EDITORIAL

Moving beyond our comfort zone

Brian Claggett[1], Lee-Jen Wei[1], and Marc A. Pfeffer[2]*

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA; and [2]Cardiovascular Division, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115, USA

This editorial refers to 'Evaluation of early percutaneous coronary intervention vs. standard therapy after fibrinolysis for ST-segment elevation myocardial infarction: contribution of weighting the composite endpoint'[†], by J.A. Bakal et al., on page 903

The randomized controlled clinical trial currently represents the gold standard by which we test whether or not a therapeutic strategy will alter a specific predefined outcome. A carefully detailed, pre-specified statistical analysis plan is established in order to test this primary hypothesis. This framework surrounding the evaluation of a trial's primary outcome is constructed to provide a level of certainty which must be achieved in order to reject the null hypothesis and declare that an observed difference between the two (or more) therapeutic strategies is not attributable to chance. As such, extensive resources (both economic and otherwise) required to conduct a clinical outcome trial are generally narrowly focused on the answer to a single clinical question, often utilizing the time to first event. We need to move beyond this current practice and use more information obtained in a trial to make more informed inferences about treatment efficacy as well as risks.

When the primary endpoint is mortality and the question is precisely focused as to whether an intervention alters the patient's time to death, the current statistical framework provides a clearly understandable answer, and if the null hypothesis is rejected and the two therapies are indeed found to be different, subsequent secondary analyses are undertaken to explore the available additional clinical information further. However, unless the target patient population is limited to those with extremely poor prognosis, trials with a primary endpoint of all-cause mortality are rarely conducted.

The obvious issues of giving equal weight to a fatal and non-fatal endpoint in the 'time to first event' approach, which ignores all events (including deaths) that occur after an initial non-fatal outcome, are well recognized. Indeed this was the subject of a recent well-written Editorial by Anker and McMurray in this journal.[1] Despite these limitations, this type of approach has become the 'comfort zone' from which many clinical trialists, regulators, and journal editors seem reluctant to step outside.[2–4]

The provocative article by Bakal et al. now raises another common limitation of our current reporting of clinical outcome trials that occurs when multiple non-fatal measures of morbidity are incorporated into the primary outcome.[5] They illustrate how the extra dimension of adding a non-fatal to a fatal outcome is further complicated, in this case by the consideration of three different types of non-fatal outcomes: shock, congestive heart failure, and recurrent myocardial infarction (MI). Through consultation with clinician-investigators, they developed a weighting scheme for these non-fatal events, concluding that each of these events should be considered to have differing magnitudes of clinical severity. Despite the obvious subjectivity of such an approach when combining different measures of morbidity, it is clearly acknowledged that, short of death, there are degrees of morbidity. It is of interest that this attempt to add clinically meaningful information into the assessment of patient outcomes resulted in a reduction of the apparent benefit of treatment. Of course, it is certainly possible that in other circumstances, two treatments, initially found to be similar with respect to a composite outcome, actually differ greatly upon consideration of more (appropriately weighted) patient outcomes.

Even within a specific type of non-fatal event, one could consider yet another layer of complexity which could add meaningful information. For example, two patients may both suffer a stroke, with one grossly debilitated and the other making a full recovery with minimal neurological residua. Similarly, there are MIs where the patient survives with marked limitations, and others where he or she remains asymptomatic and leads a full and comfortable lifestyle. Indeed, Sampson et al.[6] distinguish between episodes of congestive heart failure, assigning weights of 0.5 or 0.2, depending on whether the event required hospitalization or simply treatment with ACE inhibitors.

We must be clear that the impetus for novel methodology regarding the analysis and reporting of clinical trial results should not solely be driven by concerns of statistical power, but by a

desire to better understand how therapies influence a broader characterization of total disease burden.

To achieve this objective, we will need to leave our current statistical/regulatory comfort zone by acknowledging that the current paradigm of 'time to first' analyses ignores a great deal of clinically relevant information. We should also challenge ourselves to account further for all unfavourable ('toxic') events induced by a therapy in conjunction with those unfavourable events that may be prevented. Indeed, a proper 'risk−benefit' analysis is not possible without understanding the relationship between the adverse and beneficial effects of the investigational treatment. To strive for the ideal in clinical trial reporting and analysis, we need to capture all relevant patient outcome information, condense it, and do so in a way such that the summaries and comparisons generated are both statistically valid and clinically interpretable. A reported P-value must not be confused with an assessment of the magnitude of a treatment's effect in a way that is meaningful to the patient, the clinician, and the regulator.

A key roadblock to reaching this goal is the simple fact that, as more information about a patient's experience becomes available, the act of appropriately condensing, summarizing, and evaluating that information in a clinically meaningful manner becomes increasingly difficult. Current approaches to the analysis of clinical trial outcomes should be assessed with respect to the totality of information used as well as the clinical interpretability of the resulting summary statistics. Several current approaches are illustrated in *Figure 1*, with the axes representing these two, often competing, aims of capturing more information and allowing greater interpretability. Consider the following two extreme examples: there would be very little utility to viewing all individual patient data without any form of statistical summary (lower right, *Figure 1*); alternatively, providing a hazard ratio for all-cause death would represent a clinically understandable comparison but provides limited ability to characterize patient disease burden fully (upper left, *Figure 1*).

We, and others, have called for movement beyond this current 'time to first' paradigm.[1−4,6−13] Several proposed methods that have effectively increased the amount of information incorporated into statistical analysis are shown in *Figure 1*. One straightforward approach to better quantifying patient disease burden is to compare patients' time spent 'alive and out of the hospital',[8] though this approach implicitly equates hospitalization with death and the resulting scale of measurement (days) is clearly understandable, but its magnitude may be difficult to interpret outside of the context of the follow-up time of a particular trial. Bakal *et al.*'s weighted composite endpoint[9] and Sampson *et al.*'s total disease burden[6] additionally attach weights to various clinical outcomes, but it may be difficult to gauge whether or not a reduction of, say, 0.4 'weighted clinical events' represents a substantial clinical benefit. Extensions of the traditional Cox model to include multiple failure times have been proposed,[10,11] but the resulting
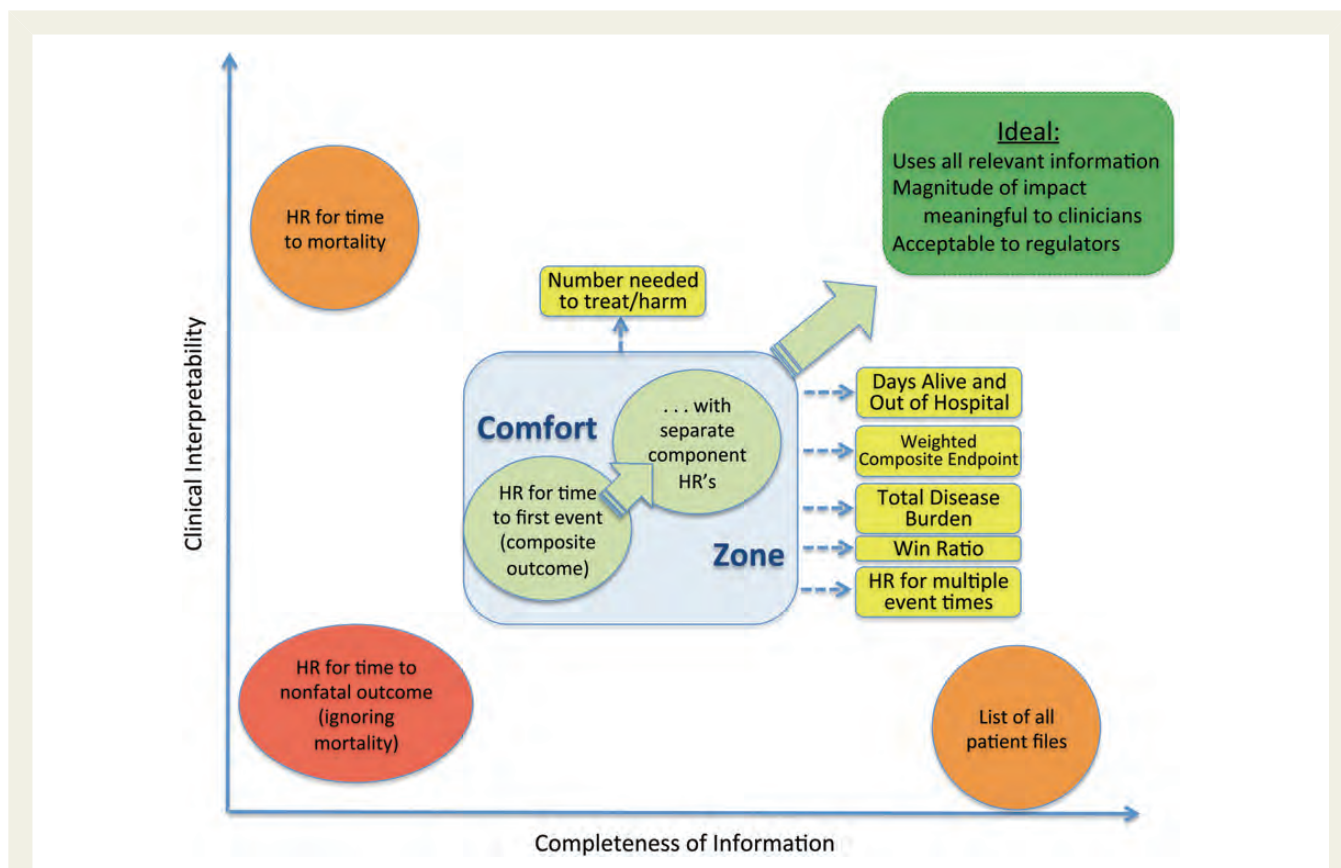
Figure 1 Statistical framework for primary analysis of major cardiovascular outcome trials.

parameter is difficult to interpret unless a common hazard ratio can be assumed to apply to each distinct failure type or recurrent event, and provides no insight into the relationship between failure times.[12] The Win Ratio,[13] by comparing patients with one another with respect to multiple event times, takes advantage of multiple patient outcomes and requires only a relative ordering of the severity of event types rather than specific weighting values. However, the resulting inference is based on the probability of one treatment group's patients being ranked above the comparator group's, and it is often difficult to recover a true clinical interpretation of a treatment's impact, even if it is shown significantly to improve its patients' relative rankings. An example of a tool that augments clinical interpretability is the number needed to treat or harm (NNT/NNH),[14] which focuses on absolute, rather than relative, reductions in event rates. However, these analyses generally do not incorporate a multiplicity of events or any weighting of severity.

Each of these approaches moves beyond the current comfort zone, utilizing more information from the trial (to the right on the x-axis) or augmenting interpretability (upwards on the y-axis), but still leave room for improvement. One reason that such vast resources are devoted to trials of novel therapies is to prove that they represent an advancement in clinical care so that the therapy may then be registered and commercialized. Few sponsors and investigators are willing to invest these vast resources required to conduct a trial without a statistical analysis plan that is within the tried and true comfort zone for clinical acceptance and regulatory approval. For now, perhaps the path forward is to develop and apply these new methods retroactively to existing data to better assess the impacts of treatments already in use, and to become more comfortable interpreting treatment effects using these new metrics.

The article by Bakal et al.[5] highlights the need to extract more information from clinical outcome trials. When we fail to do so, we run the risk of being overly optimistic about the effects of a new therapy, or, conversely, prematurely dismissing a truly effective one. Moving forward, we must be prepared to go beyond our comfort zone. We need to design clinical trials with tools to understand the impact of our experimental interventions on our patients more fully and answer the most fundamental question: 'Should this therapy change medical practice?' We owe it to the patients who have volunteered to participate in these trials, and to the future patients whose therapeutic options will be informed by the results of those trials, to develop the translational statistics to move towards this ideal.

Conflict of interest: none declared.

References

1. Anker SD, McMurray JV.Time to move on from 'time-to-first': should all events be included in the analysis of clinical trials? *Eur Heart J* 2012;**33**:2764–2765.
2. Kaul S, Diamond GA. Trial and error: how to avoid commonly encountered limitations of published clinical trials. *J Am Coll Cardiol* 2010;**55**:415–427.
3. DeMets DL, Califf RM. Lessons learned from recent cardiovascular clinical trials (part I). *Circulation* 2002;**106**:746–751.
4. Ferreira-González I, Busse JW, Heels-Ansdell D, Montori VM, Akl EA, Bryant DM, Alonso-Coello P, Alonso J, Worster A, Upadhye S, Jaeschke R, Schünemann HJ, Permanyer-Miralda G, Pacheco-Huergo V, Domingo-Salvany A, Wu P, Mills EJ, Guyatt GH.. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* 2007;**334**:786.
5. Bakal JA, Westerhout CM, Cantor WJ, Fernández-Avilés F, Welsh RC, Fitchett D, Goodman SG, Armstrong PW. Evaluation of early percutaneous coronary intervention vs. standard therapy after fibrinolysis for ST-segment elevation myocardial infarction: contribution of weighting the composite endpoint. *Eur Heart J* 2013;**34**:903–908.
6. Sampson UK, Metcalfe C, Pfeffer MA, Solomon SD, Zou KH. Composite outcomes: weighting component events according to severity assisted interpretation but reduced statistical power. *J Clin Epidemiol* 2010;**63**:1156–1158.
7. Claggett B, Tian L, Castagno D, Wei LJ. Treatment selections using risk-benefit profiles based on data from comparative randomized clinical trials with multiple endpoints. (November 2012). *Harvard University Biostatistics Working Paper Series. Working Paper 153.* **http://biostats.bepress.com/harvardbiostat/paper153**
8. Ariti CA, Cleland JG, Pocock SJ, Pfeffer MA, Swedberg K, Granger CB, McMurray JJ, Michelson EL, Ostergren J, Yusuf S. Days alive and out of hospital and the patient journey in patients with heart failure: insights from the candesartan in heart failure: assessment of reduction in mortality and morbidity (CHARM) program. *Am Heart J* 2011;**162**:900–906.
9. Bakal JA, Westerhout CM, Armstrong PW. Impact of weighted composite compared to traditional composite endpoints for the design of randomized controlled trials. *Stat Methods Med Res* 2012;in press.
10. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Stat Assoc* 1989;**84**:1064–1072.
11. Andersen PK., Gill RD. Cox's regression model for counting processes: a large sample study, *Ann Stat* 1982;**10**:1100–1120.
12. Wei LJ, Glidden DV. An overview of statistical methods for multiple failure time data in clinical trials. *Stat Med* 1997;**16**:833–839.
13. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J* 2012;**33**:176–182.
14. Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. *BMJ* 1999;**319**:1492.