Reporting of Numerical and Statistical Differences in Abstracts

Improving but Not Optimal

Eric Dryver, MD, Janet E. Hux, MD, SM, FRCPC

OBJECTIVE: The reporting of relative risk reductions (RRRs) or absolute risk reductions (ARRs) to quantify binary outcomes in trials engenders differing perceptions of therapeutic efficacy, and the merits of *P* values versus confidence intervals (CIs) are also controversial. We describe the manner in which numerical and statistical difference in treatment outcomes is presented in published abstracts.

DESIGN: A descriptive study of abstracts published in 1986 and 1996 in 8 general medical and specialty journals. Inclusion criteria: controlled, intervention trials with a binary primary or secondary outcome. Seven items were recorded: raw data (outcomes for each treatment arm), measure of relative difference (e.g., RRR), ARR, number needed to treat, *P* value, CI, and verbal statement of statistical significance. The prevalence of these items was compared between journals and across time.

RESULTS: Of 5,293 abstracts, 300 met the inclusion criteria. In 1986, 60% of abstracts did not provide both the raw data and a corresponding *P* value or CI, while 28% failed to do so in 1996 ($P < .001$; RRR of 53%; ARR of 32%; CI for ARR 21% to 43%). The variability between journals was highly significant ($P < .001$). In 1986, 100% of abstracts lacked a measure of absolute difference while 88% of 1996 abstracts did so ($P < .001$). In 1986, 98% of abstracts lacked a CI while 65% of 1996 abstracts did so ($P < .001$).

CONCLUSIONS: The provision of quantitative outcome and statistical quantitative information has significantly increased between 1986 and 1996. However, further progress can be made to make abstracts more informative.

KEY WORDS: periodicals; abstracts; statistics; format effect; quantitative results.

J GEN INTERN MED 2002;17:203-206.

The abstract has increasingly become a crucial source of information for the busy physician accessing the medical literature.¹ It is often the means by which articles to be read are selected and in cases in which the full text is unavailable, it is even the basis upon which clinical decisions are made.² Concern about the need for more adequate abstracts of research papers led an ad hoc working group of clinical epidemiologists and journal editors to publish guidelines³ for structured abstracts in

April, 1987. These guidelines have since been modified and refined,⁴ and a particular emphasis has been placed on the presentation of study results.^{5,6}

Despite these efforts to improve the informative content of abstracts, the data presented in them may still lack quantitative measures of statistical significance (*P* value and confidence interval [CI]) and may be misinterpreted by readers due to the effect of the outcome format on the appraisal of trial results.⁷ Format effects are particularly salient in the setting of a low-event rate (e.g., 2% death rate), where a large relative risk reduction (RRR) ("50% fewer deaths") corresponds to only a small absolute risk reduction (ARR) ("1% fewer persons die") and a large number needed to treat (NNT) ("need to treat 100 people to save 1 life"). A high RRR, when presented alone, will lead more physicians to recommend an intervention, while the reporting of the corresponding low ARR or high NNT will reduce enthusiasm for the intervention.⁸⁻¹² The level of clinical training and familiarity with research design and analysis have not been shown to mitigate physicians' sensitivity to the format of the outcome measure.¹⁰

The purpose of this descriptive study is to document the presentation of quantitative outcome data in intervention studies published in selected journals in the years 1986 and 1996. We hypothesized that there would be an increase in the reporting of absolute measures of risk reduction over time but that there would be differences between journals in the adequacy of quantitative data reported in abstracts.

METHODS

We examined all abstracts published in 1986 and 1996 in the following journals: The Annals of Internal Medicine, The Archives of Internal Medicine, The British Medical Journal (BMJ), Chest, Circulation, The Journal of the American Medical Association (JAMA), The Lancet, and The New England Journal of Medicine (NEJM). We strove for journal diversity by selecting general and specialty journals with differing degrees of influence. This resulted in the selection of 8 journals that covered a broad range of impact factors¹³ (see Table 1). Abstract inclusion criteria were controlled intervention studies with at least 1 binary primary or secondary outcome. We excluded abstracts relating to animal studies, abstracts presented alone without a following article, and observational studies. Abstracts were also excluded when they failed to define what the outcome measures were (e.g., those reporting that "the results were similar").

In each abstract, we chose the binary outcome most closely linked to the study question when the primary outcomes of the trial were not specified. We then

Received from the Department of Medicine, University of Toronto (ED, JEH); and the Institute for Clinical Evaluative Sciences and the Clinical Epidemiology and Health Care Research Program, Sunnybrook and Women's College Health Sciences Centre (JEH), Toronto, Ontario, Canada.

Address correspondence and requests for reprints to Dr. Hux: Institute for Clinical Evaluative Sciences, G-106, 2075 Bayview Ave., Toronto, Ontario, Canada, M4N 3M5 (e-mail: jan@ices.on.ca).

Table 1. Abstracts Meeting Inclusion Criteria and SCI Journal Impact Factors

Journal	1986	1996	Impact Factor
Annals of Internal Medicine (Annals)	8	17	9.9
The Archives of Internal Medicine (Archives)	4	14	4.1
British Medical Journal (BMJ)	17	19	4.4
Chest	6	14	1.7
Circulation	7	20	8.6
The Journal of the American Medical Association (JAMA)	10	19	6.9
The Lancet (Lancet)	23	38	17.3
The New England Journal of Medicine (NEJM)	32	52	22.7

recorded the presence or absence of several modalities used to describe difference in outcome between treatment arms, namely the "raw data," the presence of a measure of relative difference (RRR, relative risk, odds ratio, risk ratio, hazard ratio, or efficacy when relating to a relative difference), the ARR, the NNT, quantitative measures of statistical difference (P value or CI), and qualitative measures of statistical difference (i.e., a verbal statement that the outcome difference was "significant" or not). The raw data were considered to have been provided when the proportion or number of the members in each treatment group reaching the outcome was present in the abstract.

Abstracts were categorized according to the manner in which they presented numerical difference and statistical difference (Table 2). We also categorized abstracts using a 2×2 table according to whether the raw data were provided and whether they were accompanied by a corresponding P value or CI.

We used STATA (version 6.0; Stata Corp., College Station, Tex) to describe the differences in prevalence of these modalities between years and across journals. To compare the prevalence of specific modalities or combination of modalities between years, we used the χ^2 test. Fisher's exact test was used to compare the prevalence of specific modalities or combination of modalities across journals. To investigate the importance of year and journal in accounting for the prevalence of specific modalities or combination of modalities, we used analysis of variance.

RESULTS

A total of 17,281 articles were published in 1986 and 1996 in the journals under study; 5,293 of these articles had abstracts. Of these abstracts, 300 met the inclusion criteria, 107 in 1986 and 193 in 1996. Two hundred fifty-nine of these abstracts (86%) were explicitly described as randomized studies while the remainder were intervention studies with concurrent controls. The breakdown by journal and by year of the abstracts that met the inclusion criteria is provided in Table 1.

In 1986, most abstracts provided the raw data only (Table 3), and no abstract presented measures of absolute difference. In 1996, 12% (95% CI, 8% to 18%) of abstracts presented measures of absolute difference (1986 vs 1996, $P < .001$). Only 4 abstracts provided an NNT; the rest provided the ARR. Only 4 abstracts presented both a measure of relative difference and a measure of absolute difference. The journals were not statistically significantly different in the proportions reporting measures of absolute difference ($P = .4$).

In 1986, most abstracts provided either no mention of statistical significance or a P value only (Table 3). In 1996, 80% of abstracts provided a P value, a CI or both (vs 42% in 1986, $P < .001$). Thirty-five percent of abstracts in 1996 presented a CI compared to only 2% in 1986 ($P < .001$). The journals differed in the proportions reporting P values and/or CI ($P = .007$).

In 1986, 60% of abstracts failed to provide both the raw data and a quantitative measure of statistical difference (P value or CI), while this percentage shrank to 28% in 1996 ($P < .001$; RRR 53%; ARR of 32%; CI for ARR, 21% to 43%) (Table 3). Figure 1 illustrates the proportion of abstracts providing both numerical and statistical quantitative information for both years across journals. The variability between journals was highly significant ($P < .001$).

The test for interaction between year and journal was not significant ($P = .42$).

DISCUSSION

The growing body of literature demonstrating that the interpretation of trial results is sensitive to the format in

Table 2. Categories of Presentation of Difference

Numerical difference	
No data	No raw data are provided
Raw data only	Only the raw data are provided
Raw data + relative	The raw data are provided along with a measure of relative difference (e.g., RRR)
Raw data + absolute	The raw data are provided along with a measure of absolute difference (e.g., ARR)
Raw data + relative + absolute	The raw data are provided along with both a measure of relative difference and a measure of absolute difference
Statistical difference	
No statistics	No mention of statistical significance, no P value, no confidence interval
Reported significance	Statistical significance or lack thereof is mentioned verbally only
P value	P value is present
95% confidence interval	The 95% confidence interval is provided
P value + confidence interval	Both the P value and the confidence interval are provided

Table 3. Reporting of Numerical and Statistical Outcomes by Year

	Proportion of Abstracts (%)	
	1986	1996
Information provided		
Numerical and statistical	40	72
Numerical only	43	17
Statistical only	2	8
Neither	15	3
Numerical outcomes reported		
No data	17	11
Raw data only	76	55
Raw data + relative	7	21
Raw data + absolute	0	10
Raw data + relative + absolute	0	2
Statistical information reported		
No statistics	39	8
Reported significance	19	12
<i>P</i> value	39	45
95% confidence interval	0	13
<i>P</i> value + confidence interval	2	22

which the data are presented has led to calls for the reporting of absolute rather than just relative measures of therapeutic benefit.^{7,10,14,15} Our audit of abstracts published in major medical journals found a marked increase between 1986 and 1996 in the presentation of absolute measures of difference; yet only 12% of the 1996 abstracts we examined presented an ARR or NNT. As with other aspects of quality improvement in published abstracts, progress is slow.¹⁶

The appropriate or best format in which binary trials outcomes are to be presented remains a subject of debate. Some argue that the RRR may, in certain circumstances, be an attractive single measure of a therapeutic effect in a population with a heterogeneous baseline risk.¹⁴ In contrast, the NNT has been advocated as the measure that best addresses the relationship of therapeutic effort to clinical yield.¹⁷ Still others argue that any summary measure that purports to communicate the benefit of a particular treatment across a wide range of patients will be inadequate and that clinical decisions should be based on explicit considerations of individual benefits and harms.¹⁸

One potential solution to this conundrum is to report the outcomes for the intervention and control groups and allow the readers to calculate their summary measure of preference. However, work by Malenka et al.⁹ suggests that failure to explicitly report absolute measures of benefit may still bias the evaluation of a treatment. If that is the case, then providing the RRR along with baseline data in the abstract may not be sufficiently informative: the ARR or the NNT should also be explicitly reported.

In addition to assessing the summary quantitative outcome measure reported in abstracts, our work examined the reporting of the statistical precision of that result. We found a marked increase between 1986 and

1996 in the percentage of abstracts presenting either a *P* value or a CI. However, only 35% of 1996 abstracts presented CIs. As with the format of numerical difference, there is ongoing debate regarding the optimal method for reporting statistical significance as well as its importance relative to clinical significance.¹⁹ Proponents of confidence intervals²⁰⁻²³ note, for example, that concerned readers of negative trials are not simply interested in knowing whether the null hypothesis has been accepted but also want to know the range of values that have been excluded by the study so that they may infer whether a clinically significant difference might have been detected given a larger sample size.

This study does not address whether the numbers reported in the abstracts we examined were correct or whether they described a predetermined study outcome. Neither does this study purport to provide an evaluation of the quality of abstracts generalizable to all medical publications. Since journals vary widely in their mandate, readership, style, and content, the relevance of a randomly selected sample is unclear. Instead, we have evaluated a diverse sample of journals that may be of interest to the general reader.

Given the absence of substantial modifications to the structured abstract guidelines provided to authors between 1996 and the present, it is unlikely that the reporting deficiencies described here have been resolved. It may therefore fall to journal editors to actively promote further improvement in the quality of abstract results, however difficult such a venture may be.²⁴ To what extent

Proportion of Abstracts Providing Both Numerical & Statistical Information

FIGURE 1. Proportion of abstracts by year and by journal providing both the raw data and an accompanying *P* value or CI. The raw data were considered to have been provided when the percentage or number of the members in each treatment group reaching the outcome was present in the abstract. Annals, The Annals of Internal Medicine; Archives, The Archives of Internal Medicine; BMJ, The British Medical Journal; JAMA, The Journal of the American Medical Association; Lancet, The Lancet; and NEJM, The New England Journal of Medicine.

the quantitative content can be enhanced without further stripping what some consider to be an already unacceptably stilted abstract format²⁵ of whatever prosodic quality it retains is an open question. Nonetheless, further efforts to reduce format effects in data presentation may be required if readers are to gain a balanced understanding of the benefit conferred by the treatment being described.

The opinions, results, and conclusions are those of the authors, and no endorsement by the Ministry of Health or by the Institute for Clinical Evaluative Sciences is intended or should be inferred.

Dr. Hux is a Career Scientist of the Ontario Ministry of Health and receives salary support from the Institute for Clinical Evaluative Sciences in Ontario.

REFERENCES

- Pitkin RM. The importance of the abstract. *Obstet Gynecol.* 1987;70:267.
- Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. A study of the use and usefulness of online access to MEDLINE in clinical settings. *Ann Intern Med.* 1990;112:78-84.
- Ad Hoc Working Group for Critical Appraisal of the Medical Literature. A proposal for more informative abstracts of clinical articles. *Ann Intern Med.* 1987;106:598-604.
- Haynes RB, Mulrow CD, Huth EJ, Altman DG, Gardner MJ. More informative abstracts revisited. *Ann Intern Med.* 1990;113:69-76.
- LeBlond RF. Improving structured abstracts (Letter). *Ann Intern Med.* 1989;111:764.
- Altman DG, Gardner MJ. More informative abstracts. *Ann Intern Med.* 1987;107:790-1.
- Hux JE, Naylor CD. In the eye of the beholder. *Arch Intern Med.* 1995;155:2277-80.
- Bobbio M, Demichelis B, Giustetto G. Completeness of reporting trial results: effect on physicians' willingness to prescribe. *Lancet.* 1994;343:1209-11.
- Malenka DJ, Baron JA, Johansen S, Wahrenberger JW, Ross JM. The framing effect of relative and absolute risk. *J Gen Intern Med.* 1993;8:543-8.
- Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. *Am J Med.* 1992;92:121-4.
- Bucher HC, Weinbacher M, Gyr K. Influence of method of reporting study results on decision of physicians to prescribe drugs to lower cholesterol concentration. *BMJ.* 1994;309:761-4.
- Marshall KG. Prevention. How much harm? How much benefit? 1. Influence of reporting methods on perception of benefits. *Can Med Assoc J.* 1996;154:1493-9.
- Journal Citation Reports SCI. Philadelphia: Institute for Scientific Information; 1994.
- Sackett DL, Cook RJ. Understanding clinical trials: what measures of efficacy should journal articles provide busy clinicians? *BMJ.* 1994;309:755-6.
- Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ.* 1995;310:452-4.
- Pitkin RM, Branagan MA, Burmeister LF. Accuracy of data in abstracts of published research articles. *JAMA.* 1999;281:1110-1.
- Naylor CD, Chen E, Strauss B. Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Ann Intern Med.* 1992;117:916-21.
- Dowie J. The "number needed to treat" and the "adjusted NNT" in health care decision making. *J Health Serv Res Policy.* 1998;3:44-9.
- Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. *J Clin Epidemiol.* 1998;51:355-60.
- Gardner MJ, Altman DG. Confidence intervals rather than p-values: estimation rather than hypothesis testing. *BMJ.* 1986;292:746-50.
- Borenstein M. The case for confidence intervals in controlled clinical trials. *Control Clin Trials.* 1994;15:411-28.
- Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. *CMAJ.* 1995;152:169-73.
- Simon R. Confidence intervals for reporting results of clinical trials. *Ann Intern Med.* 1986;105:429-35.
- Pitkin RM, Branagan MA. Can the accuracy of abstracts be improved by providing specific instructions? A randomized controlled trial. *JAMA.* 1998;280:267-9.
- Spitzer WO. The structured sonnet. *J Clin Epidemiol.* 1991;44:729.