Journal of Clinical Epidemiology 93 (2018) 94−102

Journal of Clinical Epidemiology

Interpretation of epidemiologic studies very often lacked adequate consideration of confounding

Lars G. Hemkens[a,b], Hannah Ewald[a], Florian Naudet[c], Aviv Ladanie[a], Jonathan G. Shaw[d], Gautam Sajeev[e,f], John P.A. Ioannidis[b,c,f,g,*]

[a]*Department of Clincial Research, Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, University of Basel, Basel, Switzerland*

[b]*Department of Medicine, Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA*

[c]*Meta-Research Innovation Center at Stanford (METRICS), Stanford, CA, USA*

[d]*Division of Primary Care and Population Health, Stanford University School of Medicine, Stanford, CA, USA*

[e]*Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA*

[f]*Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA*

[g]*Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA*

Accepted 18 September 2017; Published online 21 September 2017

Abstract

Background and Objective: Confounding bias is a most pervasive threat to validity of observational epidemiologic research. We assessed whether authors of observational epidemiologic studies consider confounding bias when interpreting the findings.

Study Design and Setting: We randomly selected 120 cohort or case−control studies published in 2011 and 2012 by the general medical, epidemiologic, and specialty journals with the highest impact factors. We used Web of Science to assess citation metrics through January 2017.

Results: Sixty-eight studies (56.7%, 95% confidence interval: 47.8−65.5%) mentioned "confounding" in the Abstract or Discussion sections, another 20 (16.7%; 10.0−23.3%) alluded to it, and there was no mention or allusion at all in 32 studies (26.7%; 18.8−34.6%). Authors often acknowledged that for specific confounders, there was no adjustment (34 studies; 28.3%) or deem it possible or likely that confounding affected their main findings (29 studies; 24.2%). However, only two studies (1.7%; 0−4.0%) specifically used the words "caution" or "cautious" for the interpretation because of confounding-related reasons and eventually only four studies (3.3%; 0.1−6.5%) had limitations related to confounding or any other bias in their Conclusions. Studies mentioning that the findings were possibly or likely affected by confounding were more frequently cited than studies with a statement that findings were unlikely affected (median 6.3 vs. 4.0 citations per year, $P = 0.04$).

Conclusions: Many observational studies lack satisfactory discussion of confounding bias. Even when confounding bias is mentioned, authors are typically confident that it is rather irrelevant to their findings and they rarely call for cautious interpretation. More careful acknowledgment of possible impact of confounding is not associated with lower citation impact.

Keywords: Confounding; Bias; Observational studies; Research reporting; Bibliometrics

Conflict of interest: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf and declare no financial support for this project. L.G.H. is member of the RECORD initiative which aims to improve reporting of observational studies using routinely collected health data. He has no other relationships or activities that could appear to have influenced the submitted work. F.N. has relationships (travel/accommodation expenses covered/reimbursed) with Servier, BMS, Lundbeck, and Janssen who might have an interest in the work submitted in the previous 3 years. He has no other relationships or activities that could appear to have influenced the submitted work. All other authors declare no relationships or activities that could appear to have influenced the submitted work.

Authors' contributions: L.G.H. and J.P.A.I. conceived the study. L.G.H. analyzed the data. All authors interpreted the results. L.G.H. wrote the first draft and all authors made revisions on the article. L.G.H., H.E., A.L., F.N., J.G.S., and G.S. extracted the data. All authors read and approved the final version of the article. L.G.H. and J.P.A.I. are guarantors. All authors had full access to all 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.

Funding: This work was supported by a grant of the Laura and John Arnold Foundation to The Meta-Research Innovation Center at Stanford.

Role of the funding source: The funders had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the article or its submission for publication.

Data sharing: No additional data available.

Ethical approval: Not required for this study.

* Corresponding author. Stanford Prevention Research Center, Medical School Office Building, Room X306 1265 Welch Rd, Stanford, CA 94305, USA. Tel.: 650-725-5465; fax: 650-725-6247.

E-mail address: jioannid@stanford.edu (J.P.A. Ioannidis).

https://doi.org/10.1016/j.jclinepi.2017.09.013

What is new?

Key findings

- Many highest impact observational studies lack any discussion of confounding bias. Even when mentioned, authors are typically confident that it is rather irrelevant for their findings and they rarely call for cautious interpretation.

What this adds to what was known?

- There is no evidence that acknowledging the potential impact of confounding diminishes citation impact of epidemiological studies.

What is the implication and what should change now?

- There is a need to encourage researchers and to sensitize reviewers and editors to discuss and communicate study limitations introduced by confounding.

1. Introduction

A confounder may create spurious associations between an exposure and an outcome observed in epidemiologic studies [1]. For example, many more people drinking coffee have lung cancer than people not drinking coffee, but this is because they more often smoke [2]. Many confounders are difficult to pinpoint with certainty, many are entirely unknown, and many others are known, but are still not measured and thus cannot be considered in the analysis of epidemiologic studies. Understanding confounding and separating it from causal effects can be very difficult. For example, even smoking's causal role in cancer, and its potential to confound other observed associations in cancer studies, was not clear across many years of early epidemiologic research [3]. Bias caused by unknown confounders is directly addressable only by randomization, and thus, confounding bias can never be entirely ruled out in non-randomized studies. Consequently, in the most widely applied framework to assess quality of evidence for healthcare decisions (GRADE), evidence from observational research is initially considered low quality [4].

Because bias due to confounding is a core limitation of observational research, numerous recommendations and statements call for a careful consideration when reporting, discussing, and making conclusions from observational research [5–10]. For example, the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement, the most widely endorsed guideline for reporting of observational research, prominently emphasizes the discussion of confounding and explicitly states "*It is important not only to identify the sources of bias and confounding that could have affected results, but also to discuss the relative importance of different biases, including the likely direction and magnitude of any potential bias*" and "*due consideration should be given to confounding […]. Authors should also consider residual confounding due to unmeasured variables or imprecise measurement of confounders*" [6].

Despite these recommendations, many investigators might feel that acknowledgment of confounding will cast doubts on their findings. They might prefer to either be silent about this possibility or explicitly discredit the possibility that confounding may have affected their conclusions. Important questions can be asked: Do authors of epidemiologic studies published in major journals acknowledge confounding properly and sufficiently? Does more explicit acknowledgment of confounding as a limitation decrease the subsequent citation impact of their work? To address these questions, here we conducted a meta-epidemiologic survey of observational studies published in high-impact journals. Our primary aim was to assess whether authors of observational epidemiologic studies consider confounding bias when interpreting the findings in the Discussion sections and concluding statements of their articles. Our secondary aim was to determine whether such explicit discussion is associated with lower citation impact.

2. Methods

2.1. Data identification and eligibility

We selected 24 journals with the highest impact factors (Journal Citation Reports 2010): The top eight from the "medicine, general, and internal" category [New England Journal of Medicine, Lancet, JAMA, Annals of Internal Medicine, PLOS Medicine, BMJ, Archives of Internal Medicine (currently JAMA Internal Medicine), CMAJ], the top eight from the "public, occupational, and environmental health" category (Environmental Health, Epidemiology, International Journal of Epidemiology, American Journal of Epidemiology, Bulletin of the World Health Organization, American Journal of Preventive Medicine, European Journal of Epidemiology, Genetic Epidemiology), and the journal with highest impact factor in each of eight "medical specialty" sub-categories (cardiology and cardiovascular disease, gastroenterology, obstetrics and gynecology, oncology, pediatrics, rheumatology, surgery, urology and nephrology; i.e., Circulation, Gastroenterology, Obstetrics, and Gynecology, Journal of Clinical Oncology, Pediatrics, Annals of Rheumatic Diseases, Annals of Surgery, Journal of the American Society of Nephrology). We did not consider journals focusing exclusively on reviews (e.g., Epidemiologic Reviews) or on basic and/or preclinical research (e.g., Cancer Cell).

We searched MEDLINE for cohort and case–control studies published in these journals in 2011 and 2012 (last search on December 4, 2015; details in Webappendix 1).

The articles retrieved were stratified by journal category. Two independent reviewers (H.E. and F.N.) evaluated

randomly selected articles for eligibility until they identified 120 eligible articles (20 per journal type and year; which would allow for standard deviation of <4% for estimated proportions of 75% or 25%). The study flow is shown in Webappendix 2. We included any study clearly described as ''cohort study'' or ''case−control study'' (explicitly using these terms) and reporting any exposure−outcome association and thus being theoretically prone to confounding bias. No further eligibility criteria were applied. Any disagreements were resolved by discussion or with a third reviewer (L.G.H.). The random sample included studies published in 22 of the 24 eligible journals (exceptions were Bulletin of the World Health Organization and Genetic Epidemiology), and each journal contributed a median of four studies [interquartile range (IQR) 2−6].

2.2. Data extraction

Two independent reviewers (two of L.G.H., H.E., F.N.) extracted the reported study design (i.e., case−control, prospective, retrospective, or unclassified cohort study or nested case−control study; we applied these specific terms to categorize the study design as self-reported by the authors) and categorized the area of research for all pertinent articles. Any disagreements were resolved by discussion or with the third reviewer (L.G.H., H.E., or F.N.).

In addition to manual extractions, two independent reviewers (L.G.H. and H.E.) searched all full-texts automatically (using PDF viewer software) for terms related to propensity scores or marginal structural models anywhere in the articles and they assessed if propensity score−based methods or marginal structural models were used in the studies. There was perfect agreement (100%) between reviewers.

One reviewer (L.G.H.) extracted from Web of Knowledge bibliographic data, specifically the journal's 2010 impact factor and how often the study was cited (Web of Science Core Collection) through January 2, 2017, to calculate an annual citation rate (total citations received per years elapsed since publication).

2.3. Evaluation of confounding statements and bias consideration

We systematically evaluated the consideration of confounding bias in the Abstract and Discussion sections of included studies using six standardized prespecified questions (Table 1). We focused on the Abstract and Discussion because these are the sections readers typically focus on the most and from which they are most likely to draw bottom line conclusions on what the research means and what caveats might exist. We did not evaluate the Introduction, Methods, or Results sections of the publications.

First, we evaluated if the term ''confounding'' in any form is mentioned at all, regardless of whether it is actually used to discuss the findings of the study or not. We specifically screened Abstract and Discussion sections of the articles for the term ''confounding'' or variations thereof (Question 1). We also captured any allusions or statements referring to the concept of confounding bias without explicitly using such terms. We also specifically screened the articles for the term ''bias'' (Question 2) and explicitly perused any mentions of bias for possible relations to confounding. Details with examples are shown in Table 1.

Second, we evaluated if the authors explicitly mention specific potential confounders that were not adjusted for in the analyses (Question 3), or if the authors explicitly discuss whether confounding bias is likely, possible, or unlikely to affect their main findings (Question 4).

Third, we evaluated if confounding bias is considered when interpreting the results or drawing conclusions. Specifically, we evaluated if the authors state that their main results need to be interpreted with caution due to confounding, using the term ''caution,'' ''cautious,'' or variants thereof (Question 5). Finally, we specifically screened whether their concluding statements include any limitation or uncertainty related to confounding or bias at all (Question 6). This was evaluated in the section either headed ''conclusion,'' ''summary,'' or similar; if such heading did not exist, we evaluated all paragraphs following a concluding statement beginning with, for example, ''in conclusion,'' or ''in summary,'' or evaluated the last paragraph of the Discussion.

We developed and pilot tested the operationalization of the questions and iteratively specified the wording of the questions to arrive at detailed extraction instructions. Two reviewers (two of L.G.H., H.E., F.N., A.L.) then assessed all articles independently (unaware of any extractions in the pilot), resolving any disagreements by discussion or with a third reviewer (L.G.H. or H.E.).

2.4. Data analysis

In addition to an overall description of the study sample and the statements on confounding, we analyzed whether the consideration of confounding (Questions 1−6) differed between the journal types (general medical vs. epidemiology vs. specialty journal), study types (cohort vs. case−control), exposures (modifiable vs. nonmodifiable), and whether it was associated with journal impact factor and article annual citation rate. We tested differences between continuous variables with the Mann−Whitney U test, differences between categorical data with the Fisher's exact test. Results for continuous measures are medians with IQRs. All analyses were done with Stata 13.1. *P* values are two tailed.

3. Results

3.1. Evaluated studies

Of the 120 articles, 90 described cohort studies (75%) and 30 case−control studies (25%; Table 2; details in Webappendix 3). Case−control studies were typically published in epidemiologic journals (17 of 30; 56.7%). The 120 studies covered a wide spectrum of medical areas, and there were differences in the areas covered between general medical journals and specialty

Table 1. Assessment of consideration of confounding bias in Abstracts and Discussions

Question
1. Do the authors mention confounding using explicitly the terms "confounder(s)," "confounding," "confound," or do they allude to it without using those terms, or is confounding not considered at all? Examples for "yes": *"We caution that these associations may reflect unmeasured confounding by diet or other lifestyle factors"* [11]. Example for "alluded": *"Another potential limitation is our inability to control for age at menopause among women having a hysterectomy before natural menopause; for these women, age at menopause is unknown"* [12]. *"When we included the characteristics we could define in multivariable models the association of arm injection site with a significantly higher risk of medically attended local reactions persisted, but it is possible that bias may have influenced the findings"* [13].
2. Do the authors mention bias using explicitly the term "bias"? Example for "yes": *"Where available, we relied on HIV diagnosis based on clinical features, which may be subject to biases in assessing the factors contributing to diarrheal disease among participants since HIV infection at early stages may have been missed and not all data were routinely captured"* [14].
3. Do the authors mention specific confounders that have not been adjusted for? (If yes, what were the reasons? If not, were there unspecified unmeasured confounders without specifically stating which ones?) Example for "yes": *"We were unable to adjust for additional confounding variables with a known association with mortality (for example, blood glucose and postarrest pH) that were not collected as part of the PICANet data set"* [15].
4. Do the authors state that their main findings are likely, possibly, or unlikely affected by residual confounding? Example for "yes, likely": *"Therefore, some residual confounding with parental psychopathology seems likely"* [16]. Example for "yes, possibly": *"However, although we adjusted for severity of the initial diagnosis of depression, we could use only a crude measure as we did not have a validated depression severity score. We cannot therefore exclude the possible effect of residual confounding on our results"* [17]. Example for "yes, unlikely": *"Minimal differences were observed between the crude and adjusted odds ratios for the exposure variable, suggesting that SEIFA and ethnicity were unlikely to be major confounders in this analysis"* [18].
5. Do the authors state that their main findings need to be interpreted with caution due to confounding? We answered this question with "yes" in cases with a clear statement that cautious interpretation is required because of confounding. Example for "yes": *"Caution is needed when interpreting the results of the analyses on proportion of the association explained. First, the proportion estimates, decomposed from the total effect by adjusting for other biomarkers, may be biased if there is unmeasured confounding between the biomarkers and the outcome [Reference]. In the present study, we included a large variety of known risk factors as well as of biomarkers, thereby minimizing unmeasured confounding"* [19].
6. Do the authors call for caution or indicate limitations or uncertainty due to possible confounding or other bias in their conclusions? Example for "yes": *"We caution that these associations may reflect unmeasured confounding by diet or other lifestyle factors"* [11]. *"Given the small sample size, however, the potentially confounding effects of maternal IQ cannot be excluded and should be evaluated in a larger study"* [20]. *"In summary, notwithstanding the possibility of residual selection bias, patients who […]"* [21].

journals, with pediatrics and oncology being more common in the latter. Most studies (74; 61.7%) analyzed effects of exposures that cannot practically be investigated in experimental studies as they are either not directly modifiable or are harmful (e.g., associations of health outcomes with environmental factors, biomarkers, or demographic characteristics). Effects of potentially modifiable exposures (e.g., drugs, diets, or surgery) were analyzed in 35 studies (29.2%) and were less common in epidemiologic journals. The median impact factor of the 22 journals was 7.9 (IQR, 5.6–13.5) in 2010 and the studies received a median of 5.1 (IQR, 2.5–9.2) annual citations, with clear differences depending on journal type. Of the 120 studies, only six used propensity score methods and one used marginal structural modeling.

3.2. Mere mentioning of confounding or bias

Confounding bias was not mentioned or alluded to at all in Abstracts and Discussions of 32 of the 120 studies (26.7%; 95% confidence interval [CI]: 18.8–34.6%; Table 3); in 20 studies (16.7%; 95% CI: 10.0–23.3%), there was some allusion to the concept of confounding indirectly without using this specific term, and 68 of 120 (56.7%; 95% CI: 47.8–65.5%) mentioned the term "confounding" or some same-root variant. The term "bias" was used in 72 of the 120 studies (60%; 95% CI: 51.2–68.8%). Twenty-seven studies (22.5%; 95% CI: 15.0–30.0%) mentioned neither confounding nor bias at all in their Abstracts and Discussions.

3.3. Any mention that confounding may affect results

Among the 68 of 120 studies that used the term "confounding" or related terminology, three (2.5%; 95% CI: 0–5.3%) said that it is likely that confounding affects their main findings, 26 (21.7%; 95% CI: 14.3–29.0%) said it is possible, 11 (9.2%; 95% CI: 4.0–14.3%) said it is unlikely, and the remaining 28 did not comment in this regard.

Table 2. Characteristics of studies

Study characteristics	Total, no. (%)	Journal category			*P*-value
		General medicine, no. (%)	Epidemiology, no. (%)	Medical specialties, no. (%)	
Number of studies	120 (100)	40 (100)	40 (100)	40 (100)	–
Study design					<0.01
Case–control	22 (18.3)	3 (7.5)	13 (32.5)	6 (15.0)	–
Nested case–control study[a]	8 (6.7)	2 (5.0)	4 (10.0)	2 (5.0)	–
Cohort study, prospective	48 (40.0)	19 (47.5)	17 (42.5)	12 (30.0)	–
Cohort study, retrospective	25 (20.8)	8 (20.0)	2 (5.0)	15 (37.5)	–
Cohort study, unclassified	17 (14.2)	8 (20.0)	4 (10.0)	5 (12.5)	–
Area of disease or condition					<0.01
Cardiology, CVD	12 (10.0)	5 (12.5)	5 (12.5)	2 (5.0)	–
Obstetrics and gynecology	16 (13.3)	6 (15.0)	8 (20.0)	2 (5.0)	–
Oncology	16 (13.3)	0 (0.0)	7 (17.5)	9 (22.5)	–
Pediatrics	27 (22.5)	6 (15.0)	7 (17.5)	14 (35.0)	–
Other	49 (40.8)	23 (57.5)	13 (32.5)	13 (32.5)	–
Type of exposure					<0.01
Pathogens	4 (3.3)	2 (5.0)	0 (0.0)	2 (5.0)	
Genetics	5 (4.2)	0 (0.0)	2 (5.0)	3 (7.5)	
Diet	5 (4.2)	3 (7.5)	2 (5.0)	0 (0.0)	
Surgery	6 (5.0)	1 (2.5)	1 (2.5)	4 (10.0)	
Demographic characteristics	7 (5.8)	1 (2.5)	5 (12.5)	1 (2.5)	
Comorbidities	9 (7.5)	4 (10.0)	3 (7.5)	2 (5.0)	
Diagnostics/prediction rules	12 (10.0)	5 (12.5)	1 (2.5)	6 (15.0)	
Environmental factors	13 (10.8)	1 (2.5)	11 (27.5)	1 (2.5)	
Biomarkers	14 (11.7)	3 (7.5)	6 (15.0)	5 (12.5)	
Drug treatment	14 (11.7)	6 (15.0)	1 (2.5)	7 (17.5)	
Nonmodifiable, other, or multiple	10 (8.3)	4 (10.0)	2 (5.0)	4 (10.0)	
Modifiable, other, or multiple	17 (14.2)	8 (20.0)	4 (10.0)	5 (12.5)	
Modifiable and nonmodifiable	4 (3.3)	2 (5.0)	2 (5.0)	0 (0.0)	
Citation impact					
IF 2010 (median, IQR)	7.9 (5.6–13.5)	14.5 (13.5–30.0)	5.7 (4.5–5.7)	7.9 (5.4–12.0)	
(range), n = 120	2.5–53.5	9.0–53.5	2.5–5.9	4.4–19.0	
Citations/year (median, IQR)	5.0 (2.6–9.8)	9.1 (4.8–19.7)	3.7 (2.3–5.1)	5.1 (2.5–9.2)	
(range), n = 120	0.2–66.7	1.3–66.7	0.2–11.1	0.7–33.6	

Abbreviations: CVD, cardiovascular disease; IF, impact factor; IQR, interquartile range.
[a] Including two case–cohort studies.

3.4. Acknowledgment of unmeasured confounders

Authors of 34 studies (28.3%; 95% CI: 20.3–36.4%) acknowledged that for specific confounders, there was no adjustment, and the reason provided in the majority (28 of 34) was that these confounders had not been measured. Another eight studies mentioned unmeasured confounding in general without specifying the unmeasured confounders.

3.5. Cautious interpretation and limitations in conclusions

An explicit statement in the Discussion section (or Abstract) that the interpretation of study results should be made with caution due to possible confounding was made in only 2 of 120 studies (1.7%; 95% CI: 0–4.0%). Specifically, in a study of caffeinated beverage and soda consumption and time to pregnancy, Hatch et al. clearly stated "*We caution that these associations may reflect unmeasured confounding by diet or other lifestyle factors*" [11]. In a study of the association of different biomarkers and risk of type II diabetes, Montonen et al. stated "*Caution is needed when interpreting the results of the analyses on proportion of the association explained. First, the proportion estimates […] may be biased if there is unmeasured confounding between the biomarkers and the outcome [References]*" [19].

Only 4 of 120 studies (3.3%; 95% CI: 0.1–6.5%) mentioned any limitations related to bias or confounding in their Conclusions.

Of the three studies where the authors' discussion expressed that confounding likely affects their main results, this caution was clearly expressed in the Conclusions in one of the three. Such caution was conveyed in the Conclusion in only 2 of the 26 studies where the authors mentioned possible confounding.

Of the 42 studies where unmeasured confounders were discussed (specifically or in general terms), only one (2.4%) explicitly stated that the interpretation of the results should be made with caution and only four (9.5%) expressed in their Conclusions limitations because of confounding or any other bias.

Table 3. Statements on confounding

Question	Total, no. (%)	Journal category: General medicine, no. (%)	Epidemiology, no. (%)	Medical specialties, no. (%)	*P*-value interrater agreement
Total	120 (100)	40 (100)	40 (100)	40 (100)	
1. "Confounding" mentioned in Abstract or Discussion?					0.33
					88.2%
Yes, specific term	68 (56.7)	24 (60.0)	26 (65.0)	18 (45.0)	
Alluded	20 (16.7)	6 (15.0)	7 (17.5)	7 (17.5)	
No	32 (26.7)	10 (25.0)	7 (17.5)	15 (37.5)	
2. Term "Bias" used in Abstract or Discussion?					0.30
					93.6%
Yes	72 (60.0)	27 (67.5)	25 (62.5)	20 (50.0)	
No	48 (40.0)	13 (32.5)	15 (37.5)	20 (50.0)	
3. Specific nonadjusted confounders acknowledged?					0.50
					89.8%
Yes	34 (28.3)	11 (27.5)	14 (35.0)	9 (22.5)	
…because not measured	28 (82.4)	11 (100)	12 (85.7)	5 (55.6)	0.039
…because of other reasons	4 (11.8)	0 (0)	2 (14.3)	2 (22.2)	
…no reasons given	2 (5.9)	0 (0)	0 (0)	2 (22.2)	
No	86 (71.7)[a]	29 (72.5)	26 (65.0)	31 (77.5)	
4. Any mention that findings may be affected by confounding?					0.39
					86.5%[b]
Likely	3 (2.5)	1 (2.5)	1 (2.5)	1 (2.5)	
Possibly	26 (21.7)	10 (25.0)	8 (20.0)	8 (20.0)	
Unlikely	11 (9.2)	3 (7.5)	7 (17.5)	1 (2.5)	
No statement	80 (66.7)	26 (65.0)	24 (60.0)	30 (75.0)	
5. Cautious interpretation needed?					0.33
					99.2%
Yes	2 (1.7)	0 (0)	2 (5.0)	0 (0)	
No	118 (98.3)	40 (100)	38 (95.0)	40 (100)	
6. Conclusions include any limitations?					>0.99
					98.3%
Yes	4 (3.3)	1 (2.5)	2 (5.0)	1 (2.5)	
No	116 (96.7)	39 (97.5)	38 (95.0)	39 (97.5)	

[a] In 8 of the 86 studies, unmeasured confounding was mentioned, but no specific confounder stated.
[b] Interrater agreement calculated only for the 40 studies making a statement.

3.6. Overall assessment

The interrater agreement was very high for all assessed questions, ranging from 86.5% to 99.2%. Figure 1 shows the overlap we observed between the different ways of handling and characterizing the potential presence and impact of confounding bias.

3.7. Associations with type of journal and impact

The findings were overall the same across the types of journals (Table 3). None of the evaluated aspects of considering confounding bias were associated with journal impact factor or subsequent citation impact, with one exception (Table 4). Studies with a statement that the findings were possibly or likely affected by confounding bias were more frequently cited than those studies with a statement that the findings were unlikely affected (median 6.3 vs. 4.0 citations per year, $P = 0.04$). We found no differences between cohort and case–control studies or between studies evaluating modifiable vs. nonmodifiable exposures (data not shown).

4. Discussion

Our analysis of 120 randomly selected epidemiologic studies showed that while a narrow majority studies do mention confounding bias to some degree, very few acknowledge that it is a reason for major caution in interpreting the key findings. More than a quarter of the articles completely ignored "confounding" in the Abstract or Discussion sections, and most of them do not even mention the term "bias" in general. Despite the frequent presence and even awareness of specific unmeasured confounders and the often reported possible impact on the main findings, conclusions are almost never made with explicit caution. We found only two cases with explicit statements that cautious interpretation is required because of confounding. Interestingly, in one of them, this caution owing to unmeasured confounding is immediately diluted in the text by stating "*In the present study, we included a large variety of known risk factors as well as of biomarkers, thereby minimizing unmeasured confounding*" [19]. This illustrates the overall impression we gained during our evaluation, that

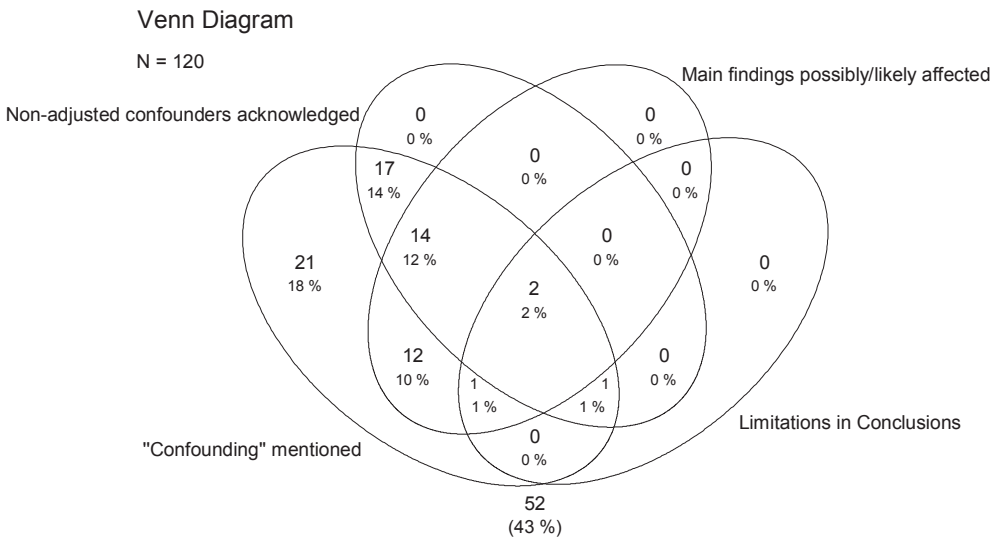

Fig. 1. Venn diagram on different aspects of consideration of confounding bias in discussions of epidemiologic research. Each ellipsoid area corresponds to one aspect of consideration of confounding bias. The numbers indicate the number of studies sharing the characteristics in the overlapping areas, for example, there are 14 epidemiologic studies (12% of 120) in which ''confounding'' is mentioned in the Abstract or Discussion, the authors deem the main findings possibly or likely affected by confounding, and nonadjusted confounders are acknowledged, but there are no limitations in the Conclusions related to confounding or any bias. Fifty-two studies are not covered by any of the areas. The percentages do not correspond to the size of circular areas.

many discussions of confounding in these top journals are superficial and appear to be attempts to negate the importance and impact of confounding in the published work.

We found no indications that this phenomenon is limited to certain areas of research, as findings were similar across types of journals, their impact factors, and study types and topics. Of note, many of the studies we evaluated were from journals that published the STROBE reporting guidelines in 2007 (i.e., Lancet, Epidemiology, Bulletin of the World Health Organization, BMJ, PLOS Medicine, Annals of Internal Medicine). The observed association of higher study citation numbers with statements acknowledging that confounding bias could exist might be just a chance finding, or be due to confounding. Nevertheless, it suggests that statements acknowledging potential methodological weaknesses have no negative citation impact.

Investigators should not worry that their observational study will be discredited if they acknowledge (as they should) that their work is subject to confounding that might affect their results. Acknowledgment and thorough discussion of the impact of confounding bias may be a marker of researchers with more epidemiologic training being involved in the study, who may have better institutional access to better, larger datasets, and work in larger research teams, all of which may also help explain higher citation rates for articles that explicitly discuss confounding. We did not adjust for any of

Table 4. Citation impact

Question	No. of studies	Journal if (median, IQR)	*P*-value	Citations per year (median, IQR)	*P*-value
1. ''Confounding'' mentioned in Abstract or Discussion?					
Yes	68	9.0 (5.7–14.4)	0.46	5.4 (2.6–9.5)	0.69
No or allude	52	6.7 (5.4–13.5)		4.6 (2.6–10.8)	
2. Term ''Bias'' used in Abstract or Discussion?					
Yes	72	8.7 (5.7–14.4)	0.24	5.2 (2.7–8.8)	0.79
No	48	6.7 (5.4–13.5)		4.9 (2.5–12.4)	
3. Specific nonadjusted confounders acknowledged?					
Yes	34	9.0 (5.7–13.5)	0.73	5.6 (2.2–9.3)	0.72
No	86	6.7 (5.4–13.5)		4.8 (2.7–10.2)	
4. Any mention that findings may be affected by confounding?					
Possibly or likely	29	13.5 (5.7–14.4)	0.07	6.4 (4.7–10.2)	0.04
Unlikely	11	5.7 (2.5–13.5)		4.0 (2.2–5.1)	
5. Cautious interpretation needed?					
Yes	2	5.2 (4.5–5.9)	0.30	2.4 (2.2–2.5)	0.15
No	118	8.3 (5.7–13.5)		5.1 (2.7–10.0)	
6. Conclusions include any limitations?					
Yes	4	7.1 (4.2–10.9)	0.59	9.7 (5.3–21.9)	0.28
No	116	7.9 (5.6–14.0)		4.9 (2.6–9.5)	

IF: Journal Citation Reports 2010 Impact Factor.

these potentially explanatory variables in our descriptive analyses as we do not aim to make any causal inferences. If anything, we observed more citations for articles that acknowledged confounding than for those that did not.

The acknowledgment of unmeasured confounding (in accordance to the STROBE reporting guideline) has been systematically assessed in previous empirical work for observational research published in five general medicine journals and five epidemiologic journals (most of them included also in our analysis) for the years 2004−2007 and 2010−2012 [22,23]. Comments on the likelihood of unmeasured confounding were present in 59−85% of the studies, but only 16−32% gave any qualitative statement about the impact on the findings, which agrees well with our overall study results. However, both of these previous empirical studies narrowly evaluated observational research specifically focusing on medical interventions, while we examined the broader landscape of observational investigation within the medical literature, only the minority of which pertained to interventions.

Some limitations of our work deserve closer attention. First, we analyzed only a small sample of the observational study literature. Perhaps, a larger sample may have allowed us to detect small differences between journal types or other factors affecting the consideration of confounding. However, large differences are unlikely to have been missed.

Second, we evaluated studies that were published 4 and 5 years ago, which was necessary for a meaningful analysis of subsequent citation impact. Previous evaluations have found that the introduction of STROBE in 2007, arguably the most influential effort to improve reporting quality, has had only modest impact on reporting quality [22,23]. No new major similar efforts have been launched in the last 5 years; therefore, we have no reason to believe that reporting of observational research would have changed substantially in the last few years.

Third, by only looking at 24 high-impact journals, it is uncertain if our findings are generalizable to the rest of the medical literature. It is quite possible that we may even underestimate the extent to which implications of confounding bias go unaddressed in the medical literature.

We also acknowledge that confounding bias might be seen by some researchers as an inevitable limitation of observational studies that is too well-known to merit discussion. However, as causal interpretations depend on the validity of the implicit assumption of no unmeasured/residual confounding, the implications of bias due to failure of this assumption should be considered. Dealing with confounding bias, understanding its impact (e.g., through qualitative discussion of the magnitude and direction of bias and more quantitative sensitivity analyses [24,25]), minimizing its influence, and acknowledging the residual uncertainty is an integral core for inference-making in epidemiology. In some situations, authors might not be much interested in causality and expressions about cautious interpretation, for example, when they explore associations for developing diagnostic rules. However, only very few studies in our sample addressed such topics.

Underreporting of limitations may exaggerate conclusions and could sometimes be perceived as sensationalism, overall diminishing trust in research. We found no evidence that considering the possibility of confounding bias diminishes citation impact. This agrees also with recent evaluations of press releases of observational studies showing that cautious interpretations and wide media coverage are well compatible [26,27]. This is reassuring for researchers and may encourage them to discuss and communicate any limitation introduced by confounders in a thorough and determined way and "*not take them as mythical or uncontrollable phantoms that destroy studies*" [28].

Overall, we believe that there is a need to encourage researchers to report more careful and determined considerations of confounding bias and to encourage peer-reviewers, journal editors, and research funders to appreciate this. Many of the journals we analyzed have published the STROBE guideline, and some explicitly refer to them in their Instructions for Authors. Recently, PLOS Medicine intensified the requirements for authors of observational studies, asking that they "*must complete the appropriate reporting checklist not only with page references, but also with sufficient text excerpted from the manuscript to explain how they accomplished all applicable items*" [29]. Our results demonstrate that such activities are well justified. Given that not much has improved over many years, facing the tsunami of big datasets with all their promises, limitations, and risks of spurious findings [30], we believe that more concerted action is needed to improve the appropriate discussion of epidemiologic findings.

5. Conclusion

Confounding bias is a pervasive threat to the validity of observational epidemiologic research. Inadequate consideration and lack of discussion of implications of confounding bias are very frequent among the highest impact observational studies. Despite reasonable cause for careful discussion and cautious interpretation, authors often convey confidence, without cause or supporting evidence, that confounding bias is largely irrelevant for their findings. We think that such confidence is not justified.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2017.09.013.

References

[1] Rothman K, Greenland S, Lash T. Modern epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
[2] Guertin KA, Freedman ND, Loftfield E, Graubard BI, Caporaso NE, Sinha R. Coffee consumption and incidence of lung cancer in the NIH-AARP Diet and Health Study. Int J Epidemiol 2016;45:929−39.
[3] Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: recent evidence and a discussion of some questions. J Natl Cancer Inst 1959;22:173−203.

[4] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401−6.
[5] Morton SC, Costlow MR, Graff JS, Dubois RW. Standards and guidelines for observational studies: quality is in the eye of the beholder. J Clin Epidemiol 2016;71:3−10.
[6] Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. PLoS Med 2007;4(10):e297.
[7] von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335:806−8.
[8] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015;12(10):e1001885.
[9] Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report−Part I. Value Health 2009;12(8):1044−52.
[10] Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Rockville MD: Agency for Healthcare Research and Quality; 2013.
[11] Hatch EE, Wise LA, Mikkelsen EM, Christensen T, Riis AH, Sorensen HT, et al. Caffeinated beverage and soda consumption and time to pregnancy. Epidemiology 2012;23:393−401.
[12] Press DJ, Sullivan-Halley J, Ursin G, Deapen D, McDonald JA, Strom BL, et al. Breast cancer risk and ovariectomy, hysterectomy, and tubal sterilization in the women's contraceptive and reproductive experiences study. Am J Epidemiol 2011;173:38−47.
[13] Jackson LA, Yu O, Nelson JC, Dominguez C, Peterson D, Baxter R, et al. Injection site and risk of medically attended local reactions to acellular pertussis vaccine. Pediatrics 2011;127(3):e581−7.
[14] O'Reilly CE, Jaron P, Ochieng B, Nyaguara A, Tate JE, Parsons MB, et al. Risk factors for death among children less than 5 years old hospitalized with diarrhea in rural western Kenya, 2005-2007: a cohort study. PLoS Med 2012;9(7):e1001256.
[15] Ferguson LP, Durward A, Tibby SM. Relationship between arterial partial oxygen pressure after resuscitation from cardiac arrest and mortality in children. Circulation 2012;126:335−42.
[16] Niederkrotenthaler T, Rasmussen F, Mittendorfer-Rutz E. Perinatal conditions and parental age at birth as risk markers for subsequent suicide attempt and suicide: a population based case-control study. Eur J Epidemiol 2012;27(9):729−38.
[17] Coupland C, Dhiman P, Morriss R, Arthur A, Barton G, Hippisley-Cox J. Antidepressant use and risk of adverse outcomes in older people: population based cohort study. BMJ 2011;343:d4551.
[18] Cook AG, deVos AJ, Pereira G, Jardine A, Weinstein P. Use of a total traffic count metric to investigate the impact of roadways on asthma severity: a case-control study. Environ Health 2011;10:52.
[19] Montonen J, Drogan D, Joost HG, et al. Estimation of the contribution of biomarkers of different metabolic pathways to risk of type 2 diabetes. Eur J Epidemiol 2011;26(1):29−38.
[20] Mazumdar M, Bellinger DC, Gregas M, et al. Low-level environmental lead exposure in childhood and adult intellectual function: a follow-up study. Environ Health 2011;10:24.
[21] Lacson E Jr, Xu J, Suri RS, et al. Survival with three-times weekly in-center nocturnal versus conventional hemodialysis. J Am Soc Nephrol 2012;23(4):687−95.
[22] Groenwold RH, Van Deursen AM, Hoes AW, Hak E. Poor quality of reporting confounding bias in observational intervention studies: a systematic review. Ann Epidemiol 2008;18:746−51.
[23] Pouwels KB, Widyakusuma NN, Groenwold RH, Hak E. Quality of reporting of confounding remained suboptimal after the STROBE guideline. J Clin Epidemiol 2016;69:217−24.
[24] Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology 2011;22:42−52.
[25] Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantitative bias analysis. Int J Epidemiol 2014;43:1969−85.
[26] 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.
[27] 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:e0168217.
[28] Vandenbroucke JP. The history of confounding. Soz Praventivmed 2002;47(4):216−24.
[29] Plos Medicine Editors. Observational studies: getting clear about transparency. PLoS Med 2014;11(8):e1001711.
[30] Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Routinely collected data and comparative effectiveness evidence: promises and limitations. CMAJ 2016;188(8):E158−64.