American Journal of Epidemiology

Vol. 00, No. 00
https://doi.org/10.1093/aje/kwac137
Advance Access publication:

Practice of Epidemiology

Causal and Associational Language in Observational Health Research: A Systematic Evaluation

Noah A. Haber*, Sarah E. Wieten, Julia M. Rohrer, Onyebuchi A. Arah, Peter W. G. Tennant, Elizabeth A. Stuart, Eleanor J. Murray, Sophie Pilleron, Sze Tung Lam, Emily Riederer, Sarah Jane Howcutt, Alison E. Simmons, Clémence Leyrat, Philipp Schoenegger, Anna Booman, Mi-Suk Kang Dufour, Ashley L. O'Donoghue, Rebekah Baglini, Stefanie Do, Mari De La Rosa Takashima, Thomas Rhys Evans, Daloha Rodriguez-Molina, Taym M. Alsalti, Daniel J. Dunleavy, Gideon Meyerowitz-Katz, Alberto Antonietti, Jose A. Calvache, Mark J. Kelson, Meg G. Salvia, Camila Olarte Parra, Saman Khalatbari-Soltani, Taylor McLinden, Arthur Chatton, Jessie Seiler, Andreea Steriu, Talal S. Alshihayb, Sarah E. Twardowski, Julia Dabravolskaj, Eric Au, Rachel A. Hoopsick, Shashank Suresh, Nicholas Judd, Sebastián Peña, Cathrine Axfors, Palwasha Khan, Ariadne E. Rivera Aguirre, Nnaemeka U. Odo, Ian Schmid, and Matthew P. Fox

* Correspondence to Dr. Noah A. Haber, 1265 Welch Road, Palo Alto, CA 94305 (e-mail: noahhaber@gmail.com).

Initially submitted October 18, 2021; accepted for publication July 26, 2022.

We estimated the degree to which language used in the high-profile medical/public health/epidemiology literature implied causality using language linking exposures to outcomes and action recommendations; examined disconnects between language and recommendations; identified the most common linking phrases; and estimated how strongly linking phrases imply causality. We searched for and screened 1,170 articles from 18 high-profile journals (65 per journal) published from 2010–2019. Based on written framing and systematic guidance, 3 reviewers rated the degree of causality implied in abstracts and full text for exposure/outcome linking language and action recommendations. Reviewers rated the causal implication of exposure/outcome linking language as none (no causal implication) in 13.8%, weak in 34.2%, moderate in 33.2%, and strong in 18.7% of abstracts. The implied causality of action recommendations was higher than the implied causality of linking sentences for 44.5% or commensurate for 40.3% of articles. The most common linking word in abstracts was "associate" (45.7%). Reviewers' ratings of linking word roots were highly heterogeneous; over half of reviewers rated "association" as having at least some causal implication. This research undercuts the assumption that avoiding "causal" words leads to clarity of interpretation in medical research.

association; causal inference; causal language; observational study

Abbreviations: CI, confidence interval; MeSH, Medical Subject Headings; RCT, randomized controlled trial; SJR, SciMago Journal Rank.

Health sciences research often investigates the relationship between an exposure and an outcome. Causal effects between these variables are often implicitly of interest, including studies based on nonrandom assignment of the exposure. Most researchers are aware that inferring causality may be fraught with difficulty, and that cautious interpretation may be warranted. However, this "caution" often manifests itself as avoiding causal language, potentially at the expense of clarity regarding study objectives and the plausibility of the underlying causal assumptions. Some author guidelines (1) explicitly prohibit the use of causal language in studies other than randomized controlled trials (RCTs), often justified by the inaccurate, but common, belief that causal inference is only possible with RCTs (2, 3).

Health scientists and editors often employ euphemisms or language workarounds (4, 5). For example, researchers may reserve use of causal language for only some parts of the manuscript (6) or use language that can pass as either causal or noncausal. Alternatively, noncausal language may be used throughout the manuscript, but suggested recommendations may imply or require a causal interpretation (7). It is not clear what "counts" as causal language, with no clear standards and few attempts (6, 8–12) to define and categorize what constitutes causal language.

The use of ambiguous language leads to potential disconnects between the authors' intentions, methods, conclusions, and the perceptions of the work by research consumers and decision-makers (4, 5, 13–16). It may also indirectly erode research quality by enabling researchers to make ambiguously causal implications without being accountable to the methodological rigor required for causal inference. Otherwise, noncausal language may morph into causal language in outlets for medical practitioners (7, 10), press releases (17–19), and media reports (16, 20). While some loss of nuance may be attributed to press officers, journalists, and news recipients, too-strong interpretation often starts from the study publications themselves (16).

Despite widespread discussions about causal language use (4, 5, 21), systematic evidence of its use in practice is limited. In a review of 60 observational studies, a fifth were judged to have inconsistencies in their use of causal language (6). Prevalence and use of causal language has been examined in studies concerning the overall medical literature (6, 16, 22), obesity (11), and orthopedics (23), noting that in the latter all uses of causal language in non-RCTs were assumed to be "misuse." To date, there have been no large-scale systematic assessments of language used to link exposures and outcomes in the medical and epidemiologic literature; existing efforts (6, 8–12) heavily focus on binary assessments of the language used (causal vs. noncausal).

This study systematically examined the linking language used in studies with a main exposure and outcome in the high-profile medical and epidemiologic literature. Our objectives were to: 1) identify the linking words and phrases used to describe relationships between exposures and outcomes, 2) generate estimates of the strength of causality stated or implied by the linking phrases and sentences using a guided subjective assessment process, 3) examine the prevalence of action recommendations that would require causal inference to have been made, and 4) examine disconnects between causal implications in linking sentences and action implications.

METHODS

Our target sample consisted of studies quantifying the relationship between a main exposure and an outcome in humans, published in high-profile general health, medicine, or epidemiology journals between 2010 and 2019. Years 2020–2021 were not included due to disproportionate focus on coronavirus disease 2019. The study was preregistered on the Open Science Framework (OSF) (24). Changes from the preregistered protocol are documented and explained in Web Appendix 1 (available at https://doi.org/10.1093/aje/kwac137).

Our search consisted of a preliminary search for appropriate journals and a secondary search for published papers within these journals.

SEARCH

Journal inclusion/exclusion criteria

The "top" journals in health, medicine, and epidemiology were determined by journal ranking from journals listed under Journal Citation Reports (JCR) (25) categories for medicine and public health and SciMago's category for Medicine. The top 200 journals from the SciMago Journal Rank (SJR) (26) and JCR's impact factor rating for medical journals, as well as the top 200 highest impact factor–rating journals for public health as extracted on May 26, 2020, were screened according to the inclusion criteria in the first column in Table 1.

Among the journals meeting these criteria, lists of the 15 highest-ranked journals by 1) impact factor, 2) h-index, and 3) SJR score were combined into a single list without duplicates.

Search terms

We searched PubMed to identify all articles published in an eligible journal between 2010 and 2019 inclusive (Web Appendix 2). Medical Subject Headings (MeSH) terms were used to eliminate articles not meeting inclusion criteria. The search was performed using R, version 4.0.5 (R Foundation for Statistical Computing, Vienna, Austria) and the easyPubMed package (27).

Articles were stratified by journal and whether they had the "randomized controlled trial" MeSH tag. Identified articles were sorted in journal/article type–stratified random order for screening. Disease areas were obtained for each article using the 2020 MeSH tag hierarchy (28).

SCREENING

Study inclusion/exclusion criteria

Studies were eligible for inclusion if they were mainly concerned with the quantitative association of a main exposure/outcome pair, as detailed in the second column of Table 1.

Studies investigating more than 1 exposure/outcome set were excluded because: 1) it would not be possible to assess a main exposure/outcome pair per study; 2) study objectives and designs could not easily be compared with other papers; and 3) it would impose additional strain on the management of the data and review.

Procedures

Articles were screened continuously for each journal until journal quotas were met, with the addition of a small buffer

Table 1. Inclusion and Exclusion Criteria for an Analysis of Causal and Associational Language in Observational Health Research

Journals	Individual Studies
Mainly serves articles that are: Peer-reviewed About health-specific topics Non–meta-analyses, review, or other secondary research designs) Mainly concerning human-level observations (e.g., not animal models or microbiology) Must be a general health, medicine, or epidemiology journal Journals focused on a narrow specialty and/or disease/area of medicine were excluded. The main study design must not be a review or meta-analysis, or other secondary study design.	Observations must be human- or at an aggregate group of humans–level of observation The main research question must be to examine the causal and/or noncausal association between one main exposure concept and one main outcome concept. One main exposure/outcome can include multiple measures of the same or similar broad exposure and/or outcome concept. Articles can include many exposures/outcomes, but focus in particular on one exposure/outcome pair as their main association of interest (e.g., in the title, in the study aims). Articles that are about more than one main concept (e.g., searching for what “risk factors” are associated with the outcome) were excluded. The main research question must be examined quantitatively. The main study design must not be a review or meta-analysis, or other secondary study design.

used for training purposes and for replacement of articles rejected during review. The journal quotas were 65 non-RCT articles and 6 RCT articles per journal, totaling 1,278 articles (1,170 non-RCTs and 108 RCTs). The sample size was informed by informal explorations of sample data sets balanced against reviewer capacity. We did not perform a formal sample-size calculation because: 1) this descriptive study does not involve substantial hypothesis testing; 2) the variance in the language to be analyzed is unknown and is one of the study objectives; and 3) the larger the sample size, the more in-depth we can explore less frequently used language, so we aimed to fully exhaust the available review capacity.

Articles were randomly assigned to 3 of 18 screening reviewers, with 2 independent reviewers and 1 arbitrating reviewer. During screening, the arbitrating reviewer made the inclusion/exclusion decision only in cases where the 2 independent reviewers disagreed.

Screening reviewers were presented with a list consisting exclusively of titles and abstracts. The order of the lists to review was sorted randomly, stratified by journal and study design type (i.e., RCT vs. non-RCT). An administrator periodically consolidated completed screening reviews and assigned articles for arbitration when disagreements occurred. Once quotas for each journal were met, further screening of articles from those journals was disabled through an automated system.

An additional decision was made during screening on June 24, 2021, to drop journals where less than 10% of articles screened met the inclusion criteria and/or that did not have sufficient remaining unscreened articles to meet the minimum quota of articles from a single journal (see Web Appendix 1).

MAIN REVIEW

Reviewer recruitment and selection

Reviewers were recruited through a combination of personal and Twitter solicitations. Reviewers were selected from those with relevant graduate school education, expertise in relevant areas (e.g., epidemiology, causal inference, medicine, econometrics, meta-science, etc.), availability, and to maximize the diversity of fields, life experiences, backgrounds, and kinds of contributions to the group. All reviewers who completed their assigned reviews are coauthors.

A plurality (n = 16/48) of reviewers were doctoral-level students, followed by postdoctoral fellows (n = 12/48) and faculty (n = 10/48). The majority listed epidemiology as one of their primary fields (n = 27/48), followed by statistics/biostatistics (n = 9/48), medicine (n = 6/48), economics (n = 4/48), and psychology (n = 4/48), among other fields. Twelve reviewers had formal clinical training, while 6 were currently practicing clinicians. A plurality of reviewers were based in the United States (n = 18/48), with others based in the United Kingdom (n = 9/48), Germany (n = 4/48), Australia (n = 4/48), and Canada (n = 4/48), among other countries. Additional details are available in Web Figure 1.

Reviewer roles and training

All reviewers received 1 hour of instruction and an additional set of training articles to review before the independent review. Reviewers were encouraged to engage in an active discussion on Slack (Slack Technologies, San Francisco, California) to clarify guidelines, discuss issues,

Table 2. Causal Implication Strength Rating Scale for an Analysis of Causal and Associational Language in Observational Health Research

Rating	Linking Sentence	Action Recommendation
N/A		No action recommendation exists.
None	The linking sentence does not imply in any way that a causal relationship was identified.	The action recommendation would be made appropriately in the absence of any causal relationship.
Weak	The linking sentence might imply that a causal relationship was identified, but it is unclear or not possible to come to that conclusion in the absence of any causal inference.	The action recommendation may be made appropriately had a causal relationship been identified, but it is unclear or not possible to come to that recommendation in the absence of any causal inference.
Moderate	The linking sentence mostly implies that a causal relationship was identified, but it is unclear or not possible to come to that conclusion in the absence of any causal inference.	The action recommendation most likely could only be made appropriately had a causal relationship been identified, but it is unclear or not possible to come to that recommendation in the absence of any causal inference.
Strong	The linking sentence clearly implies that causality had been identified.	The action recommendation could only be made appropriately had a causal relationship been identified.

Abbreviation: N/A, not applicable.

and generate community standards for ambiguous areas. Reviewers were instructed to avoid referring to specifics of a particular study and to instead keep the discussion in general terms to balance eliciting individual subjective opinions with group guidance. By design, reviewers may have changed their understanding of the guidance over time through discussion, and they were therefore allowed to make changes at any point before arbitration.

Each article was reviewed by 3 randomly selected reviewers; 2 independent reviewers, and an arbitrating reviewer. The arbitrating reviewer was given the submitted data from the independent reviewers. Rather than simply resolving conflicts, the arbitrating reviewer's task was to generate what they believed to be the best and most accurate review of each article, given the information from both independent reviewers, their own reading, and the ongoing community discussions. Arbitrating reviewers could decide in favor of one reviewer over another, consolidate and combine reviewers' responses, or overturn both independent reviewers, as they deemed appropriate. The arbitrator review data represents the main output of the review process and was used for all subsequent analyses.

Review framework and tool

The review framework and tool were designed to elicit well-guided, replicable, subjective assessments. The framing and definitions of words used (e.g., definitions and guidance for how/why language might be "causal") are provided in Web Appendix 3.

Reviewers had the option to recuse themselves from reviewing an article for any reason (e.g., conflicts of interest, connections to authors); the article was then reassigned to another reviewer. Reviewers could also request that an administrator reevaluate the inclusion of a study. If the administrator determined that the article did not meet inclusion criteria, it was replaced with one from the buffer of accepted screened reviews.

Reviewers first identified the main outcome and exposure, preferably from the title of the study. Reviewers were asked to identify and copy and paste the main linking sentence, which generally was a sentence in the conclusions section of the abstract or full text containing the primary exposure, outcome, and the linking word/phrase. A linking word/phrase is defined as a word or phrase that describes the nature of the connection between some defined exposure and some defined outcome as identified by the study analysis. This can describe the type of relationship (e.g., "associated with") and/or differences in levels (e.g., "had higher") that may or may not be causal in nature. Then, reviewers were asked to identify modifying phrases, or any words/phrases that modify the nature of the relationship in the linking phrase. This includes signals of direction, strength, doubt, negation, and statistical properties of the relationship (e.g., "may be," "positively," "statistically significant").

Reviewers assessed the degree to which the linking sentence implied a causal relationship between the exposure and outcome using a 4-point scale ("linking sentence causal strength") shown in Table 2.

Next, reviewers were asked to identify any sentences that contained action recommendations (how a consumer of the research might utilize the results and conclusions of the research). This could include recommending that some actor(s) consider changes (or no changes) in some set of procedures and actions. General calls for additional research were not considered action recommendations. After identifying this sentence (if applicable), reviewers were asked to consider the extent that this recommendation would require that a causal relationship had been identified, shown in Table 2.

In this framing, "no causal implication" does not imply "no or null effects." Reviewers were instructed to consider causal implications conceptually separately from the size (or lack thereof) of associations and correlations. Strong causal implications may be made even if the effect size measured was null, so long as the language implied that a causal relationship was being examined.

In addition to the title and abstract review given for all studies, one-third of the articles underwent full text assessment. This extended review 1) repeated the abstract review questions for the discussion section and any popout sections (i.e., sections that do not appear as part of the main text or abstract but summarize and highlight key aspects of the study), and 2) included additional questions to help indicate potential areas of causal intent (29) (see our Open Science Framework project repository (30)). Reviewers also extracted whether there was any theoretical discussion about causal relationships between the exposure and outcome in the introduction, the number of covariates controlled or adjusted for, whether confounding was explicitly mentioned by name (14), whether a formal causal model was used, and whether explicit causal disclaimer statements were made (e.g., "causation cannot be inferred from observational studies, but . . . ").

Root linking words/phrases language strength

After arbitrator reviews were completed, we compiled and curated a list of words from the linking words/phrases in the arbitrator reviews and manually stemmed into their root words. Reviewers then rated the causal implications of all root words found more than once in our sample. This was to mimic language decision processes that base their causal language assessment on selecting words that are or are not causal, and to establish our own systematic assessments of word ratings. Reviewers were presented with up to 4 randomly selected linking words/phrases that contained the root word and had been submitted by arbitrating reviewers (e.g., the root word "associate" had 4 phrases, including phrases like "associated with" or "association").

ANALYSIS

The statistical analysis was largely descriptive. Except for comparisons between RCTs and non-RCTs, all statistical analysis was performed on the arbitrated data set of the non-RCTs only.

Comparisons between 2 ordinal categorical variables were estimated by Spearman's correlation coefficients. Associations between strength ratings and key binary variables (e.g., study type, journals, topic areas, etc.) were estimated with ordinal logistic regression.

All measures of statistical uncertainty were clustered by journal and calculated using a block bootstrapping procedure, unless otherwise specified, where 95% confidence intervals (CIs) were obtained from percentiles of the bootstrapped estimate distribution. Where the journals themselves were covariates, the clustered sandwich estimator was used. For root word rating proportions, there were no journal clusters, and therefore the Wilson estimator was used. No weights were applied, with journals and articles contributing equally to the main results.

Heterogeneity between reviewers was evaluated using Krippendorff's α. For the purpose of this review, disagreement between reviewers is a key result (i.e., heterogeneity between subjective opinions) rather than error.

All data management and analyses were conducted using R, version 4.0.5 (R Foundation for Statistical Computing). Spearman correlation coefficients were determined using the pspearman package (31). Ordinal logistic regression was performed using the MASS package (32).

DATA AND CODE AVAILABILITY

All data and code are publicly available through the Open Science Framework repository (24), except for files containing personal identifying information and/or personal application programming interface keys.

RESULTS

Search and screening

Figure 1 summarizes the selection of journals and articles into the sample. Eighteen journals were identified meeting our search criteria (listed in the caption).

After searching PubMed for articles, we screened articles until 65 non-RCTs and 6 RCTs were accepted (except in the *European Journal of Epidemiology*, where only 3 RCTs were identified and included). This yielded 1,170 non-RCTs and 105 RCTs, totaling 1,275 studies reviewed. There were 10 recusals recorded during the main review. The 3 most common MeSH disease areas were "pathological conditions, signs and symptoms" (n = 377), "cardiovascular diseases" (n = 324), and "nutritional and metabolic diseases" (n = 198). See Web Figure 2 for full terms.

Linking words and phrases

After the arbitrator reviews were completed, root words were obtained through stemming the linking phrases to identify and rate the root linking words.

By far the most common root linking word identified in abstracts was "associate" (n = 535/1,170; 45.7%, 95% CI: 40.0, 51.9), followed by "increase" (n = 71/1,170; 6.1%, 95% CI: 4.7, 7.8) (Table 3). The same root word was identified in both the abstract and discussion for 48.2% cases (95% CI: 43.7, 53.6). We found 9 (0.8%, 95% CI: 0.4, 1.3) studies where the main root linking word was "cause." There were 16 (1.4%, 95% CI: 0.6, 2.3) articles that used the word "cause," when additionally including any instance of the word "cause" in either the linking or modifying phrases.

Causal implication strengths

Summary data. Reviewers rated the abstract linking sentence as having no causal implication in 13.8% (95% CI: 11.9, 15.9), weak in 34.2% (95% CI: 31.4, 36.7), moderate

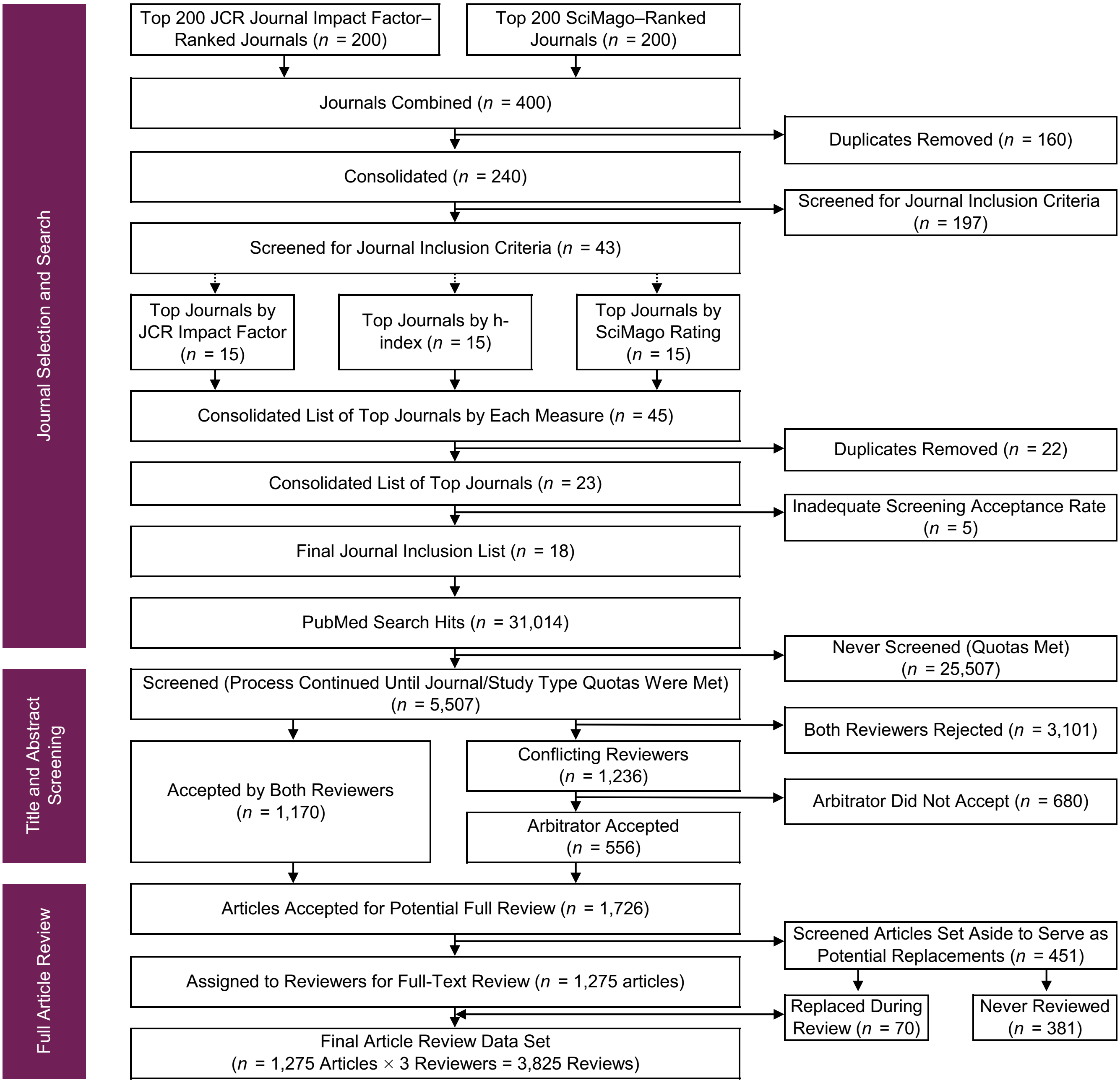

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram detailing the search and screening process to arrive at our final sample for a study of causal and associational linking language in observational research and health evaluation literature. Journals included were *American Journal of Epidemiology*, *American Journal of Medicine*, *American Journal of Preventive Medicine*, *American Journal of Public Health*, *Annals of Internal Medicine*, *BioMed Central Medicine*, *British Medical Journal*, *Canadian Medical Association Journal*, *European Journal of Epidemiology*, *International Journal of Epidemiology*, *Journal of Internal Medicine*, *Journal of the American Medical Association*, *Journal of the American Medical Association Internal Medicine*, *The Lancet*, *Mayo Clinic Proceedings*, *New England Journal of Medicine*, *PLOS Medicine*, and *Social Science and Medicine*. JCR, Journal Citation Reports.

in 33.2% (95% CI: 29.8, 36.7), and strong in 18.7% (95% CI: 15.1, 22.6) of instances (Figure 2). Eight journals had popout sections in their articles. The language used was very similar in the abstract, full-text discussion, and popout sections.

Action recommendations were identified in 34.2% (95% CI: 29.0, 39.6) of abstracts. Of these, 5.3% (95% CI: 3.5, 7.2) were rated as having a causal implication of none, 19.0% (95% CI: 15.2, 23.0) as weak, 42.8% (95% CI: 39.0, 46.4) as moderate, and 33.0% (95% CI: 29.0, 37.1) as strong.

By comparison, action recommendations were identified in 60.3% (95% CI: 52.7, 67.5) of discussion sections, about twice that in abstracts. We found negligible, if any,

Table 3. List and Frequency of Identified Root Words Used to Link the Exposure and Outcome in an Analysis of Causal and Associational Language in Observational Health Research

Root Word	No.[a]
Associate	535
Increase	71
High	36
Predict	34
Reduce	33
Likely	29
Lower	26
Relate	25
Improve	21
Effect	19
Risk	17
Different	16
Decrease	14
Influence	13
Risk Factor	13
Contribute	12
Effective	12
Affect	10
Link	10
Cause	9
Impact	9
Result	9
Benefit	7
Correlate	7
Explain	7
Attribute	6
Change	6
Decline	6
Elevate	6
Lead	6
Better	4
Compare	4
Greater	4
Protect	4
Show	4
Similar	4
Appear	3
Demonstrate	3
Determinant	3
Factor	3
Less	3
Occur	3
Prevent	3
Role	3
Achieve	2
Consistent	2
Differ	2
Due	2
Excess	2
Precede	2
Reveal	2
Twice	2
Vary	2
Worse	2
Other terms	78

Table continues

Table 3. Continued

[a] The number of times each of these root words appears in the linking phrases in the abstracts of our samples. In cases where 2 of these words are in the same phrase (e.g., "similar risk") the more common of the 2 is selected (in this case, "risk"). In cases where selected linking phrases had 2 or more words that were included in the root word list, the more common word was selected as the root word mainly associated with that study and section.

differences between the overall strength of the action implications found in discussions sections vs. abstracts (log odds for higher rank: −0.00026, 95% CI: –0.00024, 0.00013). There was also no apparent pattern in implication strength over time (Web Figure 3).

Comparison of linking sentence strength vs. action implication strength. Of the 34% (n = 400) of studies with action recommendations, 15.3% (95% CI: 11.7, 19.2) had action recommendations that implied causality less strongly than the linking sentence, 40.3% (95% CI: 35.1, 45.8) were commensurate, and 44.5% (95% CI: 39.9, 48.4) were stronger (Figure 3A). There was a weak correlation between the strength of causal implication in the linking sentence and the action recommendation (Spearman's correlation coefficient = 0.349, 95% CI: 0.256, 0.435). While stronger causal action recommendations are less likely to occur when linking sentences are weaker (Figure 3B), studies with weaker linking sentences also often make strong causal action implications. Among the 76.0% of studies with no action recommendation in the abstract, 14.5% (95% CI: 11.6, 17.6) of linking sentences were rated as having a causal implication of "none," 34.0% (95% CI: 30.3, 37.5%) as weak, 33.1% (95% CI: 29.2, 37.3) as moderate, and 18.3% (95% CI: 14.5, 22.5) as strong. We found negligible, if any, differences in the strength of the linking sentences between abstracts that did and did not contain action recommendations (log odds for higher rank: 0.087, 95% CI: –0.162, 0.320).

Words and phrases

Ratings among reviewers (n = 47) for causal implication of root words were highly heterogeneous, with the only

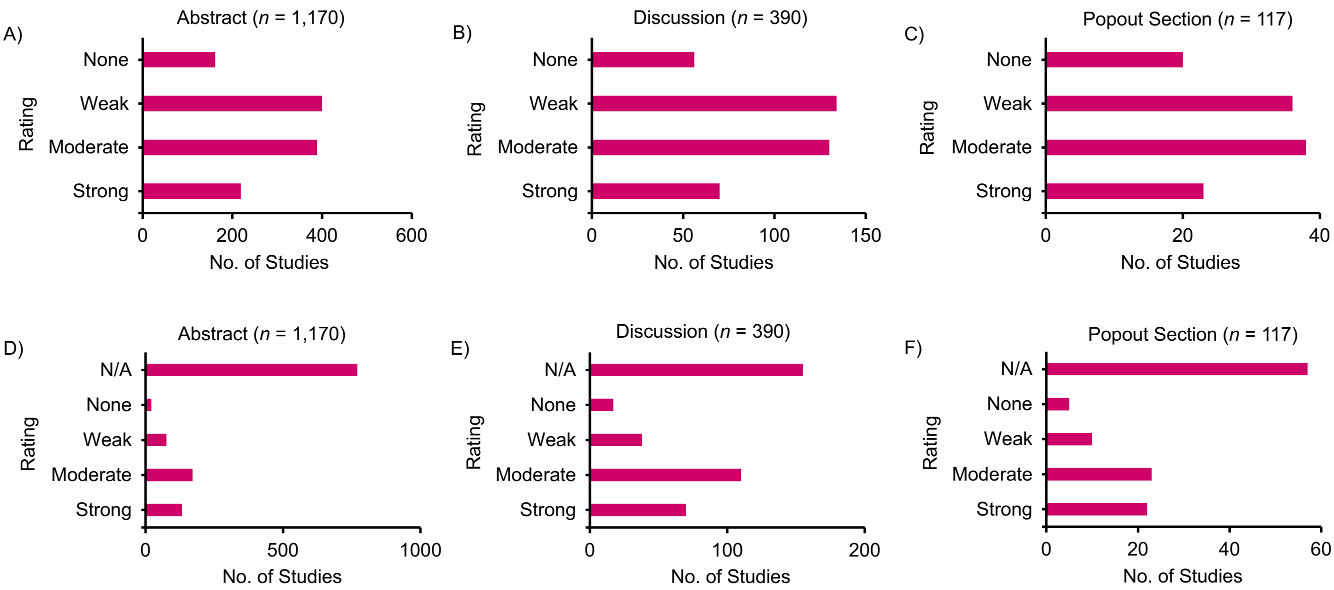

Figure 2. Summary scores for the degree of causal implication in linking sentences and action recommendations in a study of causal and associational linking language in observational research and health evaluation literature, showing the frequency of key strength of causal implication metrics for the 1,170 non–randomized-control trial studies in our sample, as indicated by the arbitrating reviewer. A–C) The strength of causal implication ratings for the language ratings in the abstract, discussion, and popout sections; D–F) the strength of causal implication ratings for the action recommendations in the abstract, discussion, and popout sections.

word to reach near consensus on causal implications being "cause" (Figure 4). Reviewers rated words such as "correlate" and "associate" generally weaker, in terms of their causal implications, than words such as "impact," "effect," "affect," and "prevent." Notably, many root words could be used in a variety of ways with potentially different meanings. For example, the root word "lower" could be used purely descriptively, as in "people with X had lower Y," or indicating X as a driving force, as in "X lowered Y."

Although the word "associate" ranked among the lowest overall in having causal implications, more than half of the reviewers judged that the word "associate" carried at least some causal implication (n = 26/47; 55.3%, 95% CI: 41.2, 68.6). For comparison, 78.6% (95% CI: 75.7, 81.2) of linking sentences containing the root word "associate" were rated as having at least some causal strength.

Modifying phrases

Common modifying phrases are identified and discussed in Web Figure 4.

Differences in strength across key strata

Non-RCTs vs. RCTs. The most common linking word identified in RCT abstracts was "associate" (n = 16/105), followed by "reduce" (n = 14/105) and "increase" (n = 11/105). Additional results for RCTs are available in Web Appendix 4.

Journals and journal policies. Differences between journals are discussed in Web Appendix 5 and Web Figures 5–7.

Indications of potential causal interest

Most studies in our sample provided at least some indication of potential causal interest (Figure 5). While only 3.8% (95% CI: 2.0, 6.0) of studies presented formal causal models, most offered some discussion of the theoretical nature of the causal relationship between exposure and outcome (80.0%, 95% CI: 75.2, 85.4). Among those that discussed theory, 58.7% (95% CI: 51.4, 64.8) moderately or strongly indicated a theoretical causal relationship between exposure and outcome; 24.6% (95% CI: 20.9, 28.0) of studies had a disclaimer statement regarding causality; and 68.7% (95% CI: 63.3, 73.7) explicitly mentioned variations of the word "confound." Finally, most studies controlled or adjusted for several variables, with 35.1% (95% CI: 30.5, 39.9) having 10 or more control variables.

Interrater comparisons

The Krippendorff's α comparing the reviewers' ratings for linking language strength in the abstract was 0.29. Independent reviewers gave the same score in 35.1% of instances; 41.2% differed by 1 category, 19.9% by 2 categories, and 3.8% by 3 categories. Agreement increased to 0.41 when including the independent and arbitrating reviewers.

For the action recommendations (where most articles were rated as "N/A" for missing) the Krippendorff's α was 0.70. The 2 independent reviewers agreed exactly in 67.6% of cases, differed by 1 category in 14.4% of cases, by 2 in 8.6%, by 3 in 5.3%, and by 4 in 4.1%. Similarly, agreement increased to 0.76 when including the arbitrating reviewers.

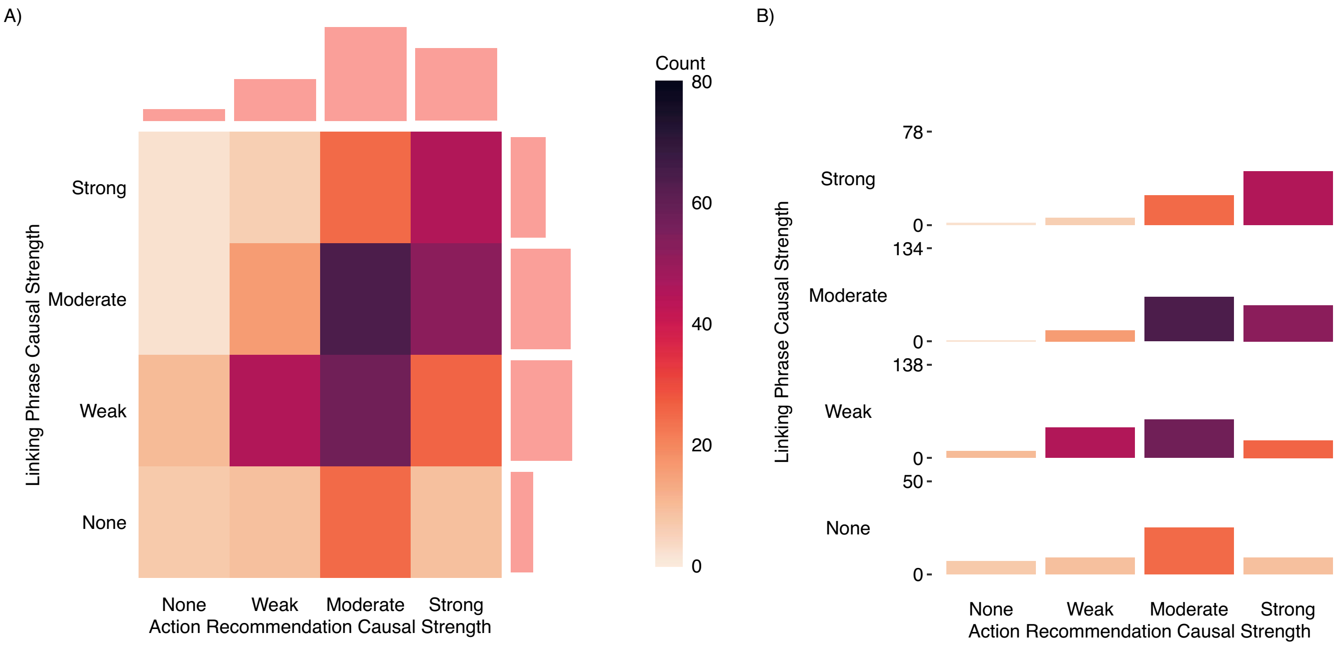

Figure 3. Comparison of the strength of causal implications in the abstracts for the linking phrase and action recommendations in a study of causal and associational linking language in observational research and health evaluation literature, showing the distribution of linking sentence and action recommendation language among the 400/1,170 non–randomized-control trial studies in which there was an action recommendation present in the abstract. A) An unconditional heatmap, with colors representing the number of articles in the strata, and histograms on the top and right showing the overall distribution of ratings for each axis; B) the distributions within each level of linking sentence causal strength.

DISCUSSION

Our systematic evaluation of the use of causal language and implications in the high-profile medical and epidemiologic literature found that: 1) by far the most common word used linking exposures and outcomes was "associate"; 2) although few studies explicitly declared an interest in estimating causal effects, the majority used language that moderately or strongly implied causality; 3) while approximately a third of articles issued action recommendations, the vast majority of these were found to imply that causality had been inferred; 4) causal language in action recommendations ratings tended to be stronger than the language in linking sentences; and 5) although many studies used disclaimers warning readers against making causal inferences, an implicit interest in causality was apparent from common discussions of causal mechanisms and widespread adjustment for confounding. Overall, we found a substantial disconnect between the causal implications used in technical linking language and research implications.

Our results suggest that "Schrödinger's causal inference" (33)—where studies avoid stating (or even explicitly deny) an interest in estimating causal effects yet are otherwise embedded with causal intent, inference, implications, and recommendations—is common in the observational health literature. While the relative paucity of explicit action recommendations might be seen as appropriate caution, it also invites causal inference since there are often no useful and/or obvious alternative (noncausal) interpretations. While we hypothesized a priori that there might be substantial differences between the language used RCTs and that in non-RCTs (e.g., that RCTs would use the linking word "cause" frequently), the words they used were very similar. "Associate," for example, was the most common linking word for both RCTs and non-RCTs. The degree of causal interpretation for common linking words may have been affected by the lack of explicitly causal language, such that the meaning of traditionally noncausal words may have broadened to include potentially stronger causal interpretations (34). It is likely that the rhetorical standard of "just say association" has meant that many researchers no longer fully believe that the word "association" just means association.

At this time, we do not know the degree to which journal editors, reviewers, authors, or academic community standards contribute to the implicit and explicit rules of causal language. While there are relatively few explicit and public rules governing language at journals, journals may employ formal internal guidelines and unspoken informal norms.

Our measures of causal implication are based on subjective assessments, which are critical to evaluating and interpreting human language. Reviewers substantially differed regarding the causal implications of many linking words, even in the presence of extensive guidance, processes, and training for how to assess causal implication in language. Different interpretations may arise from different backgrounds, experiences, and other factors affecting personal interpretations. Our reviewers, for example, are likely to have been selected into our study due to an interest in

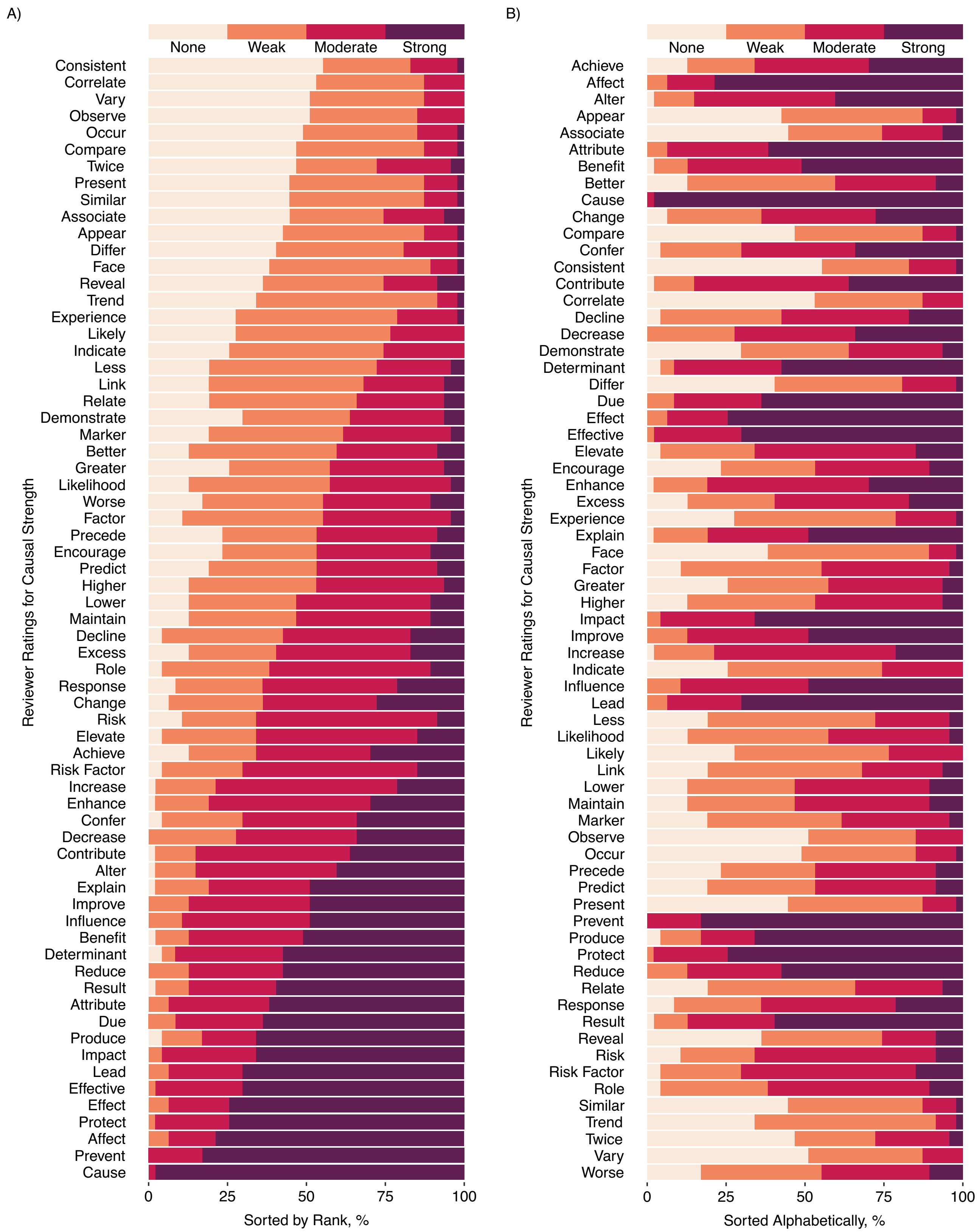

Figure 4. Strength of causal implication ratings for the most common root linking words in a study of causal and associational linking language in observational research and health evaluation literature, showing the distribution of ratings given by reviewers during the root word rating exercise. On the left side, they are sorted by median rating + the percentage of reviewers who would have to change their ratings in order for the rating to change (A). On the right, the chart is sorted alphabetically (B).

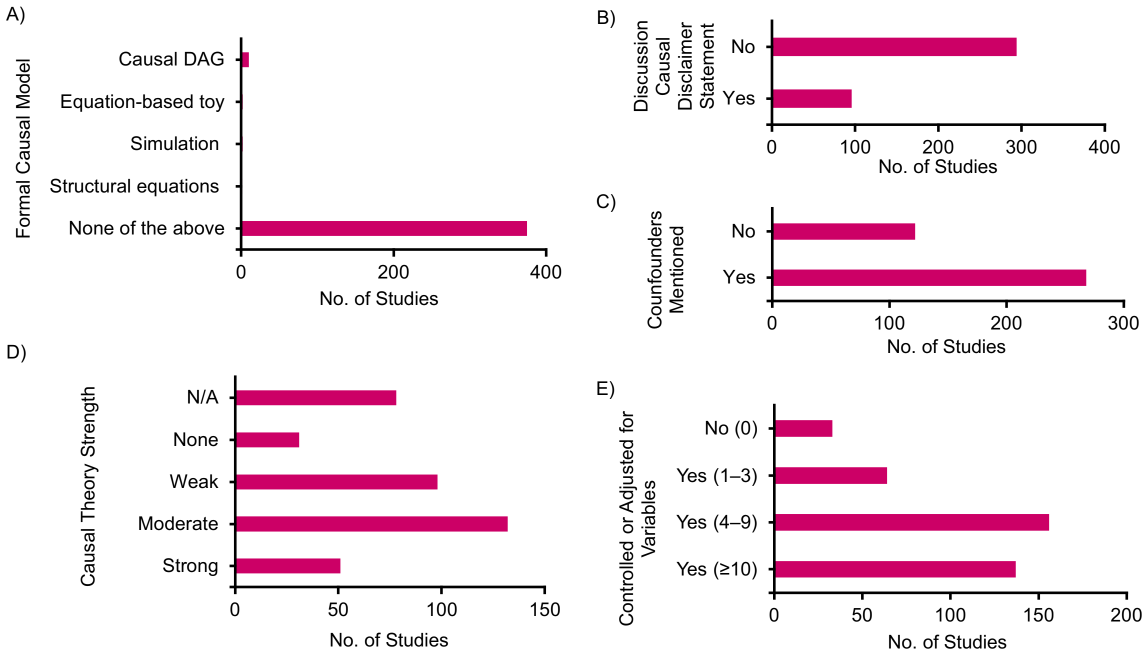

Figure 5. Frequency of indicators of potential causal interest in a study of causal and associational linking language in observational research and health evaluation literature. A) Formal causal model; B) discussion of causal disclaimer statement; C) confounders mentioned; D) causal theory strength; E) control/adjustment variables. These results are from the 390 articles reviewed in full.

and knowledge of causal inference and language. We expect that alternative potential target populations of research consumers may also interpret these words differently, whether by virtue of differing frameworks for assessing language, personal interpretations, or community standards. Rather than attempting to be fully representative of any one possible population of people who interact with this research, we chose to have a coauthor reviewer pool with representation from a wide variety of possible target populations covering a wide variety of research traditions that might interact with this type of research.

Beyond the reviewers themselves, it also matters how words are used and in what context. For example, ratings between "associate" alone in the root word rating exercise had less causal implication compared with in-context ratings of sentences with "associate" in the linking phrase. Aspects of the rating and interpretation process are also likely to be particularly challenging; for example, in reviewer discussions many reported difficulty with evaluating the degree of causal implication for sentences with null findings. Research consumers and decision-makers may have entirely different interpretations and frameworks, consciously or otherwise.

This study was designed with replicability in mind. The review process was designed to balance independent subjective assessments from skilled researchers and practitioners with explicit guidance and discussion among reviewers. Our assessment process is applicable to any number of areas of systematic evidence review and evaluation, which is often limited to shallow "objective" measures. Beyond preregistration, nearly all parts of this project were fully open and disseminated to the public to view and comment on, including documents, data, and code, resulting in a very large number of contributors, comments, and suggestions throughout the process.

Results may not be directly generalizable to other settings, alternative samples, and reviewers. Because our inclusion criteria excluded studies that were examining several potential factors or exposures and their relationships with outcome(s), our sample likely excluded many multiple-exposure articles with terms such as "risk factors," "correlates," or "predictors." Our journal selection, which included only the most prominent general medical, public health, and epidemiology journals, may not be representative of different fields, subfields, journals, and policies. We did not examine the strength of evidence, nor did we examine any information that would indicate the appropriateness of claims.

The practice of avoiding causal language linking exposures and outcomes appears to add little if any clarity. Common standards for which words and language are "causal" or when "causal" words are appropriate do not appear to match interpretation. While being careful about what we claim is critical for medical science, being "careful" is often implemented by stripping out causal language in conclusions, and therefore any hint of what question is being answered. Knowing that people with X have 42 times the relative risk of Y is not informative if we do not know what question that association attempts to answer (35). Misalignment between the research question being asked and action implications is on its own a source of confusion, which could be avoided if the causal nature of the research question were made explicit. Further, these practices may weaken methodological accountability, as studies that only indirectly imply causality can be shielded from critiques regarding lack of causal inference rigor (4).

Rather than policing which words we use to describe relationships between exposures and outcomes, we recommend focusing on how researchers, research consumers, and reviewers can better identify and assess causal inference study designs and assumptions. Quantitative empirical research should clearly state its target estimand to clarify the research question (36), including explicitly stating when such estimands are causal. Authors, reviewers, and editors should focus on being clear about what questions are being asked (37, 38), what decisions are being informed, and the degree to which we are and are not able to achieve those goals.

ACKNOWLEDGMENTS

Author affiliations:
Meta Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, United States (Noah A. Haber, Sarah E. Wieten, Cathrine Axfors); Department of Psychology, University of Leipzig, Leipzig, Germany (Julia M. Rohrer); Department of Epidemiology, Fielding School of Public Health, University of California Los Angeles, Los Angeles, California, United States (Onyebuchi A. Arah); Leeds Institute for Data Analytics, University of Leeds, Leeds, United Kingdom (Peter W. G. Tennant); Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States (Elizabeth A. Stuart, Ian Schmid); Department of Epidemiology, School of Public Health, Boston University, Boston, Massachusetts, United States (Eleanor J. Murray, Matthew P. Fox); Nuffield Department of Population Health, Big Data Institute, University of Oxford, Oxford, United Kingdom (Sophie Pilleron); Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore (Sze Tung Lam); Capital One, Chicago, Illinois, United States (Emily Riederer); Psychology, Health and Professional Development, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, United Kingdom (Sarah Jane Howcutt); Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada (Alison E. Simmons); Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, United Kingdom (Clémence Leyrat); School of Economics and Finance, University of St. Andrews, Fife, United Kingdom (Philipp Schoenegger); School of Philosophical, Anthropological, and Film Studies, University of St. Andrews, Fife, United Kingdom (Philipp Schoenegger); Epidemiology Department, Oregon Health & Science University–Portland State University School of Public Health, Portland, Oregon, United States (Anna Booman); Berkeley Public Health, University of California Berkeley, Berkeley, California, United States (Mi-Suk Kang Dufour); Center for Healthcare Delivery Science, Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States (Ashley L. O'Donoghue); Interacting Minds Center/Linguistics, Cognitive Science, and Semiotics, Aarhus University, Aarhus, Denmark (Rebekah Baglini); Department of Epidemiological Methods and Etiological Research, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany (Stefanie Do); School of Medicine, Griffith University, Nathan, Queensland, Australia (Mari De La Rosa Takashima); School of Human Sciences, University of Greenwich, London, United Kingdom (Thomas Rhys Evans); Occupational and Environmental Epidemiology and NetTeaching Unit, Institute and Clinic for Occupational, Social and Environmental Medicine, University Hospital, Ludwig Maximilian University of Munich, Munich, Germany (Daloha Rodriguez-Molina); Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany (Taym M. Alsalti); Center for Translational Behavioral Science, Florida State University, Tallahassee, Florida, United States (Daniel J. Dunleavy); School of Health and Society, University of Wollongong, Wollongong, New South Wales, Australia (Gideon Meyerowitz-Katz); Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy (Alberto Antonietti); Department of Anesthesiology, Universidad del Cauca, Cauca, Colombia (Jose A. Calvache); Department of Mathematics, University of Exeter, Exeter, Devon, United Kingdom (Mark J. Kelson); Harvard T. H. Chan School of Public Health, Harvard University, Boston, Massachusetts, United States (Meg G. Salvia); Department of Mathematical Sciences, University of Bath, Bath, United Kingdom (Camila Olarte Parra); Faculty of Medicine and Health, The University of Sydney School of Public Health, Sydney, New South Wales, Australia (Saman Khalatbari-Soltani); British Columbia Centre for Excellence in HIV/AIDS, Epidemiology and Population Health Program, Vancouver, British Columbia, Canada (Taylor McLinden); UMR Institut national de la santé et de la recherche médicale 1246 SPHERE, University of Nantes and University of Tours, Nantes, France (Arthur Chatton); Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, United States (Jessie Seiler); Faculty of Medicine, Carol Davila University of Medicine and Pharmacy, Bucharest, Romania (Andreea Steriu); College of Dentistry, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia (Talal S. Alshihayb); Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Québec, Canada (Sarah E. Twardowski); School of Public Health, University of Alberta, Edmonton, Alberta, Canada (Julia Dabravolskaj); Faculty of Medicine and Health, The University of Sydney School of Public Health, Sydney, New South Wales, Australia (Eric Au); Department of Kinesiology and Community Health, University of Illinois Urbana-Champaign, Champaign, Illinois, United States (Rachel A. Hoopsick); Community Medicine Program, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, United States (Shashank Suresh); Department of Neuroscience, Karolinska Institute, Stockholm, Sweden (Nicholas Judd); Finnish Institute for Health and Welfare, Helsinki,

Finland (Sebastián Peña); Clinical Research Department, London School of Hygiene & Tropical Medicine, London, United Kingdom (Palwasha Khan); Department of Population Health, Division of Epidemiology, New York University Grossman School of Medicine, New York, New York, United States (Ariadne E. Rivera Aguirre); Center for Health Sciences, Exponent, Inc., Oakland, California, United States (Nnaemeka U. Odo); and Department of Epidemiology, School of Public Health, Boston, Massachusetts, United States (Matthew P. Fox).

No funding was granted specifically for the support of this study. The Meta-Research Innovation Center at Stanford University is supported by Arnold Ventures LLC (Houston, Texas), formerly the Laura and John Arnold Foundation. S.P. was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant (agreement no. 842817). S.K.-S. is supported by the Australian Research Council Centre of Excellence in Population Aging Research (project number CE170100005). I.S. is supported by the National Institute of Mental Health (grant T32MH122357). E.A.S.'s time was supported by the National Institute of Mental Health (grant R01MH115487) and the Bloomberg American Health Initiative. A.L.O. is funded by a philanthropic gift from Google.org outside of the submitted work. O.A.A. is supported by the National Institute of Biomedical Imaging and Bioengineering (grant R01EB027650), National Center for Advancing Translational Sciences UCLA Clinical Translational Science Institute (grant UL1TR001881), and a philanthropic gift from the Karen Toffler Charity Trust.

Data, data analysis code, and materials are available on the Open Science Framework project https://osf.io/jtdaz/.

This work was supported by many people who made contributions to this work. Turki Althunian contributed to the screening process. Jess Rohmann contributed to the piloting process. This work was additionally supported by comments and contributions from Alyssa Bilinksi, Pascal Goldsetzer, Caroline Blaine, Otto Kalliokoski, Eero Raittio, Tanya Colyer, Tim Watkins, Alexander Breskin, Arindam Basu, Jessica L. Rohmann, Luke A McGuinness, Todd Johnson, Mario Malički, Sebastian Skejø, Scott Graham, Michael Chaiton-Murray, John Edlund, Katelyn Smalley, Danielle Newby, Anita Williams, Cord Phelps, Colleen Derkatch, Alexander Wolthon, Pallavi Rohella, Damien Croteau-Chonka, Steven Goodman, and John Ioannidis.

Presented at the Annual Meeting of the Society for Epidemiologic Research, June 14–17, 2022, Chicago, Illinois.

A preprint of this article has been published online. (Haber, NA, Wieten SE, Rohrer JM, et al. Causal and Associational Language in Observational Health Research: A Systematic Evaluation. *medRxiv.* 2021. https://doi.org/10.1101/2021.08.25.21262631).

All errors are the sole responsibility of the authors, and no funders had any role in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication. The researchers were independent from funders, and all authors, external and internal, had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Conflict of interest: none declared.

REFERENCES

1. JAMA. Instructions for authors. https://jamanetwork.com/journals/jama/pages/instructions-for-authors. Accessed May 11, 2021.
2. Chipperfield L, Citrome L, Clark J, et al. Authors' submission toolkit: a practical guide to getting your research published. *Curr Med Res Opin.* 2010;26(8):1967–1982.
3. AMA Manual of Style Committee. *AMA Manual of Style: A Guide for Authors and Editors*. 11th ed. New York, NY: Oxford University Press; 2020. https://www.amamanualofstyle.com/view/10.1093/jama/9780190246556.001.0001/med-9780190246556. Accessed May 11, 2021.
4. Hernán MA. The c-word: scientific euphemisms do not improve causal inference from observational data. *Am J Public Health.* 2018;108(5):616–619.
5. Grosz MP, Rohrer JM, Thoemmes F. The taboo against explicit causal inference in nonexperimental psychology. *Perspect Psychol Sci.* 2020;15(5):1243–1255.
6. Olarte Parra C, Bertizzolo L, Schroter S, et al. Consistency of causal claims in observational studies: a review of papers published in a general medical journal. *BMJ Open.* 2021;11(5):e043339.
7. Prasad V, Jorgenson J, Ioannidis JP, et al. Observational studies often make clinical practice recommendations: an empirical evaluation of authors' attitudes. *J Clin Epidemiol.* 2013;66(4):361–366.e4.
8. Adams RC, Challenger A, Bratton L, et al. Claims of causality in health news: a randomised trial. *BMC Med.* 2019;17(1):91.
9. Adams RC, Sumner P, Vivian-Griffiths S, et al. How readers understand causal and correlational expressions used in news headlines. *J Exp Psychol Appl.* 2017;23(1):1–14.
10. Buhse S, Rahn AC, Bock M, et al. Causal interpretation of correlational studies—analysis of medical news on the website of the official journal for German physicians. *PLoS One*. 2018;13(5):e0196833.
11. Cofield SS, Corona RV, Allison DB. Use of causal language in observational studies of obesity and nutrition. *Obes Facts.* 2010;3(6):353–356.
12. Watkins TR. Understanding uncertainty and bias to improve causal inference in health intervention research [thesis]. Sydney, New South Wales, Australia: University of Sydney; 2019. https://ses.library.usyd.edu.au/bitstream/handle/2123/20772/watkins_tr_thesis.pdf. Accessed October 18, 2021.
13. Hall MG, Grummon AH, Maynard OM, et al. Causal language in health warning labels and US adults' perception: a randomized experiment. *Am J Public Health.* 2019;109(10):1429–1433.
14. Kezios KL. Is the way forward to step back? Documenting the frequency with which study goals are misaligned with study methods and interpretations in the epidemiologic literature. *Epidemiol Rev.* 2022;43(1):4–18.
15. Alvarez-Vargas D, Braithwaite DW, Lortie-Forgues H, et al. Hedges, mottes, and baileys: causally ambiguous statistical

language can increase perceived study quality and policy relevance [preprint]. *PsyArXiv*. 2020. https://doi.org/10.31234/osf.io/nkf96. Accessed July 15, 2020.
16. Haber N, Smith ER, Moscoe E, et al. Causal language and strength of inference in academic and media articles shared in social media (CLAIMS): a systematic review. *PLoS One.* 2018;13(5):e0196346.
17. Sumner P, Vivian-Griffiths S, Boivin J, et al. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*. 2014;349:g7015.
18. Sumner P, Vivian-Griffiths S, Boivin J, et al. Exaggerations and caveats in press releases and health-related science news. *PLoS One.* 2016;11(12):e0168217.
19. Schwitzer G. Addressing tensions when popular media and evidence-based care collide. *BMC Med Inform Decis Mak.* 2013;13(S3):S3.
20. Haneef R, Lazarus C, Ravaud P, et al. Interpretation of results of studies evaluating an intervention highlighted in Google health news: a cross-sectional study of news. *PLoS One.* 2015;10(10):e0140889.
21. Thapa DK, Visentin DC, Hunt GE, et al. Being honest with causal language in writing for publication. *J Adv Nurs.* 2020;76(6):1285–1288.
22. Ramspek CL, Steyerberg EW, Riley RD, et al. Prediction or causality? A scoping review of their conflation within current observational research. *Eur J Epidemiol*. 2021;36(9):889–898.
23. Varady NH, Feroe AG, Fontana MA, et al. Causal language in observational orthopaedic research. *J Bone Joint Surg Am*. 2021;103(19):e76.
24. Haber N, Wieten S. Data repository and pre-registration. https://osf.io/jtdaz/. Accessed March 25, 2022.
25. Clarivate Analytics. Journal Citation Reports. https://jcr.clarivate.com/JCRJournalHomeAction.action. Accessed May 26, 2020.
26. SciMago. Journal and country rank. https://www.scimagojr.com/journalrank.php?area=2700&order=h&ord=desc. Accessed May 26, 2020.
27. Fantini D. easyPubMed: search and retrieve scientific publication records from PubMed. https://CRAN.R-project.org/package=easyPubMed. Accessed June 1, 2021.
28. National Library of Medicine. Download MeSH data. https://www.nlm.nih.gov/databases/download/mesh.html. Accessed August 16, 2021.
29. Han MA, Guyatt G. Systematic survey of the causal language use in systematic reviews of observational studies: a study protocol. *BMJ Open*. 2020;10(7):e038571.
30. Haber N, Wieten S. Review tool. https://osf.io/jk3bp. Accessed September 8, 2022.
31. Savicky P. pspearman: Spearman's rank correlation test. https://CRAN.R-project.org/package=pspearman. Accessed March 25, 2022.
32. Venables WN, Ripley BD. *Modern Applied Statistics With S*. New York, NY: Springer; 2002.
33. Tennant PWG, Murray EJ. The quest for timely insights into COVID-19 should not come at the cost of scientific rigor. *Epidemiology.* 2021;32(1):e2–e2.
34. de Carvalho A, Reboul AC, Van der Henst J-B, et al. Scalar implicatures: the psychological reality of scales. *Front Psychol*. 2016;7:1500.
35. Adams D. *The Hitchhiker's Guide to the Galaxy*. New York, NY: Harmony Books; 1980.
36. Tennant PW, Harrison WJ, Murray EJ, et al. Use of directed acyclic graphs (DAGs) in applied health research: review and recommendations [preprint]. *medRxiv*. 2019. http://medrxiv.org/lookup/doi/10.1101/2019.12.20.19015511. Accessed December 31, 2019.
37. Fox MP, Edwards JK, Platt R, et al. The critical importance of asking good questions: the role of epidemiology doctoral training programs. *Am J Epidemiol.* 2020;189(4):261–264.
38. Lundberg I, Johnson R, Stewart BM. What is your Estimand? Defining the target quantity connects statistical evidence to theory. *Am Sociol Rev.* 2021;86(3):532–565.