WHEN IS SIGNIFICANT NOT IMPORTANT?

Finding Clinical Meaning in Cancer Data

By John Dudley Miller

Researchers studying treatments for advanced pancreatic cancer recently showed that overall survival "was significantly prolonged" by combining erlotinib and gemcitabine rather than giving gemcitabine alone. The study was statistically significant at the 95% confidence level and may influence future treatment decisions.

But there is one small problem. Giving erlotinib and gemcitabine together added only 10 days to the average patient’s life, according to the study in the May 20 issue of *Journal of Clinical Oncology*. Plus, the combination increased the side effects patients experienced, most notably diarrhea. Those facts raise an important question—do 10 extra days of life constitute a meaningful or desirable increase, given the months of reduced quality of life (QOL) that patients might endure to get them?

This example illustrates that statistical significance and clinical importance—the degree of improvement a new treatment must make for it to be clinically meaningful—are different concepts. Statistical significance at the $P < 0.05$ level means only that there’s less than a 5% chance that the difference observed in the study, or a more extreme difference, would not have been observed if there is truly no difference between the study groups (i.e., if the null hypothesis is true).

But many doctors misinterpret this finding. Statistical significance does not imply that the difference in outcome between the treatment and control groups is large, according to statisticians. Nor does an even higher degree of statistical significance, say, $P < 0.01$, mean that there is a greater effect on patients. This counterintuitive conclusion follows because statistical significance depends largely on how many patients reach the primary endpoint in a trial. If more patients are enrolled, a smaller and less clinically important difference can be statistically significant. These facts have led statisticians and medical researchers to look for new ways to analyze data that will have more meaning in the clinic.

"Just because it’s statistically significant doesn’t mean that it correlates to importance in the clinical setting," says Amylou Dueck, Ph.D., a mathematician who works as a QOL researcher and statistician at the Mayo Clinic in Scottsdale, Ariz. A study can be statistically significant but clinically trivial, meaning that it provides such a small improvement that doctors cannot justify using it to make treatment decisions.
QOL Base

Until the mid-1980s, little work had been done to find objective ways to measure clinical improvement and test for clinical significance. Even now the work is incomplete, so that there is no general consensus about which methods are best. Clinical significance is a subtopic of the larger field of QOL indicators. The raw data for clinical significance tests is the difference in overall QOL scores between patients given a new treatment and ones given the standard treatment, according to Gordon Guyatt, M.D., a biostatistics professor at McMaster University in Hamilton, Ontario. (QOL is often measured with a test called the SF-36 tool.)

Several numerical methods have been developed over the past 20 years to determine how much of an improvement in QOL scores constitutes a clinically significant change. One reason that no consensus has emerged is that they all involve somewhat arbitrary definitions. One widely used method defines a “moderate” effect as a change of at least one-half the standard deviation of patients’ baseline QOL scores before a trial starts. The method defines a small effect as 0.2 standard deviations and a large effect as 0.8 or more standard deviations. Another popular test defines small, moderate, and large effects as ones that are at least 3%, 8%, and 13%, respectively, of the theoretical range of any QOL assessment tool.

Guyatt, a leader in measuring clinical significance, is developing other measures of clinical importance. But in the meantime, he thinks that researchers writing journal articles and presenting at meetings need to report at least two elementary measures: the number of patients needed to treat (NNT) to save one additional life and the reduction in absolute risk that a new treatment causes.

Absolute Risk and NNT Matters

Absolute risk refers to the percentage of patients who die or reach some other bad outcome when given a particular therapy. The absolute risk reduction is the percentage of patients who don’t die when given the new treatment who would have died on the standard treatment. So if 20% of patients die when given a standard treatment but only 10% who get the competing treatment die, the absolute reduction in the risk of dying for patients getting the second treatment is 10%.

NNT is defined as 1 divided by the absolute risk reduction of a new treatment compared with an old one. Given a 10% reduction in absolute risk, the NNT to save one patient’s life is 1/0.1, which equals 10. That means for every 10 patients given the newer treatment, one additional person will be alive at a given point who would have died had he or she been given the standard treatment.

The NNT concept applies not only to saving lives but also to other positive outcomes, like the number of patients whose cancers go into remission or the number who survive 1 year. All other things being equal, the lower the NNT, the more clinically significant a new treatment is. The number-needed concept can also be used to compute the number needed to harm when giving a new treatment, such as the number necessary to cause one additional patient to suffer diarrhea as a side effect.

“I’ve always viewed [NNT] as a handy way for a clinician to get a feel for how useful this particular treatment is going to be in their population,” says Richard Davidson, M.D., a professor at the University of Florida College of Medicine. Among the largest, most important clinical trials, like those comparing the effects of different cholesterol-lowering drugs in lowering heart attacks, Richardson believes that “there has been a tremendous improvement in the numbers of papers that address some aspect of clinical significance” over the past couple of decades.

But Guyatt disagrees. He recently examined many published clinical trials to see if NNT, absolute risk, or other clinical significance measures were reported in their abstracts, the only part of studies likely to be read by most busy clinicians. “Very few of the abstracts actually presented the results so that you would be able to estimate the size of the effect in terms of its importance,” he says, only about 20%. The only good news is that 10 years ago, “it would have been zero, or next to zero.” Most of the change has taken place in just the last 5 years, he says.

Unfortunately, Guyatt says, most clinical trials still focus too much on relative risk, a measure that can make new treatments look much better than they actually are. If only 1% of patients on standard therapy survive a cancer but 1.5% of patients getting a new therapy beat it, then the new drug lowers the relative risk of death by 33%, an apparently huge amount. Yet it saves only an additional 0.5% of patients, one in every 200.

Big Trials, Little Results

Dueck, Davidson, and Guyatt agree that one of the most unfortunate outcomes of overreliance on statistical significance and the underreporting of clinical significance is that doing so skews which studies get published. Unimportant but highly statistically significant studies may win out over potentially important studies that did not turn out to be statistically significant because they enrolled too few patients.

For example, take two studies comparing the same drug regimens. One enrolls 100 patients and another enrolls 400. The larger trial will be able to detect as statistically significant an observed difference...
half as large as that in the first. A third study with 900 patients will be able to detect and find significant a difference one-third as large as that in the first study. All three studies could have shown the same trend, but if the actual difference in patients’ response is small, only the 900-patient study will find it statistically significant.

Taken to its extreme, the dependence of statistical significance on sample size means that small differences will be found statistically significant if researchers are willing to enroll many patients, i.e., 1,000 or more. “Usually with such a large sample size even little teeny differences can be statistically significant,” Dueck says. These are less likely to be clinically significant than statistically significant differences in smaller studies.

The dependence of statistical significance on sample size can create the opposite problem, too: A clinically significant difference may not be statistically significant if too few patients have been enrolled in the study. In a May 2002 article in *Arthritis and Rheumatism*, researchers reported that combining methotrexate and prednisone did not decrease treatment failures in giant-cell arteritis patients relative to patients who were given prednisone alone. The *P* value was 0.26, nowhere near the somewhat arbitrary *P* < 0.05 that statistical significance demands.

But adding methotrexate to prednisone lowered treatment failures from 77.3% to 57.5%, a drop of almost 20 percentage points. According to Davidson, who uses this study to teach medical students, that clinically meaningful difference was not statistically significant most likely because the researchers enrolled only 98 patients in their trial, the minimum number they calculated would be necessary to detect a decrease of at least 50% in treatment failures. “Because the authors had set statistical significance at a 50% improvement, it did not reach statistical significance,” he says. No larger trial was later conducted to prove the positive result, so it remains uncertain whether the result is reliable.

“Negative” results that may nevertheless be potentially clinically relevant are common, even in the best medical journals, Richardson says. In a 1994 study in the *Journal of the American Medical Association*, David Moher, and colleagues examined 70 randomized clinical trials that found non–statistically significant differences between patients assigned to two different treatment options and were published in *JAMA*, the *New England Journal of Medicine*, or the *Lancet* in 1975, 1980, 1985, and 1990. Only 36% of these studies included enough patients to give them an 80% chance of detecting a 50% improvement of one treatment over the other. Only 16% could have detected a 25% improvement. Even fewer could have detected a 20% improvement like the one that adding methotrexate to prednisone produced in arteritis patients.

So if too few patients can hide clinical significance and too many can make trivial differences statistically significant, what should a researcher do? Dueck and Guyatt advocate studies that are not too big and not too small. Most studies already satisfy that requirement, because they don’t go forward if they don’t have the resources to enroll enough patients to prove statistical significance, and they don’t have enough funding to pay for many more patients than that.

**Significant, But Not Clinically**
The pancreatic cancer study, where patients were treated with erlotinib and gemcitabine or gemcitabine alone, included 569 patients, enough to find a mere 10-day improvement in survival statistically significant. Is an improvement that small clinically significant?

Not to the researcher who ran the study, Malcolm Moore, M.D., of the Princess Margaret Hospital in Toronto. In an editorial in the June 1 issue of the *Journal of Clinical Oncology*, he wrote that in pancreatic cancer studies, where the average patient lives only 6 months, “if 600 randomly assigned patients are required to demonstrate a difference, then the clinical significance of that result is debatable.” He advocates conducting smaller studies for that disease, perhaps half that size, so that more drugs and drug combinations can be evaluated. Those that show statistical significance even with modest sample sizes will be more likely to be clinically significant too, he says.

That’s not the view of Alan Sandler, M.D., an associate professor of medicine at the Vanderbilt–Ingram Cancer Center in Nashville. It’s a mistake to judge clinical significance by using only the average increase in survival time that a new drug provides, says Sandler, who is currently leading several phase I, II, and III studies of treatments for lung and esophageal cancers as the Eastern Cooperative Oncology Group’s thoracic committee cochair.

Instead of dismissing as useless studies that find small changes in median survival times, Sandler suggests looking at 1-year survival rates, which in the pancreatic cancer study increased from 16% to 23% when erlotinib was added. That difference is a big deal to those patients, he says. The NNT to allow one additional patient to live for a year is actually small, about 14 patients, so the expense of keeping one additional patient alive that long would be low relative to studies with higher NNTs. So some clinicians trying to do the best for their patients may decide that 10 days is worth it, while clinical trial researchers may not agree.

“I think that the concept of clinical significance is clearly an important one,” Sandler says, “but I don’t know that we’ve all reached a consensus as to what it really is.” To him, measures like NNT are more relevant to societywide decisions about health care costs and less useful to physicians like himself who must make decisions about individual patients.
Sandler sees a gulf between the values of some clinicians and clinical investigators. “I’m not so sure we’re all on the same page … as to what is clinically relevant or not,” he says. So pancreatic cancer studies run through the Eastern group do not report measures of clinical significance. Instead, they consider a new treatment to be important if it reduces the relative risk of death by 25%, no matter how low the absolute reduction in risk might be or how large the NNT to save one extra life turns out to be.

From Guyatt’s point of view, this widespread use of such an ad hoc measure of importance as a 25% relative risk reduction represents a failure by clinical significance researchers to present their results in ways that clinicians can easily interpret. He thinks the “gold standard” measure for easily grasped explanations, which has not yet been created, would compare the percentages of experimental and control groups surpassing some clinical significance goal, because clinicians already understand and use percentages.

“I am convinced that to really get clinicians to be able to relate to [clinical significance], we need to do that.”

© Oxford University Press 2007. DOI: 10.1093/jnci/djm283