Examining Heterogeneity in Phase II Trial Designs May Improve Success in Phase III

By Rabiya S. Tuma

The failure rate for late-stage clinical trials in oncology is higher than that for any other area of medicine. Between 1991 and 2000, for example, 55% of phase III trials in oncology failed, compared with 30% in infectious disease and 20% in cardiology, according to a study by Schering-Plough and Merck investigators. A separate study from researchers at Princess Margaret Hospital in Toronto estimated that between 1998 and 2003, 85% of the phase III trials that tested new therapies for solid tumors failed to meet their primary endpoint.

Although many factors contribute to this relatively low success rate in oncology, some experts suggest that a key problem is the tendency of researchers to ignore patient and tumor characteristics in phase II trials.

“Human beings are very heterogeneous, but for some strange reason when people design clinical trials they act as if people are peas in a pod,” said Peter Thall, Ph.D., a biostatistics professor at the University of Texas M. D. Anderson Cancer Center in Houston.

Phase II trials are designed to look for evidence that a new treatment has some antitumor activity in a given disease. Because these trials are generally viewed as exploratory and are not used as definitive evidence of effectiveness, most research teams enroll a limited number of patients. But that small sample size leaves trials particularly vulnerable to incorrect outcomes that cannot be reproduced in subsequent large randomized trials. Statisticians and clinical trial specialists are working on phase II trial designs to counter the problem, but not everyone agrees that it can be solved in phase II—it may simply be an issue that must be tackled with large phase III trials that are designed more smartly.

There are two major problems with lumping patients together. First, researchers are usually comparing the results of their phase II trials to historical controls, Thall said. Thus, intertrial variability can influence whether researchers continue to move a drug through clinical development or kill it. Second, individual patient characteristics—what statisticians call covariates—can affect the success or failure of a treatment. “You can think of treatment as an attempt to overcome covariates,” Thall said. What may appear to be a treatment difference may be due to an imbalance in covariates in trial arms—or to differences between patients in different phase II trials that are being compared.

Examples of important covariates include age and disease severity, which most researchers record and try to account for in their analyses. More problematic are the unknown variables, which Thall refers to as latent covariates. These variables may have a tremendous effect on a drug’s activity, but they may not even be recognized at the time of the trial. In large randomized trials, patient heterogeneity is less of a problem because the individuals’ differences—both known and unknown—tend to balance out between arms, as a result of the randomization. In small trials, even randomized ones, the covariates are likely to be unevenly distributed, which can skew the trial results.

Thall illustrated the problem by looking at what could happen if researchers ignored heterogeneity in a 100-patient trial in which a patient’s biomarker status affected the likelihood that she would respond to the new therapy. In the modeling experiment, the biomarker-positive and biomarker-negative patients were assumed to be equally common in the population. Computer simulations showed that, if the new treatment improved the response in biomarker-positive patients by 15% but had no effect on the response rate in biomarker-negative patients, the trial would be stopped nearly half the time because it looked like the drug had no effect. By contrast, if the two patient groups are treated separately, enrollment in the biomarker-negative group would be correctly stopped early three-quarters of the time because of a lack of response, while enrollment in the biomarker-positive group would be incorrectly stopped only 10% of the time. Overall, the researchers were much more likely to get the correct answer when they accounted for a covariate that influenced response than if they ignored it.

“Generally speaking, the trial is a disaster,” Thall said, noting that this is exactly how many phase II trials are run in the real world. “So pardon me if I point out that the way most phase II trials are conducted in settings where you have heterogeneity and you ignore it is just plain wrong.”

Other experts agree with this analysis. “Early trials often have a nonrepresentative patient population that may tend to have a better response rate,” said Rebecca Betensky, a professor of biostatistics at Harvard School of Public Health. “That would lead to an overly optimistic design for a larger trial, and you wouldn’t see the effect you thought you might.” That could lead to a negative phase III trial and rejection of a drug that actually had some activity in some patients.

Remarkably, Thall and other experts said that many phase II trials ignore patient heterogeneity in favor of moving into phase III trials as quickly as possible, despite the evidence that such a path can lead to...
failure. “Most of the phase II trials that are conducted on this planet—in the oncology community at least and a lot of other ones as well—ignore patient characteristics,” Thall said.

Mark Ratain, M.D., professor and associate director for clinical sciences at the University of Chicago Cancer Research Center, notes that there is a priority for speed over accurate phase II trials. “Everybody wants to do a 30-patient trial and proceed to phase III. They don’t care really what the result is; as long as they don’t have an excuse to kill the drug, they will go to phase III.”

**Randomized Phase II Trials**

The traditional way to minimize the effect of patient heterogeneity is to randomly assign patients to experimental treatment and control arms. But even with randomization, some patient characteristics are likely to be unbalanced in a small trial. And companies may hesitate to run randomized trials because it is harder to recruit patients for those than it is to a single-arm trial in which they know they will receive the experimental agent, said Rachel Humphrey, M.D., vice president of immunology and oncology at Bristol-Myers Squibb in Wallingford, Conn. The extra work doesn’t mean that the data will be definitive.

Despite those issues, the results of a randomized phase II trial are likely to more accurately estimate the size of the new therapy’s benefit and enable researchers to design an appropriately sized phase III trial, Humphrey said. The key, though, is to run the phase II and III trials in the right patients. Then “you are better off than with a single-arm phase II that might mislead you,” she said.

Humphrey helped lead the development of sorafenib for kidney cancer while at Bayer, and she credits the success of that project to the use of an unusual phase II trial design, called a randomized discontinuation trial (RDT). Unlike most designs, the RDT allows the researchers to start with a wide-open eligibility and let the patients’ clinical responses tell them which disease should be tested further with the new drug.

In the sorafenib trial, the team allowed patients with any type of refractory advanced solid tumor to enroll in the trial as long as they had a measurable tumor. As is typical of an RDT, all patients received a standard dose of the drug for the first 12 weeks. At that point, the patients who showed evidence of tumor shrinkage continued on the drug, and those who had documented tumor growth were taken off the study. The patients who had stable disease—which could be due to the drug or the natural history of the disease—were randomized to either 12 weeks of placebo or sorafenib. The endpoint of the trial was the rate of disease progression in the randomized patients.

Just by looking at the first 12-week assessment, Humphrey’s team had a lot of information. For example, preclinical models suggested that the drug would be useful in colorectal cancer, but less than 15% of the colorectal cancer patients were randomized. “They were almost all progressing at the end of the run-in period,” she said. So the researchers stopped enrolling these patients into the trial.

By contrast, 70% of renal cell cancer patients were either randomized with stable disease or had detectable tumor
shrinkage after the first 12 weeks. “What we learned very early—and could not have been guessed at by prior data—is that renal cell cancer is particularly sensitive to sorafenib. The fraction of patients being randomized was very high,” she said. Looking at the 24-week assessment in the randomized patients, the team could see that there was a substantial progression-free survival benefit in those who remained on the active drug.

The trial design isn’t perfect, Humphrey noted, because it can be used only when the side effects of the active drug are mild enough that the switch to placebo can remain blinded. Also, because the randomized population was enriched for likely responders, the magnitude of the effect of the drug is likely to be exaggerated over a randomization of all patients with that disease. So researchers need to be conservative in designing the subsequent phase III trial. “I would, at this time, recommend the RDT as a screening assay for a variety of tumor types and to get a feel for whether the drug has benefit,” Humphrey said. The design might also be a good way to test heterogeneity within a cancer type, she said.

Betensky agrees that the design isn’t perfect, but it does solve some of the problems, especially the challenges that come with molecularly targeted therapies. “If there is one subgroup of patients that will respond to the treatment—and that is becoming more and more likely with genetically-designed treatment—but if you don’t know that subgroup ahead of time, they may get lost in a big trial because they are only one segment of a large population.”

In fact, Betensky and her colleagues found that ignoring molecular heterogeneity in brain cancer patients could lead to exactly this type of problem—and to rejection of a drug that would work for some patients.

Creative Designs
When randomization in phase II doesn’t work, statisticians and clinical researchers are developing a variety of other new phase II trial designs to deal with patient heterogeneity. The designs vary in strategy but share the key feature that they try to tackle patient heterogeneity straight on rather than as an afterthought or ignoring it altogether.

In a recent case, Thall used a complex statistical design, called a hierarchical Bayesian model, to test the activity of imatinib in 10 different subtypes of sarcoma, which are similar but not identical malignancies. All sarcoma patients were enrolled into one trial, but the drug’s activity was evaluated separately for each subtype. The researchers set up a statistical model that changed when a patient with one subtype respond or progressed. As accrual and responses accumulated, the statisticians continuously recalculated the likelihood that the drug would be effective in each subtype. When the probability of a positive trial outcome became too low for a given subtype, the investigators stopped accruing patients with that subtype and shifted their resources to the other subtypes that looked more promising.

This “borrowing strength” approach decreased the false-positive and false-negative rates across the board, Thall said. By making a modeling assumption at the outset of the trial, the team bought themselves substantial power because as they learned about one subgroup of patients, they were learning about the others, too. “You’re not getting something for nothing; you’re just not ignoring your information,” he said.

The results of the study showed that response rates to imatinib are higher in some sarcoma subtypes than others, which can guide future trials and limit the likelihood of negative phase III trials. “In terms of resource allocation it is immensely important,” Thall said. “Not pursuing false leads is just as important as developing a drug where it does work.”

The statistical models for this approach have been around for decades, but few people use them because designing such a trial is difficult and time consuming compared with designing standard trials. “Technically, statistically, it is hard to account for patient covariates,” Thall said. For a standard phase II trial, computer software already exists and can be set up and simulations run in a few minutes. But for the sarcoma trial, the programming and modeling required took several months. That cost may not seem like much if it prevents a failed phase III trial down the road, but it is an allocation of resources that few clinical investigators or trial sponsors have been willing to accept thus far, Thall said.

Betensky and Ratain concurred that getting researchers and sponsors to accept new approaches requires some persuasion, though successful examples like sorafenib and the RDT that led up to its approval helps.

The common theme for each of these approaches is to maximize the value of a trial by planning ahead and keeping track of patient variables. Response rate is often the primary endpoint for phase II trials, but recording other information such as toxic effects and toxicity rates, progression-free and overall survival, and general patient characteristics is also worthwhile, Thall said. After all, phase II trials help identify the patient groups likely to respond, so ignoring or disposing of data that might ultimately point in the right direction is not reasonable.

“I think phase II trials, or early-phase trials, can be done much more sensibly if you are willing to work a bit harder, and you can really get a lot more information out of a trial of a given size,” Thall said. And that might just might increase the success rate for phase III trials.

© Oxford University Press 2008. DOI: 10.1093/jnci/djn006