted analysis targeting this estimand? I would say in certain circumstances we should consider weighting. A good regulator is a regulator who makes good decisions. If there is reliable external information that can be used to improve decision making, then why not use it?

8. **Question to Chang Yu: Even for IO there is in general an early onset of effect for an individual patient (e.g. early tumour shrinkage). I.e. the observed late effect could be caused by an overlay of different hazards across subgroups (responders and non-responders). Could you please elaborate on the advantages of the average HR compared to try to provide a HR for the subgroup of "responders" and subgroup of "non-responders" over control?**
   Thank you for the question. In this setting, estimation is not nearly as satisfactory. The full information is probabily provided by the KM curves for the two groups. Given that, the AHR provides an overall summary of the survival experience between the two groups. Yet, it opens to critiques: 1) it is difficult to interpret; 2) our simulaitons and calculation of its theoretical value show its value depends on the chosen weight. Your comment on HR for the subgroups of "responders" and "nonresponders" are interesting. I refer you to the recent work presented at ENAR 2021 by Dr. Zhenzhen Xu on non-responder effect on deviation from PH.

9. **Question for Chang: What's the rationale for the different weights you used?**
   We started with an intuitive weight, i.e the first weight, in the framework of weighted log-rank test. Then we found that the "optimal" weight was actually mentioned in Scheonfeld (1981) and Xu et al. (2017), i.e. the weight proporitonal to the log of the HR. Then we modified the weight toward the "optimal" weight. It turned out the two weights had similar performance in simulations.

10. **How and when are they different in applications?**
    Please see above. We have not had many applications yet. The paper just came out. Our simulations suggest the two weights would offer similar performance.

11. **In a 2-stage design, can you change the weight for the 2nd stage analysis, after seeing the 1st stage results?**
    Dominic: I think there could be a way to make this possible, using for example a conditional error function. I'm not sure how much value it would, it may be worth exploring.
    CY: This is an interesting idea and actually we are exploring it. Such adaptive designs need to be assessed carefully to ensure no type I error inflation.

12. **Can you show that these weights lead to consistent estimators?**
    Using these weights, we can get an approximate estimate of the average hazard ratio. The estimator is fairly "consistent" with its theoretical value. Here consistent does not have the usual statistical definition of consistency.