Original Research

Randomness and inference in medical and public health research

Matthew J. Hayat, PhD[1] and Thomas R. Knapp, EdD, FAAN[2]

[1]School of Public Health, Georgia State University; and [2]University of Rochester and The Ohio State University

Corresponding author: Matthew J Hayat ● School of Public Health, Georgia State University ● P O Box 3984, Atlanta, GA 30302-3894 ● mhayat@gsu.edu

ABSTRACT

Background: The purpose of this study was to provide a basis for describing the types of randomness used and statistical inferences reported in the medical and public health research literature.

Methods: A study was conducted to quantify the types of research designs and analyses used and reported in medical and public health research studies. A stratified random sample of 198 articles from three top-tier medical and public health journals was reviewed, and the presence or absence of random assignment, random sampling, p-values, and confidence intervals, as well as type of research design, were quantified.

Results: Random sampling was used in 58 (29.3%) and random assignment in 21 (10.6%) articles. Most (n=125; 63.1%) research studies did not report random assignment or random sampling; however, statistical inference was applied in more than 90%.

Conclusions: Results revealed a concerning overuse of statistical inference. Incorrectly applying statistical inference when not warranted has potentially damaging medical and public health consequences. Researchers should carefully consider the appropriateness of using statistical inference in medical and public health research.

Key words: randomization, statistical inference, research design

https://doi.org/10.21633/jgpha.7.102

BACKGROUND

Three decades have passed since a series of articles appeared in the public health literature about significance tests and confidence intervals (Walker, 1986a; Fleiss, 1986a; Fleiss, 1986b). Previously, significance tests and their corresponding p-values were the mainstay of statistical reporting in the medical and public health research literature (Walker, 1986b; Fleiss, 1986c). Controversy arose as a result of the suggestion that significance tests were over-used, p-values were too limited, and confidence intervals were better and more informative (Gardner and Altman, 1986). Since then, numerous articles have appeared generally supporting one of three views: (i) arguing the merits of significance tests and p-values, (ii) promoting the use of confidence intervals, and (iii) advocating the reporting of both (Poole, 1987; McCormack et al, 2013; Feinstein, 1998; Sterne, 2002).

A noticeable gap in the literature exists for a fourth viewpoint, namely, questioning the appropriateness of any use of statistical inference for certain studies (Feinstein, 1998, is a possible exception). The goal of statistical inference is to generalize results obtained from a study sample to some larger (and usually practically unobtainable) population from which the sample was drawn. A fundamental assumption necessary for deciding if the study sample is representative of the population is selection of a random sample, for which the valid use of classical statistical inference depends. Absence of random sampling prohibits the use of significance testing and confidence intervals. In this case, the controversy about whether to report p-values or confidence intervals is a moot point. Neither should be reported.

Randomness in a study can refer to two distinctly different aspects of a research study. Random sampling (often called random selection) is a technique appropriate for generalizing from a sample to a population, whereas random assignment (often called randomization) is an experimental design strategy enabling causal inference. Statistical inference is based upon probability sampling, a necessary condition for generalizing study results beyond the study sample. Consider a 2x2 table for random assignment (yes/no) and random sampling (yes/no). Statistical inference is warranted only in the two cells corresponding to the presence of random sampling. In the cell designating random assignment without random sampling, although classical statistical tests are not warranted, randomization (permutation) tests offer an alternative approach to analysis, although here as well, results are not generalizable beyond the study sample itself (Edgington and Onghena, 2007).

An additional consideration is the unfortunate practice of reporting p-values as indicators of the magnitude of statistical significance. These values do not reflect a trend, magnitude, or size of an effect (Slakter et al, 1991). Classical statistical inference entails specifying a level of significance (α), which is determined before the data are obtained. This level is a fixed quantity, and each study comes with only one α.

However, the practice of denoting statistically significant results in a manner that suggests differing levels has become commonplace, such as displaying * for p<0.05, ** for p<0.01, and *** for p<0.001. This practice may be misleading, since it is easily interpreted as meaning that ** is more significant than * and *** is more significant than **. Within the framework of classical hypothesis testing and applying statistical tests, all that matters is whether the p-value is less than the pre-determined study level of significance.

The purpose of the present study was to provide a basis for describing the types of randomness used and statistical inferences reported in the medical and public health research literature. The data consisted of frequencies of indicators of randomness and inference in a stratified random sample of articles selected from three top-tier public health journals for the year 2013. Here, the results are described, and a discussion regarding current practices for those journals follows.

METHODS

The present study constituted a review of published articles for the year 2013 in the *American Journal of Public Health* (AJPH), the *American Journal of Preventive Medicine* (AJPM), and *Preventive Medicine* (PM). Reviews were limited to quantitative studies and focused on the following journal sections: AJPH's Online Research & Practice and Research & Practice, AJPM's Research Articles, and PM's Regular Articles. Two reviewers (the authors) underwent a self-training process and reliability assessment in order to ensure a uniform method for data collection.

Measures

The information collected for each article included the presence or absence of random assignment, the presence or absence of random sampling, whether or not confidence intervals were reported, and whether or not p-values were reported. The use of asterisks to indicate more than one level of statistical significance (e.g., * for p<0.05, ** for p<0.01, and *** for p<0.001) was tracked. The number of authors, page length, and comments for each article were also recorded.

Random assignment was indicated if treatment allocation was randomly determined, regardless of the unit of assignment (e.g., for cluster randomized trials). Random sampling was indicated whenever a probability sample was described in the article. We considered articles utilizing large national datasets with complex samples (e.g., the National Health and Nutrition Examination Survey, NHANES) to be based upon a random sample. An article was counted as reporting a confidence interval or p-value if either appeared anywhere in the main text, the tables, or figures. If the population of interest was collected in its entirety, it was classified as no random sampling, since a sample was not drawn or needed.

Inter-rater Reliability

The process began with a training sample of 25 research articles taken from AJPH's January 2014 issue. Each reviewer reviewed all articles, followed by a comparison of results and a discussion of all observed disagreement. After a discussion of each difference and once reviewers were satisfied that clear criteria were established for assessing all study measures, a review of 22 research articles in the January 2014 issue was then completed for two other journals for purposes of reliability testing.

Sample

Articles for the main study were selected using stratified random sampling with proportional weighting of each stratum determined by the number of articles appearing in each journal. All research articles from 2013 in the three journals were eligible for review. There were 547 articles, with 280 (50.5%) published in AJPH, 103 (19.2%) published in AJPM, and 164 (30.3%) published in PM. A uniform random number was generated for each article and was used as the basis for selection.

A total of 196 articles (approximately 40% of the eligible collection) were needed to provide a sufficiently large, yet practically manageable, sample size. With rounding, the number of articles per journal were 100 for AJPH, 38 for AJPM, and 60 for PM, or 198 in total. Each reviewer assessed 99 articles.

Data Analysis

A Microsoft Excel spreadsheet was used for data collection, and the SAS software system version 9.3 (SAS Institute Inc., Cary, NC) was used for data analysis. Percent agreement was used to summarize the results of the reliability study, and percentages were used to summarize the measures employed in the main study.

RESULTS

Inter-rater Reliability

After completion of a training sample, 22 articles in two journals were reviewed by each of two reviewers. Five dichotomous yes/no measures were considered in the reliability analysis: random assignment, random sampling, confidence interval, p-value, and use of more than one level of significance. This amounted to 110 data elements. The percent agreement between the two reviewers was calculated. Overall agreement was 93% (102/110 data elements). With respect to each measure, agreement was 91% (20/22) for random assignment, 100% (22/22) for random sampling and for confidence intervals, 82% (18/22) for p-values, and 91% (20/22) for use of more than one level of significance. Disagreements were discussed for the eight data elements. The discrepancies appeared to be reviewer error for six of the eight disagreements, and were unclear or a reflection of poor reporting for the remaining two.

Table 1. Summary Statistics for Study Measures by Journal

	AJPH (n=100)	AJPM (n=38)	PM (n=60)	Total (n=198)
	Count	Count (%)	Count	Count (%)
Random Feature				
Random assignment	6 (6.0)	5 (13.2)	10 (16.7)	21 (10.6)
Random sampling	29 (29.0)	14 (36.8)	15 (25.0)	58 (29.3)
Statistical Inference				
Confidence intervals only	8 (8.0)	4 (10.5)	5 (8.3)	17 (8.6)
p-values only	17 (17.0)	17 (44.7)	9 (15.0)	43 (21.7)
Both	68 (68.0)	14 (36.8)	43 (71.7)	125 (63.1)
Neither	7 (7.0)	3 (7.9)	3 (5.0)	13 (6.6)
*,**,*** with reporting p-values	36 (36.0)	10 (26.3)	12 (20.0)	58 (29.3)
Research Design				
Random assignment, random sampling	2 (2.0)	1 (2.6)	3 (5.0)	6 (3.0)
No random assignment, random sampling	27 (27.0)	13 (34.2)	12 (20.0)	52 (26.3)
Random assignment, No random sampling	4 (4.0)	4 (10.5)	7 (11.7)	15 (7.6)
No random assignment, No random	67 (67.0)	20 (52.6)	38 (63.3)	125 (63.1)

Abbreviations: AJPH, *American Journal of Public Health*; AJPM, *American Journal of Preventive Medicine*; PM, *Preventive Medicine*

Random Assignment and Random Sampling

Table 1 displays frequency distributions for all study measures. Twenty one (10.6%) of the reviewed articles reported using random assignment, with PM reporting the most (10; 16.7%), and AJPM the least (6; 6.0%). Random sampling was used in 58 (29.3%) of the articles. Table 2 shows the types of statistical inference reported for each research design. Few studies had both random assignment and random sampling (6; 3.0%), whereas most of the articles overall (125; 63.1%) and in each of the three journals had neither. A relatively small number of experimental studies were reported, with 15 (7.6%) of these with random assignment but no random sampling. Non-experimental studies with random sampling constituted 52 (26.3%) of the articles.

Table 2. Frequencies of Statistical Inference Reporting by Research Design

	AJPH	AJPM	PM	Total
	Count	Count	Count	Count
Random assignment, random sampling	2	1	3	6
Confidence intervals only	0	0	0	0
p-values only	1	0	0	1
Both	1	1	3	5
Neither	0	0	0	0
No random assignment, random sampling	27	13	12	52
Confidence intervals only	1	2	2	5
p-values only	5	6	2	13
Both	21	5	7	33
Neither	0	0	1	1
Random assignment, no random sampling	4	4	7	15
Confidence intervals only	0	0	0	0
p-values only	1	3	2	6
Both	3	1	4	8
Neither	0	0	2	1
No random assignment, no random sampling	67	20	38	125
Confidence intervals only	7	2	3	12
p-values only	10	8	5	23
Both	43	7	29	79
Neither	7	3	1	11

Abbreviations: AJPH, *American Journal of Public Health*; AJPM, *American Journal of Preventive Medicine*; PM, *Preventive Medicine.*

Statistical Inference

Most articles in AJPH and PM reported both confidence intervals and p-values (125; 63.1%); only 14 (36.8%) in AJPM used both. A small number of articles reported neither (13; 6.6%). AJPM had nearly three times the number of articles reporting p-values without confidence intervals (17; 44.7%) than AJPH (17; 17.0%) or PM (9; 15.0%).

Use of *, **, and *** for reporting p values

Some articles (58; 29.3%) reported p-values with a hierarchy of significance. Of the three journals, AJPH reported this most often (36; 36.0%), followed by AJPM (10; 26.3%) and PM (12; 20.0%).

DISCUSSION

These study findings are alarming, as they suggest statistical inference was inappropriately applied in most of the publications reviewed. However, this is not a surprising result, as comparable work in recent years has shown similar results, suggesting that most published research may be wrong or invalid (Ioannidis, 2005). Readers of the medical and public health research literature need to trust the conclusions published in its journals. The findings presented here could be useful for editors and statistical reviewers in contemplating manuscript reviews and for researchers in deciding what type of statistical analyses to perform and to report. Statistical inference is practical and useful only when the goal is to make inferences about the population based on the sample and is warranted only for studies with a random sample. Without one, the necessary assumptions are not met and inference may not be appropriate (Smith, 1983, and Copas and Li, 1997, relate to highly technical articles attempting to provide partial defenses for the use of formal inferential techniques for some studies in which non-random sampling was employed.)

What should authors report? Statistical reporting guidelines for the three journals reviewed (AJPH, AJPM, PM) each adhere to the American Medical Association's (AMA) Manual of Style, as do many other medical and public health journals (AMA, 2009). The AMA manual sets the standard of reporting for hundreds of public health and medical journals, but it does not mention the necessity of random sampling in order for statistical inferences to be used. Regarding which statistical inference results to report, the AMA manual states:

> "While hypothesis testing often results in the P value, P values themselves can only provide information

> about whether the null hypothesis is rejected. Confidence intervals (CIs) are much more informative since they provide a plausible range of values for an unknown **parameter**, as well as some indication of the **power** of the study as indicated by the width of the CI. Confidence intervals are preferred whenever possible. Including both the CI and the P value provides more information than either alone. This is especially true if the CI is used to provide an interval estimate and the P value to provide the results of hypothesis testing." (AMA, 2009, p 888; bolding included)

Although p-values and confidence intervals are commonly reported in the medical and public health literature, many scientists lack sufficient understanding to interpret them correctly (Wulff et al, 1987). Both are inferential statistics and are meaningful only with respect to making statements about a larger population based on a random sample taken from it. If statistical inference is warranted, and there is a parameter to estimate, a sensible approach is to make use of a confidence interval. Alternatively, if there is a particular hypothesis about a parameter that needs to be tested, hypothesis testing with a statistical test and resulting p-value is appropriate.

Other observations arose from the reviews. Only one of the reviewed studies presented results that departed from the conventional 0.05 level of significance, and instead presented results for 99% confidence intervals. This speaks to how deeply entrenched the arbitrary 0.05 level of significance as the accepted threshold has become in the thinking and publishing of modern day research. A few articles reported standard errors only, leaving it to the reader to incorporate them in determining either p-values or confidence intervals, or both. Lastly, in a few articles, researchers sampled the sample (non-randomly). This was particularly noticeable for studies that used NHANES data.

CONCLUSIONS

Statistical inference is the process of using randomly selected sample data to make inferences about one or more population parameters. Without such randomness, the legitimacy of inferences is called into question. Most articles reviewed in the medical and public health journals consisted of research studies that lacked both random assignment and random sampling. Nonetheless, statistical inference was used in more than 90% of such studies. Only through consideration of the research design can researchers correctly assess whether or not inferences are warranted.

References

American Medical Association. *AMA Manual of Style: A Guide for Authors and Editors. 10th ed.* Oxford: Oxford UP; 2009.

Copas JB, Li HG. Inference for non-random samples. *J R Stat Soc Series B Stat Methodol. 1997;*59(1):55-95.

Edgington ES, Onghena P. *Randomization tests (4th ed)*. London: Chapman & Hall; 2007.

Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. *J Clin Epidemiol*. 1998;51(4):355-60.

Fleiss, JL. Significance tests have a role in epidemiologic research: Reactions to A.M. Walker. *Am J Public Health.* 1986a;76(5):559-560.

Fleiss, JL. Confidence intervals vs significance tests: quantitative interpretation. *Am J Public Health.* 1986b;76(5):587-588.

Fleiss, JL. Dr. Fleiss responds. *Am J Public Health*. 1986c;76(8):1033-1034.

Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*. 1986;292(6522):746-50.

Ioannidis JPA. Why Most Published Research Findings Are False. *PLoS Med*. 2005; 2(8): e124.

McCormack J, Vandermeer B, Allan GM. How confidence intervals become confusion intervals. *BMC Med Res Methodol*. 2013;13(134):1-6.

Poole, C. Beyond the confidence interval. *Am J Public Health*. 1987;77(2):195-199.

Slakter MJ, Wu YW, Suzuki-Slakter NS. *, **, and ***; statistical nonsense at the .00000 level. *Nurs Res*. 1991;40(4):248-9.

Smith TMF. On the validity of inferences from non-random samples. *J R Stat Soc Series A General*. 1983;146(4):394-403.

Sterne JA. Teaching hypothesis tests--time for significant change? *Stat Med*. 2002;21(7):985-94; discussion 995-999, 1001.

Walker, AM. Significance tests represent consensus and standard practice. *Am J Public Health*. 1986a;76(8):1033.

Walker, AM. Reporting the results of epidemiologic studies. *Am J Public Health.* 1986b;76(5):556-558.

Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics? *Stat Med*. 1987;6(1):3-10.