Positive-Outcome Bias and Other Limitations in the Outcome of Research Abstracts Submitted to a Scientific Meeting

Michael L. Callaham, MD; Robert L. Wears, MD; Ellen J. Weber, MD; Christopher Barton, MD; Gary Young, MD

Context.—Studies with positive results are more likely to be published in biomedical journals than are studies with negative results. However, many studies submitted for consideration at scientific meetings are never published in full; bias in this setting is poorly studied.

Objective.—To identify features associated with the fate of research abstracts submitted to a scientific meeting.

Design and Setting.—Prospective observational cohort, with 5-year follow-up of all research submitted for consideration to the major annual 1991 US research meeting in the specialty of emergency medicine.

Participants.—All research abstracts submitted for consideration at the meeting for possible presentation.

Main Outcome Measures.—Characteristics associated with acceptance for presentation at the meeting and subsequent publication as a full manuscript.

Results.—A total of 492 research abstracts were submitted from programs in emergency medicine and other specialties affiliated with 103 US medical schools. A total of 179 (36%) were accepted for presentation and 214 (43%) were published in 44 journals. Of the 179 abstracts accepted for presentation, 111 studies were published. Scientific quality of abstracts or prestige of the journal in which the study was eventually published did not predict either of these outcomes. The best predictors (by logistic regression) of meeting acceptance were a subjective “originality” factor (odds ratio [OR], 2.07; 95% confidence interval [CI], 1.13-3.89) and positive results (OR, 1.99; 95% CI, 1.07-3.84), and, for publication, meeting acceptance (OR, 2.49; 95% CI, 1.49-4.35) and large sample size (OR, 2.26; 95% CI, 1.23-4.31). Forty-nine percent (241) of abstracts did not report on blinding, and 24% (118) did not report on randomization. Acceptance and publication were both more likely for positive outcomes (P = .03). Funnel plots showed the classic distribution of positive-outcome bias at each of the submission, acceptance, and publication phases. Meeting acceptance predicted publication with a sensitivity of only 51%, specificity of 71%, positive predictive value of 57%, and negative predictive value of 66%.

Conclusions.—Positive-outcome bias was evident when studies were submitted for consideration and was amplified in the selection of abstracts for both presentation and publication, neither of which was strongly related to study design or quality.

JAMA. 1998;280:254-257

POSITIVE-OUTCOME (also known as “publication”) bias refers to the fact that research with positive outcomes is much more likely to be published than that with negative outcomes.1-4 Presentation of results in abstracts at scientific meetings is the first and often only publication for most biomedical research studies.5 However, the abstract selection process for meetings rarely has been studied. We, therefore, examined all research submitted for presentation at a national meeting to determine if positive-outcome bias was present in this process and what characteristics determined successful subsequent publication in a peer review journal.

METHODS

The Society for Academic Emergency Medicine (SAEM) meeting is comparable to the meetings of 31 other societies of the Council of Academic Societies.6 Abstracts with mandatory structured formats were submitted and, independent of our study (and similar to other specialty meetings), each submission was evaluated by 5 to 7 blinded members of the SAEM screening committee (selected for their relevant expertise) and ranked on a 5-point Likert scale. Selection was based on an average score, and no journal had right of first refusal.

Four years after the 21st annual meeting of SAEM in 1991, all SAEM data were obtained by the authors and each submitted study was categorized according to design by a blinded Delphi panel. Since no established system exists for abstract classification, we modified a previously published approach.7 In addition, the review panel ranked each study for scientific quality and “originality” (“newsworthiness”) on a Likert scale like the one previously validated.8 Institutional review board approval was obtained, and a detailed description of the methods used is available from the authors.

All authors’ names were searched in MEDLINE in late 1995 to determine if the study had been published in any listed journal.9 For papers not found, the search was repeated in early 1996; if still not found, a questionnaire was sent to the authors and EMBASE and the Cochrane Collaboration databases were also searched.10 Journal impact factor was derived from the Science Citation Index for the year of publication.11 The authors’ institutions were ranked according to a system of based on total dollars of National Institutes of Health (NIH) grant support.
There is no standardized definition of positive results. ¹⁶ We used one of the more common definitions—that results were positive if the studied variable produced positive (beneficial) results.¹,³,¹⁷,¹⁸ Some authors have defined positive results as those reporting statistically significant results (regardless of direction);⁴,⁵,¹⁶,¹⁹-²¹ so we also used this definition.

We performed the major logistic regression analysis on those studies in which the subjects were either humans or animals, and the design was a prospective interventional trial, a prospective observational study, or a retrospective observational study. We also separately examined prospective studies with controls, excluding retrospective studies. We used a general iterative model-building strategy, as suggested by Hosmer and Lemeshow²² and Harrell,²³ assessed for goodness of fit and subjected to bootstrap validation using S-Plus version 4.0, release 3 (Mathsoft Inc, Seattle, Wash) and Harrell extensions (http://www.lib.stat.cmu.edu). Correlation coefficients were calculated using JMP software 3.2.2 for Macintosh (SAS Institute, Inc, Cary, NC). We calculated the effect size of all interventions for the subgroup of prospective studies with controls in the usual fashion, except that SDs were not available.

RESULTS

Five hundred research abstracts were submitted to the SAEM selection committee from a total of 144 institutions, 103 of which had formal US medical school affiliations. Eight duplicates were deleted, leaving 492 abstracts as the basis of our study. The submitting schools averaged an NIH funding rank of 56 (of all US medical schools) (95% confidence interval [CI], 49-62).

Thirty percent of studies did not state a hypothesis, and 49% did not report on blinding, 24% on randomization, and 74% on exclusion criteria. Seventy-six percent of the studies were conducted on humans, 10% on animals, and the remainder on other models. Twenty-nine percent of the studies were retrospective, 27% prospective observational with control groups, and 26% prospective interventional trials. Sixty-six percent of all submitted interventional trials had positive outcome by our initial definition, 83% by the positive P-value definition, and 80% by effect size. Respective figures for all submitted observational prospective studies were 70%, 92%, and 80%.

Three hundred eighty studies met the criteria for logistic regression (see “Methods”). Most measures of scientific merit did not predict the decision to accept an abstract for presentation. Instead, this decision was most strongly related to positive results and the reviewers’ subjective originality score, while controlling for institutional funding, study design, randomization, blinding, controls, exclusion criteria, and sample size (Table 1). Results were similar for the subgroup of 166 prospective studies with control groups except that sample size greater than 50 was also predictive (odds ratio [OR], 2.4; 95% CI, 1.0-6.5) and similarly controlled. Full regression results are available from the authors.

One hundred seventy-nine (36%) of the submitted abstracts were accepted for presentation at the meeting (11 reports on teaching methods were excluded). The SAEM committee scores determining acceptance correlated best with our subjective quality scale (R = 0.57) and originality factor (R = 0.49). Two hundred fourteen (43%) of the 492 studies submitted were published, an average of 18 months after presentation, in 44 journals with impact factors ranging from 0.23 to 24.5. A follow-up questionnaire to authors of unpublished papers was returned by 226 authors and identified 21 publications not found in MEDLINE.³ One hundred four (49%) of the 214 studies ultimately published were rejected for presentation at the meeting. The mean impact factor of the publishing journal did not differ for those papers rejected for the meeting vs those accepted (1.48 vs 1.19; P = .47), nor did time to publication.

One hundred forty-seven studies (70%) were published in emergency medicine specialty journals. The remaining studies were published in 39 other journals, including American Journal of Public Health, Annals of Internal Medicine, JAMA, The New England Journal of Medicine, Pediatrics, and Stroke. Of all the studies published after the meeting, 38% had been published 1 year later, 68% in 2 years, 88% in 3 years, and 95% in 4 years.

Publication of a full manuscript in the main group of 380 studies was related most strongly to abstract acceptance and sample size, controlling for the previous variables by logistic regression (Table 2). Results in the 166 prospective studies with control groups were identical. Results of these analyses did not differ using either of the 2 definitions of positive outcome (“Methods”).

Data allowing calculation of effect size was reported in only 122 (66%) of 186 prospective studies. Positive-outcome bias was evident for all studies submitted before any screening (Figure). Submissions with less positive effect size were then disproportionately rejected for presentation. At the level of publication of full manuscripts, the same bias again appeared. A funnel plot (Figure) shows the absence of expected negative effect sizes at low sample size, which is the hallmark of positive-outcome bias.

Numerical testing confirms the funnel plots. The mean effect size of all submitted papers was 0.71 (95% CI, 0.40-1.01). The mean effect size of papers accepted for the meeting was 0.92 vs 0.45 for those rejected. The mean effect size of papers eventually published was 0.96 vs 0.45 for those never published (P = .03, Kruskal-Wallis analysis of variance). Effect size contributed much more to acceptance or publication than study sample size.

LIMITATIONS

Our study was limited to 1 specialty, but 103 medical schools contributed and the publication rate was comparable with 31 other academic society meetings.⁶ Research from this meeting was published in 44 journals, 39 of them outside this specialty. The emergency medicine literature, the abstracts we studied, and the general medicine literature are identical in the proportion of studies with positive outcomes.¹ Effect size data were available for only a minority of studies, and most studies were not randomized controlled trials. These subgroups might not be representative of

Table 1.—Characteristics Predicting Acceptance for Presentation in Main Group of 380 Submitted Studies*  

<table>
<thead>
<tr>
<th>Variable</th>
<th>Results</th>
<th>Adjusted Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive outcome</td>
<td>1.99 (1.07-3.84)</td>
</tr>
<tr>
<td>Originality score</td>
<td>Low, medium, high</td>
<td>2.07 (1.13-3.89)</td>
</tr>
<tr>
<td>Quality score</td>
<td>Low, medium, high</td>
<td>1.53 (0.99-2.41)</td>
</tr>
</tbody>
</table>

*Excludes surveys, simulations, and descriptions of methods. See “Methods” section for all variables tested. Logistic regression analysis vs null model χ² = 50.0, P < .001. Copas-le Cessie-van Houwelingen-Hosmer-Lemeshow goodness-of-fit test, P = .86.

Fox and Lemeshow²² and Harrell,²³ assessed for goodness of fit and subjected to bootstrap validation using S-Plus version 4.0, release 3 (Mathsoft Inc, Seattle, Wash) and Harrell extensions (http://www.lib.stat.cmu.edu). Correlation coefficients were calculated using JMP software 3.2.2 for Macintosh (SAS Institute, Inc, Cary, NC). We calculated the effect size of all interventions for the subgroup of prospective studies with controls in the usual fashion, except that SDs were not available.

RESULTS

Five hundred research abstracts were submitted to the SAEM selection committee from a total of 144 institutions, 103 of which had formal US medical school affiliations. Eight duplicates were deleted, leaving 492 abstracts as the basis of our study. The submitting schools averaged an NIH funding rank of 56 (of all US medical schools) (95% confidence interval [CI], 49-62).

Thirty percent of studies did not state a hypothesis, and 49% did not report on blinding, 24% on randomization, and 74% on exclusion criteria. Seventy-six percent of the studies were conducted on humans, 10% on animals, and the remainder on other models. Twenty-nine percent of the studies were retrospective, 27% prospective observational with control groups, and 26% prospective interventional trials. Sixty-six percent of all submitted interventions had positive outcome by our initial definition, 83% by the positive P-value definition, and 80% by effect size. Respective figures for all submitted observational prospective studies were 70%, 92%, and 80%.

Three hundred eighty studies met the criteria for logistic regression (see “Methods”). Most measures of scientific merit did not predict the decision to accept an abstract for presentation. Instead, this decision was most strongly related to positive results and the reviewers’ subjective originality score, while controlling for institutional funding, study design, randomization, blinding, controls, exclusion criteria, and sample size (Table 1). Results were similar for the subgroup of 166 prospective studies with control groups except that sample size greater than 50 was also predictive (odds ratio [OR], 2.4; 95% CI, 1.0-6.5) and similarly controlled. Full regression results are available from the authors.

One hundred seventy-nine (36%) of the submitted abstracts were accepted for presentation at the meeting (11 reports on teaching methods were excluded). The SAEM committee scores determining acceptance correlated best with our subjective quality scale (R = 0.57) and originality factor (R = 0.49). Two hundred fourteen (43%) of the 492 studies submitted were published, an average of 18 months after presentation, in 44 journals with impact factors ranging from 0.23 to 24.5. A follow-up questionnaire to authors of unpublished papers was returned by 226 authors and identified 21 publications not found in MEDLINE.³ One hundred four (49%) of the 214 studies ultimately published were rejected for presentation at the meeting. The mean impact factor of the publishing journal did not differ for those papers rejected for the meeting vs those accepted (1.48 vs 1.19; P = .47), nor did time to publication.

One hundred forty-seven studies (70%) were published in emergency medicine specialty journals. The remaining studies were published in 39 other journals, including American Journal of Public Health, Annals of Internal Medicine, JAMA, The New England Journal of Medicine, Pediatrics, and Stroke. Of all the studies published after the meeting, 38% had been published 1 year later, 68% in 2 years, 88% in 3 years, and 95% in 4 years.

Publication of a full manuscript in the main group of 380 studies was related most strongly to abstract acceptance and sample size, controlling for the previous variables by logistic regression (Table 2). Results in the 166 prospective studies with control groups were identical. Results of these analyses did not differ using either of the 2 definitions of positive outcome (“Methods”).

Data allowing calculation of effect size was reported in only 122 (66%) of 186 prospective studies. Positive-outcome bias was evident for all studies submitted before any screening (Figure). Submissions with less positive effect size were then disproportionately rejected for presentation. At the level of publication of full manuscripts, the same bias again appeared. A funnel plot (Figure) shows the absence of expected negative effect sizes at low sample size, which is the hallmark of positive-outcome bias.

Numerical testing confirms the funnel plots. The mean effect size of all submitted papers was 0.71 (95% CI, 0.40-1.01). The mean effect size of papers accepted for the meeting was 0.92 vs 0.45 for those rejected. The mean effect size of papers eventually published was 0.96 vs 0.45 for those never published (P = .03, Kruskal-Wallis analysis of variance). Effect size contributed much more to acceptance or publication than study sample size.

LIMITATIONS

Our study was limited to 1 specialty, but 103 medical schools contributed and the publication rate was comparable with 31 other academic society meetings.⁶ Research from this meeting was published in 44 journals, 39 of them outside this specialty. The emergency medicine literature, the abstracts we studied, and the general medicine literature are identical in the proportion of studies with positive outcomes.¹ Effect size data were available for only a minority of studies, and most studies were not randomized controlled trials. These subgroups might not be representative of
Our results show that acceptance and publication decisions were more strongly associated with acceptance than traditional measures of scientific quality, such as study design, randomization, sample size, and blinding (Tables 1 and 2).

Positive-outcome bias appeared again in the selection process for publication in a journal. Full publication was best predicted by acceptance at the meeting and study size, rather than study methods or quality. Our results confirm a smaller study of the cancer literature, which did not control for scientific quality.

A number of potential solutions, such as trials registries, have been proposed to remedy positive-outcome bias.21,25-27

We offer one more solution: that all studies submitted to scientific meetings be published as abstracts, indicating whether or not they were chosen to be presented. This might encourage researchers to submit studies with negative findings, and readers and researchers could more easily identify the entire spectrum of research. Journals might adopt a similar practice, publishing the abstracts of all submitted manuscripts.

Despite the mandatory structured format, 49% of SAEM abstracts failed to report adequately about blinding, 74% about exclusion criteria, 24% about randomization, and 14% about sample size. Perhaps because these deficiencies made the merit of the research difficult to evaluate, acceptance for presentation at the meeting predicted publication as a full manuscript with a sensitivity of only 51%, a specificity of 71%, a positive predictive value of 57%, and a negative predictive value of 66%.

We wish to thank Mary Ann Schropp, executive director of the SAEM, for her generous help with data, and John Gallagher, MD, for sharing the institutional research funding data.

References


cents. We agree that the methods used in our study do not provide the most accurate assessment of vigorous physical activity in children. While use of measurement devices such as heart rate recording or accelerometers is preferable, such measurements are not practical in a large national survey. Although the accuracy of self-report of vigorous activity from children does warrant examination, it is unlikely that most parents could provide data that were more reliable. The subjective assessment of vigorous physical activity by a child may also differ from what parents consider vigorous, especially among overweight children.

We continue to stress that all health care professionals should create opportunities to encourage youth to minimize sedentary behaviors and to participate in physical activity that may be continued for a lifetime of healthy living.

Ross E. Andersen, PhD
The Johns Hopkins University School of Medicine
Baltimore, Md
Carlos Crespo, DrPH, MS
American University
Washington, DC
Susan J. Bartlett, PhD
The Johns Hopkins University School of Medicine
Michael Pratt, MD, MPH
Centers for Disease Control and Prevention
Atlanta, Ga


The JAMA Patient Page

To the Editor.—Over the past few decades, the demand for consumer health-related information has increased dramatically. Easy-to-understand health-related information correlates with increased patient satisfaction.1-4 Patient education is both a science and an art. I was pleased to see the JAMA Patient Page,5 and I find it useful in educating patients. I have evaluated many patient-oriented computer software programs and databases, and I found the Patient Page to be complete, accurate, and useful.

However, in future issues, extra attention should be given to the type size and sentence spacing;6 Type size is crucial to readability; at least 12-point type is generally recommended, especially for older patients. Adequate spacing between letters, lines, and paragraphs also enhances readability. Typically, text spacing should have no more than −3 kerning (space between the letters of words) on the page. We make every effort to enhance readability through design, use of white space, and pull-out information. Although we would like to increase the size of the font, doing so would seriously limit the amount of information we can include on the page. JAMA plans to conduct focus groups to see how useful the Patient Page has been for medical professionals and patients. We will explore issues of readability, value of the information, and ways to improve the page. After more comprehensive feedback about the page, we will consider whether changes need to be made to improve the overall quality of the Patient Page.

Mi Young Hwang, MSJ
Writer, JAMA Patient Page
Richard M. Glass, MD
Deputy Editor, JAMA

CORRECTION

Incorrect Figure.—In the Peer Review article entitled “Positive-Outcome Bias and Other Limitations in the Outcome of Research Abstracts Submitted to a Scientific Meeting,” published in the July 15, 1998, issue of The Journal (1998;280:54-250), the Figure was printed with the incorrect inset graph. The inset graph should be a linear plot of the studies accepted for presentation, not a log axis plot. The Figure is correct as printed below.

Funnel plot of all 122 prospective studies submitted to the meeting that reported effect sizes (see text). Studies subsequently accepted for presentation at the meeting are plotted separately from those rejected. Inset graph similarly shows results for all submitted research with effect-size data at the journal publication level. The absence of the left (negative) base of the inverted funnel region suggests the presence of selection bias in favor of positive outcomes.