Causal and Associational Language in Observational Health Research: A systematic evaluation

Noah A. Haber[1], Sarah E. Wieten[1], Julia M. Rohrer[2], Onyebuchi A. Arah[3], Peter W.G. Tennant[4], Elizabeth A. Stuart[5], Eleanor J. Murray[6], Sophie Pilleron[7], Sze Tung Lam[8], Emily Riederer[9], Sarah Jane Howcutt[10], Alison E. Simmons[11], Clémence Leyrat[12], Philipp Schoenegger[13], Anna Booman[14], Mi-Suk Kang Dufour[15], Ashley L. O'Donoghue[16], Rebekah Baglini[17], Stefanie Do[18], Mari De La Rosa Takashima[19], Thomas Rhys Evans[20], Daloha Rodriguez-Molina[21], Taym M. Alsalti[22], Daniel J. Dunleavy[23], Gideon Meyerowitz-Katz[24], Alberto Antonietti[25], Jose A. Calvache[26], Mark J. Kelson[27], Meg G. Salvia[28], Camila Olarte Parra[29], Saman Khalatbari-Soltani[30], Taylor McLinden[31], Arthur Chatton[32], Jessie Seiler[33], Andreea Steriu[34], Talal S. Alshihayb[35], Sarah E. Twardowski[36], Julia Dabravolskaj[37], Eric Au[38], Rachel A. Hoopsick[39], Shashank Suresh[40], Nicholas Judd[41], Sebastián Peña[42], Cathrine Axfors[1], Palwasha Khan[43], Ariadne E. Rivera Aguirre[44], Nnaemeka U. Odo[45], Ian Schmid[5], Matthew P. Fox[46]

Corresponding author: Noah A. Haber, (noahhaber@stanford.edu), ORCiD 0000-0002-5672-1769

[1]Meta Research Innovation Center at Stanford (METRICS), Stanford University, 1265 Welch Rd, Stanford, CA, 94305, United States, [2]Psychology, University of Leipzig, Städt. Kaufhaus Neumarkt 9, Leipzig, 04109, Germany, [3]Epidemiology, University of California Los Angeles, 650 Charles E. Young Drive South, Los Angeles, California, 90095, United States, [4]Leeds Institute for Data Analytics, University of Leeds, Level 11 Worsley Building, Leeds, LS2 9NL, United Kingdom, [5]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N Broadway, Baltimore, MD, 21205, United States, [6]Epidemiology, Boston University, 715 Albany Street, Boston, MA 02118, Boston, Massachusetts, 2118, United States, [7]Nuffield Department of Population Health, Oxford University, Big Data Institute, Richard Doll Building, Old Road Campus, Headington, Oxford, OX3 7LF, United Kingdom, [8]Yong Loo Lin School of Medicine, National University of Singapore, 1E Kent Ridge Road, Singapore, Singapore, 119228, Singapore, [9](No affiliation data provided), [10]Psychology, Health and Professional Development, Oxford Brookes University, Faculty of Health and Life Sciences, Oxford, OX3 0FL, United Kingdom, [11]Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, 155 College St, Toronto, Ontario, M5T 3M7, Canada, [12]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom, [13]School of Economics and Finance, School of Philosophical, Anthropological, and Film Studies, University of St Andrews, The Scores, St Andrews, Fife, KY16 9AJ/KY16 9AR, United Kingdom, [14]Epidemiology, Oregon Health & Science University-Portland State University School of Public Health, 1810 SW 5th Ave, Portland, OR, 97201, USA, [15]Berkeley School of Public Health, University of California Berkeley, 2121 Berkeley Way, Berkeley, CA, 94720-7360, United States, [16]Center for Healthcare Delivery Science, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Boston, MA, 02215, United States, [17]Interacting Minds Center/Linguistics, Cognitive Science, and Semiotics, Aarhus University, Aarhus University, Jens Chr. Skous Vej 4, Aarhus, Central Denmark, 8000, [18]Department of Epidemiological Methods and Etiological Research, Leibniz Institute for Prevention Research and Epidemiology -BIPS, Achterstrasse 30, Bremen, 28359, Germany, [19]School of Medicine, Griffith University, Griffith University Nathan campus, Nathan, QLD, 4111, Australia, [20]School of Human Sciences, University of Greenwich, University of Greenwich, London, SE10 9LS, United Kingdom, [21]Occupational and Environmental Epidemiology and NetTeaching Unit, Institute and Clinic for Occupational, Social and Environmental Medicine; University Hospital, LMU Munich, Ziemssenstr. 1, Munich, Germany, Bavaria, 80336, [22]Department of Education and Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, Berlin, 14195, Germany, [23]Center for Translational Behavioral Science, Florida State University, 2010 Levy Ave Building B, Tallahassee, Florida, 32304, [24]School of Health and Society, University of Wollongong, Northfields Avenue, Wollongong, NSW, 2522, Australia, [25]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy, [26]Department of Anesthesiology, Universidad del Cauca, Carrera 5 # 13N – 36, Popayan, Cauca, 190002, Colombia, [27]Department of Mathematics, University of Exeter, Streatham Campus, Exeter, Devon, EX4 4QE, United Kingdom, [28]Harvard T.H. Chan School of Public Health, Harvard University, 677 Huntington Ave, Boston, MA, 02115, United States, [29]Department of Mathematical Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, United Kingdom, [30]Faculty of Medicine and Health, The University of Sydney School of Public Health, Sydney, New South Wales, Australia, [31]British Columbia Centre for Excellence in HIV/AIDS, Epidemiology and Population Health Program, 608–1081 Burrard Street, Vancouver, British Columbia, V6Z 1Y6, Canada, [32]UMR INSERM 1246 SPHERE, University of Nantes, University of Tours, 22 bd Benoni-Goullin, Nantes, 44200, France, [33]Department of Epidemiology, University of Washington School of Public Health, Hans Rosling Center for Population Health, Seattle, WA, 98195, United States, [34]Faculty of Medicine, UMF Carol Davila, INSP, Bucharest, sector 5, 50463, Romania, [35]College of Dentistry, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia, [36]Epidemiology, Biostatistics and Occupational Health, McGill University, Purvis Hall, Montreal, QC, H3A 1A2, Canada, [37]School of Public Health, University of Alberta, 3-50E University Terrace, Edmonton, Alberta, Canada, [38]Faculty of Medicine and Health, The University of Sydney School of Public Health, Sydney, NSW, Australia, [39]Department of Kinesiology and Community Health, University of Illinois Urbana-Champaign, 1206 S. Fourth Street, Champaign, IL, 61820, United States, [40]Community Medicine, University of Pittsburgh Medical Center, 7555 Saltsburg Rd, Pittsburgh, PA, 15206, United States, [41]Department of Neuroscience, Karolinska Institute, Solnavägen 1, Stockholm, 17177, Sweden, [42]Finnish Institute for Health and Welfare, Mannerheimintie 166, Helsinki, Finland, [43]Clinical Research Department, London School of Hygiene & Tropical Medicine, Keppel St, London, WC1E 7HT, United Kingdom, [44]Department of Population Health, Division of Epidemiology, New York University Grossman School of Medicine, 180 Madison Ave, 4-35A, New York, New York, 10024, United States, [45]Exponent, Inc., Center for Health Sciences, Exponent, Inc., 475 14th Street, Suite 400, Oakland, CA, 94612, United States, [46]Epidemiology, Boston University, 801 Massachusetts Ave, Boston, Massachusetts, 02118, United States

Keywords: causal language, association, causal inference, observational study

Abstract

Background: Avoiding “causal” language with observational study designs is common publication practice, often justified as being a more cautious approach to interpretation. We aimed to i) estimate the degree to which causality was implied by both the language linking exposures to outcomes and by action recommendations in the high-profile health literature, ii) examine disconnects between language and recommendations, iii) identify which linking phrases were most common, and iv) generate estimates by which these phrases imply causality.

Methods: We identified 18 of the most prominent general medical/public health/epidemiology journals, and searched and screened for articles published from 2010 to 2019 that investigated exposure/outcome pairs until we reached 65 non-RCT articles per journal (n=1,170). Two independent reviewers and an arbitrating reviewer rated the degree to which they believed causality had been implied by the language in abstracts based on written guidance. Reviewers then rated causal implications of linking words in isolation. For comparison, additional review was performed for full texts and for a secondary sample of RCTs.

Results: Reviewers rated the causal implication of the sentence and phrase linking the exposure and outcome as None (i.e., makes no causal implication) in 13.8%, Weak in 34.2%, Moderate in 33.2%, and Strong in 18.7% of abstracts. Reviewers identified an action recommendation in 34.2% of abstracts. Of these action recommendations, reviewers rated the causal implications as None in 5.3%, Weak in 19.0%, Moderate in 42.8% and Strong in 33.0% of cases. The implied causality of action recommendations was often higher than the implied causality of linking sentences (44.5%) or commensurate (40.3%), with 15.3% being weaker. The most common linking word root identified in abstracts was “associate” (n=535/1,170; 45.7%) (e.g. “association,” “associated,” etc). There were only 16 (1.4%) abstracts using “cause” in the linking or modifying phrases. Reviewer ratings for causal implications of word roots were highly heterogeneous, including those commonly considered non-causal.

Discussion: We found substantial disconnects between causal implications used to link an exposure to an outcome and the action implications made. This undercuts common assumptions about what words are often considered non-causal and that policing them eliminates causal implications. We recommend that instead of policing words, editors, researchers, and communicators should increase efforts at making research questions, as well as the potential of studies to answer them, more transparent.

Introduction

Health sciences research often investigates the relationship between a particular exposure and an outcome. Causal effects between these variables are often implicitly of interest, including studies based on non-random 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, rather than cautious examination of methodological strength of inference and uncertainty. Some author guidelines (e.g., Journal of the American Medical Association [1]) explicitly prohibit the use of causal language in any study that is not a randomized controlled trial (RCT), 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 non-causal. Alternatively, non-causal language may be used throughout the manuscript, but practical recommendations may still be offered that suggest or require a causal interpretation.[7] It is not entirely 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 perceptions of the work by research consumers and decision-makers.[4,5,13] 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 non-causal language may morph into causal language in outlets for medical practitioners,[7,10] press releases,[14–16] and media reports.[17,18] Ambiguous language may also imply greater support for any practical recommendations that require causal interpretation.[19] While some loss of nuance may be attributed to press officers, journalists, and news recipients, too-strong language often starts from the study publications themselves.[17] Most importantly, choice of language impacts research consumers' and decision-makers' perceptions,[13] which in turn impacts health decisions.

Despite widespread discussions about causal language use,[4,5,20] systematic evidence of its usage in practice is limited. In a review of 60 observational studies that were published in The BMJ, 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,[6,17,21] obesity,[11] and orthopedics,[22] 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 epidemiological literature; existing efforts[6,8–12] heavily focus on binary assessments of the language used (causal vs. non-causal).

This study systematically examined the linking language used in studies with a primary exposure and outcome in the high-profile medical and epidemiological literature. Our objectives were to (i) identify the linking words and phrases used to describe relationships between

exposures and outcomes, (ii) generate estimates of the strength of causality stated or implied by the linking phrases and sentences using a guided subjective assessment process, (iii) examine the prevalence of action recommendations that would require causal inference to have been made, and (iv) examine disconnects between causal implications in linking sentences and action implications.

Methods

Our target sample consists of studies that primarily quantified the relationship between a primary exposure and an outcome in humans and were 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 the coronavirus disease of 2019 (COVID-19). The study was pre-registered on the Open Science Framework (OSF): https://osf.io/jtdaz/. Changes made to the protocol after preregistration are documented and explained in Appendix 1.

Search

Our search was structured in two steps: a preliminary search for appropriate journals and a secondary search for published papers within these journals.

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)[23] categories for medicine and public health and SciMago’s category for Medicine. The top 200 journals from the SciMago Journal rank (SJR)[24] and JCR’s impact factor rating for medical journals, and the top 200 highest impact factor rating journals for Public Health as extracted on May 26, 2020 were screened according to the following inclusion criteria: (1) primarily serves articles that are peer-reviewed, about health-specific topics, reporting primary data (e.g., the journal cannot be one which primarily serves reviews, meta-analyses, and other secondary data), primarily concerning human-level observations (e.g., not animal models or microbiology); (2) must be a general health/medicine/epidemiology journal (i.e., journals which are focused on a narrow speciality and/or disease area of medicine were excluded); (3) the journal must have been founded in 2010 or earlier.

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. An additional decision was made during screening on June 24, 2021 to drop journals that where fewer 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 Appendix 1).

Search terms

We searched PubMed to identify all articles published in an eligible journal between 2010 and 2019 inclusive (details in Appendix 2).Medical Subject Headings (MeSH) terms were used to eliminate articles not meeting inclusion criteria. The search was performed in R[25] using the EasyPubMed package[26].

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[27] for disease area headings.

Screening

Study inclusion/exclusion criteria

Studies were eligible for inclusion if they were primarily concerned with the quantitative association of a primary exposure/outcome pair, as assessed by reviewers as below:

- Observations must be human- or at an aggregate group of humans-level of observation
    - The primary research question must be to examine the causal and/or non-causal association between one primary exposure concept and one primary outcome concept
    - One “primary” 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 primary association of interest (e.g., in the title, in the study aims)
        - Articles that are about more than one primary concept (e.g., searching for what ‘risk factors’ are associated with the outcome) were excluded.
- The primary research question must be examined quantitatively using primary data (i.e., not a review or meta-analysis)

Studies investigating more than one exposure/outcome set were excluded because (1) it would not be possible to assess a primary 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 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, totalling 1,278 articles (1,170 non-RCTs and 108 RCTs). The sample size was informed by informal explorations of sample datasets 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 analysed in this study is unknown and is one of the key objectives of this study, 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 three of 18 screening reviewers, with two primary reviewers and one arbitrating reviewer. During screening, the arbitrating reviewer made the inclusion/exclusion decision only in cases where the two primary reviewers disagreed.

Main review

Reviewer recruitment and selection

Reviewers were recruited through a combination of personal and Twitter solicitations. Reviewers were selected from those who expressed interest based on 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.

Reviewer roles and training

All reviewers received one hour of instruction and an additional set of training articles to review before the primary review. During the training and main review periods, reviewers were encouraged to engage in an active discussion on Slack to clarify guidelines, discuss issues, and generate community standards for review areas that may be more ambiguous. 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 were therefore allowed to make changes at any point before arbitration.

Each article was reviewed by three unique randomly selected reviewers; two independent primary reviewers and an arbitrating reviewer. The arbitrating reviewer was given the submitted data from the primary 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 primary reviewers, their own reading, and the ongoing community discussions. Arbitrating reviewers were free to decide in favor of one reviewer over another, consolidate and combine reviewer responses, or overturn both primary 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 of the key questions for our study. The framing and definitions of words used (e.g., definitions and guidance for how/why language might be “causal”) are provided in Appendix 3.

Reviewers had the option to recuse themselves of reviewing each article for any reason (e.g., conflicts of interest, connections to authors, etc.); 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 primary outcome and exposure, preferably from the title of the study. Reviewers were asked to identify and copy and paste the primary 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 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”).

Table 1: Causal implication strength rating scale

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 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 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 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 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.

Reviewers assessed the degree to which the linking sentence implied a causal relationship between the exposure and outcome using a four point scale ("linking sentence causal strength") shown in Table 1..

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 may 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 1.

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.

All articles received a review of the titles and abstracts. In addition, one-third of the articles underwent full text assessment. This extended review 1) repeated the abstract review questions for the discussion section and any pop-out 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,[28] (see supplementary data). 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, 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 that were 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 four randomly selected linking words/phrases that contained the root word and had been submitted by arbitrating reviewers (e.g., the root word "associate" had four phrases, including phrases like "associated with" or "association").

Analysis

The statistical analysis was largely descriptive (e.g., describing the distributions of key extracted variables). Except for comparisons between RCTs and non-RCTs, all statistical analysis was performed on the arbitrated dataset of the non-RCTs only.

Comparisons between two ordinal categorical variables (e.g., strength ratings for causal implications of linking sentence vs. action implications) 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 as such the Wilson estimator was used. No weights were applied (i.e., journals and articles respectively contribute equally to our main results).

Heterogeneity between reviewers was evaluated using Krippendorf's alpha. 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 v4.0.5.[25] Spearman correlation coefficients were determined using the pspearman package.[29] Ordinal logistic regression was performed using the MASS package.[30]

Data and code availability

All data and code are publicly available through our OSF repository: https://osf.io/jtdaz, except for files containing personal identifying information and/or personal API keys.

Patient and Public Involvement statement

No patients or participants were involved with this research. All data were obtained from academic literature sources.

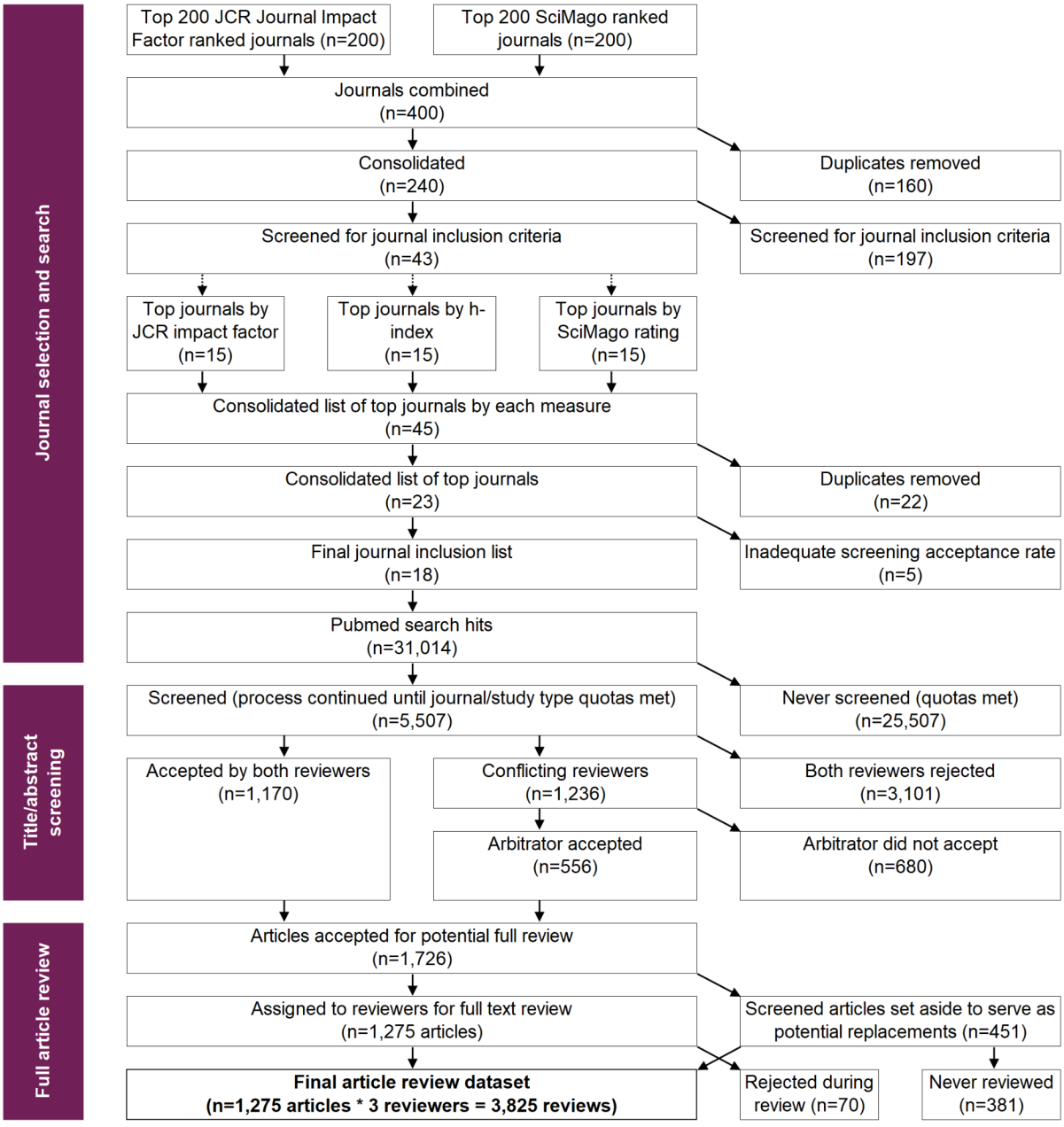
Ethics approval

This research is not human subjects research, and as such no ethical approval was required. This research complies with the Declaration of Helsinki.

Results

Search and screening

Figure 1: PRISMA diagram showing the selection of journals and articles into the study sample



Caption: This chart shows the PRISMA diagram detailing the search and screening process to arrive at our final sample.

Figure 1 summarises the selection of journals and articles into the sample. Eighteen journals were identified meeting our search criteria: 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.

After searching PubMed for articles published in these journals from 2010-2019, we screened articles until 65 non-RCTs and 6 RCTs were accepted from each journal except the European Journal of Epidemiology, where only 3 RCTs were identified and included. This yielded 1,170 non-RCTs and 105 RCTs, totalling 1,275 studies reviewed. There were 10 recusals recorded during the main review. The three 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 Appendix 4 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.

Figure 2: List and frequency of identified root words used to link the exposure and outcome

Root linking words

Root word	n	Root word	n	Root word	n	Root word	n
associate	535	risk factor	13	elevate	6	prevent	3
increase	71	contribute	12	lead	6	role	3
high	36	effective	12	better	4	achieve	2
predict	34	affect	10	compare	4	consistent	2
reduce	33	link	10	greater	4	differ	2
likely	29	cause	9	protect	4	due	2
lower	26	impact	9	show	4	excess	2
relate	25	result	9	similar	4	precede	2
improve	21	benefit	7	appear	3	reveal	2
effect	19	correlate	7	demonstrate	3	twice	2
risk	17	explain	7	determinant	3	vary	2
different	16	attribute	6	factor	3	worse	2
decrease	14	change	6	less	3		
influence	13	decline	6	occur	3	Other	78

Caption: This chart shows the number of times each of these root words appears in the linking phrases in the abstracts of our samples. In cases where two of these words are in the same phrase (e.g., "similar risk") the more common of the two is selected (in this case "risk"). In cases where selected linking phrases had two or more words which were included in the root word list, the more common word was selected as the root word primarily associated with that study and section.

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%) (Figure 2). The same root word was identified in both the abstract and discussion for 48.2% cases (95% CI 43.7, 53.6%). We found only 9 (0.8%, 95% CI 0.4, 1.3%) studies where the primary 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(s) strengths

Summary data

Figure 3: Summary scores for the degree of causal implication in linking sentences and action recommendations

Strength of causal implication in linking sentence

Abstract		Discussion		Popout section	
Rating	**n**	**Rating**	**n**	**Rating**	**n**
None	162	None	56	None	20
Weak	400	Weak	134	Weak	36
Moderate	389	Moderate	130	Moderate	38
Strong	219	Strong	70	Strong	23
Total	1170	Total	390	Total	117

Strength of causal implication in action recommendation

Abstract		Discussion		Popout section	
Rating	**n**	**Rating**	**n**	**Rating**	**n**
N/A	770	N/A	155	N/A	57
None	21	None	17	None	5
Weak	76	Weak	38	Weak	10
Moderate	171	Moderate	110	Moderate	23
Strong	132	Strong	70	Strong	22
Total	1170	Total	390	Total	117

Caption: This chart shows the frequency of key strength of causal implication metrics for the 1,170 non-RCT studies in our sample, as indicated by the arbitrating reviewer.

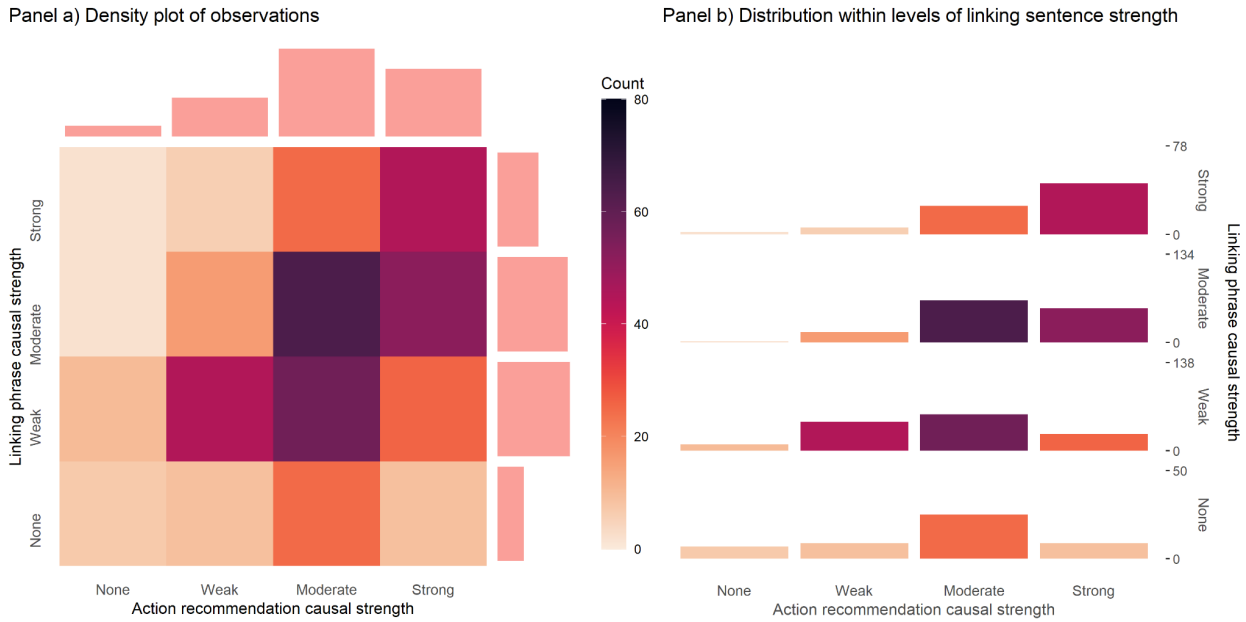
Reviewers rated the abstract linking sentence as having no causal implication in 13.8% (95% CI 11.9, 15.9%), weak in 34.2% (95% CI 31.4, 36.7%), moderate in 33.2% (95% CI 29.8, 36.7%), and strong in 18.7% (95% CI 15.1, 22.6%) of instances Figure 3). The language used was very similar in the abstract, full-text discussion, and pop-out 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%) Weak, 42.8% (95% CI 39.0, 46.4%) Moderate, and 33.0% (95% CI 29.0, 37.1%) 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, 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 (Appendix 5).

Comparison of linking sentence strength vs. action implication strength

Figure 4: Comparison of the strength of causal implications in the abstracts for the linking phrase and action recommendation

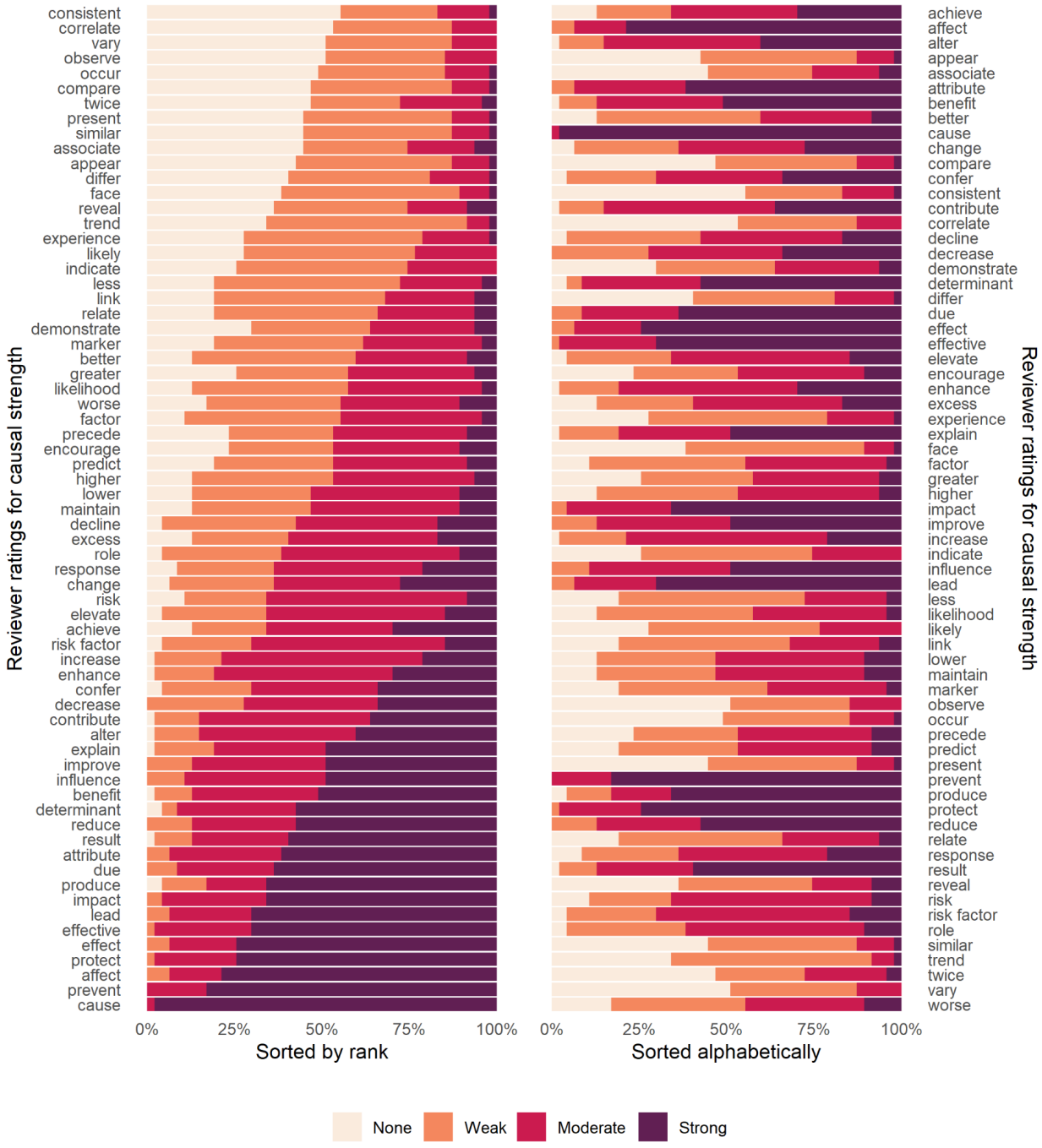


Caption: This chart shows the distribution of linking sentence and action recommendation language, among the 400/1,170 non-RCT studies in which there was an action recommendation present in the abstract. Panel A shows an unconditional heatmap, with colors representing the number of articles in the strata, and histograms on the top and right showing the overall distribution of ratings for each axis. Panel B shows the distributions within each level of linking sentence causal strength.

Figure 4 shows the strength of causal implication for linking sentences and action recommendations among those studies with an action recommendation in the abstract (n=400, 34.0%). Of studies with action recommendations, 15.3% (95% CI 11.7, 19.2%) had action recommendations that less strongly implied causality 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 (Panel A). 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 (Panel B), 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%) Weak, 33.1% (95% CI 29.2, 37.3%) Moderate, and 18.3% (95% CI 14.5, 22.5%) 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

Figure 5: Strength of causal implication ratings for the most common root linking words



Caption: This chart shows the distribution of ratings given by reviewers during the root word rating exercise. On the left side, they are sorted by median rating + the number of reviewers who would have to change their ratings in order for the rating to change. On the right, the chart is sorted alphabetically.

As shown in Figure 5, ratings among reviewers (n=47) for causal implication of root words were highly heterogeneous, with the only word to reach near consensus on causal implications being “cause” (Figure 5). Reviewers rated words such as “correlate” and “associate” generally rated 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."

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 words and phrases

Modifying phrases were identified in the abstracts of 72.1% of studies (95% CI 69.0, 75.6%). 11.2% (95% CI 8.6, 14.2%) of studies had a modifying phrase with variations on "statistical" and/or "significant." Phrases expressing caution (e.g., "may be," "could," "potentially") or strength (e.g. "strongly," "substantially,") were both fairly common in the modifying phrases extracted. However, given the wide variety of phrases extracted and the lack of a pre-established framework for doing so, no formal categorization of modifying phrases was performed or quantified. The frequency of modifying words and phrases identified three or more times are shown in Appendix 6.

Differences in strength across key strata

Non-RCTs vs. RCTs

For the RCTs, reviewers rated the causal implication of the abstract linking sentence as being None for 9.5% (95% CI 4.8, 15.2%), Weak for 6.7% (95% CI 2.8, 11.4%), Moderate for 27.6% (95% CI 19.0, 36.4%), and Strong for 56.2% (95% CI 46.4, 65.8%). This is overall much stronger than for the non-RCTs [log-odds of RCTs having higher rank: 1.63 (95% CI 1.26, 2.04)].

Overall, 75.2% (n=79/105; 95% CI 66.7, 82.9%) of RCTs in our sample had no action recommendation. Of the 26 that did, 0 were rated as having a causal implication of None, 6.7% (95% CI 2.8, 11.4%) Weak, 27.6% (95% CI 19.4, 36.6%) Moderate, and 56.0% Strong (95% CI 45.7, 64.8%). The action recommendation causal implications in the RCTs appeared stronger than in the non-RCTs [log odds for higher rank: -0.398 (95% CI -0.916, 0.009), noting that this is underpowered due to insufficient RCTs with action recommendations to make reasonable inference about differences.

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).

Journals and journal policies

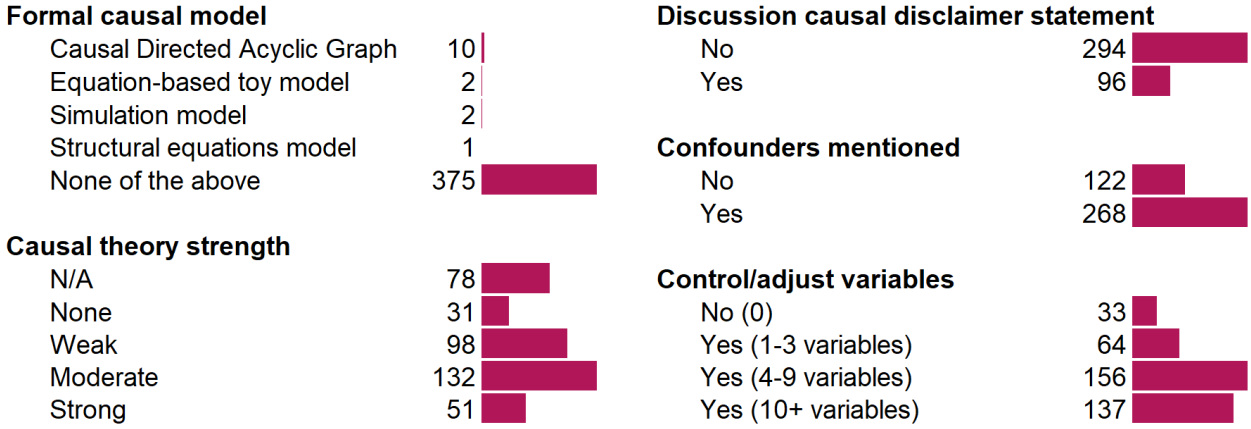
The strength of causal linking language and action recommendations was generally similar between journals (Appendix 7). Three journals have publicly posted policies regarding causal language. The Journal of the American Medical Association (JAMA) and JAMA Internal

Medicine explicitly restrict the use of “causal” language to RCTs, while the American Journal of Epidemiology (AJE) discourages the use of the word “effect” exclusively, giving guidance as to when it should be used. Compared with the other 15 journals, the strength of the causal language in these three journals was weaker, especially for JAMA which had the lowest rank of any journal [log odds of having higher rank: -0.627 (95% CI -0.771, -0.483) for JAMA, -0.083 (95% CI -0.235, 0.069) for JAMA Internal Medicine, and -0.080 (95% CI -0.229, 0.069) for AJE. The difference in the causal language strength in named journals of epidemiology compared to other journals appears to be small , with the log odds of having a higher linking language strength being -0.140 (95% CI -0.447, 0.166)].

There was little difference between journals in causal strength of action recommendations when present, but there were some differences in their prevalence (Appendix 8 and Appendix 9). Papers in JAMA, JAMA Internal Medicine, and AJE were less likely to include actions recommendations than papers in the other 15 journals [log odds for having action recommendations: -0.624 (95% CI -0.885, -0.363 for JAMA, -0.090 (95% CI -0.351, 0.170) for JAMA Internal Medicine, and -0.806 (95% CI -1.067, -0.546) for AJE]. Articles from epidemiology journals (n=3) were less likely to include action recommendations than the remaining (n=15) journals [log odds for having action recommendations -0.516 (95% CI -0.870, -0.163)].

Indications of potential causal interest

Figure 6: Frequency of indicators of potential causal interest



Caption: These results are from the 390 articles reviewed in full.

Most studies in our sample provided *some* indication of potential causal interest (Figure 6). 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%; 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 the two. 24.6% (95% CI 20.9, 28.0%) of studies had a disclaimer statement regarding causality. 68.7% (95% CI 63.3, 73.7%) mentioned “confounding” explicitly (i.e. using the word “confound” or

variations of it). Finally, most studies in our sample controlled or adjusted for several variables, with 35.1% (95% CI 30.5, 39.9%) having 10 or more control variables.

Inter-rater comparisons

The Krippendorff's alpha comparing the reviewers' ratings for linking language strength in the abstract was 0.29. Primary reviewers gave the same score in 35.1% of instances; 41.2% differed by one category, 19.9% by two categories, and 3.8% by three categories. Agreement increased to 0.41 when including the primary and arbitrating reviewers.

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

Discussion

Our systematic evaluation of the use of causal language and implications in the high-profile medical and epidemiological 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 only about a third of articles issued action recommendations, the vast majority of these were found to imply that causality had been inferred, and 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 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,"[31] - 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 (non-causal) interpretations. To our surprise, we found that the RCTs in our sample used similar linking words to the non-RCTs. Our review suggests that the degree of causal interpretation for common linking words has been impacted by the unavailability of explicitly causal language, such that the meaning of traditionally non-causal words has broadened to include potentially stronger causal interpretations.[32] 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. Outside of this study, we expect that a more general set of consumers of health research (clinicians, policy-makers, and others) may also interpret these words differently, whether by virtue of differing frameworks for assessing language, personal interpretations, or community standards. It also matters how words are used and in what context. For example, we found that 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 pre-registration, nearly all parts of this project were fully open and disseminated to the public to view and comment, 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 multi-exposure articles with terms such as “risk factors,” “correlates,” or “predictors.” Our journal selection also included only the most prominent general medical, public health, and epidemiology journals, and 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 the association between X and Y is 42 is not informative if we do not know what question that association attempts to answer.[33] Further, these practices may weaken methodological

accountability, as studies that only indirectly imply causality can be shielded from critique on the grounds of 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 designs and assumptions; and for authors and editors to focus on being clearer about what questions we are asking,[34,35] what decisions we are trying to inform, and the degree to which we are and are not able to achieve those goals.

Works cited

1. Instructions for Authors | JAMA | JAMA Network [Internet]. [cited 2021 May 11]. Available from: https://jamanetwork.com/journals/jama/pages/instructions-for-authors
2. Chipperfield L, Citrome L, Clark J, David FS, Enck R, Evangelista M, et al. Authors' Submission Toolkit: A practical guide to getting your research published. Curr Med Res Opin. 2010 Aug;26(8):1967–82.
3. AMA Manual of Style Committee. AMA Manual of Style: A Guide for Authors and Editors [Internet]. 11th ed. Oxford University Press; 2020 [cited 2021 May 11]. Available from: https://www.amamanualofstyle.com/view/10.1093/jama/9780190246556.001.0001/med-9780190246556
4. Hernan MA. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. Am J Public Health. 2018 May;108(5):616–9.
5. Grosz MP, Rohrer JM, Thoemmes F. The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. Perspect Psychol Sci. 2020 Sep;15(5):1243–55.
6. Olarte Parra C, Bertizzolo L, Schroter S, Dechartres A, Goetghebeur E. Consistency of causal claims in observational studies: a review of papers published in a general medical journal. BMJ Open. 2021 May;11(5):e043339.
7. Prasad V, Jorgenson J, Ioannidis JP, Cifu A. Observational studies often make clinical practice recommendations: an empirical evaluation of authors' attitudes. J Clin Epidemiol. 2013 Apr;66(4):361-366 e4.
8. Adams RC, Challenger A, Bratton L, Boivin J, Bott L, Powell G, et al. Claims of causality in health news: a randomised trial. BMC Med. 2019 May 16;17(1):91.
9. Adams RC, Sumner P, Vivian-Griffiths S, Barrington A, Williams A, Boivin J, 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, Mühlhauser I. Causal interpretation of correlational studies – Analysis of medical news on the website of the official journal for German physicians. Berner ES, editor. PLOS ONE. 2018 May 3;13(5):e0196833.
11. Cofield SS, Corona RV, Allison DB. Use of causal language in observational studies of obesity and nutrition. Obes Facts. 2010 Dec;3(6):353–6.
12. Watkins TR. Understanding uncertainty and bias to improve causal inference in health intervention research [Internet]. [Australia]: UNSW Sydney; 2019. Available from: https://ses.library.usyd.edu.au/bitstream/handle/2123/20772/watkins_tr_thesis.pdf
13. Hall MG, Grummon AH, Maynard OM, Kameny MR, Jenson D, Popkin BM. Causal Language in Health Warning Labels and US Adults' Perception: A Randomized Experiment. Am J Public Health. 2019 Oct;109(10):1429–33.
14. Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Venetis CA, Davies A, et al. The association between exaggeration in health related science news and academic press releases: retrospective observational study. BMJ. 2014 Dec 9;349:g7015.
15. Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Bott L, Adams R, et al. Exaggerations and Caveats in Press Releases and Health-Related Science News. PLoS One. 2016;11(12):e0168217.
16. Schwitzer G. Addressing tensions when popular media and evidence-based care collide. BMC Med Inform Decis Mak. 2013 Dec;13(S3):S3.
17. Haber N, Smith ER, Moscoe E, Andrews K, Audy R, Bell W, 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.
18. Haneef R, Lazarus C, Ravaud P, Yavchitz A, Boutron I. 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.
19. Alvarez-Vargas D, Braithwaite DW, Lortie-Forgues H, Moore MM, Castro M, Wan S, et al. Hedges, mottes, and baileys: Causally ambiguous statistical language can increase perceived study quality and policy relevance [Internet]. PsyArXiv; 2020 May [cited 2020 Jul 15]. Available from: https://osf.io/nkf96
20. Thapa DK, Visentin DC, Hunt GE, Watson R, Cleary M. Being honest with causal language in writing for publication. J Adv Nurs. 2020 Jun;76(6):1285–8.
21. Ramspek CL, Steyerberg EW, Riley RD, Rosendaal FR, Dekkers OM, Dekker FW, et al. Prediction or causality? A scoping review of their conflation within current observational research. Eur J Epidemiol [Internet]. 2021 Aug 15 [cited 2021 Aug 19]; Available from: https://link.springer.com/10.1007/s10654-021-00794-w
22. Varady NH, Feroe AG, Fontana MA, Chen AF. Causal Language in Observational Orthopaedic Research. J Bone Jt Surg [Internet]. 2021 Apr 22 [cited 2021 Aug 19];Publish Ahead of Print. Available from: https://journals.lww.com/10.2106/JBJS.20.01921
23. Clarivate Analytics. Journal Citation Reports [Internet]. 2018 [cited 2020 May 26]. Available from: https://jcr.clarivate.com/JCRJournalHomeAction.action
24. SciMago. SciMago Journa and Country Rank [Internet]. 2020 [cited 2020 May 26]. Available from: https://www.scimagojr.com/journalrank.php?area=2700&order=h&ord=desc
25. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/
26. Fantini D. easyPubMed: Search and Retrieve Scientific Publication Records from PubMed [Internet]. 2019. Available from: https://CRAN.R-project.org/package=easyPubMed
27. NLM. Download MeSH Data [Internet]. U.S. National Library of Medicine; [cited 2021 Aug 16]. Available from: https://www.nlm.nih.gov/databases/download/mesh.html
28. Han MA, Guyatt G. Systematic survey of the causal language use in systematic reviews of observational studies: a study protocol. BMJ Open. 2020 Jul 28;10(7):e038571.
29. Savicky P. pspearman: Spearman's rank correlation test [Internet]. 2014. Available from: https://CRAN.R-project.org/package=pspearman
30. Venables WN, Ripley BD. Modern Applied Statistics with S [Internet]. Fourth. New York: Springer; 2002. Available from: https://www.stats.ox.ac.uk/pub/MASS4/
31. Tennant PWG, Murray EJ. The Quest for Timely Insights into COVID-19 Should not Come at the Cost of Scientific Rigor. Epidemiology. 2021 Jan;32(1):e2–e2.
32. de Carvalho A, Reboul AC, Van der Henst J-B, Cheylus A, Nazir T. Scalar Implicatures: The Psychological Reality of Scales. Front Psychol [Internet]. 2016 Oct 25 [cited 2021 Aug 20];7. Available from: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01500/full
33. Adams D. The Hitchhiker's Guide to the Galaxy. New York: Harmony Books; 1980.
34. Fox MP, Edwards JK, Platt R, Balzer LB. The Critical Importance of Asking Good Questions: The Role of Epidemiology Doctoral Training Programs. Am J Epidemiol. 2020 Apr 2;189(4):261–4.
35. Lundberg I, Johnson R, Stewart BM. What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. Am Sociol Rev. 2021 Jun;86(3):532–65.

Acknowledgements

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.

All errors are the sole responsibility of the authors.

Author roles

Protocol design: NAH, SW, MPF, JMR, OAA, PWGT, EJM, EAS
Study administration: NAH, SW
Data management: NAH
Statistical analysis: NAH
Graphical design: NAH
Study design (piloting): SW, SPi, ER, CL, ALO, RB, SD, MDLRT, TSA, DJD, MS, TM, SPe
Data analysis (screening): NAH, SPi, STL, SJH, AES, PS, AB, MSKD, SD, TRE, DRM, TMA, GMK, AA, JAC, MJK, COP
Data analysis (main review): NAH, SW, MPF, JMR, OAA, PWGT, EAS, SPi, STL, ER, SJH, AES, CL, PS, AB, MSKD, ALO, RB, SD, MDLRT, TRE, DRM, TMA, DJD, GMK, AA, JAC, MJK, MS, COP, TM, AC, JS, AS, TSA, SET, JD, EA, RAH, SKS, SS, NJ, SPe, CA, PK, AERA, NUO, IS
Manuscript writing: NAH, SW, JMR
Manuscript editing: NAH, SW, MPF, JMR, OAA, EJM, PWGT, EAS, SPi, STL, ER, SJH, AES, CL, PS, AB, MSKD, ALO, RB, SD, MDLRT, TRE, DRM, TMA, DJD, GMK, AA, JAC, MJK, MS, COP, TM, AC, JS, AS, TSA, SET, JD, EA, RAH, SKS, SS, NJ, SPe, CA, PK, AERA, NUO, IS

NAH serves as the primary guarantor of all aspects of the study and takes full responsibility for the work.

Competing interests

The authors declare no competing interests.

Transparency statement

The lead author, Noah A. Haber, affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Funding

No funding was granted specifically for the support of this study, 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 that 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 is also required.

The Meta-Research Innovation Center at Stanford University is supported by Arnold Ventures LLC (Houston, Texas), formerly the Laura and John Arnold Foundation.

Sophie Pilleron was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 842817.

Saman Khalatbari-Soltani is supported by the Australian Research Council Centre of Excellence in Population Ageing Research (Project number CE170100005).

Ian Schmid is supported by National Institute of Mental Health grant T32MH122357.

Elizabeth Stuart's time was supported by National Institute of Mental Health grant R01MH115487 and the Bloomberg American Health Initiative.

Ashley O'Donoghue is funded by a philanthropic gift from Google.org outside of the submitted work.

Onyebuchi A. Arah is supported by 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.

Appendices

Appendix 1: Changes from pre-registered protocol

Major changes:

1. The primary measure of linking language causal implication strength indirectly through the root words to direct reviewer ratings of the sentences themselves.
    a. In the original protocol, the primary method of generating causal implication for the linking language was through the root word rating system, where those ratings would then be applied back to the studies from which they came. No question was asked regarding the causal implications of the linking sentence in context.
    b. During piloting, we added the question to the review tool which had reviewers directly assess the causal implications of the linking sentences as a whole in order to better rate and review the language in context.
    c. During the primary/independent review, but before the arbitrator review phase, we changed our primary linking language measure from the root word exercise to the direct ratings of the sentences themselves.
    d. This decision was made for three reasons
        i. This greatly simplified the estimation of the main results, negating the need to back-apply causal language from the root word ratings
        ii. The full sentence context would be a more direct and contextually sensitive assessment of causal language than the root word exercise.
        iii. During an interim data quality check of the reviewers' extracted linking phrases, we found that the extracted data were much more heterogeneous than initially anticipated, lending some doubt whether the original strategy was viable and interpretable.
2. Journals with very low rates of screening acceptance were retroactively excluded from the list of journals
    a. This decision was made partway through the screening process itself.
    b. Because the protocol specified that we would have the same number of articles accepted from each of the journals, during the screening process we found that screeners would have to work vastly more to meet journal quotas among the journals which had very low rates of screening acceptance.
    c. On June 24, journals which had screening acceptance rates of below 10% or journals in which there were not enough unscreened articles remaining to meet quotas were excluded, and quotas were increased to compensate among the remaining journals.
    d. This decision was made for two primary reasons:
        i. Keeping these journals would have created an infeasible amount of screening required to complete the screening process.

ii. Journals with such low rates of screening acceptance were likely less relevant to meet our stated objectives and journal inclusion criteria.

Minor changes:

1. We selected 18 journals, rather than the initial expected 20 from the protocol.
    a. The expected number of journals in the protocol (20) was made in error. We chose to follow the process, rather than aim for a specific number of journals. This initially yielded 24 journals, 6 of which were later removed due to low screening acceptance rates (see above)
2. The sample size target changed to 1,170 non-RCTs (61 per journal) and 90 RCTs (6 per journal)
    a. The protocol was initially stated to be 1,525 articles accepted, with 61 non-RCTs per journal and 6 RCTs.
    b. This was reduced due to lower than expected screening acceptance rates in order to ensure that screening logistics were feasible and that schedules would be met.
3. The data extraction form received a large number of minor tweaks to the language, phrasing, and guidance.
    a. These changes were made as part of the protocol-specified piloting process.
4. The root word extraction process was performed on the linking phrases collected from the arbitrator reviews, rather than the primary reviews.
    a. This ensured a cleaner dataset of linking words and phrases from which to extract root linking words
5. Root words were only included in the root word linking exercise if there were two or more instances of them from the arbitrator reviews, and a light curating process was performed afterwards.
    a. This was performed due to clean up highly heterogeneous extracted linking words and phrases
6. The root word rating exercise was changed to being performed after the arbitrator reviews.
    a. In the original protocol, the root word exercise occurred during the arbitrator reviews.
    b. This change was made in order to accomodate extracting the root words from the arbitrator-extracted linking phrases
7. The reviewers were assigned to review all of the words in the root word list
    a. The original protocol specified that the reviewers would only review 20 randomly selected root words
    b. This was performed in order to maximize the power of our sample.
8. Spearman's correlation coefficients were added to directly examine the correlation of rankings between ordinal categories
    a. This was not originally specified in the protocol due to oversight, and was added later.
9. The population weighted tertiary analysis was removed

a. This was omitted due to lack of clear value of targeting an alternative “population” of studies, to simplify the breadth of analyses, and due to lack of space

Appendix 2: Search terms

Our search was performed and pulled from PubMed to extract title, abstract, MeSH keywords, and citation data, using the following terms:

> ((<*year*>[PDAT]) AND (<*journal ISSN*>[Journal])
> AND
> (Humans[mesh] AND "Journal Article"[PT] AND English [la] AND hasabstract))
> NOT
> (("Meta-Analysis"[Publication Type] OR "Review"[Publication Type] OR "Case Reports"[Publication Type] OR "Editorial"[Publication Type] OR "Letter"[Publication Type]))"

Where <*year*> is the years from 2010 to 2019, and <*journal ISSN*> is the journal in question. The above search was performed for every year/journal combination and combined.

Appendix 3: Definitions and frameworks

Exposure: For this project, "Exposure" refers to the independent variable of interest (in a regression sense) or the primary or antecedent variable being investigated for a possible (non-)causal link to the study outcome, or resulting or end-point variable. It may be labelled by terms such as treatment, factor, risk factor, protective factor, determinant, intervention, correlate, predictor, agent, cause, causative agent, or other terms.

Outcome: For this project, "Outcome" refers to the dependent or effect variable of interest that is being investigated for a possible link to the exposure (surrogate measures or clinical events). It is typically a post-exposure variable i.e. assumed or known to be preceded by the exposure. It is sometimes called the study endpoint variable, consequence, result, or other terms.

Linking word/phrase: A linking word/phrase describes the nature of the connection between some defined exposure and some defined outcome, generally used in a sentence containing both exposure and outcome. 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. For our purposes, the phrase may contain 1-3 words, where one of the words is a preposition to link the exposure and outcome. Some examples may include constructions such as "associated with," "effect of," "increased," "was higher than," "correlated with," "caused," "harms," "predicts," "risk factor for," "determined," "impacts," "decreased," "linked to," etc.

Modifying word/phrase: A modifying word/phrase is a word or phrase that modifies the linking word/phrase describing the nature of the relationship between the exposure and outcome. This includes adding signals of direction, strength, doubt, negation, and statistical properties to the relationship. This may include phrases like "may be," "positively," "strongly", "potentially", "is likely to," "does/is not," "statistically significant," etc.

Causal language: Causal language implies that one entity influences (or does not influence) another. We define language as being causal if that language implies that movement (or lack thereof) in the outcome was either 1) impelled by the exposure of interest (i.e. a change in the exposure drives or does not drive a change in the outcome, e.g., increase, decrease, improve, change), or 2) implies attribution of the outcome to the exposure (i.e. assigns the responsibility for the change or lack of change in the outcome to the exposure, e.g. "due to," "since," "attributable to").

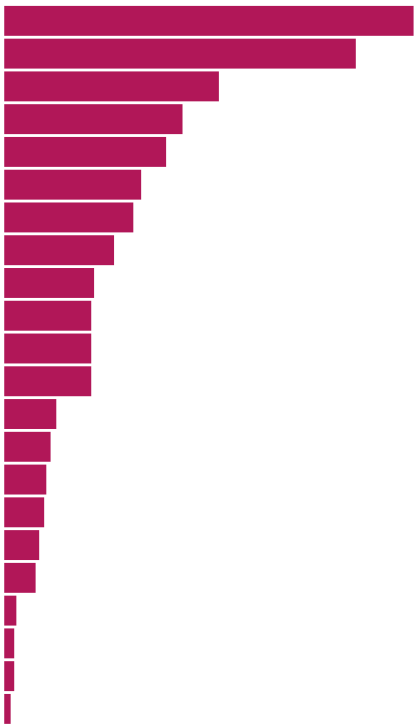
Action recommendation: This is a description of how a consumer of the research in question might utilize the results and conclusions of the research. This may include recommending that some actor consider changes (or no changes) in some set of procedures and actions. Action recommendations concern what to do with the research. For our purposes, we do not count calls for additional research as action recommendations.

Causal implication of recommendations: Recommendations may often imply a causal interpretation of a finding. For example, authors may suggest that it could be beneficial to

change the amount of an exposure, which rests on the assumption that the exposure has a causal effect on the outcome. As a variation, it may also be suggested that an exposure need not be changed, which rests on the assumption that the absence of a causal effect has been established.

Appendix 4: MeSH disease areas

MeSH disease area	n
Pathological Conditions, Signs and Symptoms	377
Cardiovascular Diseases	324
Nutritional and Metabolic Diseases	198
Neoplasms	164
Female Urogenital Diseases and Pregnancy Complications	149
Nervous System Diseases	126
Infections	119
Endocrine System Diseases	101
Immune System Diseases	83
Digestive System Diseases	80
Respiratory Tract Diseases	80
Male Urogenital Diseases	80
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	48
Skin and Connective Tissue Diseases	43
Chemically-Induced Disorders	39
Wounds and Injuries	37
Musculoskeletal Diseases	32
Hemic and Lymphatic Diseases	29
Stomatognathic Diseases	11
Occupational Diseases	9
Otorhinolaryngologic Diseases	9
Eye Diseases	6



Appendix 5: Causal strength over time

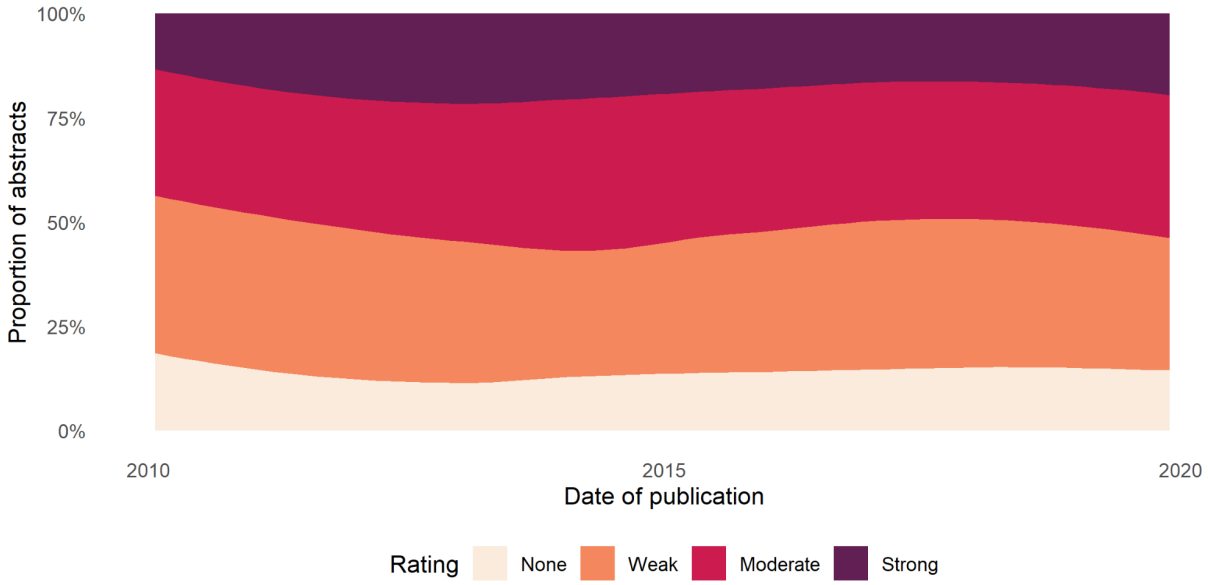


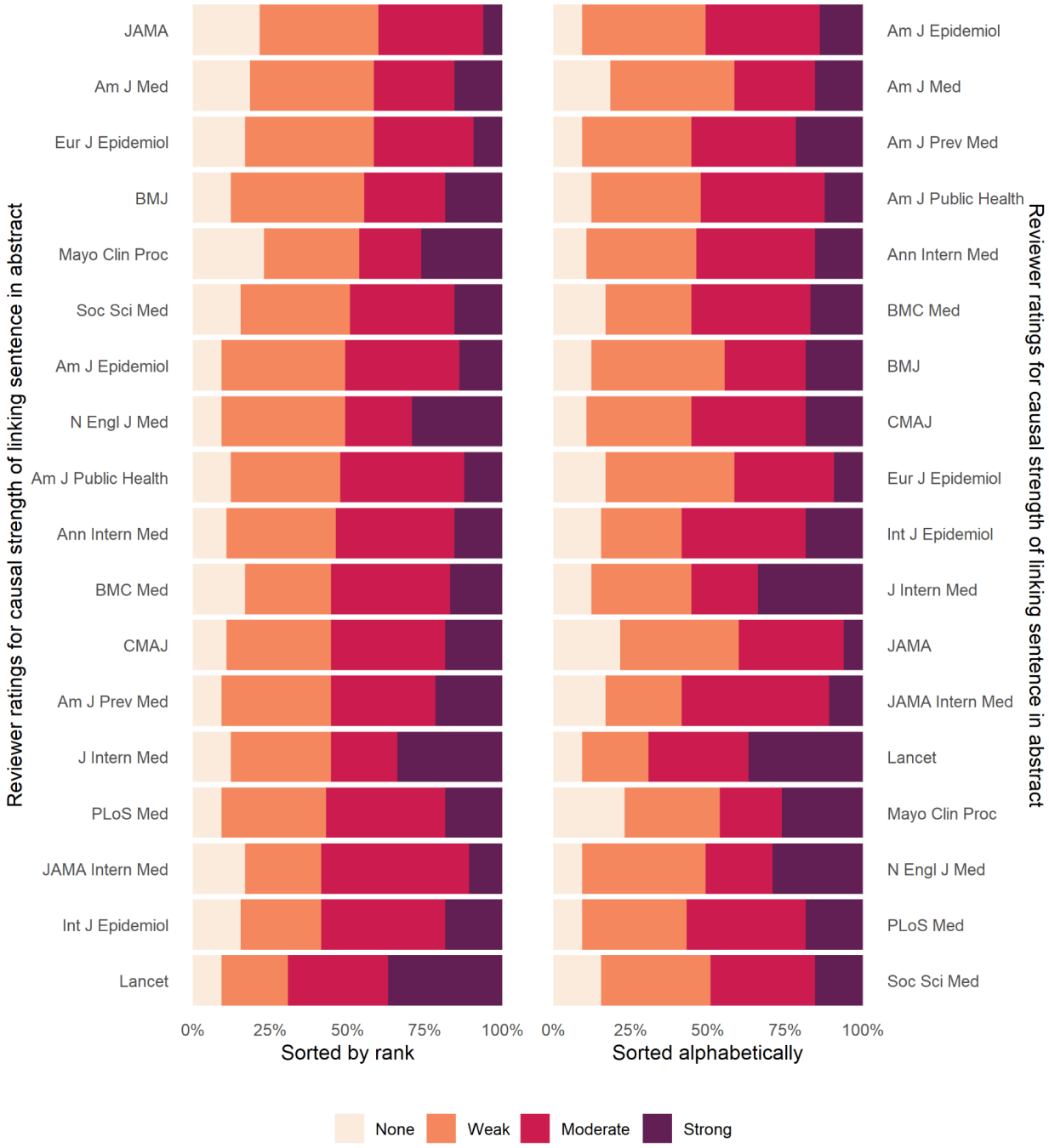
Chart is generated through LOESS smoothing the proportions in each category over time.

Appendix 6: Modifying phrases

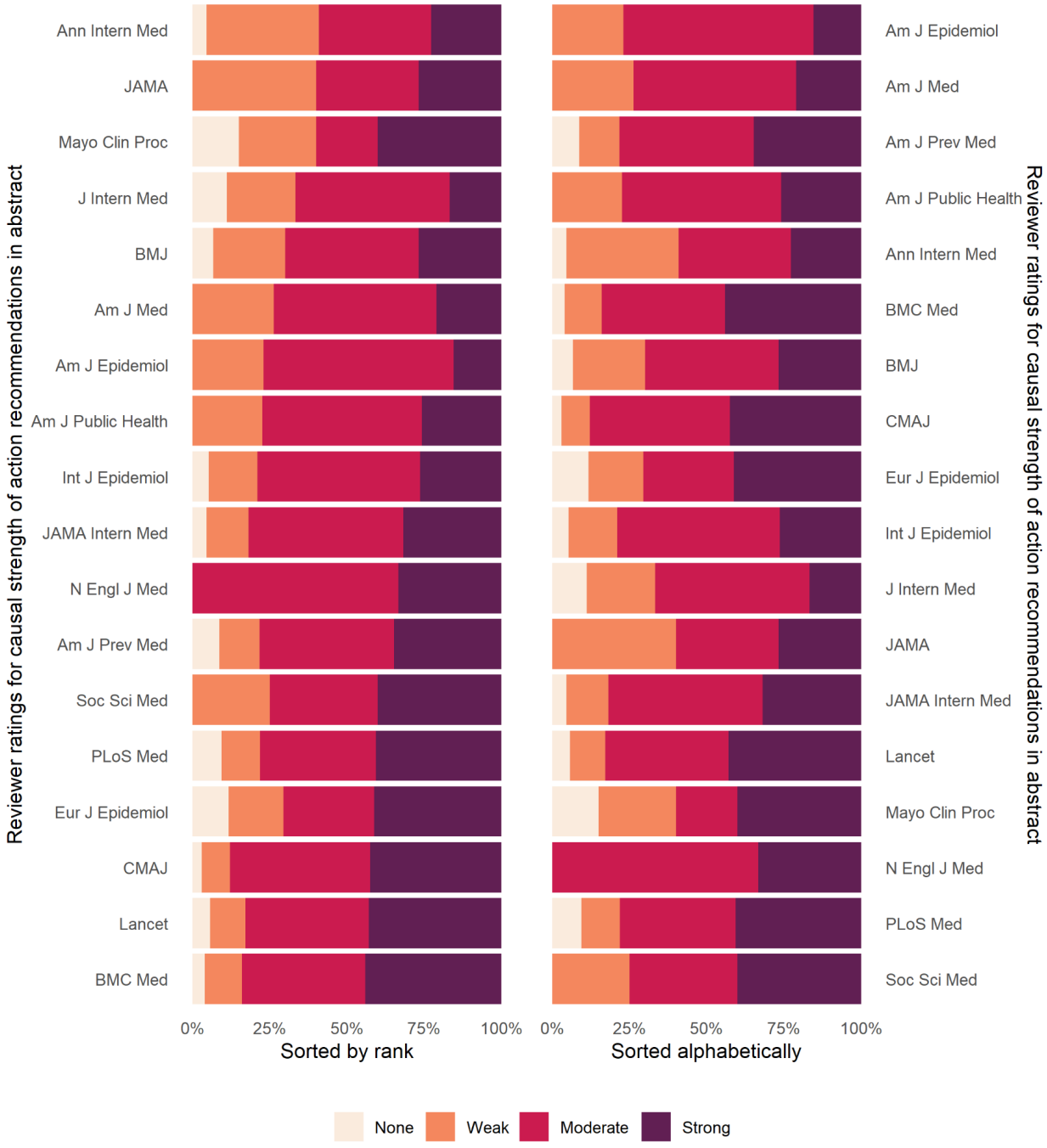
Modifying phrases

Phrase	n	Phrase	n	Phrase	n	Phrase	n
significantly	60	positively	10	important	5	generally	3
increased	52	strong	10	inverse	5	higher risk	3
not	46	greater	9	positive	5	independent of	3
lower	43	can	8	adverse	4	longer	3
significant	33	less	8	high	4	lower risk	3
higher	26	substantial	8	highly	4	major	3
independently	23	better	7	increases	4	might be	3
may	22	did not	7	may have	4	modest	3
increased risk	19	increase	7	reductions	4	negative	3
no	16	markedly	7	statistically	4	potentially	3
reduced	15	substantially	7	appear	3	reduction	3
independent	14	suggest	7	appears to be	3	risk	3
strongly	12	decreased	6	clinically	3	significantly incre	3
may be	11	improved	6	could	3	similar	3
more	11	appears	5	decreasing	3	statistically signifi	3
inversely	10	consistently	5	do not support	3	U-shaped	3

Appendix 7: Causal strength of linking sentences in abstract, by journal



Appendix 8: Causal strength action recommendations in abstract, by journal



Appendix 9: Causal strength action recommendations in abstract including NAs, by journal