Eur J Epidemiol (2010) 25:843–845
DOI 10.1007/s10654-010-9507-8

LETTER TO THE EDITOR

Re: The ongoing tyranny of statistical significance testing in biomedical research

Tyler J. VanderWeele

Received: 31 August 2010 / Accepted: 1 September 2010 / Published online: 20 November 2010

I would like to thank Stang et al. [1] for an excellent commentary on the limitations and problems with significance testing as currently practiced in biomedical research. I would, however, like to also put in a good word for the *P*-value by way of defending its being reported, along with confidence intervals, in statistical analyses. It is of course the case that the confidence interval generally renders the *P*-value redundant. From the point estimate, along with a confidence interval and knowledge of the model, one could perform the necessary transformations to calculate the standard error estimate and thence the *P*-value. I do nevertheless believe that, in spite of its redundancy, the *P*-value is generally worth reporting. As noted by Stang et al., the *P*-value, at least under the Fisherian interpretation, gives a measure of evidence against the null. The difference between a *P*-value of 0.02 and 0.0002 is of consequence and I do not believe the average reader of the biomedical literature is in general capable of distinguishing between these two given the confidence interval alone.

Suppose a reader encountered an odds ratio from a logistic regression of 1.19 with a confidence interval of (1.08, 1.31). I think it is not very apparent from this information whether the *P*-value in this case is closer to 0.02 or 0.0002, in spite of the 100-fold difference between the two! In fact, in this example, the *P*-value comes to 0.0004. A reader with information that the confidence interval was (1.08, 1.31) would likely not dismiss the possibility of a chance finding; a reader who saw a *P*-value of 0.0004 would be far more likely to do so.

T. J. VanderWeele (✉)
Harvard School of Public Health, Boston, MA, USA
e-mail: tvanderw@hsph.harvard.edu

As readers are unlikely to make the necessary calculations to obtain *P*-values while reading, I believe authors do readers a service by reporting both confidence intervals and *P*-values. Refraining from doing so hinders interpretation. Some journals have gone so far as to insist in their Online Instructions to Authors that contributors refrain from giving *P*-values entirely. While I would oppose not reporting confidence intervals and, like Stang et al., believe we should give up the 0.05 rule, because of the above considerations, I do not see any grounds for dismissing the *P*-value entirely.

Of course, as pointed out by Stang et al., a highly significant *P*-value does not, in observational research, necessarily give much evidence for an effect because of the possibility of a plethora of potential biases. What a highly significant *P*-value does allow an investigator to do, however, is more or less dismiss sampling variability as a potential explanation for the finding and rather focus on the arguably more important sources of error such as confounding, selection, measurement error and model misspecification. Indeed, recent literature in statistics has begun to explore how test statistics in observational studies should be chosen on the basis of robustness to these other sources of bias rather than simply on concerns of statistical power [2].

References

1. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol. 2010;25:225–30.
2. Rosenbaum PR. Design sensitivity and efficiency in observational studies. J Am Statist Assoc. 2010;105:692–702.

The Authors Reply

On a use of the null *P*-value

We thank Tyler Vander Weele [1] for his kind words about our recent paper on statistical significance [2]. We will thank him even more in the future if he stops suggesting that we are intent on "dismissing the *P*-value entirely." We oppose some uses of *P*-values, not *P*-values themselves. If any of us does issue a call for *P*-values to be banished, it will require no reading between the lines to discern.

Long ago, one of us [3] suggested that interpretation might be enhanced by inspecting the graph of all *P*-values for all hypothetical values of the measure one is estimating. Our support for the continued reporting of point estimates and confidence intervals [2] was tantamount to encouraging the reporting of three *P*-values: $P = 1$ for the point estimate and $P = 0.05$ for each of the interval's limits (if the confidence level is 95%). How these recommendations can be viewed as aligning us with those who would instruct authors "to refrain from giving *P*-values entirely" is a mystery.

To be fair, Vander Weele does not defend the reporting of all *P*-values. He has a specific *P*-value in mind, one so special it has become known as "the" *P*-value. It is the *P*-value for the null hypothesis. We have no objection to the reporting of this or any other *P*-value. What matters to us at present is what Vander Weele would have us do with it.

As we have tried to stress, we believe it is wise for epidemiologic researchers to focus on estimates: point estimates, interval estimates or entire *P*-value or likelihood functions. Estimates are what systematic reviewers seek when they review a literature. When they do not find them and find null *P*-values instead, they are understandably disappointed, for they know their review will not reach as deep an understanding of the state of epidemiologic research on the topic as it might have.

When we look at the forest plot in the Figure [4], we see 13 estimates. Apparently, Vander Weele sees something else: up to 13 "findings." We are not sure how he defines this key term in his framework. It would appear that a finding cannot be a null *P*-value, as he urges us to use null *P*-values in coming to judgments about findings. It would also seem that findings, as the objects of those judgments, cannot be the judgments themselves. Our best guess is that Vander Weele's "findings" are estimates, though we suspect that not every estimate qualifies.

Vander Weele seems to view chance and bias as competing, alternative explanations for findings. Chance is to be considered first, for reasons that are obscure to us, with the aim of "dismissing" it, for reasons that seem just as obscure. Given a very low null *P*-value, he would have us "rather focus" on bias than on chance, as though each finding must have one or the other (or, presumably, causality) as its explanation.

Our working model is that chance and bias are always part of the explanation for every estimate and, of course, that causality often plays a role as well. In our view, the ever-present influences of chance and bias on estimates should always be considered, in no fixed order of priority. Our aim is not to "dismiss" any of these influences, but to gauge them as to their magnitude and, in the case of biases, their direction.

Of the 13 estimates in the Figure [4], we would say that the estimate of Leon (2005) is the least influenced by chance. It has the narrowest confidence interval. We are not even sure that Vander Weele would consider this estimate a "finding." We suspect not.

We would say that chance played the greatest role in helping produce the estimate of Liao (2001), which has the widest confidence interval. Vander Weele, in contrast, would be least inclined to "dismiss" chance as "a potential explanation for th[is] finding," as it has the lowest of all the 13 null *P*-values ($P_{\text{null}} = 0.006$).

The difference in outlooks could not be sharper. We would encourage epidemiologists to focus on estimates, to assess the precision and validity of every estimate, and to use the width of a confidence interval to gauge an estimate's precision. We would discourage epidemiologists from focusing on findings, from considering chance and bias as competing explanations for findings, and from using

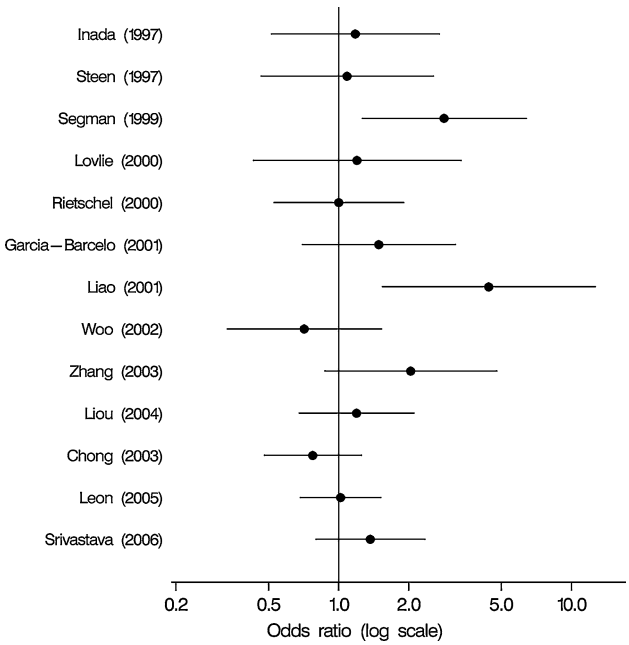

Fig. 1 Odds ratio estimates and 95% confidence intervals from case–control studies of tardive dyskinesia prevalence contrasting the Serine/Glycine and Serine/Serine rs6280 polymorphisms in the dopamine receptor 3 gene [4]

null *P*-values to dismiss chance before considering bias (Fig. 1).

We have expressed some critical thoughts on Vander Weele's considered views of these matters. We hope no one, most of all he, will impute to us a suggestion that his views should be banned or "dismissed entirely."

References

1. Vander Weele TJ. Re: "The ongoing tyranny of statistical significance testing in biomedical research." Eur J Epidemiol.
2. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol. 2010;25:225–30.
3. Poole C. Beyond the confidence interval. Am J Public Health. 1987;77:195–9.
4. Tsai HT, North KE, West SL, Poole C. The DRD3 rs6280 polymorphism and prevalence of tardive dyskinesia: a meta-analysis. Am J Med Genet Part B 2010;153B:57–66.

C. Poole
Department of Epidemiology,
UNC Gillings School of Global Public Health,
University of North Carolina,
Chapel Hill, NC, USA
e-mail: chaspoole@gmail.com

O. Kuss
Institut für Medizinische Epidemiologie,
Biometrie und Informatik,
Univerisätsklinikum und Medizinische Fakultät,
Martin-Luther-Universität Halle-Wittenberg,
Halle, Germany
e-mail: oliver.kuss@medizin.uni-halle.de

A. Stang
Institut für Klinische Epidemiologie,
Medizinische Fakultät,
Martin-Luther-Universität Halle-Wittenberg,
Halle, Germany
e-mail: andreas.stang@medizin.uni-halle.de