Are Bayes’ Days Upon Us? Statistical Methods Could Change the Conduct of Clinical Trials

A group of statisticians who dislike the null hypothesis as much as struggling graduate students do are teaming up with clinical researchers to redefine the questions that can be asked and answered in clinical cancer trials, using statistical methods first put forth by an 18th century minister.

Rev. Thomas Bayes’ theorem had lofty goals from the time it first arose in the 1760s—a mathematician and theologian who published it after Bayes’ death thought the theorem could be used to prove the existence of God. Now it is inching into vogue for some medical research. Proponents say Bayesian methods can tell physicians how much better a drug is than another, rather than merely if it works better, and it will ensure better treatments for trial volunteers. But some skeptics argue that pure Bayesian methods aren’t appropriate for drug trials because they explore the effectiveness of drugs rather than confirm that the drugs work.

The U.S. Food and Drug Administration is developing guidelines that allow researchers to use Bayesian methods in trials for medical devices such as stents or pacemakers. Although the comment period ended in late August, the complete guidelines probably won’t be issued for several more months. Bayesian enthusiasts hope such rules for cancer studies aren’t far behind, though there are no immediate plans for relevant guidelines.

Bayesian analysis is designed to test the likelihood that a prior hypothesis is true given new evidence, and it gives you a probability based on that prior hypothesis. This future probability, proponents say, is rife with useful information such as how much better one therapy might be than another, and it can allow researchers to watch the study in progress and adjust it. Bayesian statistics requires considerable computing know-how and transparency of each assumption, however. For these reasons, it has remained squarely outside general use by clinicians.

The “frequentist” methods currently used in clinical trials set up a problem with a yes-or-no answer, but they have trouble with the more subtle “how well” question (though the methods can answer whether a treatment works better than can a preassigned number).

Nothing’s Easy

The allure of Bayesian-based clinical trials is not their simplicity. They require substantial planning, are often more work, and don’t always mean you can use fewer subjects. Advocates say the real draw is that such trials are intuitive, can answer real-world questions by using all available information, and can prevent trial subjects from getting inferior treatment.

“Physicians are all Bayesians,” says oncologist Robert Benjamin, M.D., of the University of Texas M. D. Anderson Cancer Center in Houston, which has embraced the methods. Each time doctors get a new piece of information—how well a patient responds to an antibiotic, for example—they incorporate that knowledge into their brains and use it the next time they make a decision.

Traditional studies, using frequentist statistics, sometimes use contorted logic and answer one primary question at a time. To set up such a study, the researcher assumes that the “null hypothesis” is true—for example, that a new drug quells pain no better than an old drug—and hopes that the data will show that the null hypothesis is unlikely to be true. If the new drug has a positive effect, frequentist statistics indicate how likely it is that the effect occurred by chance, the P value. If the probability that chance was to blame is less than, say, 5% (a P value of 0.05), the researcher has a positive answer, a certain degree of confidence, and possibly a new, effective drug.

On the other hand, Bayesian methods start with the data that have already been gathered, called prior information, which can be gleaned from a variety of sources. Prior information is the starting point, and data gathered during the trial either support it or contradict it. For example, previous studies might have shown that compound A is very good at clearing tumors and so it has a high likelihood of success in certain patients. A Bayesian computer model takes that prior information and churns out the probability that the treatment works. If statisticians or clinicians decide that the prior information is sound based on, say, decades of experimental data, then they can weigh the information heavily in the computer model. Strong data from the trial will be required to change it. If the experts think the prior information is little more than a best guess, then they can give very little weight to the starting point. The Bayesian model can also calculate probabilities midstream and determine how likely it is that the treatments will perform the same until the end of the trial.

The key benefit of Bayesian methods, says statistician Donald Berry, Ph.D., also of M. D. Anderson, is the ability to change the study’s course if the subjects’ welfare is at stake—using so-called adaptive randomization. In a multiyear frequentist study, new patients will have the same chance of being enrolled in either group, regardless of whether the new or old drug is performing better. This approach can put patients at a disadvantage. “This is dumb, to subject
patients to really ineffective therapy,” Berry says.

A Bayesian model, on the other hand, can periodically show researchers that one arm is outperforming the other and then put more new volunteers into the better arm. Also, advocates say Bayesian methods get medical researchers more involved. For example, a Bayesian trial on a cancer drug for children might pull information from how well the drug performs in adults, which requires formalizing an opinion about how similar the disease is between adults and children.

“These are interesting questions that are not answered by statisticians,” says statistician Steven Goodman, M.D., Ph.D., of Johns Hopkins School of Medicine in Baltimore. “Imagine if I asked instead, ‘Okay we know something about the adults—what should be our error rate on the child’s trial?’ The oncologist is sitting there wondering, ‘The what?’”

“Bayesian approaches empower the people with the real expertise to be at the table.”

**Priors, False Positives, and P values**

But some researchers argue that Bayesian methods aren’t rigorous enough to base regulatory decisions on. Robert O’Neill, Ph.D., director of the FDA’s Office of Biostatistics, says that Bayesian methods are great—and already in use—for exploratory studies, for example, to determine drug dosing. But he has two concerns with using the method with large confirmatory studies such as phase III trials: predicting false positives and deciding what data to use as prior information.

An important part of frequentist statistics is that it can show the probability that a drug looks like it’s working but is not actually working, so-called false positives. A P value of 0.05 means that 5% of the time a positive result will actually be wrong. “This is an important index of uncertainty,” says O’Neill. You don’t get a similar number from Bayesian studies.

But frequentist properties such as false positives and false negatives can be pulled from Bayesian analyses, says biostatistician Peter Thall, Ph.D., with the M. D. Anderson group. “The false-positive rate is a property decided on before the trial is ever run,” he says.

And prior information, according to Bayesian hardliners, should either be derived from previous research or not weighted heavily. But O’Neill says that’s not a given, especially if the prior information is based on a strong opinion or intuition, which can vary from person to person. “The prior information may or may not be empirically determined.” In fact, how to handle prior information is just the sort of guideline researchers need from the FDA, says statistician and Bayesian proponent Stacy Lindborg, Ph.D., of Eli Lilly, especially since two Bayesian models can churn out different results from the same data if the prior information is different enough. What’s worse, the prior information can sabotage the outcome or increase the study length if it is wrong and weighted heavily. “There’s been a long-standing concern about the relevance of prior information,” O’Neill says.

But the alternative is not acceptable to some researchers who think P values are overrated. For example, researchers might throw out valid treatments that don’t measure up to an arbitrary, predetermined point, such as a therapy that improves survival by 45% instead of a predetermined 50%. “Using the traditional approach, we have overlooked important treatment advances,” says Benjamin.

And hematologist Elihu Estey, M.D., at M. D. Anderson says that many researchers misunderstand P values and might be undercutting their results with their eagerness to see how things are going. He says many people don’t realize that every time you look at frequentist data before the study’s over, you are devaluing your P value and will require more data than originally planned. “If you look 10 times, you have to divide your P value by 10,” he says.

**Bayes on Trial**

Despite the reservations, researchers are trying out Bayesian statistics in clinical trials. Oncologist Robert Maki, M.D., Ph.D., of Memorial Sloan-Kettering Cancer Center in New York, and colleagues have undertaken one of the first oncology phase III studies using Bayesian statistics and adaptive randomization. Their work on soft-tissue sarcomas, part of which has been presented at American Society of Clinical Oncology conferences, compared a two-drug combination to one of the drugs given over a new dose schedule.

“Our concern was that a large number of patients would receive inferior treatment” under a traditional study, Maki says.

The study started out by dividing the volunteers into two treatment groups and then monitoring them for any tumor growth. Stable disease or any shrinkage was considered good. Every 6 weeks, the researchers would review the data and if one treatment showed an advantage, more patients would be steered to the better treatment. “If both treatments were equal, the number of patients in each arm would shift back and forth, and at the end of the day it would be a wash,” says Maki.

But that’s not what happened with their sarcoma patients. By the end of the study, the team was shunting more people toward the combination therapy at a ratio of 3 to 1. Maki estimates that the team prevented about 20% of the volunteers from getting the inferior treatment. “That’s better for patients, for physicians to be able to tell patients when recruiting them, and we get answers more quickly,” he says.

In the midst of such fervor, Lindborg reminds us that “No one’s advocating altering the paradigm.”

O’Neill and others agree—there’s a place for both approaches. And for now, the FDA has no plans to develop guidelines for Bayesian methods in clinical trials, though they will be looking at the adaptive-designs offshoot. But Lindborg is encouraged by the openness she sees the agency is having with the discussion and understands the FDA’s prudence.

“While we may be convinced that this is truly in the best interest of the patient, we need to understand that the FDA is concerned about how a change on any level of statistical methodology might affect their ability to make the right decisions,” she says. Although interest is growing, for now Bayes’ ways will have to settle for the easy questions—those theological rather than biological.

—Mary Beckman

© Oxford University Press 2006. DOI: 10.1093/jnci/dj456