Journal of Andrology, Vol. 22, No. 3, May/June 2001

Interpreting Positive Studies

Andrology Lab Corner

STEPHEN D. SIMON

From the Office of Medical Research, Children's Mercy Hospital, Kansas City, Missouri.

In a previous column (Simon, 2001), I discussed the interpretation of negative studies. For negative studies, one needs to assess the adequacy of the sample size and verify that the resulting confidence interval lies entirely inside the range of clinical indifference.

Interpreting positive studies requires a different perspective. Don't worry about sample size being too small; the researchers obviously had enough power to reject the null hypothesis. Still, it is important to examine the clinical relevance of the findings. Was the outcome measure clearly related to an outcome of interest to the typical patient? Did the outcome measure change enough to have an impact?

Is There a Clear Link to a Measurement of Direct Interest to Patients?

Much medical research relies on surrogate measurements. A surrogate is a measurement that is not of direct interest to the typical patient, but one that is clearly related to an outcome that is interesting (to the patient). Blood cholesterol is a surrogate measurement that is related to cardiovascular mortality; an outcome of direct interest to patients. Researchers use surrogate measurements because research into outcomes of direct patient interest is often difficult to perform. The practical importance of the research findings, however, may be unclear when there is a weak and uncertain link between the surrogate measure and the measure of direct interest.

In andrology, the outcome of interest to the patient is often a measure of fertility, such as time to pregnancy. Such studies are especially expensive and difficult to conduct, so the research often focuses on sperm and semen quality measurements (surrogate measurements) that we believe are related to fertility. Interpretation of this research requires a thorough understanding of the relationship between these surrogate measurements and fertility. In particular, it helps a lot to be able to quantify how much of a change in fertility is associated with a change in the surrogate measurement.

An example of where the link is somewhat tenuous is in a recent and controversial study of plastic disposable diapers (Partsch et al, 2000). In this study, children wearing these diapers had a mean scrotal temperature that was higher than when they were wearing cotton diapers. The difference was 0.6 to 1.1 degrees Celsius, depending on the group studied. This finding is valuable for identifying future directions for research examining possible explanations for the decline in the quality of human semen over time that some researchers have found (which is a controversial issue in itself).

One needs to be cautious, however, with this research finding because it is difficult to establish a firm link between increased scrotal temperatures in infancy and decreased fertility in adulthood. In particular, it is unclear how much of an increase in scrotal temperature is needed to cause a decline in fertility. An editorial (Hughes, 2000) that appears in the same issue as the study by Partsch et al provides an excellent discussion of the nature of the link between these measurements.

Did Things Change Enough to Have Clinical Relevance?

Not every change that achieves statistical significance is worthy of clinical concern. Patients receiving a certain drug were 50% more likely to experience a side effect (upper gastrointestinal bleeding) within 30 days of prescription (Carson et al, 1987). But the background rate of the side effect is so rare that one would have to prescribe the drug more than 22 000 times before seeing one additional side effect on average.

When examining the magnitude of change in a research study, one first needs to define a range of clinical indifference. This is the same range discussed in my previous column (Simon, 2001). If the difference lies inside the range of clinical indifference, then the research findings have statistical significance without clinical significance. Better yet, examine the confidence interval. When the full confidence interval lies entirely inside the range of clinical indifference, then there is a definitive negative finding even though the results are statistically significant.

Correspondence to: Steve Simon, PhD, Research Biostatistician, Office of Medical Research, Children's Mercy Hospital, Room HHC-600, 2401 Gillham Road, Kansas City, MO 64108 (e-mail:ssimon@cmh.edu).

Received for publication December 14, 2000; accepted for publication December 14, 2000.

Generally, the sample sizes have to be pretty large before there is statistical significance without practical significance. In the study by Carson et al (1987), the researchers had to study the Medicaid records of more than 90 000 patients. There are exceptions, however, especially for measurements that are highly precise. For example, laboratories can now quantify pesticides to extremely low concentrations. It is not always clear, however, whether a small change in these concentrations is clinically important.

How much of a change in sperm count would be so small as to be considered clinically unimportant? That is a difficult question to answer. Perhaps one way of looking at it is to estimate the fraction of men who would, if exposed, see their sperm count drop to less than 20 million per milliliter or some other clinically relevant cutoff (see Welch et al, 1988 for an interesting analysis along these lines).

Not every study needs to demonstrate clinical relevance. In particular, when the interest is in mechanisms, the presence of an effect of any size may be enough to support or eliminate certain mechanisms of action.

Summary

When examining the results of a positive research study, one needs to consider two issues of clinical relevance. First, is the outcome measure of direct interest to patients, or is it clearly linked to a measure of direct interest to patients? Second, is the difference large enough to lie outside the range of clinical indifference? Don't accept the findings of a positive research study without a careful consideration of clinical relevance.

References

Carson JL, Strom BL, Soper KA, West SL, Morse ML. The association of nonsteroidal anti-inflammatory drugs with upper gastrointestinal tract bleeding. *Arch Intern Med.* 1987;147:85–88.

Hughes IA. How vulnerable is the developing testis to the external environment? *Arch Dis Child.* 2000;83:281–282.

Partsch C-J, Aukamp M, Sippell WG. Scrotal temperature is increased in disposable plastic lined nappies. *Arch Dis Child.* 2000;83:364–368.

Simon SD. Interpreting negative studies. *J Androl.* 2001;22:13–16.

Welch LS, Schrader SM, Turner TW, Cullen MR. Effects of exposure to ethylene glycol ethers on shipyard painters: I. Male Reproduction. *Am J Ind Med.* 1988;14:509–526.