Surrogate Endpoints: Wishful Thinking or Reality?

Stuart G. Baker

More than 100 years ago, the noted French mathematician Henri Poincaré quoted the following remark about the assumption of a normal distribution: “Everybody firmly believes in it because the mathematicians imagine it is a fact of observation, and observers that it is a theory of mathematics” (1). A similar generalization could be made about methods to validate surrogate endpoints: Biostatisticians believe that the methods they propose are useful because clinicians adopt them, and clinicians believe that the methods proposed by biostatisticians are useful because they have the “imprimatur” of mathematical statistics. Given this state of affairs, a critical examination of methods to validate surrogate endpoints is needed.

However, before delving further it is necessary to precisely state the role of surrogate endpoints. The purpose of a surrogate endpoint is to draw conclusions about the effect of intervention on true endpoint without having to observe the true endpoint. If this purpose could be achieved, clinical research would be greatly accelerated. Unfortunately it is a tall order, and many proposed surrogate endpoints have subsequently been shown to have led to incorrect conclusions about the effect of intervention on the true endpoints (2). Therefore, before a surrogate endpoint can be used with confidence, it must be validated. Part of the controversy with the use of surrogate endpoints is that there is no agreed-upon definition of a validated surrogate endpoint. Essentially, validation of a surrogate endpoint consists of whatever the investigators think will make them and others feel confident about the use of the surrogate endpoint in a future trial. Validation measures must ensure that this confidence is grounded more in reality than wishful thinking.

In recent years, biostatisticians have proposed a wide variety of measures for validating surrogate endpoints (3). One of the earliest and best known validation measures is the set of criteria proposed by Prentice (4,5) to ensure that rejection of the null hypothesis of no effect of intervention on the surrogate endpoint implies rejection of the null hypothesis of no effect of intervention on the true endpoint. These criteria are 1) treatment affects the surrogate endpoint, 2) treatment affects the true endpoint, 3) the distribution of the true endpoint conditional on the surrogate endpoint is the same for both arms of randomized trial (i.e., the association of the surrogate endpoint with the true endpoint is independent of treatment arm), and 4) a subtle and often overlooked mathematical condition to ensure that the null hypothesis for the true endpoint implies the null hypothesis for the surrogate endpoint, rather than vice versa, as would be obtained if only the first three criteria were satisfied. [For binary surrogate endpoints, condition 4 is simply an association between the surrogate and true endpoints (5)]. The most important condition is 3, which is sometimes called the Prentice criterion.

To formally check the Prentice criterion for a candidate surrogate endpoint in a trial with a surrogate and true endpoint, Freedman et al. fit a model for true endpoint as function of treatment and the candidate surrogate endpoint (6). They noted that if treatment has a statistically significant effect on the true endpoint in this model, the Prentice criterion does not hold for the candidate surrogate endpoint. Thus a statistically significant effect of treatment on true endpoint in this model indicates a poor surrogate endpoint unless the sample size were so large that even a tiny deviation from the Prentice criterion would be statistically significant. However, if there is no statistically significant effect of treatment on true endpoint in the model, it does not constitute strong evidence that the Prentice criterion holds (6). (Similarly, in standard hypothesis testing, not rejecting a null hypothesis does not constitute strong evidence that a null hypothesis holds.) For this situation, Freedman et al. (6) proposed, as a validation measure, the proportion of treatment effect explained by surrogate endpoint (PTE), which equals one minus the ratio of the estimated treatment effect on true endpoint adjusted for the surrogate endpoint and the estimated treatment effect on true endpoint not adjusted for the surrogate endpoint. Freedman et al. (6) proposed PTE in the context of a binary true endpoint, and Lin et al. (7) extended PTE to survival data.

In this issue of the Journal, Petrylak et al. (8) analyzed several changes in prostate-specific antigen (PSA) as possible surrogate endpoints for survival in a retrospective analysis of data from a single clinical trial of chemotherapy for men with androgen-independent prostate cancer. They examined whether the candidate surrogate endpoints met several criteria for validating a surrogate endpoint, but the centerpiece of their analysis is the PTE. For various candidate surrogate endpoints (various PSA declines and PSA velocity at several time points after the beginning of treatment), Petrylak et al. computed 95% confidence intervals for the PTE (although one might argue that these confidence intervals should have been larger to adjust for multiple comparisons). PTE is a controversial measure because the confidence intervals are typically large and because PTE may lie outside the range of 0 to 1, indicating a logical difficulty (3,5,9). But perhaps the major drawback is deciding on a good target value for PTE (9). Petrylak et al. claim that a surrogate endpoint is good if the lower bound of the 95% confidence interval for the PTE is at least 0.5. Does a lower bound of 0.5 for the PTE convince us that, when the null hypothesis of no effect of intervention on the surrogate endpoint is rejected, the null hypothesis of no effect of intervention on the true endpoint will be rejected, as is the goal of the Prentice criteria? It is hard to say. So, this brings us back to the subjective measure of investigator expectations.
A more fundamental limitation of the PTE is that, by definition, it is based on data from a single trial. The growing view is that measures to validate surrogate endpoints should be based on multiple trials involving surrogate and true endpoints because these measures (called meta-analytic) better capture the uncertainty in relating surrogate and true endpoints than measures derived from a single trial. Although various meta-analytic validation measures for surrogate endpoints have been developed (3), one simple validation approach for binary surrogate endpoints (10) provides a clear contrast with the PTE. In this approach a set of trials with surrogate and true endpoints is successively split into “previous” trials and a “new” trial. For example if there are 10 trials, the first split might consist of the first nine trials as “previous” and the 10th as “new”, the second might consist of the first eight trials and the 10th as “previous” and the ninth as “new”, and so forth for all 10 possible splits. Using data from each “previous” trial, a prediction is made of the effect of intervention on true endpoint in the “new” trial, pretending that only data on the surrogate endpoint are available in the “new” trial. The predictions from each “previous” trial are then combined to obtain a single prediction of the effect of intervention on true endpoint in the “new” trial, again pretending that only data on the surrogate endpoint are available in the “new” trial. The validation measure is the average prediction error of the estimated predicted effect (APEP), which is the absolute difference between the predicted intervention effect (using the surrogate and true endpoints in the “previous” trials and the surrogate endpoint in the “new” trial) and the observed intervention effect (using the true endpoint in the “new” trial), averaged over all splits into “previous” trials and a “new” trial. Unlike PTE, APEP has a natural benchmark—namely, the average clinically meaningful difference (ACMD), which is implicit in the sample size of the trials. If APEP is much smaller than ACMD, then the error from using the surrogate endpoint to predict the effect of the intervention on the true endpoint will be relatively small compared with the effect of intervention on the true endpoint that the designers of the trials implicitly hoped to detect. Validation of the surrogate also requires that APEP be smaller than the average prediction error from a standard meta-analysis of true endpoints.

Of course, no validation measure can guarantee that, in a new trial, the surrogate endpoint will yield the same conclusions about the effect of intervention as the true endpoint. Petrylak et al. (8) correctly emphasized that their validation may be applicable to only one class of drugs. The same caveat applies to a meta-analytic validation approach.

So where does this leave us? Given that only a single trial was discussed by Petrylak et al., their validation should be considered tentative and viewed with caution. If data from other prostate cancer trials with the same surrogate and true endpoints as in the SWOG Trial that Petrylak et al. analyzed were available, one could perform a meta-analytic validation of the surrogate endpoint. If such a meta-analytic validation determined that the surrogate endpoint was good, it would be a much more convincing evidence of a good surrogate endpoint than the evidence in Petrylak et al. Thus if Petrylak et al. serves as an impetus for other prostate cancer clinical trialists to collect data on both surrogate and true endpoints and to make these data available for meta-analytic validation, it will have served a very useful purpose. Only with such additional data will it be possible to determine if surrogate endpoints are closer to wishful thinking or reality.

References

(9) Baker SG. A simple meta-analytic approach for using a binary surrogate endpoint to predict the effect of intervention on true endpoint. Biostatistics 2006;7:57–70.