Common Statistical and Research Design Problems in Manuscripts Submitted to High-Impact Public Health Journals

Alex H.S. Harris[1], Rachelle N. Reeder[*,1] and Jenny K. Hyun[2]

[1]*Center for Health Care Evaluation, VA Palo Alto Health Care System and Stanford University School of Medicine, 795 Willow Road (MPD-152), Menlo Park, CA 94025, USA*

[2]*National Center for PTSD, VA Palo Alto Health Care System, 795 Willow Road, Menlo Park, CA 94025, USA*

Abstract: *Introduction*: Journal editors and statistical reviewers are often in the difficult position of catching statistical and research design problems after data have been collected and analyzed. The authors sought to learn from editors and reviewers of major public health journals what common statistical and design problems they find in submitted manuscripts and what they wished to communicate to authors regarding these issues.

Materials and Methodology: Editors and statistical reviewers of 55 high-impact public health journals were surveyed to determine what statistical or design problems they encounter most often. The authors analyzed text responses using content analysis to identify major themes.

Results: Editors and reviewers (n = 25) who handle manuscripts from 26 high impact public health journals responded to the survey. The most commonly cited problems included failure to adequately describe statistical models and map them onto research questions, inadequate consideration of sample size, poor control of confounding, and inappropriate reliance on parametric tests.

Conclusions: The scientific quality of public health research and submitted reports could be greatly improved if researchers addressed frequently encountered methodological and analytic issues.

INTRODUCTION

Editors and statistical reviewers for public health journals are often in the difficult position of catching serious problems in submitted manuscripts after data have been collected and analyzed. Sometimes these problems can be remedied by requiring a revision in which the author conducts additional analysis or adjusts the interpretation of the results. However, if the study has fundamental design flaws or if the gap between actual and appropriate analysis strategy is perceived by the editors to be too large, manuscripts are rejected. Typically, authors then submit the manuscript to progressively less selective journals until the manuscript is accepted. This process is time-consuming, expensive, frustrating, and perhaps avoidable.

With the aim of improving the scientific quality of public health research and research reports, the authors of this study conducted a survey of the editors and statistical reviewers of high-impact public health journals to learn about the most common and serious statistical/design problems they encounter in submitted manuscripts. Although individual public health journals or methodological experts have occasionally published statistical and research advice intended for psychiatry researchers [1-7], to our knowledge, no survey like this has previously been conducted in the public health field. Armed with information from editors and reviewers of public health journals, researchers might avoid statistical and design problems in their own manuscripts and may be able to recognize them in others' research. By addressing these issues, researchers might dramatically increase the odds of making important scientific contributions and getting published in more prestigious journals.

*Address correspondence to this author at the Center for Health Care Evaluation (MC:152), VA Palo Alto Health Care System, 795 Willow Road, Menlo Park, CA 94025, USA; Tel: 650-493-5000, ext. 27814; Fax: 650-617-2690; E-mail: Rachelle.Reeder@va.gov

MATERIALS AND METHODOLOGY

Sample and Procedure

The 40 public health journals with the highest impact factor reported in both the 2007 Science Journal Citation Report and the 2007 Social Science Journal Citation Report [8] were culled of duplicates, yielding 55 unique high-impact journals. The Editor-in-Chiefs and known statistical editors of these journals (n = 58; 53 editor-in-chiefs, 4 statistical editors, 1 managing editor) were mailed a letter informing them of the online survey. Within a week of receiving the letter, the editors and reviewers were sent an email invitation and electronic link to the survey. One week after the first email invitation, they were sent a reminder email. All communication stressed that participation was strictly voluntary, that all responses were confidential, and that published results of the survey would not link specific comments or responses to specific editors/reviewers or to specific journals. The authors of this study also asked Editors-in-Chiefs to forward the invitation to statistically- and methodologically-oriented editors or reviewers upon which they rely, in addition to or in lieu of completing the

1874-9445/09

survey themselves. This research was part of a larger study conducted from January to June of 2008 that also surveyed psychiatry, psychology, health services, and medical journals.

Survey Content

The survey contained three parts:

(a) Brief questions about the journals for which the respondent served, how many manuscripts they handled in a typical month, and areas of methodological or research design expertise;

(b) The main open-ended question which asked the following: "As an editor-in-chief or a statistically oriented reviewer, you provide important statistical guidance to many researchers on a manuscript-by-manuscript basis. If you could communicate en masse to researchers in your field, what would you say are the most important (common and high impact) statistical issues you encounter in reviewing manuscripts? Please describe the issues as well as what you consider to be adequate and inadequate strategies for addressing them."; and

(c) One to four follow-up questions based on the respondents' self-identified primary area of statistical or methodological expertise. These questions were developed by polling a convenience sample of over 50 researchers regarding what methodological or statistical questions they would want to ask the editors or statistical reviewers of major journals.

Analysis

Responses to open-ended questions were analyzed by the authors using content analysis to identify major themes. For ease of presentation, themes are divided into two categories:

(a) Conceptual and Reporting Problems;

(b) Statistical and Sampling Issues. The authors of this study present the issues raised by respondents as well as their suggested remedies. Where appropriate, other resources are included about the research problem and ways to address it.

RESULTS

Characteristics of Responders

Respondents were 25 editors and statistical reviewers who handle manuscripts for 26 of the 55 journals in the sampling frame. Respondents reported handling a median of 10 manuscripts per month (range: 1 to 120), and a primary expertise of epidemiology (38.9%), design and analysis of clinical trials/ experiments (27.8%), general statistics (16.7%), quasi-experimental methods/ observational studies (5.6%), analysis of longitudinal research (5.6%), and measurement (5.6%). Five of the respondents reported handling manuscripts from the *American Journal of Epidemiology*, more than any other journal, 4 other journals had 2 respondents each handling manuscripts for them, and the other 20 journals had 1 respondent each.

Respondents' Advice on Conceptual and Reporting Problems

Survey respondents emphasized that research questions, study design, sampling strategies, analysis, and conclusions need to be congruent and appropriately described, as well as noted the surprising frequency with which manuscripts fail to meet these basic requirements. Specifically, they noted the following issues:

(a) Research questions and hypotheses are often not clearly operationalized and stated.

(b) Statistical models are commonly not directly related to the hypotheses and/or are not completely described.

(c) Results (test statistics, parameter estimates, degrees of freedom, etc.) are only partially presented and often not checked for accuracy. According to survey respondents, omissions and errors in numerical results are common.

(d) Steps in analyses (choice of assessed factors, outcomes, collapsing groups, etc.) are often not made explicit so that the presented analysis appears to be the pre-specified one. Especially when the presented analyses are just one of many conducted, the other analyses should be described and the sensitivity of the results to analytic choices discussed. (e) A clear separation should exist between exploratory (post-hoc) hypothesis-testing and primary, pre-planned hypothesis-testing. The sin of "Cherry Picking" occurs when researchers present the results of a fishing expedition as if it were a test of a single pre-specified hypothesis. Exploratory analyses are valuable but need to be identified as such.

(f) Authors need to understand the hypotheses their statistical models are testing, and match interpretations and conclusions directly to the results of these tests. Respondents noted a generally poor understanding of how various research designs affect the ability to make causal claims. A common over-interpretation of statistical results is attribution of causality in observational studies [9]. However, even very sophisticated statistical models (e.g., instrumental variable analysis) cannot completely address residual confounding and allow unreserved causal claims based on observational data [9].

(g) Authors often fail to adequately describe their sampling strategy, including procedures and inclusion/ exclusion criteria, or discuss how their sampling strategy affects the scope of inference.

(h) The description of the results also should clearly describe the extent of missing data. (i) According to respondents, authors sometimes fail to connect statistical significance with epidemiological importance. In studies with large samples, it is possible to detect statistically significant associations with no real-world meaning.

Respondents' Advice On Statistical and Sampling Issues

In addition to the design- and analysis-related problems mentioned, the following statistical and sampling issues were

highlighted by respondents as common and important problems in submitted manuscripts:

(a) Statistical power is often not calculated a priori so that sample sizes are adequate to detect clinically meaningful effects. Researchers often fail to pre-specify the magnitude of clinically meaningful differences in outcomes [10,11].

(b) Often the research question or data are not matched to an appropriate statistical method. For example, repeated t-tests without alpha corrections are sometimes used to compare groups of observations made at several time points instead of a proper longitudinal analysis (e.g., modeling trajectories with mixed effects regression) [12]. Another example is using odds ratios instead of relative risks when incidence rates are low, which exaggerates the magnitude of the association [13,14].

(c) The underlying assumptions of the chosen statistical methods are not checked or properly addressed. Standard large sample statistical methods are often used for small samples and in instances where the model assumptions are obviously violated. Dependencies in data (e.g., repeated measures of individuals, individuals grouped within multiple organizations) need to be analyzed with models (e.g., mixed-effect regression models aka HLM or generalized estimating equations) that do not have assumptions of independence and adjust for correlation within clusters or across time [12,15,16]. Linear models are used without considering if the relationship has non-linear forms. Several respondents lamented the underutilization of non-parametric, robust, and Bayesian modeling, which can avoid, or at least model uncertainty regarding, some of the untenable assumptions of classic parametric analyses [17-19].

(d) Researchers often transform numeric variables (body weight) into categories with justification, thereby discarding information. Categorization of numeric variables may be justified when natural or clinical thresholds exist, but should not be done merely to fit data into familiar categorical statistical frameworks.

(e) A statistically sound method of addressing the missing data should be used (e.g., multiple imputation) and sensitivity analyses should be conducted to evaluate the impact of the missing data assumption on the results [20-23].

(f) Many public health studies are quasi-experimental, that is concerned with estimating the association between an exposure variable (e.g., living within 1 mile of a major pollution source) and an outcome (e.g., respiratory disease) without the benefit of randomization. Choosing and controlling for potential confounders requires judgment and a sound statistical approach. Knowing what data to collect is a critical step in such research, because once data on an important confounder is omitted, this problem is not easily remedied with sophisticated analyses. Even when data on important confounders is available, the ideal approach for analyses is debatable. Favored quasi-experimental strategies are propensity score analysis [24-26] instrumental variable analysis [9,27,28] Regardless of statistical approach, authors should understand the limitations of the quasi-experimental design and not attribute causality to exposure variables of interest in discussing results of quasi-experimental studies, and should consider other explanations of results such as self-selection or confounding [29].

(g) Another common misinterpretation stems from the fact that many researchers do not understand the distinction between an equivalence trial and a difference trial [30]. Especially in small trials, failure to reject the null hypothesis that two treatments are equivalent does not establish the equivalence of the treatments. Unless the researchers specifically designed an adequately powered equivalence or non-superiority trial, the proper interpretation of a null result is "we failed to find a difference but that does not mean one does not exist."

CONCLUSIONS

Journal editors and statistical reviewers of high impact public health journals identified several common problems that significantly impact the quality, and therefore acceptability, of submitted manuscripts. However, this study was not intended to provide tutorials or even extensive guidance on the identified problem areas. Often the respondents called for a return to fundamental research practices that are already familiar to all researchers (e.g., checking model assumptions, clearly and fully describing the analysis and results), a call for greater transparency in the analytic process, fidelity to the scientific process that calls for a priori specification of hypotheses, and understanding the limitations to causality and generalizability inherent in decisions regarding sample selection, research design, and statistical analyses.

In other cases, the responses refer to new analytic strategies or standards that permeate the research community unevenly although familiarity with these methods may be increasingly necessary in the public health realm [31]. For instance, not all researchers know that mixed effects regression models (aka HLM) are now de rigor for analyzing non-independent data, such as longitudinal or clustered data. Likewise, the standards for handling missing data are rapidly evolving, but many researchers continue to employ old and statistically discredited approaches (e.g., last observation carried forward). With no definitive source for researchers to consult to learn about new and evolving standards, they often learn about them in the form of manuscript reviews. By surveying editors and reviewers directly, the authors of this study identified research design and statistical issues that authors can address before they submit their manuscripts. A worthy strategy for upgrading the scientific quality of the research reports submitted for publication is for researchers to enlist the input of a knowledgeable statistician during the planning, analysis, and report preparation phases of research projects [32]. This study's reference list also provides accessible citations for author-researchers that can serve as complimentary resources to further improve the quality of manuscripts submitted to scientific journals.

Although statistical controversies do exist and were reflected somewhat in the survey responses (e.g., propensity scores analysis *vs*. instrumental variable analysis, GEE *vs*. mixed-effects regression), these debates concern which of several reasonable strategies are optimal. However, the main recommendations of respondents were uniform and uncontroversial in that they warned against indefensible strategies (e.g., ignoring issues of missing data) and offered reasonable alternatives.

Several limitations of this survey study are noteworthy. The sampling strategy was to invite Editor-in-Chiefs and statistical editors to complete the survey and/or forward the invitation to statistically-oriented reviewers upon whom they rely. We did not collect data on the position of the anonymous respondents, therefore we do not know the composition of the response sample or how this may have influenced the results. Also, the response rate, although similar to other surveys of journal editors, raises questions about response bias and generalizability. Although we cannot know why editors did not respond, some indicated they were simply too busy. Although we have no reason to suspect substantial response bias, it is possible that the opinions and experiences of non-responders are markedly different from responders. This limitation needs to be considered in the interpretation of our findings.

Findings from this study are limited by the expertise of the sample. Given the broad scope of public health research and practice from microbiologist to policymaker, specific emphasis on certain statistical methods or research designs may or may not be appropriate in different public health arenas. However, as stated in the Institute of Medicine's report *The Future of the Public's Health in the 21st Century*, public health initiatives and policy will become increasingly dependent on data-driven, empirical results [33]. Improvements in research design, data analysis, interpretation and reporting of results are part the larger role of public health in the future.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to report.

ACKNOWLEDGEMENTS

This study was supported by the VA Office of Research and Development, Health Services Research and Development Service (MRP-05-168-1). As employees of the U.S. government, we cannot transfer the copyright for this work.

REFERENCES

[1] Gail M, Byar D, Pechacek T, Corle D. Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT). Control Clin Trials 1992; 13(1): 6-12.

[2] Donner A. Statistical methodology for paired cluster designs. Am J Epidemiol 1987; 126(5): 972-9.

[3] Little R, Yau L. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. Psychol Methods 1998; 3: 147-59.

[4] Wijeysundera DN, Austin PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. J Clin Epidemiol 2009; 62: 13-21.

[5] Johansson PM, de Leon AP, Sadigh S, Tillgren PE, Rehnberg C. Statistical modelling needed to find the effects from a community-based elderly safety promotion program. Eur J Public Health 2009; 19(1): 100-5.

[6] Heo M, Leon AC. Statistical power and sample size requirements for three level hierarchical cluster randomized trials. Biometrics 2008; 64(4): 1256-62.

[7] Blance A, Tu YK, Baelum V, Gilthorpe MS. Statistical issues on the analysis of change in follow-up studies in dental research. Commun Dent Oral Epidemiol 2007; 35(6): 412-20.

[8] Institute for Scientific Information. Journal Citation Report: Thompson Scientific 2007.

[9] Kim MH, Do YK. Strengthening causal inference in studies using non-experimental data: an application of propensity score and instrumental variable methods. J Prev Med Public Health 2007; 40(6): 495-504.

[10] Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. Arch Gen Psychiatry 2006; 63(5): 484-9.

[11] Zumbo B, HUbley A. A note on misconceptions concerning prospective and retrospective power. The Statistician 1998; 47(2): 385-8.

[12] Singer JD, Willett JB. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. London, England: Oxford University Press 2003.

[13] Simon SD. Understanding the odds ratio and the relative risk. J Androl 2001; 22(4): 533-6.

[14] Schechtman E. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat--which of these should we use? Value Health 2002; 5(5): 431-6.

[15] Raudenbush SW, Bryk AS. Hierarchical Linear Models : Applications and Data Analysis Methods. 2nd ed. Thousand Oaks, CA: Sage 2001.

[16] Hardin JW, Hilbe JM. Generalized Estimating Equations. Boca Raton, FL: Chapman & Hall/CRC 2003.

[17] Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. Int J Epidemiol 2006; 35(3): 765-75.

[18] Greenland S. Bayesian interpretation and analysis of research results. Semin Hematol 2008; 45(3): 141-9.

[19] Sprent P, Smeeton N. Applied Nonparametric Statistical Methods. 3rd ed. Boca Raton, FLA: Chapman &Hall/CRC 2001.

[20] Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychol Methods 2002; 7(2): 147-77.

[21] Schafer JL. Analysis of Incomplete Multivariate Data. New York, NY: Chapman & Hall 1997.

[22] Rubin DB. Inference and missing data. Biometrika 1976; 63: 581-92.

[23] Daniels M, Hogan J. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Senitivity Analyis. New York, NY: Chapman & Hall 2008.

[24] Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. Stat Med 2005; 25(12): 2084-106.

[25] McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 2004; 9(4): 403-25.

[26] Luellen JK, Shadish WR, Clark MH. Propensity scores: an introduction and experimental test. Eval Rev 2005; 29(6): 530-58.

[27] Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007; 297(3): 278-85.

[28] Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. Am J Epidemiol 2008; 169: 273-84.

[29] Shadish R, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA: Houghton Mifflin Company 2001.

[30] Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006; 295(10): 1152-60.

[31] Diez-Roux AV. Multilevel analysis in public health research. Annu Rev Public Health 2000; 21: 171-92.

[32] Herbison P. How to make your article more acceptable for the statistical reviewer. Neurourol Urodyn 2007; 26(3): 318-22.

[33] Institute of Medicine (U.S.). Committee on Assuring the Health of the Public in the 21st Century. The Future of the Public's Health in the 21st Century. Washington, DC: National Academies Press 2003.

Received: April 06, 2009 Revised: October 02, 2009 Accepted: October 08, 2009