PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf)** and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Selective citation in the literature on the hygiene hypothesis: a citation analysis on the association between infections and rhinitis
AUTHORS	Duyx, Bram; Urlings, Miriam; Swaen, Gerard; Bouter, Lex; Zeegers, Maurice

VERSION 1 – REVIEW

REVIEWER	Veronica Ivey Sawin Independent contractor (epidemiology), United States
REVIEW RETURNED	04-Oct-2018

GENERAL COMMENTS	• Please correct editorial errors, especially in abstract (i.e., "publications with supportive [results] were cited more often…, odd[s] ratio adjusted for study design…") • Please define variables more clearly in main report (e.g., specificity, authority, conclusive title, etc.). • Please be consistent in the use of "positive" vs. "supportive" findings. • Although some technical aspects of the hygiene hypothesis are important to provide as background, this dominates the introduction. The stated objective of this research is to determine whether citation bias exists in this field and to evaluate factors that may be associated with citation, rather than the validity of the hygiene hypothesis, though little is mentioned of the importance of citation bias in the introduction. • In the final paragraph of the introduction, the authors justify the selection of key variables for their analysis by their inclusion in another citation network. Is this the driving reason to examine these factors? It would be better to describe why these determinants are relevant to this research question. • Is the list of search terms for hay fever exhaustive? • The potential for an odds ratio to overestimate relative risk is a limitation and perhaps belongs in the discussion section. The explanation that citation is not a common outcome in this dataset is a finding and should not be included in the methods. Also, please specify a common outcome (i.e., >10%). • Please describe how the specificity of the publication was determined and the importance of this variable. Evaluation of specificity seems highly subjective. One might assume that if a publication presents findings that are relevant to subsequent publications, then it would not matter if other, less relevant findings are also presented. • Please describe how missing data were managed. In footnotes of several tables, authors specify variables included in models "if possible"—if these data were missing, were publications included in the analysis in question? It appears that the only statement on

	the treatment of missing values is buried in a paragraph on explained variance. Please specify n in these cases. • Although the finding that narrative reviews and editorials are far more likely to cite supportive studies is interesting, these types of publications are arguably less important in advancing scientific knowledge. Why include these types of publications in this analysis? • Throughout the paper, please emphasize what this research adds to what is already known about citation bias.

REVIEWER	Yujia Zhai Tianjin Normal University, China
REVIEW RETURNED	06-Oct-2018

GENERAL COMMENTS	Based on the health hypothesis proposed by Strachan, this study constructed a citation network and studied the selective citation in the literature. The selection and limitations of the data are described in detail in the article, but interpretation of the research method is still rather vague. In addition, the conclusion of the results does not explain the contribution of the research which needs to be further clarified. Detailed comments: 1 Research background This study focuses on the influencing factors of selective citation, but most part of the background is about the health hypothesis and its adaptation. The literature review does not involve selective citation studies at all. There is no clear explanation and foundation to raise these research questions for this study. Therefore, I suggest the author to strength the description and research review of the selective citation phenomenon. 2 The process of text analysis “The search output was then limited to publications that investigated exposures related to the original hygiene hypothesis.” Regarding the analysis of publications, did the researchers read the fulltext of 5,551 publications? The authors should specifically indicate the process of their analysis. How many people read and analyzed? Have the results been compared and validated? 3 Concept definition On page 6, the authors describe the difference between “cited” and “citing”, but it makes me even more confused. They need to add an example to explain what is called potential cited, potentially citing and potential citation path. What is the difference between the potential citation path and the real citation path? If A refers to B, then there is a citation path between A and B. What does “potential” and “realised” mean? 4 In the ‘Discussion’ section The section needs major revision: firstly, the results should be discussed in light of extant literature. Then, there is a need to absolute mention whether this paper helps advance theoretical understanding of the phenomena studied. Page 4, lines 49-50, "This latter step was data-driven because we realised during the analysis that prospective cohort studies were cited less often." What is the causal relationship of this sentence?

	Page 12, 33-34 “Surprisingly, publications with a conclusive title were less likely to receive citations.” What is the judgment of a conclusive title?

VERSION 1 – AUTHOR RESPONSE

Reviewers' Comments to Author

Reviewer: 1, Veronica Ivey Sawin

2. • Please correct editorial errors, especially in abstract (i.e., “publications with supportive [results] were cited more often…, odd[s] ratio adjusted for study design…”)

We thank the reviewer for pointing this out. The editorial errors have been corrected.

3. • Please define variables more clearly in main report (e.g., specificity, authority, conclusive title, etc.).

We decided to move a large part of the explanation of the variables from the supplementary text to the main document. We believe that the main document is now detailed enough to read without checking the supplement. However, some relatively unimportant details can still be found in the supplement.

The Methods section on data extraction consequently has been rewritten as follows:

Data extraction

A range of characteristics were extracted or derived from each included publication. These characteristics are described below and were all tested as determinant of citation in the statistical analysis.

Publication characteristics – content-related. The following variables were in this subcategory: type of exposure, publication type, sample size, specificity, and study outcome.

Type of exposure refers to the type of exposure that is being studied or reviewed: only number of siblings, only infection history, or both.

Publication type was classified into empirical and non-empirical publications. Empirical publications were further classified into the following *study designs*: cross-sectional, case-control, retrospective cohort, prospective cohort, and experimental studies. Non-empirical publications were further classified into: narrative reviews, systematic reviews, and other (editorials, leading articles, commentaries).

Sample size concerned the number of participants in the publications. Non-empirical publications had no sample size. The sample size of the empirical publications was classified into three equal categories based on tertiles.

The *specificity* of the publications varied. Some publications only deal with Strachan's hygiene hypothesis, others are broader. Specificity ranges from 1 (very broad) to 3 (very specific). For instance, an empirical publication that only investigates the association between number of siblings and rhinitis would be classified as '3'; if it also investigates the impact of heliminth infections and growing up on a farm, and if it also includes other health outcomes such as asthma or auto-immune diseases, it would be classified as '1'.

Study outcome was scored as follows: 1. supportive of the hygiene hypothesis; 2. mixed or unclear results; 3. non-supportive of the hygiene hypothesis. An inverse relationship between past exposure and rhinitis is considered to be supportive for the hygiene hypothesis, while a neutral or positive relationship was scored as non-supportive. The scoring was based on the authors' interpretation of the results, as it was stated in the text of the publication. (See also Text S2 for more details.)

Publication characteristics - not content-related. The following variables were in this category: conclusiveness of the title, funding source, number of authors, number of affiliations, and number of references. *Title conclusiveness* was coded as yes if in the title a conclusion was stated that included the direction of the relationship (e.g. "Inverse relation between infections and allergies"), otherwise as no (e.g. "Infections, rhinitis, and their relationship"). *Funding source* was coded as non-profit (e.g. government or university), for-profit, both, or not reported.

Journal characteristics. The following variables were in this category: publisher and journal impact factor. *Journal impact factor*, in the publication year of the cited publication, was retrieved from the Journal Citation Reports (JCR) database. *Journal publisher* was also retrieved from JCR.

Author characteristics. The following variables were in this category: *gender* of the corresponding author (see also Text S2), *country* of the corresponding author, and affiliation of the corresponding author. *Affiliation* was classified as government, university, industry or other.

Citation characteristics. There were some variables that depend on the cited publication as well as the citing publication: self-citation and within-network authority. A *self-citation* was defined as a citation between two publications that have at least one author in common.

Authority was a measure for the authority of the authors within the network. It was calculated for each author and each year separately, by counting the number of within-network citations to all publications in which the author had been involved. As the number of citations is likely to increase each year, so does the author's authority. Because we were interested in the authority at the moment of citation, the authority value of a cited publication also depends on the publication year of the citing publication. In case of multiple authors, we used the authority value of the author with the highest authority in that year.

4. • Please be consistent in the use of "positive" vs. "supportive" findings.

We thank the reviewer for pointing this out. Our use of 'positive' and 'supportive' was indeed somewhat inconsistent and confusing. Positive results should only refer to the direction of an association, and supportive results should only refer to results that support the hypothesis. (In the case of the hygiene hypothesis, this implies that negative results are supportive, making our inconsistency even more confusing.) We have changed the wording accordingly.

5. • Although some technical aspects of the hygiene hypothesis are important to provide as background, this dominates the introduction. The stated objective of this research is to determine whether citation bias exists in this field and to evaluate factors that may be associated with citation, rather than the validity of the hygiene hypothesis, though little is mentioned of the importance of citation bias in the introduction.

We agree with the reviewer that, in retrospect, the introduction had too much focus on the hygiene hypothesis itself. We deleted some paragraphs with detailed background information, and added a paragraph on the relevance of citation bias:

> The number of publications in the research on the hygiene hypothesis is large. It is therefore not feasible for authors to cite every relevant publication in the network and some kind of selection needs to take place. If this selection is based on study outcome, we speak of citation bias (3, 7). The consequences of citation bias can be similar to those of publication bias and reporting bias: disregard of counter-evidence leading to unfounded consensus (8) or polarisation (9), ill-advised research programmes and research waste (8, 10), distorted information in the media (11), and misguided medical decisions (12). Citation bias has been studied in many disciplines. Our systematic review gives an overview of these studies (13) . Many of these studies showed evidence for citation bias in their field, with supportive publications being cited about twice as often as non-supportive ones.

6. • In the final paragraph of the introduction, the authors justify the selection of key variables for their analysis by their inclusion in another citation network. Is this the driving reason to examine these factors? It would be better to describe why these determinants are relevant to this research question.

The variables included in our research design had been shown to be related to citation in previous studies, or have otherwise often been studied. The rationale for their inclusion was not clearly explained in the previous version of our paper, and we added a paragraph to fill this gap.

> Factors other than study outcome may also have an impact on citation, as was recently shown by Onodera and Yoshikane (14). Measures for journal status (impact factor), author status (number of citations, country of affiliation), and collaboration (number of authors, number of affiliatons) were often found to be related to citation count. The same was consistently found for the number of references of the cited publication. Furthermore, the reporting (15) and source (16, 17) of funding were shown to be related to citation, but the impact of author's affiliation (18) and gender (19-21) is less clear. On the other hand, sample size and study design - both markers of study quality, and as such legimate reasons to base a citation on - often seem unrelated to citation (17, 18, 22-24). In our previous citation networks, we also found associations with self-citation and the specificity of a publication, but not with the title of a publication (25, 26).

7. • Is the list of search terms for hay fever exhaustive?

Based on the reviewer's comment, we double-checked if any words or synonyms for hay fever were missing in our search query. We checked MESH terminology, keywords of included articles and performed a general search on the internet, but could not identify any words that may have been missing. Therefore, we indeed believe that the list of search terms for hay fever is exhaustive.

8. • The potential for an odds ratio to overestimate relative risk is a limitation and perhaps belongs in the discussion section. The explanation that citation is not a common outcome in this dataset is a finding and should not be included in the methods. Also, please specify a common outcome (i.e., >10%).

We followed up on these suggestions, changed the wording of the paragraph and moved it to the Discussion:

> Another limitation is our use of odds ratios to assess the likelihood of citation. The odds ratio may overestimate the true relative risk in studies where the outcome is common (i.e. occurs in more than 5% of all cases, (31)). In our network, citation is not a common outcome (7%) and consequently the overestimation of the true relative risk will be relatively small.

9. • Please describe how the specificity of the publication was determined and the importance of this variable. Evaluation of specificity seems highly subjective. One might assume that if a publication presents findings that are relevant to subsequent publications, then it would not matter if other, less relevant findings are also presented.

The reviewer asks how our variable for specificity was determined, and wonders whether this variable is important. Specificity is determined by the number of predictors and health outcomes that are not directly relevant for the original hygiene hypothesis. The determination is indeed, as the reviewer suggests, somewhat subjective, but it was scored by two raters who had a high agreement. This description was added to the Methods section:

> The *specificity* of the publications varied. Some publications only deal with Strachan's hygiene hypothesis, others are broader. Specificity ranges from 1 (very broad) to 3 (very specific). For instance, an empirical publication that only investigates the association between number of siblings and rhinitis would be classified as '3'; if it also investigates the impact of helminth infections and growing up on a farm, and if it also includes other health outcomes such as asthma or auto-immune diseases, it would be classified as '1'.

The reviewer further points out that specificity should not have an influence on citation; after all, all publications that are included in our network generate (or synthesise) evidence that is relevant for the original hygiene hypothesis, and this is independent of how many other variables were included in their studies. This is a valid point. Nevertheless, we believe that the importance of this variable is suggested by the results of our analysis. The most specific publications are much more likely to be cited than the least specific publications (odds ratio adjusted for study design: 5.0, 95% confidence interval: 3.1-7.9). In fact, one might argue that specificity behaves in exactly the same way as most of the other publication characteristics, showing an impact on citation where none is

warranted. (We believe that an impact on citation is only warranted in the case of study design and sample size.)

10. • Please describe how missing data were managed. In footnotes of several tables, authors specify variables included in models “if possible”—if these data were missing, were publications included in the analysis in question? It appears that the only statement on the treatment of missing values is buried in a paragraph on explained variance. Please specify n in these cases.

Missing data were managed by excluding them from the analyses. This is always stated in the tables. For instance, we could not retrieve the gender for three of the corresponding authors, and there were 159 potential citations to these three publications, leaving us 5392 instead of 5551 potential citations in our analysis on the impact of gender. We have stated it in the following way in the tables: ‘Gender (female vs male, n = 5392)’.

We had added ‘if possible’ to ‘adjustment for study design and log sample size’ because some determinants, such as sample size, were only adjusted for study design. This text was probably confusing and we removed it. (It is still stated in the tables that sample size is ‘only adjusted for study design’.)

11. • Although the finding that narrative reviews and editorials are far more likely to cite supportive studies is interesting, these types of publications are arguably less important in advancing scientific knowledge. Why include these types of publications in this analysis?

Although narrative reviews and editorials may indeed be less important for the generation of evidence, they could impact the development of knowledge in a different way. Greenberg (2009), for instance, has argued that reviews play an important role in the propagation and amplification of evidence accumulated by empirical studies, potentially leading to unfounded belief systems in the case of citation bias. Reviews draw attention to topics and specific studies, and they may be useful in identifying remaining research questions in a certain field. These types of articles are often read by people who are new to the field, be it students, policy-makers, laymen or researchers from another research field. Also, narrative reviews and editorials may be used by authors to fill a gap in their reasoning, or to identify the most important literature, especially if it concerns a research field they are less familiar with. Our findings confirm and corroborate the notion that people should be wary by relying on narrative reviews. (see also 17. below)

12. • Throughout the paper, please emphasize what this research adds to what is already known about citation bias.

See the revision in point 17.

Reviewer 2: Yujia Zhai

13.

Based on the health hypothesis proposed by Strachan, this study constructed a citation network and studied the selective citation in the literature. The selection and limitations of the data are described in detail in the article, but interpretation of the research method is still rather vague. In addition, the conclusion of the results does not explain the contribution of the research which needs to be further clarified.

We agree with the reviewer that the method on data extraction was not very clear. It has been extended (see point 3. above). The discussion has also been amended (see point 17, below).

Detailed comments:

14: Research background

This study focuses on the influencing factors of selective citation, but most part of the background is about the health hypothesis and its adaptation. The literature review does not involve selective citation studies at all. There is no clear explanation and foundation to raise these research questions for this study. Therefore, I suggest the author to strength the description and research review of the selective citation phenomenon.

We agree with the reviewer that the introduction was too much focused on the hygiene hypothesis. We removed some of the more detailed background information, and elaborated instead on the relevance of citation bias and previous research on selective citation. See also points 5 and 6 above.

15: The process of text analysis

“The search output was then limited to publications that investigated exposures related to the original hygiene hypothesis.” Regarding the analysis of publications, did the researchers read the fulltext of 5,551 publications? The authors should specifically indicate the process of their analysis. How many people read and analyzed? Have the results been compared and validated?

The article selection was performed by two of the authors, in three steps. In the last step of the article selection, 110 articles were selected from 196 full-text articles that were all read by both authors. (See also Figure 1.) Data extraction of the 110 selected articles was performed by the same two authors Results were compared and consensus was always reached. The description in the Methods has been extended:

> Article selection (first based on title, then on abstract and finally on full-text; Figure 1) and data extraction were performed independently by MJEU and BD. Results were compared after each step, and disagreements were resolved in consensus meetings.

More importantly, we realise that our previous description of publications and citations was not unambiguously clear. There were 110 publications in our network, with 5551 potential citations between these 110 publications.

METHODS: For clarification: a publication in our network can both cite and be cited by other publications *in the network*, leading to a multitude of citation paths.

RESULTS: This network of 110 publications had a total of 5551 potential and 392 actual citation paths (7%) *between these publications*.

16: Concept definition

On page 6, the authors describe the difference between “cited” and “citing”, but it makes me even more confused. They need to add an example to explain what is called potential cited, potentially citing and potential citation path.

What is the difference between the potential citation path and the real citation path? If A refers to B, then there is a citation path between A and B. What does “potential” and “realised” mean?

Our description of the citation relationships was indeed somewhat ambiguous. We only looked at citations within the network of 110 publications. So we first assessed which citations within the network were possible (a potential citation), and then whether these citations had occurred (a realised or actual citation). This implies that a publication in our network can serve two roles: it can be cited and it can cite itself. We believe that the descriptions in the Methods and Results are now clearer:

For clarification: a publication in our network can both cite and be cited by other publications in the network, leading to a multitude of citation paths. Not all citation paths are possible as one can only cite articles that were published before. In our study, a citation is considered possible if the cited publication is published before the citing publication is submitted. If such *potential* citation occurred, we call it an *actual* citation. (See also supplementary Text S2.)

Our binary dependent variable was citation *within the network* (or, more precisely, whether a potential citation had occurred or not).

17: In the ‘Discussion’ section

The section needs major revision: firstly, the results should be discussed in light of extant literature. Then, there is a need to absolute mention whether this paper helps advance theoretical understanding of the phenomena studied.

The reviewer raises a valid point. The aim of our research is mostly descriptive rather than advancing theoretical understanding. We believe that this is of value in itself, since the phenomenon that we describe has been shown to be disruptive and harmful. Still, the reviewer is right to point out that our findings should be discussed in the light of the existing evidence, and we revised this section accordingly.

With regard to study outcome, supportive publications are cited more than three times more often than non-supportive publications, while publications with mixed results are cited more than two times as often. This is a clear sign of citation bias, and corroborates previous findings (13). Similarly, publications are more likely to refer to other publications with the same study outcome rather than to those that provide counter-evidence to their conclusion. This type of citation bias (based on concordance) has not been studied frequently. In our previous network analyses, on trans fatty acids – cholesterol, and on chlorinated water – asthma, we found no evidence for increased citations between publications with the same study outcomes (25, 26), but three other studies, all related to cardiovascular disease, did find evidence for

this type of citation bias (9, 29, 30).

The magnitude of citation bias even increases if we focus on how empirical publications are cited by reviews and editorials. Reviews and editorials in our network are up to 8 times more likely to cite supportive publications rather than non-supportive ones. As reviews are generally assumed to give an unbiased summary of the existing evidence, this is a worrying finding. It confirms the notion that people should be cautious to rely on narrative reviews.

Greenberg states that reviews play an important role in the development and acceptance of belief systems (8). According to him, reviews can amplify the impact of empirical studies because their evidence is propagated when these reviews are cited themselves. Trinquart et al. showed that reviews (including systematic reviews) on the health impact of salt intake display signs of citation bias, and that the conclusions of these reviews are in the same direction as the evidence they include (9). A similar link between the selective citation to supportive evidence and supportive conclusions of reviews was found by Leng (29). This mechanism might explain how reviews can amplify the effect of citation bias. If reviews draw supportive conclusions based on selective citation of supportive evidence, then support for a hypothesis will be propagated while counter-evidence will fade from the literature.

18

Page 4, lines 49-50, "This latter step was data-driven because we realised during the analysis that prospective cohort studies were cited less often." What is the causal relationship of this sentence?

First, during exploration of the data, we saw that prospective and retrospective cohort studies behaved differently with regard to citations. Therefore, we decided to take this difference into account in our analyses, with study design as determinant and as covariate. The following phrase is added to the Supplement (Text S1):

> During exploration of the data, we noticed a big different in the citation behaviour of retrospective and prospective cohorts studies. We therefore decided to amend our preregistered data analysis plan (http://hdl.handle.net/10411/ZKGGOG). We differentiated between these research designs (by assigning them a different categorical value), and took this difference into account in our analyses with study design as determinant or as covariate.

19

Page 12, 33-34 “Surprisingly, publications with a conclusive title were less likely to receive citations.” What is the judgment of a conclusive title?

This is a good question that might shed light on our results. But there were 11 publications with a conclusive title, 7 of which were supportive of the hygiene hypothesis. These numbers are too low to conduct stratified analyses.

VERSION 2 – REVIEW

REVIEWER	Veronica Ivey Sawin Independent contractor (epidemiology), United States
REVIEW RETURNED	16-Nov-2018

GENERAL COMMENTS	The authors addressed my comments on the prior draft.

REVIEWER	Yujia Zhai Tianjin Normal University, China
REVIEW RETURNED	16-Nov-2018

GENERAL COMMENTS	In the Author characteristics section, the author determines the gender through the first name and www.genderchecker.com. What is the accuracy rate? Please do a random sample test to verify that this method is indeed accurate. The author has answered the questions I asked before. In this version, I have no other questions besides the above uncertainty that needs to be supplemented.

VERSION 2 – AUTHOR RESPONSE

Here we address the last remaining issue, with regards to the validity of our gender assessment. As suggested by the reviewer, we preformed a validity check on a random sample (based on a random number generator), and checked our original assessment against that of another online tool for gender assessment: https://gender-api.com/. The following text was added to Text S2 in the supplementary file:

We performed a validity check of the gender assessment on a random sample of 20 publications. We checked our original assessment (described above) against the results from another gender assessment tool: Gender-API.com. This tool takes into account the person's country while assessing gender based on his or her first name Additionally, it gives an accuracy score for each assessment. The results can be found in the Table below. Our reference assessment reached the same results as the original one, with a 100% accuracy.

Table. Validity check for gender assessment on random sample (N=20).

ID [1]	First Name of Corresponding Author	Country of Corresponding Author	Gendercheck er Assessment	Our Data-extraction Assessme nt	Gender-API Assessment [2] (validity check)	Accuracy of Gender-API Assessme nt [2]
1	David	UK	male	male	male	99%

5	Barbara	UK	female	female	female	98%
13	Nick [3]	UK	male	male	male	98%
14	Sarah [3]	UK	female	female	female	98%
18	Juha	Finland	unisex	male [4]	male	100%
24	Anthony	UK	male	male	male	99%
25	Erika	Germany	female	female	female	98%
26	Mustafa	Turkey	male	male	male	100%
31	Johannes	Germany	male	male	male	99%
35	Paolo	Italy	male	male	male	99%
73	Anne-Louise [5]	Australia	no match	female	female	100%
	Anne		unisex			
	Louise		female			
75	Keiko	Japan	female	female	female	99%
78	Aarif	Turkey	male	male	male	75%
79	Sharad	India	male	male	male	100%
87	Woei Kang [5]	Singapore	no match	male		
	Woei		male		male	67%
	Kang		unisex		male	88%
94	Jonathan	USA	male	male	male	99%
95	Ahmet	Turkey	male	male	male	100%
97	Chun-Yuh [5]	Taiwan	no match	male [4]		
	Chun		unisex		male	53%
	Yuh		no match		male	60%
103	David	UK	male	male	male	99%
109	Katherine	USA	female	female	female	99%

Notes. 1. See Text S3 for the references. 2. Based on combination of first name and country. 3. First name was not stated in publication, but retrieved via ResearchGate.net; match between profile and correpsonding author based on surname, initials, affiliation and research topic. 4. Web search revealed

a man with same name and afflilation. 5. Composite names that could not be assessed as a whole were assessed by its composites.