Effect size reporting among prominent health journals: a case study of odds ratios

Brian Chu, Michael Liu, Eric C Leas, Benjamin M Althouse, John W Ayers

ABSTRACT

Background The accuracy of statistical reporting that informs medical and public health practice has generated extensive debate, but no studies have evaluated the frequency or accuracy of effect size (the magnitude of change in outcome as a function of change in predictor) reporting in prominent health journals.

Objective To evaluate effect size reporting practices in prominent health journals using the case study of ORs.

Design Articles published in the American Journal of Public Health (AJPH), Journal of the American Medical Association (JAMA), New England Journal of Medicine (NEJM) and PLOS One from 1 January 2010 through 31 December 2019 mentioning the term ‘odds ratio’ in all searchable fields were obtained using PubMed. One hundred randomly selected articles that reported original research using ORs were sampled per journal for in-depth analysis.

Main outcomes and measures We report prevalence of articles using ORs, reporting effect sizes from ORs (reporting the magnitude of change in outcome as a function of change in predictor) and reporting correct effect sizes.

Results The proportion of articles using ORs in the past decade declined in JAMA and AJPH, remained similar in NEJM and increased in PLOS One, with 6124 articles in total. Twenty-four per cent (95% CI 20% to 28%) of articles reported the at least one effect size arising from an OR. Among articles reporting any effect size, 57% (95% CI 47% to 67%) did so incorrectly. Taken together, 10% (95% CI 7% to 13%) of articles included a correct effect size interpretation of an OR. Articles that used ORs in AJPH more frequently reported the effect size (36%, 95% CI 27% to 45%), when compared with NEJM (26%, 95% CI 17.5% to 34.7%), PLOS One (22%, 95% CI 13.9% to 30.2%) and JAMA (10%, 95% CI 3.9% to 16.0%), but the probability of a correct interpretation did not statistically differ between the four journals (χ²=0.56, p=0.90).

Conclusions Articles that used ORs in prominent journals frequently omitted presenting the effect size of their predictor variables. When reported, the presented effect size was usually incorrect. When used, ORs should be paired with accurate effect size interpretations. New editorial and research reporting standards to improve effect size reporting and its accuracy should be considered.

Introduction

In response to concerns about the accuracy of statistical reporting that informs medical and public health practice, standardised reporting guidelines (eg, Strengthening the Reporting of Observational Studies in Epidemiology [STROBE], Consolidated Standards of Reporting Trials, Preferred Reporting Items for Systematic Reviews and Meta-Analyses and others) have been developed. Increasingly, describing the effect size (the magnitude of change in outcome as a function of change in predictor) is prioritised over reporting p-values. Effect size forms the basis of both clinical importance (eg, treatments achieving statistically, but not clinically, significant improvements in outcomes are unlikely to be implemented) and practice guidelines (eg, treatments with larger effect sizes, with all else equal, are generally preferred). However, few studies have systematically quantified the
accuracy of statistical reporting in practice, and none have evaluated how articles in prominent health journals report the effect sizes of their results.

To this end, we conducted a case study of how effect sizes were reported among studies using ORs from prominent health journals. "Odds" are defined as the ratio of the probability of an outcome occurring to the probability of that outcome not occurring; an OR is the ratio of odds of an outcome between two groups. ORs do not easily translate into colloquial effect size interpretations, but numerous commentaries on how to interpret the magnitude of change in outcome as a function of change in predictor have been published.

As a result, we evaluated if articles published in prominent health journals using ORs reported the effect size of their results and whether these reports were accurate.

**Methods**

**Eligible articles**

Articles published in the American Journal of Public Health (AJPH), Journal of the American Medical Association (JAMA), New England Journal of Medicine (NEJM) and PLOS One from 1 January 2010 through 31 December 2019 mentioning "odds ratio" were obtained using PubMed by searching all available fields, including the title, abstract and keywords. These journals represent prominent medical (JAMA being the most circulated, NEJM having the highest impact factor) and public health (AJPH being published by the largest public health society) journals, and the highest-volume publisher (PLOS One publishing >200,000 articles in the past decade), and help set standards for the quality of scientific reporting in their domains. The search term "odds ratio" was introduced by incorrectly reporting the effect size of ORs, we chose because we sought to characterise statistical reporting in relation to ORs and alternative keywords provided low precision for discovering ORs (eg, only 31% of articles mentioning "odds ratio" also mentioned "logistic regression"). Next, three authors (BC, ML, JWA) randomly sampled articles to identify original research reports that included covariates. We excluded bivariable studies without covariates (nearly always trials) where authors may rely on descriptive results to portray effect size and use ORs as statistical tests rather than to estimate effect sizes. A quota of 100 articles was obtained per journal. Because only 94 NEJM articles met inclusion criteria, 6 articles were obtained from 2009.

**Qualitative assessment of random samples**

The same authors independently coded whether articles reported the effect sizes of ORs by describing the magnitude of change in the outcome as a function of change in the predictor (κ=0.90, overlapping sample n=28). Solely reporting the OR or its directionality was insufficient (table 1). Both the results reported in the abstract and complete text were read to identify interpretations. Articles with at least one effect size interpretation were labelled accordingly, even if other ORs were uninterpreted.

Subsequently, we evaluated whether the reported effect sizes were correct using criteria for the reporting of effect size from ORs developed from Davies et al and Norton et al (table 1). A correct effect size interpretation accurately reflects the definition of the ratio of odds. For example, if authors reported an OR of 1.5, the interpretation "50% increase in odds" is correct, while "50% more likely" is incorrect. Correct interpretations could also include other interpretations resulting from logistic regression (eg, change in probability, marginal risk).

Interpretations were independently reviewed by two authors (BC, ML; κ=0.76). Disagreements were discussed with a third author (JWA, ECL) until consensus was reached. Articles with any incorrect reporting of effect size were labelled as incorrect, even if other ORs were interpreted correctly.

**Analysis**

We computed the per cent of matching articles, including bootstrapped CIs, for the primary outcomes: (1) reporting the effect size from an OR and (2) providing a correct effect size interpretation. To evaluate differences in OR reporting across journals, we performed χ² tests. We evaluated if reporting practices were different among case–control studies (those described by authors as “case–control” or “case control” among our random sample) compared with all other studies using χ² tests. To quantify errors introduced by incorrectly reporting the effect size of ORs, we...
calculated correct interpretations using methods and results reported in the article. 8 We used R 3.6.1 (R Foundation) for all analyses.

**Results**

A total of 6124 articles with abstracts published in *AJPH*, *JAMA*, *NEJM* and *PLOS One* between 2010 and 2019 used ORs. At 13% (n=320), *JAMA* had the highest percentage of articles using ORs, followed by *AJPH* at 9% (n=361), *NEJM* at 8% (n=204) and *PLOS One* at 2% (n=5239). The proportion of articles using ORs in the past decade has declined in *JAMA* and *AJPH*, remained similar in *NEJM* and increased in *PLOS One* (figure 1).

Twenty-four per cent (94/400, 95% CI 20% to 28%) of sampled articles reported the effect size of an OR. Among articles making any effect size interpretation, 57% (54/94, 95% CI 47% to 67%) made an incorrect interpretation. Taken together, 10% (40/400, 95% CI 7% to 13%) of articles included a correct effect size interpretation of an OR. The proportion of articles reporting the effect size of an OR, and accuracy of such reporting has remained stable in the past decade (figure 2). Examples of omitted, incorrect and correct interpretations are included in table 2. Most articles incorrectly reporting the effect size of an OR did so by misinterpreting ORs as risk ratios. For instance, one study found a result “8.3 times more likely” with an OR of 8.32, but when we accounted for the baseline prevalence of the non-exposed group (78.2%) the estimated risk ratio was 1.24. Another study interpreted that “risk ... is 100.7 times as high” based on a calculated OR of 100.7, yet we estimated the risk ratio to be 45.3.

Articles in *AJPH* more frequently (χ²=19.30, p<0.001) reported effect sizes (36%, 95% CI 27 to 45), than *NEJM* (26%, 95% CI 17.5 to 34.7), *PLOS One* (22%, 95% CI 13.9 to 30.2) and *JAMA* (10%, 95% CI 3.9 to 16.0). However, the probability of correctly presenting the effect size did not statistically differ by journal (χ²=0.56, p=0.90).

Articles reporting case–control design (n=42) were slightly less likely to report effect sizes (14%, 95% CI 4% to 24%) and had similar rates of reporting an incorrect effect size (50%, 95% CI 9% to 90%), compared with articles that did not report case–control design.

**Discussion**

Articles in prominent journals often use ORs, but frequently omit reporting the effect size. Reported effect size interpretations were usually incorrect, in some cases by substantial margins. Articles that omitted presenting the effect size typically reported only the associations implied by ORs and their statistical significance (eg, language such as “associated with”). Making informed interpretations thus falls onto readers, who may lack access to sufficient data to make actionable effect size interpretations. Many incorrect interpretations used language such as “probability” or “likely” when referring to the raw OR values, which could incorrectly portray ORs as risk ratios or likelihood differences. Interpreting ORs as risk ratios overstates effect sizes as events become more common; an OR of 2 could arise from event probabilities of 0.01 vs 0.005, 0.5 vs 0.33 or 0.8 vs 0.67, but the corresponding risk ratios would be 2, 1.5 and 1.2.
Because our findings only included prominent journals and required only one interpretation of effect size, omissions and errors may be more common in the literature at-large. Indeed, misuse of ORs has been discussed in several specialty journals.11–13 Such reports may have gone unheeded because they were construed as problems specific to subject specialties. Similarly, opinion pieces on the challenges of effect size reporting, despite publication in prominent journals,7 may not have affected the necessary changes in practice without empirical data to support arguments. Our study is the first to systematically study how effect sizes are reported among studies in general health journals. Our findings are limited by the use of PubMed to search for mentions of “odds ratios.” Because PubMed does not allow for searching through full texts of articles, randomly sampled articles may not represent all articles using ORs. While repositories like PubMed Central archive the full texts of articles, only open-access articles are indexed which excludes many recent articles in the selected journals. Furthermore, we only studied original research that included covariates, thereby excluding some articles with a single predictor variable, such as some randomised controlled trials. We limited our selection because studies with single predictors can report the effect size using descriptive statistics and use ORs to test statistical significance. In contrast, measures including relative risk (probability of outcome in exposed relative to unexposed) or marginal effect (difference in outcome, given change in predictor) are simpler to interpret and compare across studies thereby potentially leading to more frequent and accurate effect size reporting.14 While certain study designs, such as case–control, must use ORs, the argument that reporting ORs alone sufficiently represent interpretable effect sizes is challenged by our findings of frequent erroneous interpretations, even in articles reporting case–control studies.

The consequences of not reporting or misinterpreting effect sizes may be compounded as results move beyond journals. Decision-makers (eg, legislative staff) and research disseminators (eg, news reporters) may lack training to describe the effect size from raw statistical results or to correct misinterpretations.15 Policies and disseminations, once made, are often unobserved by researchers or editors and cannot be easily modified. This is particularly concerning for journals representing best practices in medicine (NEJM, JAMA) and public health (AJPH) that directly reach decision-makers and news reporters.

New standards to improve effect size reporting and its accuracy should be considered. For example, editors could require interpretations of effect size and evaluate their accuracy when evaluating submissions, including requesting supplemental or alternative metrics facilitating ease of effect size interpretation like relative and absolute risks.8–10 We urge developers of reporting guidelines to consider the importance of effect size interpretations, as current guidelines such as STROBE do not require them16 and develop best practices for their use. Making health research more interpretable will make it more actionable for the benefit of public health.

Table 2 Example uses of ORs

<table>
<thead>
<tr>
<th>Example quotes*</th>
<th>Effect size interpretation†</th>
<th>Correct interpretation‡</th>
<th>n (%; 95% CI)§</th>
</tr>
</thead>
<tbody>
<tr>
<td>“In adjusted analyses, gay and bisexual men were more likely than heterosexual men to have poor physical health (AOR = 1.38), disability (AOR = 1.26), and poor mental health (AOR = 1.77).”</td>
<td>No</td>
<td>–</td>
<td>306 (76.5, 72.5 to 80.6)</td>
</tr>
<tr>
<td>“Both 30-day mortality (OR, 0.76; 95% CI, 0.73 to 0.80) and 90-day mortality (OR, 0.73; 95% CI, 0.70 to 0.76) were significantly lower for azithromycin users.”</td>
<td>Yes</td>
<td>No</td>
<td>54 (13.5, 10.1 to 16.9)</td>
</tr>
<tr>
<td>“This represented a 53% reduction in the risk of coronary heart disease among carriers of inactivating NPC1L1 mutations (odds ratio for disease among carriers, 0.47; 95% confidence interval [CI], 0.25 to 0.87; P=0.008).”</td>
<td>Yes</td>
<td>Yes</td>
<td>40 (10, 7.0 to 13.0)</td>
</tr>
<tr>
<td>“Households with a member who received information about chickens from a seminar are 8.3 times more likely to be aware of the vaccine (AOR: 8.32, 95% CI: 2 to 39, p&lt;0.01).”</td>
<td>No</td>
<td>–</td>
<td>156 (39.0, 32.5 to 46.0)</td>
</tr>
<tr>
<td>“Patients in the highest tertile of risk using the combined clinical and genetic model had a 7-fold increased risk of early stent thrombosis vs patients in the lowest tertile (OR, 7.63; 95% CI, 4.18 to 13.91).”</td>
<td>Yes</td>
<td>Yes</td>
<td>54 (13.5, 10.1 to 16.9)</td>
</tr>
</tbody>
</table>

*Each quote is an exact quote from an article meeting the inclusion criteria. Citations have been omitted to avoid singling out any given research group.
†Yes/No indicates if the study made an explicitly written interpretation of the magnitude of any reported OR.
‡Yes/No indicates if the study made a correct interpretation of the magnitude of any reported OR.
§Reports the corresponding n and prevalence with 95% CIs in parentheses.
AOR, adjusted OR.

Effect size interpretation is the measure of the magnitude of change in the probability of outcome and predictor is added to the model, the OR changes.7 14 In contrast, measures including relative risk (probability of outcome in exposed relative to unexposed) or marginal effect (difference in outcome, given change in predictor) are simpler to interpret and compare across studies thereby potentially leading to more frequent and accurate effect size reporting.14 While certain study designs, such as case–control, must use ORs, the argument that reporting ORs alone sufficiently represent interpretable effect sizes is challenged by our findings of frequent erroneous interpretations, even in articles reporting case–control studies.

The consequences of not reporting or misinterpreting effect sizes may be compounded as results move beyond journals. Decision-makers (eg, legislative staff) and research disseminators (eg, news reporters) may lack training to describe the effect size from raw statistical results or to correct misinterpretations.15 Policies and disseminations, once made, are often unobserved by researchers or editors and cannot be easily modified. This is particularly concerning for journals representing best practices in medicine (NEJM, JAMA) and public health (AJPH) that directly reach decision-makers and news reporters.

New standards to improve effect size reporting and its accuracy should be considered. For example, editors could require interpretations of effect size and evaluate their accuracy when evaluating submissions, including requesting supplemental or alternative metrics facilitating ease of effect size interpretation like relative and absolute risks.8–10 We urge developers of reporting guidelines to consider the importance of effect size interpretations, as current guidelines such as STROBE do not require them16 and develop best practices for their use. Making health research more interpretable will make it more actionable for the benefit of public health.
Acknowledgements  We thank Bryan E Dowd, PhD; Richard S Garfein, PhD, MPH; Theo N Kirkland, MD; Matthew L Maciejewski, PhD; John McGready, PhD; Edward C Norton, PhD; Davey Smith, MD, MAS; and Daniel Werb, PhD, for helpful feedback on earlier versions of our study.

Contributors  BC, ML and JWA initiated the project and led the design. BC led the data collection, and all authors participated in the data analysis. All authors participated in the drafting of the manuscript, read and agreed to the final submission. BC, ML and JWA initiated the project and led the design. BC led the data collection, and all authors participated in the data analysis. All authors participated in the drafting of the manuscript, read and agreed to the final submission.

Funding  The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests  None declared.

Patient consent for publication  Not required.

Provenance and peer review  Not commissioned; externally peer reviewed.

Data availability statement  The data used in the study are public in nature. The strategy to replicate our database is available in the text and the data are available on reasonable request from the authors.

Open access  This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iD
Brian Chu http://orcid.org/0000-0002-0803-5529

References