Practical Biostatistics

Size, follow-up, data analysis—good; post hoc analysis, interpretation—not so much

David L. Streiner, PhD, CPsych,[1,2] and Geoffrey R. Norman, PhD[2]

[1] Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada, and [2] Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

It's easy to know whether a critique of some article or other was written by a statistician or a methodologist—it states how badly the study was done and how incompetently the data were analyzed. Indeed, it is extremely easy to criticize any study, no matter how well it was conducted, because all applied research involves compromises of one sort or another. Well, be prepared for a surprise. In this column, we will be discussing a study that we believe was carried out well and analyzed correctly. That's not to say that we agree with their conclusions (we don't), but at least the study yields data that people can argue about without dismissing the paper as a whole.

Impressive study size and follow-up

Crawford et al[1] reported on the results of a large randomized controlled trial to determine the potential benefits of screening for prostate cancer. They conclude that "Selective use of PSA [prostate-specific antigen] screening for men in good health appears to reduce the risk of PCSM [prostate cancer-specific mortality] with minimal overtreatment." So, why do we shine our countenance upon this paper, and why do we disagree with the conclusions? Let's start off with the positives.

First, we are impressed by the study's size. Although it is not always true that bigger is better (we can indulge in all sorts of off-color jokes at this point, but we will restrain ourselves), it definitely is the case when we are studying the natural history of a disorder. This is especially true for diseases such as prostate cancer, which (thankfully) have a low prevalence. In the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial,[2] whose data were reanalyzed in the Crawford study, the sample size was 76,693 men, of whom 60 died of prostate cancer. If the study were much smaller, it still would have come up with estimates of mortality in the screened and unscreened groups, but the confidence intervals around those estimates[3] would have been much wider, meaning that we would be less sure of the numbers (after all, that's why they're called confidence intervals).

The second aspect of the study that impresses us is the follow-up rate. Research that tries to determine the natural history of a disorder has to balance two competing demands. On the one side, its duration must be long enough to allow time for the outcome to appear. The study would have been of limited usefulness if the men were followed for only 1 or 2 years, because most of the cases of cancer would have developed long after the study had ended. On the other side, the longer the study, the more difficult it is to have complete data on everyone who entered the study. People lose interest in the study and drop out, they may move or die without notifying the researchers beforehand, and so forth.

Moreover, as we've said in a previous article,[4] people do not drop out of studies for trivial reasons. In a study of this sort, they may quit because they were assigned to one group but wanted to be in the other, could not be bothered with filling out forms, or a host of other reasons. The result would be that those who remain in the study become a biased, unrepresentative sample. In this study, though, 96% of the men were available for mortality analysis 10 years after they were enrolled. Even if the other 4% were different from the completers in some substantive way, their numbers are not large enough to seriously bias the results.

Finally, the data were competently analyzed, using "competing risks regression." Without going into the details (and they are messy), what this means is that despite what we're told by health food and exercise experts, everyone is going to die of something; if you don't get knocked off by one thing, you'll be done in by something else. The problem (at least from the statistician's viewpoint) is that if

Manuscript received May 17, 2011; accepted May 18, 2011.

Correspondence to: David L. Streiner, PhD, CPsych, Department of Psychiatry and Behavioural Neurosciences, McMaster University, West 5th Campus, 100 West 5th, Box 585, Hamilton, Ontario, Canada, L8N 3K7; e-mail: streiner@mcmaster.ca.

Disclosures: The authors have no conflicts of interest to disclose.

Commun Oncol 2011;8:379–380

a person in this study died because of an infarct, this changes his probability of dying of prostate cancer (to zero, to be precise). Any analysis of outcomes must take this into account, and this study did.

The dangers of post hoc analyses

So, with all of this going for it, why are we sitting at our desks, writing this review, rather than running out getting PSA tests and biopsies? There are a number of reasons. First, as we mentioned previously, this paper is a reanalysis of the PLCO study,[2] which concluded that the risk of dying of prostate cancer was equally low in both the screened and unscreened groups. In essence, it is a post hoc analysis, dividing the groups by comorbidity status and finding benefit only among men with minimal comorbidity but no difference in the group with at least one significant comorbidity.

It is a well-known dictum in statistics that any unplanned, after-the-fact analyses are *hypothesis-generating* only, and the results should not be taken as definitive. The prime example of this is the famous ISIS-2 (Second International Study of Infarct Survival).[5] After being pressured by the journal editor to perform some post hoc analyses, the authors stipulated that they would do those suggested by the editor as long as the journal published all of the subgroup analyses. They then went to town and divided the groups according to when they were born, concluding that "subdivision of the patients in ISIS-2 with respect to their astrological birth sign appears to indicate that for persons born under Gemini or Libra, there was a slightly adverse effect of aspirin on mortality (9% increase, standard deviation [SD] 13; nonsignificant), whereas for patients born under all other astrological signs, there was a striking beneficial effect (28% reduction, SD 5; 2p < 0.00001)." The moral of the story? Don't trust post hoc analyses—if you do enough of them, something is bound to show up.

The second reason we're somewhat dubious is that our interpretation of the results is different from theirs. The authors found that to prevent one death from prostate cancer at 10 years, 723 men would have to be screened and 5 treated, and they concluded that screening was worthwhile. We look at those same numbers and draw a different conclusion. For one thing, doing 723 PSA tests to find 5 cases reflects a tremendous cost to the system, which may be better spent in other ways.

More important, though, it ignores the fact that there will be a large number of false-positive findings. Let's take the best available version of the PSA test—the complexed PSA. According to one study,[6] using the ideal cut-point results in a sensitivity of 0.85 (ie, 85% of cases are detected by the test) and a specificity of 0.35 (35% of those who are cancer-free are correctly identified).

One of the best estimates of the incidence of prostate cancer comes from a review, in which the figure was 61.8 per 100,000 white men.[7] Now let's use the lessons from previous articles[8,9] with these data and assume we screen 100,000 men. What we find is shown in Table 1. There will be 65,013 positive test results, of which only 53 will be from men who actually have cancer—a false-positive rate of over 98%! And, the test will still miss nine men who have cancer. These findings mean that nearly 65,000 men will have unnecessary follow-up tests, probably including biopsies, with all of the associated costs, risks, and side effects. Our take on things? Thanks, but no thanks.

So, the bottom line is that the study was well done. We have some reservations about the findings, because they were based on post hoc analyses, and we read the results in a different light than do the authors. (Then again, we don't make our living ordering PSA tests.) However, as we said earlier, we can argue about the interpretation, but we have a solid basis for the numbers we use in our arguments.

TABLE 1

Hypothetical results of screening for prostate cancer

	Actually have prostate cancer[a]		
Test results	**Yes**	**No**	**Total**
Positive	53	64,960	65,013
Negative	9	34,978	34,987
Total	62	99,938	100,000

[a] Based on a reported incidence of 61.8 cases per 100,000 adult white males in the United States[7]

References

1. Crawford ED, Grubb R 3rd, Black A, et al. Comorbidity and mortality results from a randomized prostate cancer screening trial. J Clin Oncol 2011;29:355–361.
2. Andriole GL, Crawford ED, Grubb RL 3rd, et al. Mortality results from a randomized prostate-cancer screening trial. N Engl J Med 2009;360:1310–1319.
3. Norman GR, Streiner DL. *P* less than 0.05: statistical inference. Commun Oncol 2009;6:284–286.
4. Streiner DL, Norman GR. Randomized controlled trials. Commun Oncol 2009;6:83–85.
5. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Lancet 1988;2:349–360.
6. Partin AW, Brawer MK, Bartsch G, et al. Complexed prostate specific antigen improves specificity for prostate cancer detection: results of a prospective multicenter clinical trial. J Urol 2003;170:1787–1791.
7. Quinn M, Babb P. Patterns and trends in prostate cancer incidence, survival, prevalence and mortality. Part I: international comparisons. BJU Int 2002;90:162–173.
8. Streiner DL, Norman GR. Diagnostic testing. Commun Oncol 2009;6:428–430.
9. Streiner DL, Norman GR. Mass screening: when does it make sense? Commun Oncol 2010;7:93–95.