RANDOM BY METHOD

Dr. Thérèse Stukel takes a cue from econometrics for a new approach to epidemiology

Early in 2008, editorial staff at *The Lancet*, one of the oldest and most prestigious general medical journals in the world, asked its international advisory board to nominate the research papers that made the greatest contribution to clinical research in 2007. To preempt the inevitable criticism of this "paper-of-the-year" exercise, a feature of the journal since 2003, the editors took care to note in a jocular preamble to the list of 12 finalists that the ambitiousness of the process "invites incredulity," given the extraordinary breadth and complexity of research published each year. On the other hand, they pointed out, in lieu of workable judging criteria the process is "an opportunity to celebrate research and researchers, enriched by the passion of colleagues about the papers that excited them most." Considerable if informal grounds, in other words, for some serious global boasting should one's publication be nominated for the award.

Dr. Thérèse Stukel

THE 16% REDUCTION SHE SAW INSTEAD, WHICH SQUARELY MATCHED RCT RESULTS FOR THE SAME PROCEDURE, WAS EVIDENCE THAT HER APPROACH AND INSTRUMENT WERE VALID.

News of precisely that honour traversed the grapevine to Dr. Thérèse Stukel a few weeks later, for her *The Journal of the American Medical Association (JAMA)* paper, "Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods." Stukel, a senior scientist in clinical epidemiology at Sunnybrook Research Institute (SRI), as well as at the Institute for Clinical Evaluative Sciences, was shocked. "It was a huge honour and a big surprise," she said the following month in her office at SRI. An honour, because *The Lancet* is a high-impact journal; a surprise, because it's unusual for any medical research journal to place a statistical methods paper alongside valued medical advances. "It was a good day for biostatisticians everywhere," she said with a smile.

It has been a decent 18 months for biostatisticians since Stukel published her article. At the time this magazine went to press, the paper had been cited 34 times in peer-reviewed journal articles. Some of those citations were in methodology and epidemiology journals, but several were in general medical research publications, suggesting that the paper is having an impact where Stukel says it's most needed — among a broad audience of clinical researchers whose specialty is not statistical methods, and who rely on methodologists to ensure the validity of their observational studies.

Stukel and her team showed that conventional statistical analyses of cardiac catheterization (an invasive test that measures blood flow and heart function, results of which often lead to surgery) indicate the procedure reduces mortality by 50%. Randomized controlled trials (RCTs) — considered a much more reliable method of testing efficacy than observational studies — show a decrease of just 8% to 21%. Using an econometric technique new to epidemiology, instrumental variable analysis, Stukel came up with a decrease in mortality that strikingly matched RCT results — 16%. "We caught the perfect example," said Stukel. "Nobody believes that an invasive diagnostic test reduces mortality by 50%, so almost nobody could dispute those findings. The only response could be, 'Do we have to pay attention to this?' and 'What's going on?'"

What's going on is that many observational studies have a problem called treatment selection bias, and it's not an easy problem to fix. Treatment selection bias is the inadvertent tendency to attribute to a particular treatment the improved outcomes in a group that received that treatment, when compared to a group that didn't get it, even though the underlying reason for those improvements is better preexisting health in the group that fared better. Numerous factors can affect interpretations of baseline health, including age, education, income and other medical problems. Many studies have shown that, in some instances, physicians will perform more aggressive treatment — especially surgery — on healthier patients who are at lower risk for complications from surgery or for dying, than on patients who are at higher risk, even though the latter group may benefit more from the treatment. Reasons for this phenomenon include physicians' perceived potential for causing more harm in high-risk patients, and devotion of time and resources to patients' other problems in place of the treatment in question. As a result, those who get a certain therapy don't always have the same risk profile or prognosis for improvement and survival as those who don't get it, and this discrepancy can skew the results of a study that compares treatment efficacy in the two groups.

Randomized controlled trials are one solution to this problem. In an RCT, randomization determines which patients are selected for treatment and which aren't, so the two groups should theoretically have similar baseline characteristics and health, eliminating selection bias. This approach works well (although not without its own recruitment challenges — patients must be free of certain comorbid conditions to be eligible, resulting in underrepresentation of these), and RCTs are considered the best means of assessing therapies and procedures. The trouble is, they're expensive to run, time-consuming, potentially unethical and for some treatments, impossible. A pill is easily given as a placebo, but surgeons, for example, can't perform sham operations, and a trial to test for linkage between cancer and smoking would clearly have ethical problems. For these reasons, most studies are observational, not RCTs.

Researchers use several strategies to deal with selection bias in observational studies. They ensure good study design, which can minimize opportunities for bias, and use statistical models to adjust for baseline differences between patient groups, to make outcome comparisons in those groups more valid. One common approach, standard regression analysis, incorporates all available covariates (patient risk factors) into the model to control for patient differences. This works, provided those risk factors are the only ones at play. If there are unmeasured factors, however, for example, diet, environmental toxicity or genetics, then regression can't control for them.

Epidemiologists therefore use another set of techniques called propensity score methods. Propensity score methods classify patients into deciles of risk to determine their probability of getting treatment, then look at outcomes among patients within the resulting strata to ensure similar risk and health baselines in treated and untreated patients. The problem with this approach, as Stukel explained, is "it's subject to the same caveats as regression models — if you don't control for a certain risk factor because you can't see it, you can't measure it in the propensity score."

Propensity-based matching is a third method that restricts the analysis to a group of patients who are much closer in baseline risk and then only looks at them. But this technique has the same problem as propensity score, and as Stukel expected, it produced a similarly misleading (50%) decrease in mortality when assessing the effect of cardiac catheterization.

Stukel's instrumental variable analysis, in contrast, produced a significantly more realistic measure of catheterization's benefit, and for what appears to be a compelling reason: it mimicked the randomization of an RCT. In an earlier *JAMA* paper in 2005, Stukel showed that although regional rates of catheterization vary dramatically

Dr. Thérèse Stukel

(from 29% to 93%, depending on availability of cath labs and services, among other factors), patients themselves were largely the same across regions in their baseline risk and other potentially confounding influences. Confident that patients were similar, and then comparing outcomes in regions with high rates to those with low rates, Stukel used geography as an instrument of randomization and control: if catheterization does reduce mortality significantly, she reasoned, one would expect to see a much greater reduction in areas where it's performed regularly. The 16% reduction she saw instead, which squarely matched RCT results for the same procedure, was evidence that her approach and instrument were valid.

Dr. Muhammed Mamdani, director of the Applied Health Research Centre (AHRC) of the Li Ka Shing Knowledge Institute at St. Michael's Hospital in Toronto, read Stukel's paper and found it compelling. Mamdani has extensive training in pharmacology, epidemiology and economics, and was the director of outcomes research at Pfizer Global Pharmaceuticals in Manhattan before moving to Toronto. "I think this method of approaching observational studies in a different way, using instrumental variable analysis, is fantastic, because it tends to get at this issue of selection bias better than the traditional designs we're used to seeing," he said by telephone this September.

Mamdani is comfortable with instrumental variable analysis in part because he's seen it used extensively in econometrics, where it originated some 30 or 40 years ago. Although medical researchers have only turned to the technique in the last 10 to 15 years, Mamdani says its use is growing as technology enables more sophisticated statistical modeling and journals demand that studies pay more rigorous attention to selection bias. "I'd now believe a good instrumental analysis over a decent cohort study done in the traditional way," he said.

There are drawbacks to instrumental variable analysis, and a critical one that Mamdani, Stukel and other epidemiologists note is that it's rare to find a good instrument. Stukel admits this was her biggest challenge, and that her team considered several instruments before settling on regional catheterization rate, in part because it produced the best clinical interpretation.

Another problem is the difficulty of testing an instrument to be sure it's really functioning as it appears to be. Dr. Donald Redelmeier, Canada Research Chair in Medical Decision Sciences, and senior scientist and director of clinical epidemiology at SRI, was intrigued by Stukel's study, but not unreservedly. "An instrumental variable analysis has a big assumption about independence which is impossible to test," said Redelmeier in his office at SRI this past summer. "Sometimes it may give you an answer that corresponds with things you already know, but that may be due to offsetting errors. It may be the analysis has two errors built into it that just by a fluke cancel out, so it ends up pointing in the right direction." Regional catheterization rates, Redelmeier noted, may appear unrelated to the baseline health of study patients, but perhaps a community with a large amount of smoking, diabetes or high blood pressure will eventually bear a greater disease burden, leading to more cardiologists, more cath suites and more catheterizations, thereby confounding the apparent randomization of an instrumental variable analysis. Still, said Redelmeier, "Any methodological tool can cause problems if misapplied," and whether instrumental variable analyses will garner further attention as an epidemiological technique is "certainly a frontier question."

That this question has already found traction among clinical researchers is a testament to the credibility of Stukel's work. And while Stukel's assessment of her paper's impact is cautious—"One robin doesn't make a spring," she said—several calls from clinical researchers asking if instrumental variable analysis might work for them have boosted her optimism. Some of those inquiries didn't evolve into studies, but several are moving forward, including two at the University Health Network on blood transfusion and colonoscopy screening. Mamdani is also seeing more instrumental variable analyses, and expects that trend to continue over the next few years. As for Stukel and her paper, Mamdani said, "She's brilliant. It's one of the better methods papers that's been written in a very long time." JO

Stukel's work was supported by the Canadian Institutes of Health Research, the U.S. National Institute on Aging and the Robert Wood Johnson Foundation.