Duke | Robert J. Margolis, MD
Center *for* Health Policy

Public Workshop: Oncology Clinical Trials in the Presence of Non-Proportional Hazards

The National Press Club • Washington, DC
February 5, 2018

Meeting Summary

Background

It is a remarkable time in oncology drug development. The pace of development and approvals for novel cancer treatments has been steadily increasing, representing a shift in potential treatment options that could transform patient care. In the last decade alone, the pipeline of oncology drugs in clinical development has expanded by 45 percent[1] with 68 novel cancer therapies launched globally between 2011 and 2016. In 2017, the U.S. Food and Drug Administration (FDA) approved 17 new cancer agents and over 30 efficacy supplements.[2] At the same time, the types of cancer therapies developed and approved are changing, representative of new mechanisms of action or entirely new classes of drugs. Immunotherapies, for example, use novel agents to target cancer, including immune checkpoint inhibitors, monoclonal antibodies, and vaccines, among others. These new types of cancer therapies are an increasing share of the drug development pipeline,[3] with first-in-class products representing approximately 85 percent of the current cancer pipeline.[4]

While many of these new treatments show great promise, there are ongoing challenges with oncology clinical trial designs and analysis of treatment effects for cancer therapies that may impact our understanding of investigational drugs.

Traditionally, randomized clinical trials for cancer therapies have considered time-to-event endpoints such as progression-free survival (PFS) and overall survival (OS) as the primary outcome measure in trial design. There are a number of commonly used statistical tools for standard time-to-event analyses. The log-rank test, for example, compares the survival curves of two treatment groups. The Kaplan Meier survival plot illustrates the totality of time-to-event kinetics, including the estimated median survival time. The Cox-proportional hazards model provides the estimated relative effect between treatment arms. The performance of these methods largely depends on the proportional hazards (PH) assumption – that the hazard ratio is constant over time. In other words, the hazard ratio provides an average relative treatment effect over time.

However, researchers have increasingly encountered scenarios where the proportionality assumption does not hold true. These include situations where there is delayed treatment effect, which manifests as a time lag before the separation of the survival curves; diminishing treatment effect, where the curves separate but then come back together after a period of time; or crossing hazards, where the curves actually cross each other. When the proportional hazard assumption is violated, the Cox-proportional hazard model may no longer be the optimal approach to determine treatment effect and the Kaplan-Meier estimate of median survival may not be the most valid measure to summarize the results. These distinctive characteristics in the survival curves have been observed in many clinical trials across various cancers and drug candidates, and are especially prevalent in trials of immunotherapeutics. New clinical trial design paradigms are needed to better capture and characterize these properties, including those that define potential new endpoints, modify traditional endpoints, and employ new statistical methods.

Recognizing a collaborative need to address these issues, FDA initiated a working group with pharmaceutical companies to: 1) systematically review assumptions under different statistical methods used in time-to-event analyses; 2) identify appropriate statistical tests under different non-proportionality (NPH) conditions; and 3) identify summary measures for describing treatment effect in randomized clinical trials. As a result of the cross-pharma working groups' efforts, they have proposed the "max-combo" test for analyzing time-to-event data in the presence of non-proportional hazards.

Meeting Objectives

The Duke-Margolis Center for Health Policy, under a cooperative agreement with the FDA, convened a public event on February 5, 2018, entitled, "Oncology Clinical Trials in the Presence of Non-Proportional Hazards". This workshop was the first opportunity for the group to publicly present their work in progress and receive feedback on the proposed max-combo method. The merits and drawbacks of this approach compared to other analytical methods were a major focus of the day's discussion.

The workshop provided an opportunity for representatives from across academia, industry, health care delivery, and government to explore and discuss alternative statistical methods for evaluating treatment effects of time-to-event endpoints; potential outcome measures that may more accurately capture treatment effect; and considerations for the development and design of future clinical trials. The following represents a summary of the meeting, including key feedback for the working group and areas for future research.

The Max-Combo Method: An Alternative Approach for Addressing Non-Proportional Hazards

Presenters discussed how they determined that an alternative analytical approach might be needed. The cross-pharma working group first evaluated widely used methods for hypothesis testing and estimation in the presence of non-proportional hazards, such as rank-based tests, restricted mean survival time, and Kaplan-Meier based tests. This evaluation surfaced two primary concerns that need to be addressed in any potential solution. The first was that all tests exhibited a substantial loss of statistical power under conditions of non-proportional hazards. The second was difficulty knowing not only whether non-proportional hazards might arise in new trial settings or for products, but also what type, which causes challenges in the design stage.

The group therefore concluded that an alternative method would need to be considered that had both adequate power and could be included in early-phase analytical plans. This resulted in the development of the "max-combo" test put forward at this workshop (Figure 1 in Appendix), which is based on Fleming-Harrington (FH) weighted log-rank statistics. The max-combo test tackles some of the challenges mentioned above as it is able to robustly handle a range of non-proportional hazard types, can be pre-specified at the design stage, and can choose the appropriate weight in an adaptive manner (i.e. is able to address the control of family-wise Type I error).

To illustrate the robustness of the test, presenters showed the results of a simulation exercise that included a wide range of non-proportional hazard scenarios where the method might apply. These scenarios varied across the following: presence of different types of non-proportional hazard (e.g. delayed treatment effect, diminishing treatment effect, and crossing hazards), different degrees of censoring observations, and different alternatives for treatment effect (e.g. Null, PH, and different types of non-proportional hazard). Based on the overall simulation results, the max-combo test performs reasonably well under many different scenarios and is agnostic to the type of non-proportional hazard. The max-combo test appears to have greater advantage in terms of power for delayed treatment effect and crossing hazards, and has an acceptable loss of power for diminishing effect.

Presenters also demonstrated how the test performs in real-world settings through the use of case studies, applying the test retrospectively to completed studies and comparing the results to those from the standard statistical tools used in the original trials. The log-rank test, the max-combo test, the Fleming-Harrington class of weighted log-rank tests, and restricted mean survival time were used to compare the two treatment arms in each case study. The objective in using the max-combo test is to reduce the false negative rate, as well as achieve the smallest p-value. Table 1 in the Appendix shows the case studies that were covered, the kind of non-proportional hazard scenario the trials encountered, and the key takeaways from each study. The p-value for the max-combo test is the smallest in most cases, and the test showed robustness under different types of non-proportional hazard scenarios.

Speakers also discussed how to design a clinical trial utilizing the proposed method and issues a researcher might need to anticipate. In order to achieve a robust design, it was recommended that researchers should include in their analysis plans two scenarios for which a trial is well-powered. In these analysis plans, researchers should carefully weigh the timing of analyses, as performing one too early may not account for treatment effect changes over time. In terms of summarizing treatment effect, researchers should continue to present Kaplan-Meier curves and Cox model hazard ratio estimates. However, additional summaries may be needed under conditions of non-proportional hazards. There are a variety of estimation procedures that can be employed such as piecewise Cox estimates, restricted mean survival time, weighted Cox estimates, and milestone estimates, but each method has its own limitations under non-proportional hazards. The downstream utility of the estimation technique for providing treatment effect information in the product label should be a key consideration throughout.

Outstanding Statistical and Design Considerations for the Max-Combo Test

While there was a great deal of enthusiasm around the proposed max-combo test, participants highlighted a few key areas the working group could explore further or that could use additional refinement.

Using Weights for Analyzing Time-to-Event Data

The max-combo test requires the use of weights to analyze time-to-event data. However, some participants expressed hesitation with weighting certain events more heavily than others. There was uncertainty over how to justify why early or late events in a survival analysis were receiving more emphasis. Moreover, others raised concerns about how to justify the use of weights from a patient perspective, particularly as weighting may differentially impact treatment decisions for patients and providers. A task for the working group is to determine whether there are potential problems in weighting some events more heavily than others and, if so, how to mitigate those concerns.

Pre-Specification of Clinical Trial Design

It is not clear at the design stage if any particular study is likely to encounter non-proportional hazards. In one FDA analysis of non-small cell lung cancer randomized clinical trials, only half of the studies submitted encountered non-proportional hazards.[5] Because of this uncertainty at the design stage, participants cautioned that the max-combo test may not be suitable as a default in all immuno-oncology trials. While utilizing the test as a default would enable a researcher to be prepared for any type of proportional or non-proportional hazards situation, some participants were reluctant to employ that approach as the more traditional methods currently in use (such as the log-rank test) are often more robust than the max-combo test in situations where the hazards are proportional.

If the max-combo test is not used as a default, an open question is how to determine when to pre-specify the use of the test in clinical trial design. One potential approach is by looking to previously conducted trials in the same or similar disease area or drug class. However, as was mentioned previously in the case of lung cancer, there is not always a consistent pattern of proportionality. Moreover, if a study is the first for a specific disease area, condition or mechanism of action, past knowledge may not be available to guide the study design. The working group may need to explore other tools that can help determine how to pre-specify the design in those circumstances. Overall, more discussion and research on the pre-specification of the trial design are needed.

Additional Examples Utilizing Max-Combo Test

Participants noted that that there is not enough information relating to when the max-combo test does not provide robust results. The speakers all presented examples in which the max-combo test proved mostly successful. However, participants desired a clearer understanding of when the test might not be suitable and, importantly, the steps they should take to remediate the trial design in those circumstances.

Ongoing Collaborative Efforts

Lastly, it was emphasized that the working group should continue to collaborate with colleagues from FDA, the European Medicines Agency, the National Cancer Institute, and other members of industry to address some of these remaining issues. While progress to date has been promising, it was evident that there might still be confusion over the proposed max-combo approach and that future work should therefore be transparent and informative for all stakeholders. It is important to ensure that there is a clear understanding of these methods in order for FDA to appropriately review submissions that include them, and continued public engagement will help to both strengthen the max-combo proposal itself and achieve buy-in on its use.

Key Takeaways and Areas for Future Research

In addition to the specific feedback workshop participants had for the max-combo working group, a number of broader statistical analysis and trial design challenges were discussed that may be worth further research and stakeholder dialogue.

Interpretation of Treatment Effect

In standard trials that have proportional hazards, the median can be used to describe and represent the treatment effect. However, in scenarios where there are non-proportional hazards, the median may not adequately describe treatment effect or reflect the survival pattern. Some participants suggested additional summary measures through estimation procedures such as the piecewise Cox model, restricted mean survival time, and milestone estimates, but there was no consensus from the group on which to use across the range of non-proportional hazard scenarios. Relatedly, participants also emphasized the need to examine heterogeneous populations and sub-groups. Quantile stratifications or sub-group analyses may be one path to better capturing treatment effect in certain non-proportional hazard situations.

Treatment Effect, Labeling, and Clinical Decision-Making

The ability to interpret treatment effect also has implications for product labeling and a clinician's ability to properly communicate the potential impact of treatment to patients. Often, a graph of a Kaplan-Meier survival curve is included on a label to assist with interpretation of treatment effect. However, if it is not clear that Kaplan-Meier survival curves are useful under conditions of non-proportional hazards, then their use in labeling may potentially be misleading for patients and providers. Therefore, a significant consideration is what information should be included in the label when non-proportional hazards are present. Should the Kaplan-Meier curve still be included? If so, what additional information is needed to provide adequate context? Tackling these questions may help to facilitate both better interpretation of evidence from the clinical trial and improved clinical decision-making.

Timing of Analyses

Participants also highlighted the need for more research on interim and futility analyses. There was some debate on the appropriate timing for these analyses because of the potential risks of performing them too early or too late in a study. Performing an analysis too early may misrepresent the treatment impact, because the treatment effect could change significantly after the analysis. However, performing an analysis too late may mean that some patients may be unnecessarily harmed by a treatment. On the other hand, a late analysis may mean an efficacious treatment is not getting to market fast enough to help patients. There was not a clear consensus on the timing and frequency of these analyses, and participants expressed interest in a more systematic set of guidelines especially when in the presence of non-proportional hazards.

Analysis of Crossing Hazards

Another open question is how to address the non-proportional hazard scenario of crossing hazards. Some participants wondered if this scenario should be evaluated using the same methods as those for delayed treatment effect or diminishing treatment effect because crossing hazards can present different interpretation challenges. Specifically, the reason crossing hazards occurred should be carefully examined before any conclusions are drawn. Furthermore, the replication of results may be needed to understand any underlying subgroup effect. Others argued, however, that evaluating all three types of non-proportional hazard scenarios in the same manner could better facilitate use of the max-combo test as it could be consistently applied and pre-specified. Still, many felt that crossing hazards present a different set of challenges and could merit additional work to elucidate how and why differing analytical approaches should be utilized across different non-proportional hazard scenarios.

Non-Proportional Hazards in Other Disease Areas

While the focus of the day's meeting was oncology, non-proportional hazards are encountered in other disciplines. Participants noted that these efforts could have implications for other therapeutic areas. It will be important to communicate the lessons learned to other stakeholders so as not to duplicate efforts.

Educating Key Decision-Makers

An important consideration when pursuing novel statistical approaches is the need to fully educate all stakeholders, including researchers, clinicians, patients, and payers, so that they understand the available alternatives. This is particularly important for senior decision-makers within industry, where there are internal scientific and cultural barriers that may inhibit the implementation of new designs. There will need to be efforts to translate and facilitate the use of this new statistical approach for those who are early clinical decision-makers as well as any future changes to means to interpret the trial results given HR and median have been used for so long in these settings. Furthermore, these new methods will need to be explained to clinicians and patients to ensure that clinicians can adequately interpret the evidence and can communicate to patients the potential treatment impact. Payers will require education on these novel methods as there may be a need to alter reimbursement decisions based on these new approaches.

Additional Design and Methods Work

Finally, participants noted that some challenges related to non-proportional hazards might be ameliorated as additional studies are conducted in the immunotherapy space. As more therapies within a class come on the market, the use of standard of care comparator arms within new trials may result in more typical proportional hazards and the use of now-standard analytical tools. This is not to say that tests like the max-combo are not needed in the interim, but that they may be a necessary stopgap as additional design and methods work is pursued.

Appendix

Figure 1: Max-Combo Test Design

Thomas R. Fleming and David P. Harrington proposed a class of weighted log-rank tests based on the $G^{\rho,\gamma}$ family:

Assign weight to events: $W_n(t)= (S_n(t))^{\rho}(1- S_n(t))^{\gamma}$

Values of ρ and γ imply:

- $\rho > 0, \gamma = 0$: early difference
- $\rho = 0, \gamma > 0$: late difference
- $\rho > 0, \gamma > 0$: mid difference
- $\rho = 0, \gamma = 0$: log-rank test

Max-Combo Test:

- Let, $Z_1 = G^{0,0}$, $Z_2 = G^{0,1}$, $Z_3 = G^{1,1}$, and $Z_4 = G^{1,0}$
- $Z_{max} = \max(| Z_1 |, | Z_2 |, | Z_3 |, | Z_4 |)$

Table 1: Application of Max-Combo Test in Four Cancer Clinical Trials

Case Study	NPH Pattern	Key Takeaways
The INO-VATE study comparing inotuzumab versus standard chemotherapy in patients treated with relapsed or refractory acute lymphoblastic leukemia	delayed effect and long-term remission	• Restricted mean survival time has the smallest p-value • Max-combo tests are significant, but have higher p-values than the log-rank test potentially due to small number of events after 15 months, possible crossing-hazard pattern in the first 12 months, and the multiplicity adjustment
Phase 3 study comparing Ipilimumab 10 mg/kg vs ipilimumab 3 mg/kg in patients with unresectable or metastatic melanoma	delayed effect and long-term remission	• The p-value of the max-combo test is more significant than that of the log-rank test for OS • For interim analysis, only p-value of the max-combo test and FH(0,1) are statistically significant, though FH(0,1) is more significant
A two-arm randomized phase 2 study of mitoxantrone and prednisone (MP) plus cixutumumab or ramucirumab in patients with metastatic castration-resistant prostate cancer (mCRPC)	diminishing treatment effect	• The p-value of the max-combo test is more significant than that of the log-rank test for both PFS and OS • The max-combo has the second smallest p-value, compared to FH(1,0) which is expected to work well in this scenario
The IPASS study comparing gefitinib versus carboplatin plus paclitaxel in patients with advanced pulmonary adenocarcinoma	crossing survival curves	• The p-value of the max-combo test is more significant than that of the log-rank test for both PFS and OS • The max-combo has the second smallest p-value for OS and the smallest p-value for PFS compared to FH(1,1)

References

[1] https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/global-oncology-trends-2017.pdf?_=1516825034305
[2] https://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm279174.htm
[3] https://www.cancer.org/treatment/treatments-and-side-effects/treatment-types/immunotherapy/what-is-immunotherapy.html
[4]http://www.analysisgroup.com/uploadedfiles/content/insights/publishing/the_biopharmaceutical_pipeline_report_2017.pdf
[5] Presentation, Lijun Zhang, U.S. Food and Drug Administration, "Session III: Retrospective Application of Novel Analysis Methods in Completed Trials". *Public Workshop: Oncology Clinical Trials in the Presence of Non-Proportional Hazards,* February 5, 2018, Slide 4.

Funding for this workshop was made possible by the Food and Drug Administration through grant (U19FD004971). Views expressed in the written materials and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services nor does any mention of trade names, commercial practices, or organization imply endorsement by the United States Government.

The Robert J. Margolis, MD, Center for Health Policy is part of Duke University, and as such it honors the tradition of academic independence on the part of its faculty and scholars. Neither Duke nor the Margolis Center take partisan positions, but the individual members are free to speak their minds and express their opinions regarding important issues. For more details on relevant institutional policies, please refer to the Duke Faculty Handbook, including conflicts of interest, the Code of Conduct, and other policies and procedures. In addition, regarding positions on legislation and advocacy, Duke University policies are available at http://publicaffairs.duke.edu/government.