

# Writing for Epidemiology

Kenneth J. Rothman

Scientists communicate the fruits of their labor mostly in writing, and mostly in scientific journals. Conferences and other forms of verbal communication, including the evening news, play an important role, but the written word reaches the widest audience and constitutes the archival message. Futurists may question how long this system will endure. They imagine stodgy journals, with their archaic texts and tables, being replaced by a vast network of electronic information, including animated graphics and multimedia clips of scientists at work. But whether you see the Internet as the future of scientific communication or cringe at the prospect of the unfiltered cacophony that could result, the fact is that the future is not here yet. The published work is still our fundamental end product.

Consequently, getting published, in good journals, remains important to scientists. A well written paper carries an advantage through the review process and beyond. It is not only more likely to be accepted and published, but also more likely to be read in its entirety once published. On the other hand, the jargon-laced paper that no referee wants to wade through is already headed toward oblivion. The research report that makes a legal brief look like interesting reading has placed its author at a serious but avoidable disadvantage.

In the spirit of helping prospective contributors to **Epidemiology**, I here offer some writing suggestions for you to consider in preparing a submission. These suggestions comprise strategic planning and a few editorial and stylistic nuances that together will enhance your submission and the published result.

## How to Write the Paper

An important advantage comes from beginning to write early. Attempting to draft a research paper before the analysis is conducted may seem premature, as there will be holes for the results and you cannot be sure which analytic trails need to be followed. Still, an early draft as the analysis commences, or even a sketchy outline of the manuscript, will assist you to think through the data analysis in a logical sequence, and thus may prevent wasted effort. This early draft is not meant for others, except possibly your collaborators. It is rather a model for your work plan as the research is brought to a conclusion.

If writing a commentary, letter, or critique that does not call for gathering data, writing can begin as soon as the message first takes shape. If you are invited to write an editorial comment, there may be little time to lose, as

the publication deadline could be looming. The best approach is to begin writing even if the ideas are not yet fully formed. Writing aids thinking and *vice versa*. Too much ruminating without putting the thoughts to paper, or to computer, is less efficient than setting the thoughts into full sentences early in the creation process.

## Getting Criticism and Rewriting

Part of the advantage of an early start to writing is gained from rewriting. With the luxury of more time to rewrite, and more drafts to go through, the finished product is bound to be better. It is not unusual among careful scientific writers for a well written paper to have evolved through more than 20 drafts. Those manuscripts that are submitted for consideration after only two or three drafts will generally fare poorly in competition with a finely polished and well crafted paper.

What happens between drafts? The author should be the sharpest critic. He or she must evaluate the work from the perspective of the interested but skeptical reader. Throughout the manuscript, ask, "What questions will occur to the reader at this point?" Then answer them. To do so successfully, the author should know the readership. The most common error is to assume that readers know more about your topic than they actually do. Few readers will share your level of interest in the topic, or your knowledge. The better you explain it, the wider the audience that can appreciate the work. Co-authors should serve as mutual critics. Colleagues not involved directly in the work are another important resource. Ask them not to spare your feelings, because praise of the draft will not improve it. You should keep your feelings at bay, because criticism is what you need. Incorporate any suggestion that improves the work, keeping in mind that if a reader misinterprets something you wrote, there must be an improvement that will help. Remember that resources must not be wasted: you may wear out your colleagues' good will if you ask the same people repeatedly for criticism on a succession of drafts. Although you need their advice, leaning on your colleagues should not be a substitute for your own efforts at rewriting. It works well to approach different colleagues for different drafts, making sure to keep your own critical eye active and to consult your co-authors frequently.

Drafts that are closer to the end product should be sent farther afield for more criticism. E-mail works well for this purpose. Ask for criticism from nonscientists or from those who work in peripheral fields, to learn what obstacles exist for readers from other disciplines. You

should instruct critics to appraise not only general aspects of the work, but sentence structure, word choice, and other small details as well. Try not to become so invested in the organization or the prose of the work that you are unwilling to make major changes if they are suggested and seem appropriate. There is nearly always a better way to present a finding or express a thought.

Include in your list of critics the one or two individuals who are sure to bristle with disdain as they read your report or your commentary. You know who they are, those who are sure to disagree. These antagonists are another great resource. Disarm them by sending your draft with a polite letter that states, "Although I'm certain you will take issue with the main thrust of this work, I would value your criticisms of what I have written (or done)." It is far better to get the inevitable blast of criticism before submitting the work than later as a published letter to the editor. You can deal with the criticism when you are less defensive, in private, and while action is still possible on valid points. Even if you disagree with the criticism, you can mention it in your paper and give the rebuttal.

### Choosing What Data and Analyses to Present

An unfortunate result of the technologic revolution is the ease with which multivariate analyses can be spun out. But just as word processing does not ensure better writing, multivariate analyses do not ensure better analyses. When multivariate analyses were expensive, more thought supported them. Nowadays, interest in multivariate analysis runs so high that many researchers appear to believe that it should constitute the primary means of analyzing epidemiologic data. This thinking is a serious mistake. It is the rare epidemiologic finding that cannot be demonstrated in a straightforward table of frequencies or rates.<sup>1</sup> If multivariate analysis needs to be presented at all, it should usually be a supplement to the basic tables from which readers can view your data and check your calculations. Most often, the multivariate analysis will not be needed, as it will only confirm what the tabular analysis already indicates clearly. These basic tables should not only show the crude relation between the main study variables, but include stratification for the one or two most important confounders or effect-measure modifiers.

Presenting these tables will overcome another problem that occurs in papers describing cohort studies. Too often we become so focused on describing rate ratios that we forget that cohort studies provide measurements of actual rates or risks. These values are more informative than rate ratios. A rate difference, which can be inferred easily from the bare rates, is often of greater interest than the rate ratio, because it offers a measure of the absolute effect of the study exposure. A clear tabular analysis should overcome these difficulties in a cohort study.

Another problem that stems from the indiscriminate reporting of multivariate models is the substitution of statistical interaction for a more meaningful biological

interaction. Here, I refer to the common procedure of evaluating a product term in a multivariate model and considering it an assessment of interaction. If the model is based on a logarithmic transformation, the product term measures departures from a multiplicative relation. Generally, it is the departure from an additive relation, however, that is of primary epidemiologic interest.<sup>2</sup> If assessing interaction is one of the research objectives, you should look at departures from an additive relation; otherwise, you might be better off to omit the issue entirely.

When writing for *Epidemiology*, you can also enhance your prospects if you omit tests of statistical significance. Despite a widespread belief that many journals require significance tests for publication, the Uniform Requirements for Manuscripts Submitted to Biomedical Journals<sup>3</sup> discourages them, and every worthwhile journal will accept papers that omit them entirely. In *Epidemiology*, we do not publish them at all. Not only do we eschew publishing claims of the presence or absence of statistical significance, we discourage the use of this type of thinking in the data analysis, such as in the use of stepwise regression. We also would like to see the interpretation of a study based not on statistical significance, or lack of it, for one or more study variables, but rather on careful quantitative consideration of the data in light of competing explanations for the findings. For example, we prefer a researcher to consider whether the magnitude of an estimated effect could be readily explained by uncontrolled confounding or selection biases, rather than simply to offer the uninspired interpretation that the estimated effect is "significant," as if neither chance nor bias could then account for the findings.

Many data analysts appear to remain oblivious to the qualitative nature of significance testing. Although calculations based on mountains of valuable quantitative information may go into it, statistical significance is itself only a dichotomous indicator. As it has only two values, "significant" or "not significant," it cannot convey much useful information. Even worse, those two values often signal just the wrong interpretation. These misleading signals occur when a trivial effect is found to be "significant," as often happens in large studies, or when a strong relation is found "nonsignificant," as often happens in small studies. *P*-values, being more quantitative, are preferable to statements about statistical significance tests, and we do publish *P*-values on occasion. We do not publish them as an inequality, such as  $P < 0.05$ , but as a number, such as  $P = 0.13$ . By giving the actual value, one avoids the problem of dichotomizing the continuous *P*-value into a two-valued measure. Nevertheless, *P*-values still confound effect size with study size,<sup>4</sup> the two components of estimation that we believe need to be reported separately. Therefore, we prefer that *P*-values be omitted altogether, provided that point and interval estimates, or some equivalent, are available.

One arena in which *P*-values are the usual analytic tool is in the assessment of trends, such as the trend in rate across dose categories. Even here, we believe that they should be avoided. Slope estimates are better, and

smoothed trend evaluations, such as kernel smoothing or spline regression, are better yet<sup>5</sup>; these presentations should ideally include some assessment of statistical precision to accompany the trend estimate.

Some authors conduct case-control studies and present their key findings in terms of the difference in exposure proportions between cases and controls. For example, they might state that 52% of cases were smokers, compared with only 40% of controls. In some papers, authors attempt to compare mean or median levels of a continuous exposure measure. The comparison of exposure proportions, means, or medians should be avoided, as a difference in these quantities has no direct epidemiologic interpretation. Furthermore, these exposure measures will vary among studies simply as a function of variation in the exposure prevalence in the source population for the studies. True, the case-control study is designed to obtain exposure information from cases and controls, but this information represents an intermediate step in the analysis. To fix on exposure proportions or average exposure levels is to think backward, rather than to think forward from cause to effect. Author and reader should both be thinking in terms of the level of risk that relates to a given level of exposure, a measure that will not depend on exposure prevalence. In a case-control study, you should calculate and report the odds ratios at various dose levels. These odds ratios are estimates of rate or risk ratio and are of inherently greater epidemiologic interest than a report of exposure proportions for cases and controls. It is reasonable to report exposure proportions as intermediate information, but it is even better to report the actual frequencies of exposed and unexposed cases and controls in the tables. Similarly, you should report the age distribution for cases and controls in preference to the mean age of cases and of controls, and the best way to do so is to give the actual frequencies of case and control subjects by age and exposure, so that readers can conduct their own stratified analysis of your data.

### Keeping It Brief

If you would like the freedom to expound at great length on your subject, you will have to write a book. Most journals have limited page space and wish to publish a variety of papers. Aside from excellence in writing, the best way to improve the prospects of your manuscript is to keep it brief. To do so will demand that some details be left out, or summarized in broad strokes. You may have to trim the objectives, although a relatively comprehensive paper will have more appeal than a "least publishable unit." Rather than skimp on important information, trim the fat. Unless the manuscript is a review paper, keep the literature review focused and brief, using citations to point readers toward key information. Avoid reiterating all the results in the discussion section; instead, just highlight the principal findings and move directly to the interpretation, considering plausible competing explanations, causal and noncausal, for what you found.

A cliché to avoid is the inevitable call for further research. When is more research not needed? You also do not need to remind the reader how strong the data are and how carefully they were collected, analyzed, and considered. When you state that you did something carefully, it implies that on occasion you are careless. Similarly, avoid the bromide that a given finding should be interpreted cautiously. It implies that other interpretations are reckless. Another candidate for omission is the concluding paragraph recommending some specific public health action as a consequence of your research report. A superficial recommendation based only on your results is presumptuous and unnecessary. If important, your results will be heeded without your recommendation. This is not to say that you should not follow through on the public health implications of your research. The scholarly evaluation of policy recommendations should consider a broad range of effects of an intervention, the financial and social costs, as well as policy alternatives. These elements are not inherent components of most epidemiologic studies. Instead, they represent a separate contribution to public health, much too important to toss off casually as a tag line in an epidemiologic research study. Take the time to write a separate paper with your policy evaluation instead of appending it as an afterthought to the discussion of your findings.

### Making Tables and Figures

Organize your tables efficiently into a borderless grid of rows and columns. Use horizontal rules sparingly, and no vertical rule, for separation. A good plan for the table title is to begin with a description of the main cell entries of the table, followed by appropriate qualifications or additional description. For example, "Frequencies of breast cancer cases and controls by level of alcohol consumption and 5-year age category, and age-specific relative risk estimates with 95% confidence intervals" is a title that tells the reader what to expect as the reader looks down to inspect the contents of the table. Strive for simplicity in organizing tables, so that their information can be quickly absorbed. Some journals prefer details in footnotes below the table; we prefer more information in the title and fewer footnotes below the table. Remember that somewhere in your tables you should present some actual data, as opposed to derived statistics, such as the results from mathematical models. A paper that presents only derived results with no distribution of the data robs the reader of essential information.

Whereas tables are useful to present detailed data or findings, well drawn figures can convey complicated patterns more effectively. For example, you might use a figure to convey the pattern of the relation between key study variables, especially if the pattern is derived from a locally smoothed function that cannot easily be presented in a table. A graph of a single line that connects only four or five values takes too much space for the message. Use space more efficiently by putting the in-

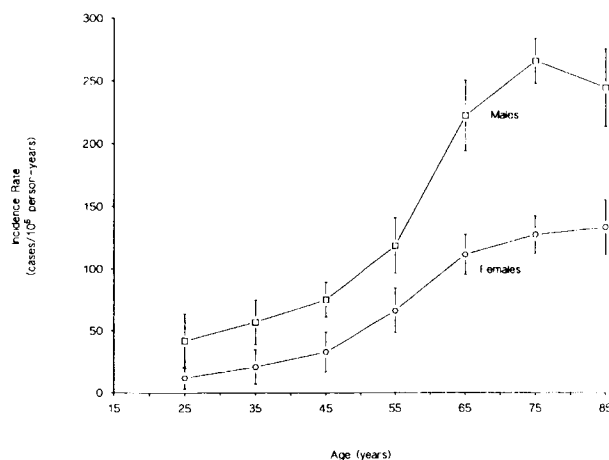


FIGURE 1. Figure with type and line strokes that are too small.

formation into the text or a table instead, or including several curves in the graph, but not so many as to clutter it. Bar and pie charts usually are inefficient uses of space; the information should be presented in a table instead. You should consult the book by Tufte<sup>6</sup> for a scholarly and definitive discussion of the use of graphical aids.

Many software packages produce figures with thin lines that are unsuitable for publication, such as those displayed in Figure 1. Either override the software defaults or get another software package. Figure lines should be broad and black, and labeling text should be large enough to be legible after reduction to a single column. Figure 2 shows a better version of the same figure, in which the strokes for all characters and lines are thicker, tick marks are longer, the lines and data points used for the male and female data are different, and arrows have been added. Both versions illustrate properly labeled axes with units of measure and tick marks that go out rather than in, to avoid encroaching on the graph.

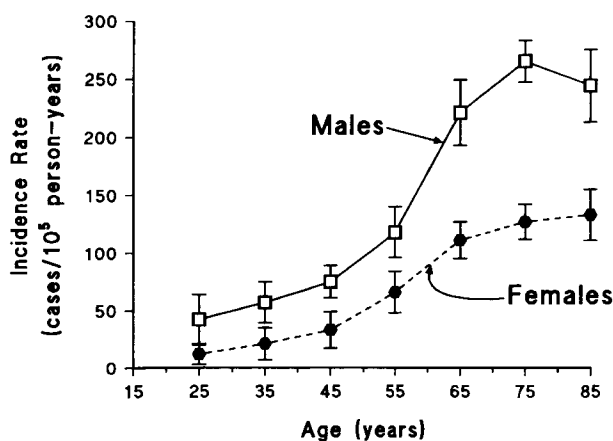


FIGURE 2. Figure with more easily read characters and lines, and other improvements.

## The Abstract

The abstract gets a wider distribution than the paper itself, through MEDLINE and other information sources. It should present the purpose of the study, a basic description of the method, the main results, and a conclusion. Be sure to include your main quantitative findings, with confidence intervals, but be selective. In **Epidemiology**, we do not use a structured abstract; instead, we hope to see a single well written paragraph that reads like a mini-presentation of the paper.

## Grammar and Style

Read Strunk and White<sup>7</sup> to learn about writing style. Colloquial language, technical jargon, and excessive use of abbreviations are best avoided for any piece that one hopes will endure. Scientists too often write in the passive voice out of habit or with the misguided notion that the passive voice is more objective. Passive voice is more boring, and less informative, than active voice. Rather than say "the diagnosis was defined in terms of a three-plus response to the Rumpelmeier test," tell the reader who decided to define it that way. Writing primarily in the active voice will make the paper easier to read and convey more information.

A few other problems to keep in mind:

1. The modifiers "any" or "no" imply a singular noun. Say "No inference is possible," not "No inferences are possible."
2. Use a noun after the word "this." Say, "This finding shows . . ." instead of "This shows . . ."
3. "Prior to" is pretentious jargon. Use "before" instead.
4. "Relationship" is usually by blood or marriage. Use "relation" instead.
5. "Minimize" means to reduce to a minimum. Do not use it when you mean simply "to reduce." For example, in the phrase "We asked detailed questions to minimize misclassification," you should use *reduce* in place of *minimize* unless you think that the approach achieved the maximum possible reduction.
6. One legacy of thinking in terms of statistical significance testing is to claim that two values "do not differ" when they do, even if not by much. Be accurate. You should say that two values do not differ only when they are identical. Similarly, you should say that variables *x* and *y* are not associated only if the measure of association is precisely zero. More generally, you should simply report the difference, along with a comment on its practical or clinical importance.

## Authorship

Authorship can be a contentious issue. Rules for authorship exist but are often ignored.<sup>3</sup> Gift authorship is frowned upon but nevertheless a common practice, as are despotic practices such as the team leader inserting himself or herself inappropriately as an author, or deny-

ing authorship to a colleague who merits it. Our view is that the primary writer should be listed as the first author, regardless of other roles on the project. Hilsinger<sup>8</sup> suggests that the first author has the responsibility to determine the identity and sequence of the other authors. Not all agree,<sup>9</sup> but we do, although ideally the determination will come from mutual agreement.

Following these suggestions should help your submission to **Epidemiology**, or any journal. Although good writing takes effort, it is an acquired skill.

---

### Acknowledgments

I am grateful to Cristina I. Cann, Cheryl Iverson, and Janet M. Lang for helpful comments.

---

### References

1. Walker AW. Reporting the results of epidemiologic studies. *Am J Public Health* 1986;76:556-558.
2. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott-Raven, 1998.
3. Uniform Requirements for Manuscripts Submitted to Biomedical Journals. *N Engl J Med* 1997;336:309-315.
4. Lang J, Rothman KJ, Cann CI. That confounded *P*-value. *Epidemiology* 1998;9:7-8.
5. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology* 1995;6:356-365.
6. Tufte ER. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press, 1983.
7. Strunk W Jr, White EB. *The Elements of Style*. 3rd ed. Boston: Allyn & Bacon, 1995.
8. Hilsinger RL Jr. Solution. In: Mundy D, ed. Lead author responsibility. *CBE Views* 1997;20:200.
9. Hedges JR. Solution. In: Mundy D, ed. Lead author responsibility. *CBE Views* 1997;20:200.