Journal of Clinical Epidemiology 121 (2020) 55−61

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice

Bas B.L. Penning de Vries[a,*], Maarten van Smeden[a], Frits R. Rosendaal[a], Rolf H.H. Groenwold[a,b]

[a]*Department of Clinical Epidemiology, Leiden University Medical Center, PO Box 9600, 2300 RC, the Netherlands*
[b]*Department of Biomedical Data Sciences, Leiden University Medical Center, PO Box 9600, 2300 RC, the Netherlands*

Accepted 18 January 2020; Published online 23 January 2020

Abstract

Objective: Article full texts are often inaccessible via the standard search engines of biomedical literature, such as PubMed and Embase, which are commonly used for systematic reviews. Excluding the full-text bodies from a literature search may result in a small or selective subset of articles being included in the review because of the limited information that is available in only title, abstract, and keywords. This article describes a comparison of search strategies based on a systematic literature review of all articles published in 5 top-ranked epidemiology journals between 2000 and 2017.

Study Design and Setting: Based on a text-mining approach, we studied how nine different methodological topics were mentioned across text fields (title, abstract, keywords, and text body). The following methodological topics were studied: propensity score methods, inverse probability weighting, marginal structural modeling, multiple imputation, Kaplan-Meier estimation, number needed to treat, measurement error, randomized controlled trial, and latent class analysis.

Results: In total, 31,641 Hypertext Markup Language (HTML) files were downloaded from the journals' websites. For all methodological topics and journals, at most 50% of articles with a mention of a topic in the text body also mentioned the topic in the title, abstract, or keywords. For several topics, a gradual decrease over calendar time was observed of reporting in the title, abstract, or keywords.

Conclusion: Literature searches based on title, abstract, and keywords alone may not be sufficiently sensitive for studies of epidemiological research practice. This study also illustrates the potential value of full-text literature searches, provided there is accessibility of full-text bodies for literature searches. © 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Keywords: Systematic literature review; Bibliometrics; Text mining; Statistical methods; Epidemiological methods

1. Introduction

Rigorous reviews of the scientific literature are essential for determining the current state of knowledge on a specific topic, to identify research areas where evidence is lacking, and as a starting point for guidance development. Although a majority of systematic reviews in epidemiology represents reviews of research findings on a specific substantive medical research topic, such as the occurrence of a particular disease or the effectiveness of a medical treatment, an important category of systematic reviews is concerned primarily with epidemiological research practice and reporting [1−5].

A variety of strategies exist to identify and screen articles for eligibility for systematic reviews [6−9]. Often, a staged search and screening approach is implemented in which the eligibility criteria for articles are made more stringent or more text fields are scrutinized with each step. In the earlier steps of the process, articles are typically excluded from the review on the basis of a small portion—for example, title, abstract, and keywords (TIABKW)—of all the available information. The goal of a search and screening approach is to identify all or a representative sample of the relevant literature on the topic of inquiry. However, excluding a selective set of articles from further study may ultimately result in a false impression of state of the literature being conveyed [7,9,10].

Conflicts of interest: There are no financial, personal, political, academic, or other relations that could lead to a conflict of interest.

* Corresponding author. Department of Clinical Epidemiology, Leiden University Medical Center, PO Box 9600, 2300 RC, the Netherlands. Tel.: +31 (0)71 526 5639; fax: +31 (0)71 526 6994.

E-mail address: B.B.L.Penning_de_Vries@lumc.nl (B.B.L. Penning de Vries).

https://doi.org/10.1016/j.jclinepi.2020.01.009
0895-4356/© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

What is new?

Key findings

- For various methodological topics, at most 50% of articles in which the topic was mentioned in the full text also mentioned the topic in the title, abstract, or keywords.

What this adds to what was known?

- A systematic literature search preferably identifies all or a representative sample of the relevant literature on the topic of inquiry. Title, abstract, and keywords are commonly used as a screening instrument for a systematic review. We found that these text fields often do not identify studies in which one of the methodological topics was mentioned.

What is the implication and what should change now?

- This bibliographic study highlights that literature searches based on title, abstract, and keywords alone may not be sufficiently sensitive for studies of epidemiological research practice.
- This study also illustrates the potential value of full-text literature searches.

Reviews of methods often begin searching for relevant literature in the same way as reviews on a substantive research topic. However, compared with substantive topics, the epidemiological and statistical methods used are likely less well documented in the small portion of information that is typically accessed in the first stage(s) of a systematic literature search, notably TIABKW. In this article, we investigate whether the traditional approach to systematic literature searching is appropriate for reviews of epidemiological practice.

2. Methods

We identified and downloaded all articles (in Hypertext Markup Language [HTML] format) published in the period 2000−2017 on the websites of five top-ranked epidemiological journals; Epidemiology (EPI), Journal of Clinical Epidemiology (JCE), European Journal of Epidemiology (EJE), International Journal of Epidemiology (IJE), and American Journal of Epidemiology (AJE).

All retrieved HTML pages were analyzed with R Statistical Software, version 3.5.1 [11]. First, we sought to extract for each article its publication date, title, abstract, keywords, and text body, in a largely automated fashion using R base regular expression algorithms (see e.g., Crawley [12], or Supplementary R Code). In-text references and reference lists were removed from the text bodies before analysis. The following methodological topics were selected for investigation: propensity score (PS) methods, inverse probability weighting (IPW), marginal structural modeling (MSM), multiple imputation (MI), Kaplan-Meier (KM) estimation, number needed to treat, measurement error, randomized controlled trial (RCT), and latent class (LC) analysis. This set of topics reflects a range of classical and modern methodological topics relevant to epidemiologic research. We subsequently determined for each of these topics whether there was any mention of the topic (see Supplementary Material for details on the search terms) and in which text field (title, abstract, keywords, and text body).

The results of the previous step were used to quantify sensitivities of fixed combinations of text fields for identifying a mention of the method in any of the article's text fields (title, abstract, keywords, or text body). For any fixed topic, we refer to the sensitivity of a particular combination of text fields (e.g., TIABKW) as the fraction of articles with a mention of the topic in any of these text fields among articles with a mention in the full text (i.e., in the title, abstract, keywords, or body). We computed sensitivities stratified by journal and by publication date (year of publication). In a sensitivity analysis, the set of articles was limited to those articles containing at least 2500 words with the aim of focusing on original research articles. In addition, we examined all articles with a mention of PS methods to determine the article type and whether or not the article described an empirical application of PS methods. Finally, we performed a post hoc analysis, designed to ignore 'irrelevant' topic mentions (e.g., mention of a topic in the introduction or discussion of an article only). In this analysis, we considered only topics mentioned in the Methods Section 2 and Results Section 3 sections, provided these sections could be readily identified. Sensitivities pertaining to this post hoc analysis are understood to refer to the fraction of articles with a mention of the topic in any of a given set of text fields among articles with a mention in the title, abstract, keywords, methods, or results text fields.

3. Results

We downloaded 31,641 HTML files from the journals' websites; 10,580 from EPI, 4,187 from JCE, 2,251 from EJE, 6,249 from IJE, and 8,374 from AJE. These files include (but are not limited to) what is published in HTML format of (indexed) articles, issue index pages, and conference abstracts. Here, we present results based on those 31,641 files. In Supplementary Material, results are presented on the subset of publications with at least 2500 words, for which results are comparable with what is presented here (Supplementary Figures 1 and 2).

Figures 1 and 2 present the sensitivities of TIABKW stratified by journal and by publication date, respectively. At most 50% of articles with a topic mention in any text field had a mention in the title, abstract, or keywords. Figures 3 and 4 depict the results for our post hoc analysis. For some topics (e.g., PS, MSM, and RCT), TIABKW mentions were considerably more sensitive for a topic mention in the full text excluding rather than including introduction and discussion. For other topics (e.g., MI, KM, and LC), TIABKW identified fewer than half the number of articles with a topic mention anywhere in the full text, regardless of whether introduction and discussion were excluded. Some methodological topics had a constant, low sensitivity throughout the study period (e.g., KM), whereas the sensitivity of TIABKW for the other topics gradually declined over time (e.g., MI, PS, IPW). There were no relevant differences in sensitivities of the reporting of topics across the different journals.

Focusing on the articles that mention PS in the full text, 247 of 378 articles mentioned PS in the text body but not in the title, abstract, or keywords. Almost a third (72/247, 29%) of these described an empirical application of the method. This rate was more than doubled after we selected only those articles that, based on the nature of their main conclusion, were deemed predominantly applied research (60/87, 69%). Of the 131 articles that mentioned PS in the title, abstract, or keywords, 82 (63%) described an empirical application. The positive predictive value of TIABKW for an empirical application was higher among predominantly empirical/applied original articles (58/60, 97%).

4. Discussion

Search engines that limit the searching of scientific articles to TIABKW, such as PubMed or Embase, are established starting points for systematic reviews of substantive epidemiological study questions (e.g., systematic reviews of the effects of a medical treatment). Our study illustrates that in systematic reviews of research

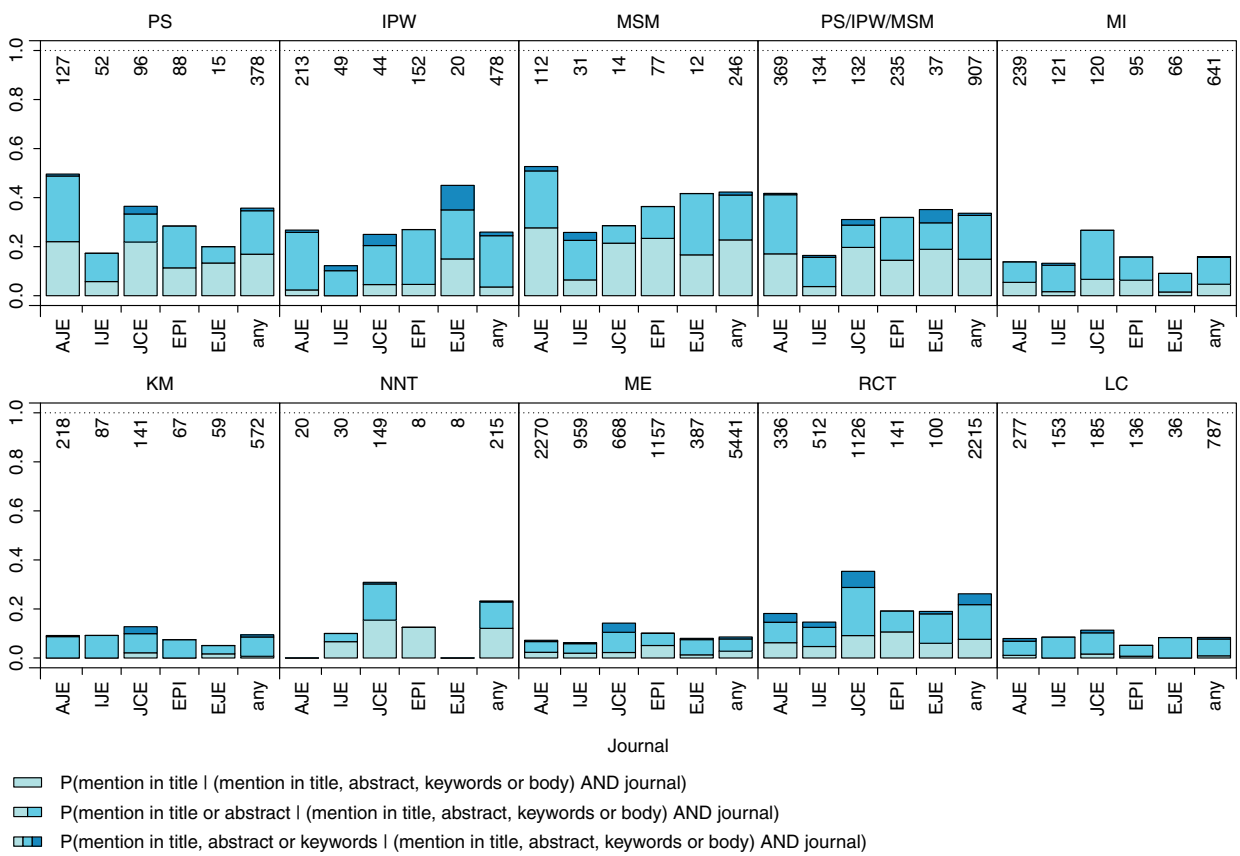

Fig. 1. Sensitivities of topic mentioning in various text fields stratified by journal. Colors relate to text fields as follows: light blue areas give the proportion of articles with a topic mention in the title among all articles published in the indicated journal with a mention in the title, abstract, keywords, or body; light blue and blue areas together give the proportion of articles with a topic mention in the title or abstract; and light blue, blue, and dark blue areas together give the proportion of articles with a topic mention in the title, abstract, or keywords. PS, propensity score; IPW, inverse probability weighting; MSM, marginal structural modeling; MI, multiple imputation; KM, Kaplan-Meier; NNT, number needed to treat; ME, measurement error; RCT, randomized controlled trial; LC, latent class; AJE, American Journal of Epidemiology; IJE, International Journal of Epidemiology; JCE, Journal of Clinical Epidemiology; EPI, Epidemiology; EJE, European Journal of Epidemiology. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

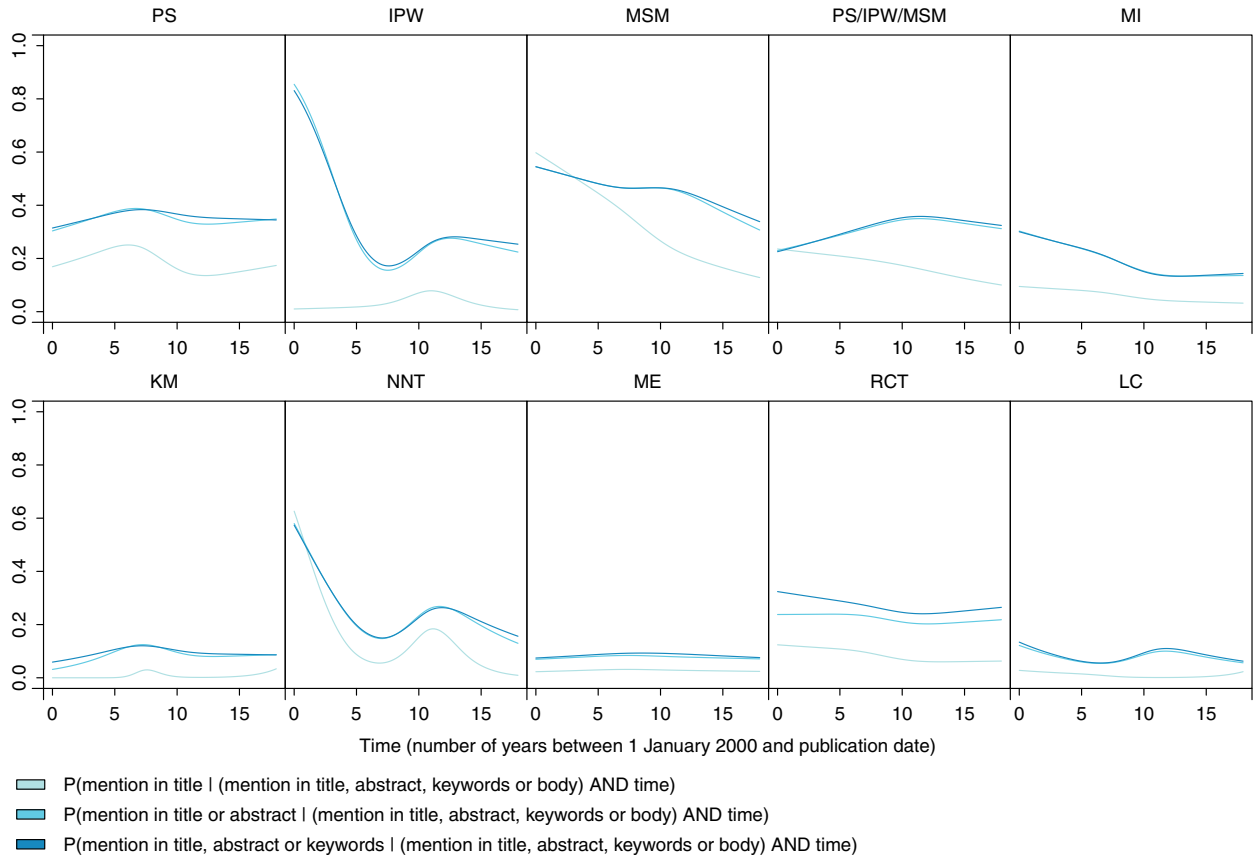

Fig. 2. Sensitivities of topic mentioning in various text fields over time. Bullets give year-specific sensitivities with bullet size being proportional to number of publications in the given year with a mention of the topic in any text field (title, abstract, keywords, or body). Solid lines reflect logistic regression fits with cubic spline transformations of publication date with four knots placed equidistantly within [2000, 2017]. Colors relate to text fields as follows: for any given journal, light blue lines give the year-specific sensitivities of a topic mention in the title for a mention in the title, abstract, keywords, or body; blue lines indicate the year-specific sensitivities of a topic mention in the title or abstract; and dark blue areas give the year-specific sensitivities of a topic mention in the title, abstract, or keywords. PS, propensity score; IPW, inverse probability weighting; MSM, marginal structural modeling; MI, multiple imputation; KM, Kaplan-Meier; NNT, number needed to treat; ME, measurement error; RCT, randomized controlled trial; LC, latent class. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

practice and reporting, searches that rely only on these tools may lead to a small and possibly selective subset of articles to be included in the review. We found a large discrepancy in terms of the number of articles identified (as potentially eligible) between searches that include text bodies and those that are restricted to TIABKW. Moreover, methodological topics tended to be documented in less detail in the title, abstract, or keywords as methods become more mainstream, contributing to a possibly selective subset of articles to be identified over time.

Reviewers are faced with the challenge of adequately handling increasingly large volumes of literature, and ignoring certain text fields may help mitigate this problem, but it may come at the cost of giving an inaccurate reflection of the state of knowledge/practice on the topic of interest. The decision to automate the selection of articles in systematic reviews using readily available search engines is usually made on practical grounds. Full-text mining may however be a promising alternative. As noted by O'Mara-Eves et al. [7], there are at least two (not necessarily distinct) ways of using data and text mining in selecting articles for further review: by reducing the list of items to be screened manually or by manually assigning articles in a (development) subset of articles to include/exclude categories to 'train' an algorithm to apply such categorizations automatically. Depending on the complexity of the task for which the algorithm is to be trained and the desired properties the trained algorithm should possess, the second (supervised-learning) approach may actually be more cumbersome than going through all articles manually. For the current analysis, we used text mining only to prune articles that would be deemed related to the topic of interest had we manually evaluated the paper. In some settings, for example, where diverse or nonspecific terminology is used, it may be difficult to find a rule that allows for relevant articles to be identified with high sensitivity and manageable specificity. In such cases, the adopted text mining approach may still leave an intractably large amount of articles to screen manually.

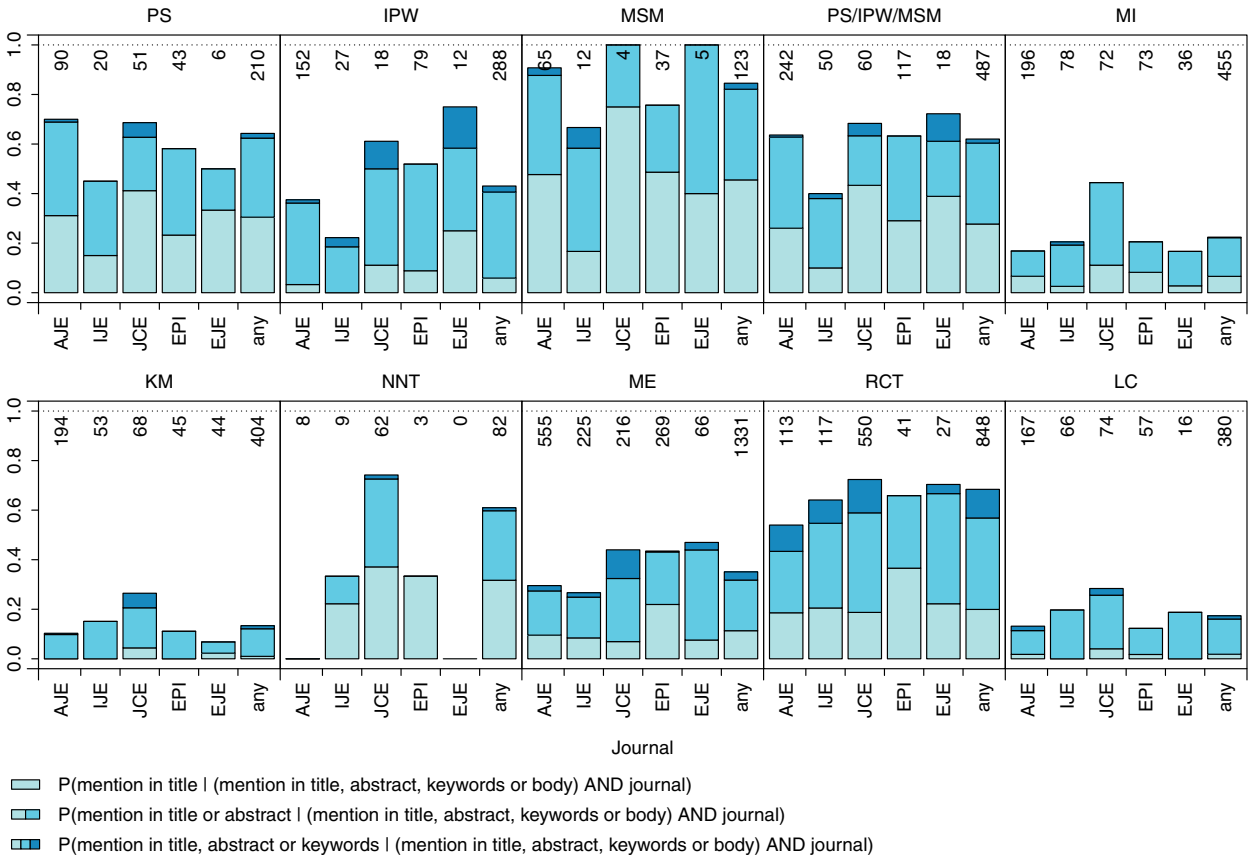

Fig. 3. Sensitivities of topic mentioning in various text fields stratified by journal, according to post hoc analysis. Colors relate to text fields as follows: light blue areas give the proportion of articles with a topic mention in the title among all articles published in the indicated journal with a mention in the title, abstract, keywords, methods, or results text fields; light blue and blue areas together give the proportion of articles with a topic mention in the title or abstract; and light blue, blue, and dark blue areas together give the proportion of articles with a topic mention in the title, abstract, or keywords. PS, propensity score; IPW, inverse probability weighting; MSM, marginal structural modeling; MI, multiple imputation; KM, Kaplan-Meier; NNT, number needed to treat; ME, measurement error; RCT, randomized controlled trial; LC, latent class. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Although our review clearly shows a possibly large difference between TIABKW searching versus full-text searching, the discrepancies we found in this review in the number of pruned articles need not always translate into the two approaches, giving a different impression of the state of research practice for any given methodological topic in epidemiology. This may depend on the review goals. Also, even if articles are missed by limiting the research to TIABKW, an important question remains whether the articles that would be omitted if we ignored the full text should have actually been included. The large discrepancy that we found for the topic RCT, for example, is likely largely explained by many articles only briefly addressing the study design in the discussion or introduction, that is, studies that may not be relevant to the reviewer (depending on the review goals) (see Fig. 3). That is, incorporating all available text fields in the screening is likely to decrease the specificity for relevant articles, resulting in a possibly much larger number of articles to be further screened on relevance. It may therefore sometimes be appropriate to restrict oneself to certain text fields. Of note, for the topic of PS, many studies that would be omitted by restricting the search to TIABKW actually detailed an empirical application of the method. Therefore, for reviews of research practice regarding PS, many relevant articles would be missed if the search/screening had been restricted to TIABKW only, especially the more recently published articles.

A limitation of this study is that it was limited to only five high-ranking epidemiological journals and nine (partly related) methodological topics. Each of these journals has a strong methodological focus, publishing on applied as well as methodological topics. Consequently, we may expect that our results do not directly translate to other fields, particularly to applied biomedical journals with a less methodological focus.

There are several operational and legal challenges to consider for full automated text data literature searches. Clearly, if researchers do not have access to the full text of articles, initial screening based on title and abstract

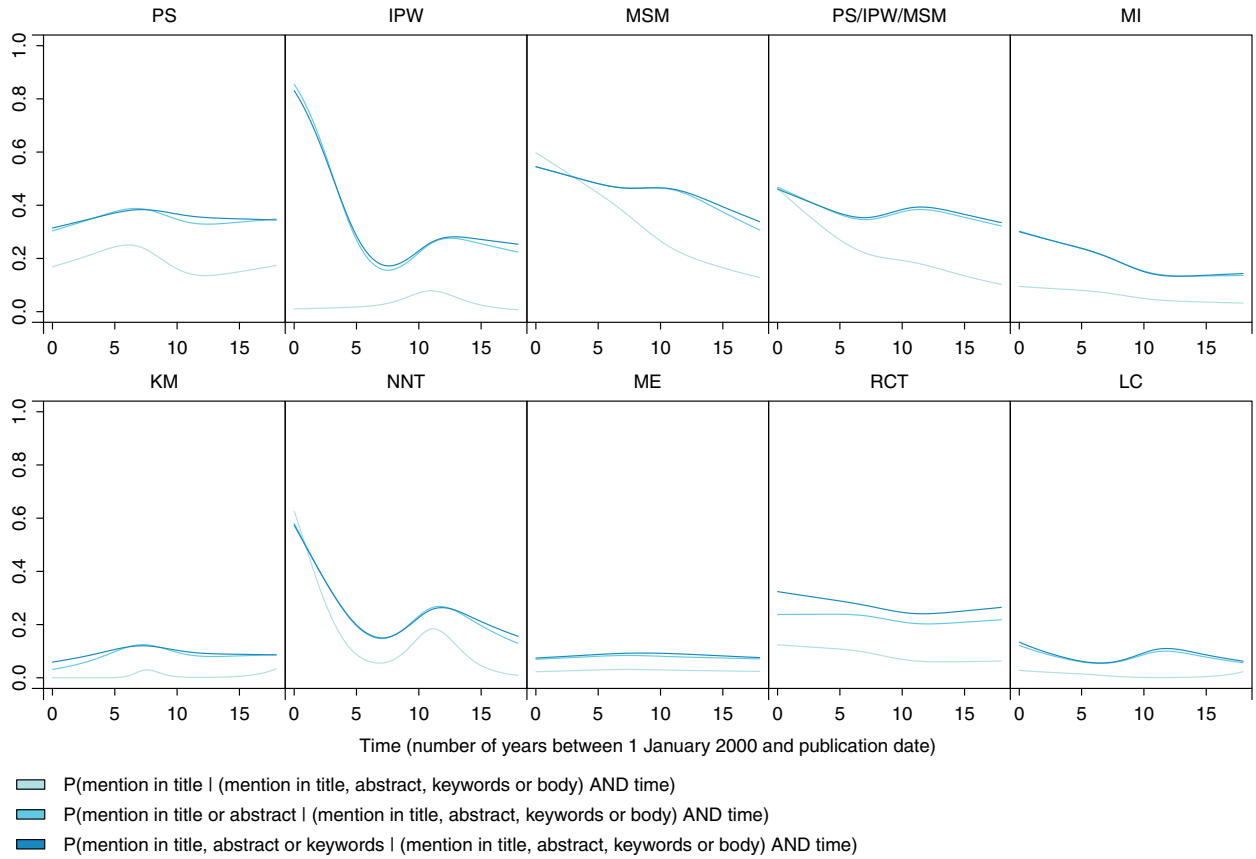

Fig. 4. Sensitivities of topic mentioning in various text fields over time, according to post hoc analysis. Bullets give year-specific sensitivities for a mention in the title, abstract, keywords, methods, or results text fields, with bullet size being proportional to number of publications in the given year with a mention of the topic in title, abstract, keywords, or methods or results (provided the text field was identified and extracted). Solid lines reflect logistic regression fits with cubic spline transformations of publication date with four knots placed equidistantly within [2000, 2017]. Colors relate to text fields as follows: for any given journal, light blue lines give the year-specific sensitivities of a topic mention in the title for a mention in the title, abstract, keywords, or body; blue lines indicate the year-specific sensitivities of a topic mention in the title or abstract; and dark blue areas give the year-specific sensitivities of a topic mention in the title, abstract, or keywords. PS, propensity score; IPW, inverse probability weighting; MSM, marginal structural modeling; MI, multiple imputation; KM, Kaplan-Meier; NNT, number needed to treat; ME, measurement error; RCT, randomized controlled trial; LC, latent class. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

might be the only viable option. Furthermore, in case of hundreds of thousands of full-text articles to be searched, downloading of the articles needs to be automated which is currently prohibited by some publishers. An alternative approach could be to restrict the search to open access articles only, but whether this is a suitable alternative depends on the objective of the review. Furthermore, there are practical barriers to perform full-text searches because this is not possible via commonly used search engines such as PubMed.

Given the various challenges to automated searches, in current practice, there probably exists a trade-off between automated full-text literature searching in a small number of journals or TIABKW searching in large databases. Although not used in this study, both approaches could be supplemented with pearl growing strategies such as MeSH terms and snowballing in an effort to increase the sensitivity [13,14].

To conclude, searches that are based on TIABKW only may not be appropriate for systematic reviews of research practice and reporting. Provided access to full-text bodies for literature searches, full-text mining is ideally incorporated also in the first stages of a systematic literature review of epidemiological practice.

CRediT authorship contribution statement

Bas B.L. Penning de Vries: Conceptualization, Methodology, Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Maarten van Smeden:** Conceptualization, Methodology, Data curation, Writing - review & editing. **Frits R. Rosendaal:** Conceptualization, Data curation, Writing - review & editing. **Rolf H.H. Groenwold:** Conceptualization, Methodology, Data curation, Writing - review & editing.

Acknowledgments

R.H.H.G. was funded by the Netherlands Organisation for Scientific Research, Netherlands (NWO-Vidi project 917.16.430). The views expressed in this article are those of the authors and not necessarily of any funding body.

Authors' contributions: All authors have made substantial contributions to the study. B.B.L.P.d.V., R.H.H.G., and M.v.S. contributed to conception and design; B.B.L.P.d.V. contributed to acquisition of and analysis of data; B.B.L.P.d.V., R.H.H.G., M.v.S., and F.R.R. contributed to interpretation of data; B.B.L.P.d.V. contributed to drafting the article; B.B.L.P.d.V., R.H.H.G., M.v.S., and F.R.R. contributed to revision and final approval.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2020.01.009.

References

[1] Ali MS, Groenwold RH, Belitser SV, Pestman WR, Hoes AW, Roes KC, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. J Clin Epidemiol 2015;68:122−31.

[2] Mendes D, Alves C, Batel-Marques F. Number needed to treat (NNT) in clinical literature: an appraisal. BMC Med 2017;15(1):112.

[3] Brakenhoff TB, Mitroiu M, Keogh RH, Moons KG, Groenwold RH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. J Clin Epidemiol 2018;98:89−97.

[4] Copsey B, Thompson J, Vadher K, Ali U, Dutton S, Fitzpatrick R, et al. Sample size calculations are poorly conducted and reported in many randomised trials of hip and knee osteoarthritis: results of a systematic review. J Clin Epidemiol 2018;104:52−61.

[5] Alfian SD, Pradipta IS, Hak E, Denig P. A systematic review finds inconsistency in the measures used to estimate adherence and persistence to multiple cardiometabolic medications. J Clin Epidemiol 2019;108:44−53.

[6] Conn VS, Isaramalai S-A, Rath S, Jantarakupt P, Wadhawan R, Dash Y. Beyond MEDLINE for literature searches. J Nurs Scholarship 2003;35:177−82.

[7] O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015;4(1):5.

[8] Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Med 2016;13:e1002028.

[9] Lefebvre C, Manheimer E, Glanville J. Chapter 6: searching for studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions: Version 5.1.0. Oxford: The Cochrane Collaboration; 2018.

[10] Egger M, Smith GD. Meta-analysis bias in location and selection of studies. BMJ 1998;316:61−6.

[11] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available at https://www.R-project.org/. Accessed July 2, 2018.

[12] Crawley MJ. 2.12: text characters strings and pattern matching. In: Crawley MJ, editor. The R Book. Chichester: John Wiley & Sons; 2013:86−101.

[13] Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ 2005;331:1064−5.

[14] Ramer SL. Site-ation pearl growing: methods and librarianship history and theory. J Med Libr Assoc 2005;93:397.