STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2020, VOL. 12, NO. 4, 520–521
https://doi.org/10.1080/19466315.2020.1825522

Taylor & Francis
Taylor & Francis Group

Rejoinder to Letter to the Editor "The Hazards of Period Specific and Weighted Hazard Ratios"

Ray S. Lin[a], Ji Lin[b], Satrajit Roychoudhury[c], Keaven M. Anderson[d], Tianle Hu[e], Bo Huang[f], Larry F. Leon[g], Jason J. Z. Liao[d], Rong Liu[h], Xiaodong Luo[i], Pralay Mukhopadhyay[j], Rui Qin[k], Kay Tatsuoka[l], Xuejing Wang[m], Yang Wang[n], Jian Zhu[o], Tai-Tsang Chen[p], Renee Iacona[j], and Cross-Pharma Non-Proportional Hazards Working Group*

[a]Genentech/Roche, South San Francisco, CA; [b]Sanofi US, Cambridge, MA; [c]Pfizer Inc., New York, NY; [d]Merck & Co., Inc., North Wales, PA; [e]Sarepta Therapeutics, Cambridge, MA; [f]Pfizer Inc., Groton, CT; [g]Celgene Co., San Diego, CA; [h]Celgene Co., Summit, NJ; [i]Sanofi US, Bridgewater, NJ; [j]Astra Zeneca, Washington, DC; [k]Janssen Research & Development, LLC, Raritan, NJ; [l]Bristol-Myers Squibb, Lawrenceville, NJ; [m]Eli Lilly and Company, Indianapolis, IN; [n]Teclison Ltd., Montclair, NJ; [o]Servier Pharmaceuticals, Boston, MA; [p]GSK, Collegeville, PA

We would like to thank the authors of the letter (Bartlett et al. 2020) for sharing their concerns regarding reporting treatment effect under nonproportional hazards (NPH), and we respect their position regarding the limitations of hazard ratio (HR) as a summary measure of treatment effect. We acknowledge that each effect measure has its values and limitations. As we highlighted in our article (Lin et al. 2020), a single effect measure is unable to comprehensively describe the magnitude of the treatment effect over time because the effect changes over time under NPH.

Various NPH patterns with a mixture of delayed effect, long-term survival or crossing effect have been observed in recent studies (e.g., immuno-oncology trials). The alternative primary analysis strategy under NPH was proposed to prospectively address the challenges of unknown NPH patterns at the stage of study design (Roychoudhury et al. 2019; Lin et al. 2020). We recommended a stepwise approach which separated the confirmatory testing from the reporting of treatment effect. The confirmatory testing step included the use of a robust MaxCombo test and was intended to establish the initial statistical difference between the two arms. Once such difference is established, further evaluation of the treatment effect is needed. To enable a comprehensive evaluation of the effect based on the totality of the data, we recommended multiple measures to be pre-specified in the analysis plan, including piecewise HRs, milestone survival probabilities (Chen 2015) and the restricted mean survival time (RMST) difference or ratio (Huang, Wei, and Ludmir 2020; Tian et al. 2020). We agree that the differences of specified quantiles may also provide additional values.

Each of these measures provides its unique values to data interpretation yet also has its own limitations, especially when its assumptions are violated. For example, piecewise HRs may be useful to describe the delayed treatment effect whereas the RMST difference or ratio may be more appropriate to describe the treatment effect in the cure fraction example illustrated in the letter. However, without prior knowledge of the NPH patterns at the design stage, we recommended that these multiple measures be prespecified in the analysis plan because this spectrum of measures can collectively enable a more comprehensive assessment of the treatment effect.

HRs have been commonly reported as a measure for the magnitude of treatment effect and well accepted for its established mathematical connection with the rank-based tests. However, we recognize its limitations discussed in the letter and agree that in the presence of unknown disease modifiers, HRs (both piecewise and average HRs) face the challenge of interpretability in the causal inference framework. Note that in a randomized controlled trial, if the proportional hazards assumption holds and there are no unobserved confounders (i.e., the patient population is homogeneous and there is no differential treatment effect across different subpopulations), a HR generated by Cox regression with treatment alone as a single covariate does have a causal interpretation. The hazard functions in this case do not depend on any confounder. If the delayed effect is expected (e.g., in immuno-oncology) and the hazard functions depend solely on treatment (i.e., homogeneous patient population), piecewise HRs could be a useful summary measure. On the other hand, we acknowledge that the patient population may be heterogeneous and there may exist known or unknown predictive factors (which predicts different levels of treatment benefit among subpopulations) or prognostic factors (which affects survival regardless of treatment assignment). Therefore, we have emphasized the use of multiple measures to establish the totality of evidence given the unknown nature and the challenge of prespecifying the confounders at the design stage.

We recommend minimizing and mitigating such heterogeneity through study design and prespecified analysis based on prior knowledge. If strong predictive factors (such as biomarkers) are expected, the study should leverage these factors to identify a population that could potentially benefit from the

CONTACT Ray S. Lin *lin.ray@gene.com* Genentech/Roche, South San Francisco, CA 94080.
*The Cross-Pharma NPH Working Group includes all the authors of this article as listed above and the following members who have contributed tremendously to this work: Prabhu Bhagavatheeswaran. Julie Cong, Margarida Geraldes, Dominik Heinzmann, Yifan Huang, Zhengrong Li, Honglu Liu, Yabing Mai, Jane Qian, Li-an Xu, Jiabu Ye, Luping Zhao.

treatment and enroll this relatively homogeneous population into the study. If strong prognostic factors are expected, randomization should be stratified by these factors and stratified analysis should be conducted accordingly (Kong and Slud 1997). For potential predictive factors, subgroup analysis can be conducted to evaluate the benefit in each subpopulation. For potential prognostic factors, exploratory analysis can be performed to identify these factors and the adjusted analysis (e.g., stratified Cox model, inverse probability weighting (Cole and Hernan 2004)) can be conducted to reduce the bias. Even though unmeasured predictive or prognostic factors may still exist, these steps will help minimize the heterogeneity and reduce the bias.

We would also like to clarify that the strategies described in Section A.3.2 in the ICH E9 (2019) estimand addendum is designed for handling post-baseline *intercurrent events*, which are defined as the events that affect either the interpretation or the observation of the endpoint. In contrast, the death or disease-progression events in our context *are* the endpoint itself and are not post-baseline intercurrent events. It is also our opinion that the principle stratum strategies and our proposed approach attempt to address different clinical questions. The principle stratum strategies were developed to eliminate or reduce the confounding and selection bias caused by the comparison of two or more noncomparable subsets identified in different treatment arms. On the other hand, the primary focus of our approach is to estimate the time-varying treatment effect by presenting a sequence of conditional piecewise treatment effects in a longitudinal manner. Therefore all the pieces in the piecewise HR sequence need to be reviewed and interpreted collectively (as well as the overall HR estimated by the Cox model) in order to adequately represent the totality of the data. We would also like to emphasize that each individual piecewise HR should not be interpreted alone without proper context of the other piecewise HRs.

It is evident that reporting the treatment effect based on a single measure is inadequate in the presence of NPH, and different measures may be more appropriate in specific NPH patterns. As HRs have been well established and commonly reported in clinical research, we believe that reporting HRs, as well as other measures such as RMST and milestone survival probabilities, provide values in quantifying the effect at this stage. We appreciate the authors of the letter for the comments, and we appreciate the Editor for the opportunity for us to further elaborate on our considerations for effect estimation under various NPH patterns. We believe this discussion highlights the value of our recommended approach: in the absence of thorough prior knowledge about the NPH patterns, reporting multiple measures enables a more comprehensive assessment of the treatment effect and will facilitate the clinical evaluation of the treatments.

References

Bartlett, J. W., Morris, T. P., Stensrud, M. J., Daniel, R. M., Vansteelandt, S. K., and Burman, C. F. (2020), "The Hazards of Period Specific and Weighted Hazard Ratios," *Statistics in Biopharmaceutical Research*, DOI: 10.1080/19466315.2020.1755722. [520]

Chen, T. T. (2015), "Milestone Survival: A Potential Intermediate Endpoint for Immune Checkpoint Inhibitors," *Journal of National Cancer Institute*, 107, djv156, DOI: 10.1093/jnci/djv156. [520]

Cole, S. R., and Hernan, M. A. (2004), "Adjusted Survival Curves With Inverse Probability Weights," *Computer Methods and Programs in Biomedicine*, 75, 45–49. [521]

Huang, B., Wei, L. J., and Ludmir, E. B. (2020), "Estimating Treatment Effect as the Primary Analysis in a Comparative Study: Moving Beyond P Value," *Journal of Clinical Oncology*, 38, 2001–2002, DOI: 10.1200/JCO.19.03111. [520]

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*. [521]

Kong, F. H., and Slud, E. (1997), "Robust Covariate-Adjusted Logrank Tests," *Biometrika*, 84, 847–862. [521]

Lin, R. S., Lin, J., Roychoudhury, S., Anderson, K. M., Hu, T., Huang, B., Leon, L. F., Liao, J. J., Liu, R., Luo, X., Mukhopadhyay, P., Qin, R., Tatsuoka, K., Wang, X., Wang, Y., Zhu, J., Chen, T. T., Iacona, R., and Cross-Pharma Non-Proportional Hazards Working Group (2020), "Alternative Analysis Methods for Time to Event Endpoints Under Non-Proportional Hazards: A Comparative Analysis," *Statistics in Biopharmaceutical Research*, 12, 187–198. [520]

Roychoudhury, S., Anderson, K. M., Ye, J., and Mukhopadhyay, P. (2019), "Robust Design and Analysis of Clinical Trials With Non-proportional Hazards: A Straw Man Guidance From a Cross-Pharma Working Group," arXiv no. 1908.07112. [520]

Tian, L., Jin, H., Uno, H., Lu, Y, Huang, B., Anderson, K. M., and Wei, L. J. (2020), "On the Empirical Choice of the Time Window for Restricted Mean Survival Time," *Biometrics*. [520]