Editorial

Targeted Survival Improvements in Clinical Trials: Are You an Absolutist or Relativist?

James Paul, BSc

The design of a clinical trial is a complex business that requires careful consideration of several factors, with one of the most crucial being the assessment of the targeted improvement in the primary outcome measure, which is also one of the principal determinants of sample size. The article by Castonguay et al in the current issue of *Cancer*[1] reviews the accuracy with which the median progression-free survival (PFS) and overall survival (OS) in the control arm have been estimated in studies of epithelial ovarian cancer over the past 10 years, and in doing so they have highlighted that the design process is even more challenging than was perhaps believed. Their primary observation is that the median OS in the control arm was underestimated by >25% in 80% of the trials they examined (12 of 15 trials); by way of contrast, the median PFS or time to disease progression (TTP) was underestimated by 25% in 20% of the trials studied (4 of 20 trials). In those studies in which significant underestimation occurred, the authors noted the detrimental impact on the power to detect absolute differences in these endpoints.[1]

Conventionally, the overall sample size for a clinical trial with a time-to-event endpoint depends on:

1. the targeted improvement in the hazard ratio (**relative** improvement in the median OS, PFS, or TTP); and
2. the type I and type II error (equivalently significance level and power)

Together, these 2 factors determine the number of events required (deaths/disease progressions). The following factors then determine the sample size required to observe the required number of events:

1. the median time to the event in the control arm (median OS, PFS, or TTP);
2. the accrual rate; and
3. the minimum follow-up period.

Modest variations in the accrual rate and median time to outcome in the control arm are to be expected and can be accommodated (within reason) by adjustment of the recruitment period or minimum follow-up time or both to achieve the required number of events specified in the study design. This ensures the correct type II error/study power for the **relative** improvement in median time to outcome.

However, it must be noted that the sample size calculation operates within the context of a targeted improvement that is a **relative** one. Although it is the relative improvement that is used by the statistician in the calculation of the sample size, the **absolute** difference is also an important and interlinked consideration, possibly even the primary focus from a clinical perspective and perhaps slightly neglected by the statistician.

The statistician takes comfort in the fact that by attaining the original number of target events specified in the design, power is maintained to detect the relative difference between the treatment arms whatever the underlying median time to outcome. Expanding on the point made by Castonguay et al[1] in Table 3 of their article, Table 1 presented herein shows how the absolute difference that can be detected with the stated study power varies as the median survival is underestimated by varying amounts; it also demonstrates how the power to detect the original target **absolute** difference declines. It should be emphasized that the power to detect the relative difference (specified in the hazard ratio) is maintained in all cases.

As can be seen in Table 1, as the control arm median is increasingly underestimated, only larger absolute differences can be detected with the power stated in the design. Equivalently, the more the control arm median is underestimated, the

Corresponding author: James Paul, BSc, Cancer Research UK Clinical Trials Unit, Institute of Cancer Sciences, University of Glasgow, The Beatson West of Scotland Cancer Centre, Level 0, 1053 Gt. Western Rd, Glasgow, G12 0YN UK; Fax: (011) 44 (0) 141 301 7189; james.paul@glasgow.ac.uk

Cancer Research UK Clinical Trials Unit, Institute of Cancer Sciences, University of Glasgow, Glasgow, United Kingdom.

See referenced original article on pages 413-22, this issue.

DOI: 10.1002/cncr.29031, **Received:** August 10, 2014; **Accepted:** August 14, 2014, **Published online** October 2, 2014 in Wiley Online Library (wileyonlinelibrary.com)

TABLE 1. Impact of Variation in Observed Versus Estimated Control Arm Median on 1) Absolute Difference That Can Be Detected With Designed Study Power and 2) the Power to Detect Target Absolute Improvement

Estimated Control Arm Median, Months	Target Hazard Ratio (HR)	Target Absolute Improvement, Months	Designed Study Power to Detect Target HR	Observed Control Arm Median (Ratio to Original Estimate)	Absolute Difference That Can Be Detected With Designed Study Power, Months	Power to Detect Target Absolute Improvement
15	0.833	3	80	15 (1)	3.0	80
15	0.833	3	80	16.875 (1.125)	3.375	71
15	0.833	3	80	18.75 (1.25)	3.75	63
15	0.833	3	90	15 (1)	3.0	90
15	0.833	3	90	16.875 (1.125)	3.375	83
15	0.833	3	90	18.75 (1.25)	3.75	75
30	0.833	6	80	30 (1)	6.0	80
30	0.833	6	80	33.75 (1.125)	6.75	71
30	0.833	6	80	37.5 (1.25)	7.5	63
30	0.833	6	90	30 (1)	6.0	90
30	0.833	6	90	33.75 (1.125)	6.75	83
30	0.833	6	90	37.5 (1.25)	7.5	75

Abbreviation: HR, hazard ratio.

less the power to detect the original absolute target improvement. The effect on the absolute difference that can be detected is less when dealing with smaller medians, such as would typically apply for an endpoint such as PFS. It is interesting to note that starting with a study design with a higher power (90% rather than 80%) means that even at the higher level of underestimation given in Table 1 (a 25% difference), reasonable power is maintained to detect the original absolute target (75%).

What is a sufficient change in the absolute difference that can be reliably detected that would require study modification? One would argue perhaps that a change from 6 months to 7.5 months might be worth considering starting to make a modification for, but not a change from 3 months to 3.75 months? It is difficult to make a definitive judgment without understanding more about how the difference to be detected in the study has been determined in the first place. A study is designed around a single important target difference, but in reality there are shades of gray around what is an "important" difference. The DELTA project (Difference ELicitation in TriAls) conducted within the United Kingdom[2] examined how trialists determined the targeted difference in clinical trials. This study concluded that there was a need for more formal methods with which to determine the target difference and improved reporting of its specification. In particular, reporting should "state the underlying basis used for specifying the target difference:

- An *important* difference as judged by a stakeholder *or*
- A *realistic* difference based on current knowledge *or*
- Both an *important* and a *realistic* difference."[2]

In addition, for survival outcomes, reporting should "**. . .**.state the target difference as an absolute and/or relative difference**. . .. . .** If both an absolute and a relative difference are provided, clarify if either takes primacy in terms of the sample size calculation."[2]

Careful consideration of what is an important target difference has been an issue of continued debate and is not a straightforward process.[3] It is also a process at which historically the oncology community has been poor.[4] Explicit consideration of this in published study reports would help investigators and ultimately consumers of their research understand the extent to which changes in underlying assumptions about median outcomes in the control arm have impacted the study objectives.

Why is so much emphasis placed on relative improvement in outcome, when to many the absolute difference is a more natural way of thinking? This emphasis has really come from statisticians.

The fact that statisticians approach survival in terms of relative improvement and the associated hazard ratio is largely a consequence of the ubiquitous use of the Cox proportional hazards model as the primary means for analyzing survival data. This statistical approach was developed to cope with the analytical difficulties of dealing with censored data (ie, the fact that we only observe the time to study outcome for a percentage of the patients taking part in the trial; for the remainder we only observe the follow-up time without outcome).

The Cox model provides an elegant solution to this difficulty, but as a model it is framed in terms of relative

event rates, as are the associated power and sample size calculations; hence the emphasis on relative improvements by statisticians.

It should also be noted that the Cox model is based on the assumption of a **constant** relative event rate (relative hazard) over the follow-up period: the "proportional hazards" assumption. When this assumption is not fulfilled, the estimate of relative improvement derived from the Cox model is not a meaningful measure of the difference in outcome between the study arms.

In recent high-profile trials,[5-7] this requirement of a constant relative rate has not been fulfilled. This has led some to propose alternative measures,[8] including an absolute change in survival as measured by restricted mean survival times (RMSTs).[9] The use of this measure requires a prespecification of a follow-up time, but requires no assumption such as a constant relative hazard (proportional hazards) to be valid. The RMST itself can be interpreted as the life expectancy between randomization onto the trial and a prespecified follow-up time. Some early work has been done on trial design and analysis based around the comparison of RMST. However, it should be noted that the statistical design still rests heavily on the accuracy of initial assumptions about time to event outcome and much of the design methodology is yet to be fully developed and applied in practice.

In any study, it is the role of the data monitoring committee to monitor how closely the assumptions used in the study design are being met as the data emerge. It is usually within their remit to make appropriate adjustments to ensure that the original objectives are fulfilled and methods exist to allow this without affecting the type I error.[10] These methods are focused on adjustments made to allow for inaccurate assumptions about the recruitment rate or event rate within the context of relative hazards and the Cox model; they do not extend to nonproportional hazards or studies in which the focus is on an absolute difference in the survival outcome. To my knowledge, there are no formal methods that allow for an adjustment to be made to the sample size in which the focus is on an absolute difference in survival, and this is a difficulty.

The problems with underestimation that Castonguay et al[1] observed in studies with OS as the primary endpoint did not appear to the same extent in studies using PFS or TTP as a primary endpoint. The authors speculate as to the reasons for this in their article and, as they suggest, the introduction of more effective treatments and clinical management into second-line and subsequent treatment must be considered to play a role.[1]

This again raises the issue concerning the most appropriate primary endpoint for phase 3 studies. The use of OS is likely to become increasingly less relevant as the influence of subsequent treatments dilutes the effect of the intervention of primary interest and thus impacts on the sensitivity of the trial to detect differences. To these arguments we can now add evidence of the difficulty of reliably planning studies around a primary endpoint of OS.

Saad and Buyse[11] recently argued that "....PFS or other tumor-based end points be considered in public health decisions when multiple lines of effective therapies are available in clinical practice, as long as the intervention of interest displays a favorable toxicity profile and no untoward effect on OS." Adoption of PFS as a primary endpoint is becoming an increasingly attractive option in several settings, although strong arguments to the contrary have been made.[12]

In the context of randomized phase 2 studies in which PFS is ubiquitous, some comfort can be taken from the fact that the shorter median PFS times indicate that the impact of underestimation is less on absolute targets; this is particularly important within the context of biomarker-enriched studies in which there may be no reliable historical information regarding the background PFS distribution on which to base the study design.

The article by Castonguay et al[1] highlights an important practical issue in the study design of recent phase 3 studies among patients with epithelial ovarian cancer. In addition, their findings have broader implications in that they:

- Suggest that more careful reflection is required on the primary focus (relative or absolute improvement in the time to even a primary endpoint) and how clinically relevant differences are determined.
- Highlight further issues with the selection of OS as a primary endpoint.
- Provide further motivation for questioning the prevalent standard statistical approach used to design and analyze studies with time to event outcomes.

FUNDING SUPPORT

Mr Paul is supported by Cancer Research UK grant C1348/A15960.

CONFLICT OF INTEREST DISCLOSURES

The author made no disclosures.

REFERENCES

1. Castonguay V, Wilson M, Diaz-Padilla I, Wang L, Oza A. Estimation of expectedness: predictive accuracy of standard therapy outcomes in randomized phase 3 studies in epithelial ovarian cancer. *Cancer.* 2015;121:413-422.

2. Cook JA, Hislop J, Adewuyi TE, et al. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technol Assess.* 2014;18: v–vi, 1-175.
3. Ellis LM, Bernstein DS, Voest EE, et al. American Society of Clinical Oncology perspective: raising the bar for clinical trials by defining clinically meaningful outcomes. *J Clin Oncol.* 2014;32:1277-1280.
4. Gan HK, You B, Pond GR, Chen EX. Assumptions of expected benefits in randomized phase III trials evaluating systemic treatments for cancer. *J Natl Cancer Inst.* 2012;104:590-598.
5. Perren T, Swart A, Pfisterer J, et al; ICON7 Investigators. A phase 3 trial of bevacizumab in ovarian cancer. *N Engl J Med.* 2011;365: 2484-2496.
6. Allegra CJ, Yothers G, O'Connell MJ, et al. Bevacizumab in stage II-III colon cancer: 5-year update of the National Surgical Adjuvant Breast and Bowel Project C-08 trial. *J Clin Oncol.* 2013;31:359-364.
7. Thoren FB, Anderson H, Strannegard O. Late divergence of survival curves in cancer immunotherapy trials: interpretation and implications. *Cancer Immunol Immunother.* 2013;62:1547-1551.
8. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol.* 2014;32:2380-2385.
9. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol.* 2013;13:152.
10. Todd S, Valdes-Marquez E, West J. A practical comparison of blinded methods for sample size reviews in survival data clinical trials. *Pharm Stat.* 2012;11:141-148.
11. Saad ED, Buyse M. Overall survival: patient outcome, therapeutic objective, clinical trial end point, or public health measure? *J Clin Oncol.* 2012;30:1750-1754.
12. Booth CM, Eisenhauer EA. Progression-free survival: meaningful or simply measurable? *J Clin Oncol.* 2012;30:1030-1033.