Choosing Phase II Endpoints and Designs: Evaluating the Possibilities

Michael LeBlanc and Catherine Tangen

Selecting the phase II design and endpoint to achieve the best possible chance of success for a confirmatory phase III study in a particular disease and treatment setting is challenging but critical. Simulating from existing clinical trial data sets and from mathematical models can be useful tools for evaluating statistical properties. *Clin Cancer Res; 18*(8); 2130–2. ©2012 AACR.

In this issue of *Clinical Cancer Research*, Sharma and colleagues (1) describe the properties of alternative endpoints to progression-free survival (PFS) and clinical response to evaluate inhibitory agents in which high levels of tumor regression are not anticipated. Data sets representing hypothetical phase II studies are simulated (or sampled) from 2 prior phase III trials in metastatic renal cancer: a positive study (sorafenib vs. placebo; ref. 2) and a negative study (AE941 vs. placebo; ref. 3). Sharma and colleagues conclude that, in this particular setting, a randomized phase II design, with an endpoint based on continuous measures of tumor size, yields the greatest power, but at the cost of a higher false-positive rate than the other design options they consider.

The authors show resampling, often called bootstrap sampling (4), from existing data sets to study the properties of new designs, and they are careful to limit their design and/or endpoint conclusions to phase II evaluations of the growth inhibitory agent sorafenib. However, good trial designs will depend on the true associations between specific treatments and patient outcomes, and it is difficult to assess their impact from only 2 clinical trials in 1 advanced disease setting. To make more general recommendations, we believe it is prudent to extend the assessment to evaluate inhibitory agents in which high levels of tumor regression are not anticipated. Data sets representing hypothetical phase II studies are simulated (or sampled) from a sufficiently large number of patients prior to undertaking a phase III study. Although the rough sample size calculations above did not require simulation, for a more complete assessment (potentially including futility monitoring), one could use sampling to study the impact of the varying effect sizes by drawing samples from a model that approximates the phase III outcome data, but with a parameter that ranges across interesting therapeutic effect sizes.

**The Magnitude or Size of the Therapeutic Effect**

The positive study chosen for simulation (sorafenib vs. placebo) showed a very large difference in PFS over the first several months. Based on visual inspection of the plots presented by Escudier and colleagues (2), the 3-month PFS rates for both arms were approximately 68% versus 42%; assuming exponentially distributed PFS, this corresponds to an HR of approximately 2.25. The difference in log-tumor size was also impressive, which is somewhat striking for a supposed cytostatic agent. For cases in which the therapeutic effect is still clinically important but more modest in magnitude, a randomized phase II study with good power would need to have a significantly larger sample size for testing new agents. For instance, for an effect size of only 60% on the log scale seen in this study (HR = 1.6), it would transform the 25-patient-per-arm phase II study, to an approximately similarly powered study of 70 patients per arm. In addition, even if only very large effect sizes are of clinical interest, one would still want adverse event data on a sufficiently large number of patients prior to undertaking a phase III study. Although the rough sample size calculations above did not require simulation, for a more complete assessment (potentially including futility monitoring), one could use sampling to study the impact of the varying effect sizes by drawing samples from a model that approximates the phase III outcome data, but with a parameter that ranges across interesting therapeutic effect sizes.

**The Statistical Relationships between Phase II and Phase III Endpoints**

The authors show the potential gain in power by using a continuous endpoint over the discrete endpoint of 90-day PFS (yes vs. no). The statistical associations between this endpoint and the endpoint used for the subsequent phase III study are critical for determining the performance of the phase II study. For instance, suppose small changes in tumor size are related to treatment but not related to overall...
PFS (primary objective of the phase III study). Such an association can lead to false positives with respect to selecting promising agents for a phase III study; this may be the situation described by the authors for the null AE941 study. More generally, the class of agents under consideration may influence the nature of association of tumor size with other endpoints such as PFS and survival and may even affect the measurement-error properties of imaging methods. Although it is important to study these issues on the basis of realistic historical data, it is very difficult to sort out the sensitivity of these assumptions without additional modeling. Mathematical studies, including simulating or sampling from models that probabilistically link tumor size, PFS, and overall survival, and how those relationships may vary with respect to the actions of specific agents, are needed to appreciate the impact of using a novel endpoint prior to implementation of that endpoint in clinical trials.

The Probability of Generating a Positive Phase III Study

The authors note that, for this example, the randomized phase II design and log-tumor ratio endpoint leads to declaring a regimen promising when, ultimately, the phase III study will be negative, with a false-positive rate of approximately 25%. A goal is to achieve a high probability of identifying regimens at phase II that will be effective in phase III trials in the presence of a potentially relatively low prevalence of truly effective regimens. Rubinstein and colleagues (5) note that it is not just high power, but rather the balance between power and false-positive rates that guides the chance that a phase II trial will ultimately lead to a positive phase III study. For instance, one could generate hypothetically effective and ineffective drugs from each of the studies (e.g., 10% effective), sample as before, and tabulate the fraction of treatments declared positive at phase II that are truly effective [denoted as the trial positive predictive value (PPV)]. In this case, one can use arithmetic rather than simulations and see that a design with 92% power and a false-positive rate of 0.25 leads to a PPV of 29% for an effective agent. However, another design, with a lower 75% power and a false-positive rate of 0.08 (corresponding to the randomized phase II on PFS in ref. 1), leads to a higher chance of effective treatment for a phase III study; PPV = 51%. Interestingly, a single-arm study with 55% power, but a false-positive rate of 0.01, would have substantially higher PPV, but with the downside of missing substantially more good agents. Figure 1 gives more general results. Although we are not suggesting that the 1-arm response design is the best choice in this setting, it emphasizes not only the importance of power for phase II studies, but also that of type I error. Furthermore, it highlights the need for increasing the fraction of promising agents for improving the PPV of phase II studies. For instance, some single-arm testing, in which appropriate historical data are

![Figure 1. The proportion with respect to phase III endpoint of positive phase II trials that are truly positive (i.e., not false positive) as a function of the percentage of active agents undergoing phase II testing, the false-positive rate (fp), and the true positive rate (tp) of the design. The properties fp and tp depend on the chosen phase II design and how the phase II and phase III endpoint models jointly depend on the treatment assignment for the specific disease.](image-url)
available, could be an effective filter prior to undertaking a randomized phase II trial.

Sharma and colleagues provide further motivation for statistical modeling and simulations to assess phase II designs (1). Ultimately, the optimal design for a particular disease and treatment setting, including whether it is single arm, randomized, or uses alternative endpoints, depends on many assumptions, which can be evaluated with the appropriate statistical strategies.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Grant Support
This work was supported, in part, by the U.S. NIH through R01-CA90998 and P01 CA53996.

Received February 19, 2012; accepted March 2, 2012; published Online-First March 8, 2012.

References
Choosing Phase II Endpoints and Designs: Evaluating the Possibilities

Michael LeBlanc and Catherine Tangen


Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-12-0454

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2012/04/11/1078-0432.CCR-12-0454.DC1

Cited articles
This article cites 4 articles, 2 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/18/8/2130.full.html#ref-list-1

Citing articles
This article has been cited by 2 HighWire-hosted articles. Access the articles at:
/content/18/8/2130.full.html#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.