

Commentary

(Unlearned) Lessons from John Graunt and Kenneth Rothman: A “CLASSic” Example

Felix M. Arellano, MD,¹ and Jordi Castellsague, MD, MPH²

¹Drug Safety Surveillance, Pfizer Inc, Peapack, New Jersey, and ²Global Epidemiology, Pfizer, Barcelona, Spain

ABSTRACT

This commentary reviews the work of John Graunt and Kenneth J. Rothman on statistical significance and the need for prespecification of study end points. The authors argue that it is dangerous to substitute oversimplifications based exclusively on whether a result has reached statistical significance for a rational process of causal inference. An example is given based on the Celecoxib Long-term Arthritis Safety Study. (*Clin Ther.* 2003;25:2891–2896) Copyright © 2003 Excerpta Medica, Inc.

Key words: epidemiology, confidence interval, significance testing, statistical testing, causal inference, CLASS, calculation errors.

INTRODUCTION

In his 1996 *Lancet* article “Lessons from John Graunt,”¹ Kenneth J. Rothman compared the investigative methodology used by Graunt, a 17th-century scientist, with that of modern-day epidemiologists. Unfortunately, some of the points Rothman makes are as valid today as they were in 1996, as will be illustrated in the present article using several examples from contemporary research, including data from the Celecoxib Long-term Arthritis Safety Study (CLASS).²

Accepted for publication September 12, 2003.

Printed in the USA. Reproduction in whole or part is not permitted.

0149-2918/03/\$19.00

STATISTICAL SIGNIFICANCE TESTING

In his article Rothman wrote, “Our present-day preoccupation with statistical testing stems from a tropism to have clear-cut, black-and-white, formula-driven answers to the complicated questions that we study. Huge and informative bodies of data have been debased into dichotomous categories because so many have been trained to ask only, ‘Is it significant?’”¹ To illustrate his point, Rothman described a meta-analysis by Hommes et al³ that compared the efficacy of subcutaneous injections of heparin with that of a continuous intravenous infusion in preventing the development of deep vein thrombosis. The data yielded an odds ratio (OR) of 0.62 (95% CI, 0.39–0.98; $P < 0.05$). Messori et al⁴ repeated the meta-analysis and obtained similar results (OR, 0.61; 95% CI, 0.298–1.251). Despite the essentially identical results (0.61 vs 0.62), the latter authors disregarded the effect because it was not statistically significant. Elsewhere, Rothman et al⁵ published the CI function (also known as P value function) data for both results (Figure 1), showing that the 2 were essentially the same and that “no sensible person should draw different conclusions from [the] two curves.”¹

Although CI functions add meaning and clarity to reported results, they are only rarely presented in scientific papers, whereas data are dichotomized into true

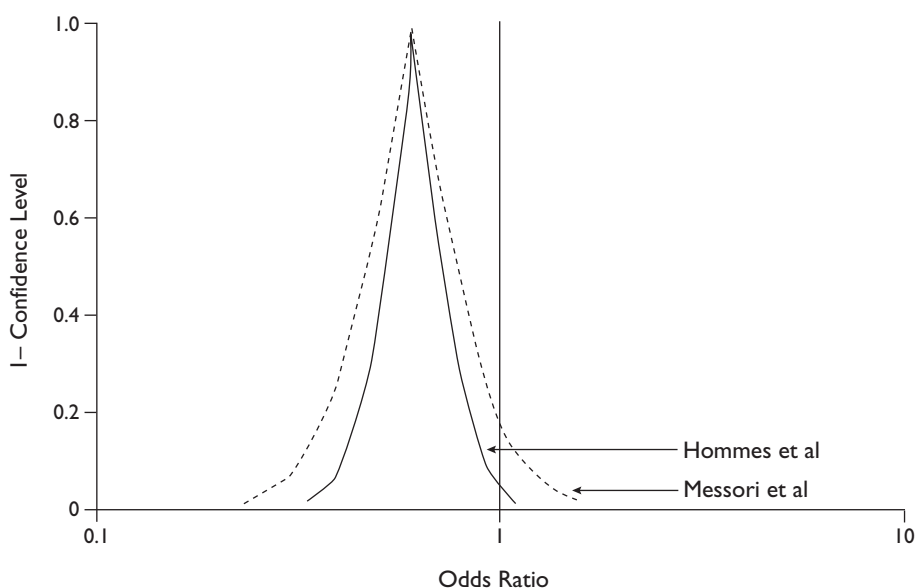


Figure 1. CI functions for the meta-analysis by Hommes et al³ as recalculated by Messori et al,⁴ showing that the 2 calculations are consistent with respect to a causal effect. Reproduced with permission.⁵

or false results according to the “magic” threshold of significance ($P < \text{or} > 0.05$). This practice is a holdover from the precomputer era, when epidemiologists had to refer to tables that were available for selected values only. The meaning behind the data from Hommes et al³—that is, the OR of 0.62—is that subcutaneous administration of heparin is better than intravenous administration for the prevention of deep vein thrombosis. The authors were 95% confident that if their results were applied to any population likely to undergo heparin therapy, the “real” OR would be between 0.39 and 0.98. The CI function provides the entire set of CIs^{6,7} as well as a graphic description of the real point estimate. Using this example, we know that the real OR cannot be 0.62, since this is the point compatible with the 0% CI. We also know that the real result will more likely be around 0.62 rather than close to either 0.39 or 0.98.

The irony of all this “enduring vexation of statistical testing”¹ is that such testing is based on the ability to reject the null hypothesis (H_0), whereas in most biological experiments, a treatment effect, however small, is almost always present. Thus, the hypothesis that there is no difference between treatments (H_0) is almost always false. Since the H_0 will almost never be true, rejection of a true H_0 , or type 1 error, will almost never occur (because 5% of almost 0 is almost 0). Because type 1 errors essentially never occur, the present-day agonizing over significance level is irrelevant from a scientific point of view.⁸ The real problem is to estimate the size of the differences between treatments, not whether the H_0 is true or false.

In CLASS,² the primary end point was ulcer complications (upper gastrointestinal bleeding, perforation, or obstruction), and the secondary end point was ulcer complications/symptomatic ulcers. The relative risks and 95% CIs in patients receiving celecoxib versus the comparator nonsteroidal anti-inflammatory drugs were 0.53 (0.23–1.16) and 0.59 (0.36–0.95) for the primary and secondary end points, respectively (all patients at 6 months). The CI interval functions (Figure 2) show that the results were essentially the same. As in the previous example, a stronger effect (0.53 vs 0.59) was disregarded because the P value crossed the threshold of significance.⁹ Thus, interpreting these findings as missing the primary end point but making the secondary end point⁹ quite simply misses the point.

“NONPRESPECIFIED” END POINTS

As described by Rothman,¹ Graunt admitted to revising his thinking in the face of his data and included in his results a finding that was not prespecified in his study design (ie, more male than female births). “Such an admission,” Rothman suggests, “might sink a proposal or a submitted manuscript today, when we are likely to be advised that we should not study a question that was not specified in gory detail before we began to collect our data.”¹ Today, some would consider this “nonprespecified” finding a “hypothesis-generating” result, to be confirmed

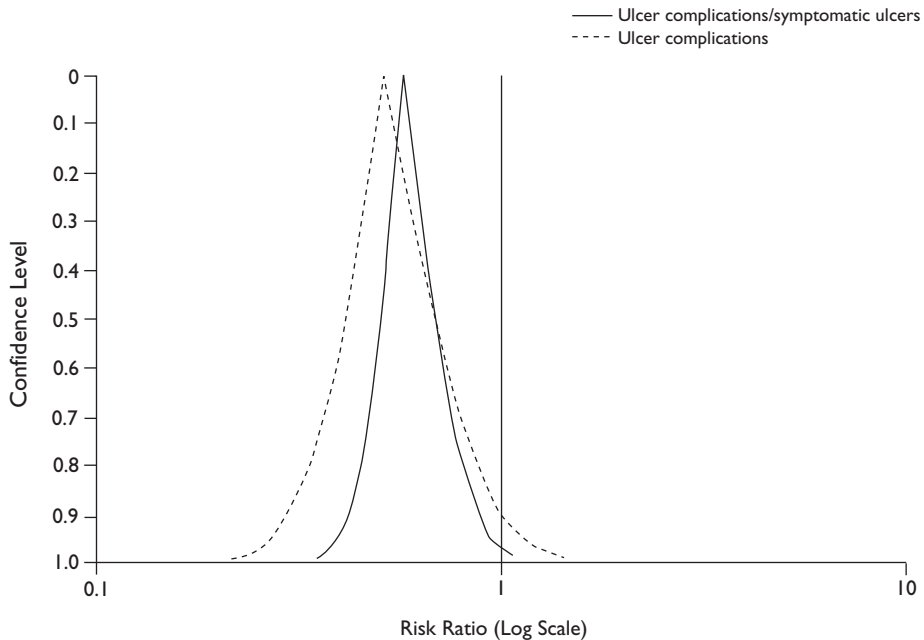


Figure 2. CI functions for the primary end point (ulcer complications) and secondary end point (ulcer complications/symptomatic ulcers) in the Celecoxib Long-term Arthritis Safety Study (CLASS),² all patients at 6 months, showing that the 2 calculations are consistent with respect to a causal effect.

by a “hypothesis-testing” study. It is difficult to understand why such findings carry less importance than ones originally included in the study design. Authors frequently list the fact that study findings were not prespecified as one of the limitations of their studies. In fact, Mukherjee et al¹⁰ listed this as the main limitation of their meta-analysis, which included CLASS, while overlooking the need to adjust for different factors before comparing raw numbers of spontaneous reports. Sometimes researchers suggest that the absence of prespecified end points would invalidate the study results due to the existence of “unintentional” selection bias, but they do not explain how the unintentional selection could bias the study or what could be done to correct it. Results are also dismissed because a trial was “not powered to study” the nonprespecified result. Because power is based on the ability to reject the H_0 , which is almost always false, most studies will be powered to detect the nonprespecified finding. The presence of sufficient power, however, does not make the nonprespecified finding any more or less relevant.

CONCLUSIONS

Determining the effect of a drug is a complex process of causal inference. To substitute statistical significance testing for this process of causal inference can only lead to errors in interpretation, as illustrated by the examples from Hommes et al, Messori et al, and CLASS. These basic conceptual errors are repeated consistently in current research. We need a better system for estimating the size of the statistical differences observed in clinical research so that the data may be properly interpreted and understood.

ACKNOWLEDGMENTS

The authors thank Dr. Kenneth J. Rothman, Boston University, Boston, Massachusetts, and Dr. Richard O. Day, University of New South Wales, Sydney, Australia, for their comments, and Celia Arellano for her editorial assistance.

The authors are employees of Pfizer Inc and may own securities of the company. However, they have done their best to avoid the effect on this work of any potential conflict of interest. The ideas expressed in this article are the authors' and are not necessarily shared by Pfizer Inc.

REFERENCES

1. Rothman KJ. Lessons from John Graunt. *Lancet*. 1996;347:37–39.
2. Silverstein FE, Faich G, Goldstein JL, et al, for the Celecoxib Long-term Arthritis Safety Study. Gastrointestinal toxicity with celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis: The CLASS study: A randomized controlled trial. *JAMA*. 2000;284:1247–1255.
3. Hommes DW, Bura A, Mazzolai L, et al. Subcutaneous heparin compared with continuous intravenous heparin administration in the initial treatment of deep vein thrombosis. A meta-analysis. *Ann Intern Med*. 1992;116:279–284.
4. Messori A, Scroccaro G, Martini N. Calculation errors in meta-analysis. *Ann Intern Med*. 1993;118:77–78.
5. Rothman KJ, Lanes S, Robins SJ. Causal inference. *Epidemiology*. 1993;4:555–556.
6. Rothman KJ, Greenland S. Approaches to statistical analysis. In: Rothman KJ, Greenland S, eds. *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins; 1998:183–199.
7. Ahlbom A. The p-value, the p-value function and the confidence interval. In: Ahlbom A, ed. *Biostatistics for Epidemiologists*. Boca Raton, Fla: Lewis Publishers; 1993:35–53.
8. Oakes MA. Critique of significance tests. In: Oakes MA, ed. *Statistical Inference*. Chestnut Hill, Mass: Epidemiology Resources Inc; 1990:22–74.
9. Jüni P, Rutjes AW, Dieppe PA. Are selective COX 2 inhibitors superior to traditional non steroidal anti-inflammatory drugs [published correction in *BMJ*. 2002;324:1538]. *BMJ*. 2002;324:1287–1288. Editorial.

10. Mukherjee D, Nissen SE, Topol EJ. Risk of cardiovascular events associated with selective COX-2 inhibitors. *JAMA*. 2001;286:954–959.

Address correspondence to: Felix M. Arellano, MD, Chief Safety Officer, Pfizer Inc, 100 Route 206 North, Peapack, NJ 07977. E-mail: felix.m.arellano@pfizer.com