Methodological obstacles in causal inference: confounding, missing data, and measurement error

Penning de Vries, B.B.L.

2

A comparison between full text mining and searching in title, abstract and keywords for systematic reviews of epidemiological practice

Bas B. L. Penning de Vries
Maarten van Smeden
Frits R. Rosendaal
Rolf H. H. Groenwold

Journal of Clinical Epidemiology 2020; 121: 55–61

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 manuscripts published in 5 top-ranked epidemiology journals between 2000 and 2017. *Study Design and Setting.* Based on a text-mining approach, we studied whether 9 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 modelling, 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 each topic, 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.

2.1 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. While 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 (Ali et al., 2015; Mendes and Batel-Marques, 2017; Brakenhoff et al., 2018; Copsey et al., 2018; Alfian et al., 2019).

A variety of strategies exist to identify and screen articles for eligibility for systematic reviews (Conn et al., 2003; O'Mara-Eves et al., 2015; Page et al., 2016; Lefebvre et al., 2018). 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—e.g., 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 enquiry. However, excluding a selective set of articles from further study may ultimately result in a false impression of state of the literature being conveyed (O'Mara-Eves et al., 2015; Lefebvre et al., 2018; Egger and Smith, 1998).

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.

2.2 Methods

We identified and downloaded all articles (in 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 R Core Team (2018). 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, or Supplementary R Code). In-text references and reference lists were removed from the text bodies prior to analysis. The following methodological topics were selected for investigation: propensity score methods (PS), inverse probability weighting (IPW), marginal structural modelling (MSM), multiple imputation (MI), Kaplan-Meier estimation (KM), number needed to treat (NNT), measurement error (ME), randomized controlled trial (RCT), and latent class analysis (LC). 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 Table S2.1 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. Additionally, we examined all articles with a mention of propensity score methods to determine the article type and whether or not the article described an empirical application of propensity score 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 and results 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.

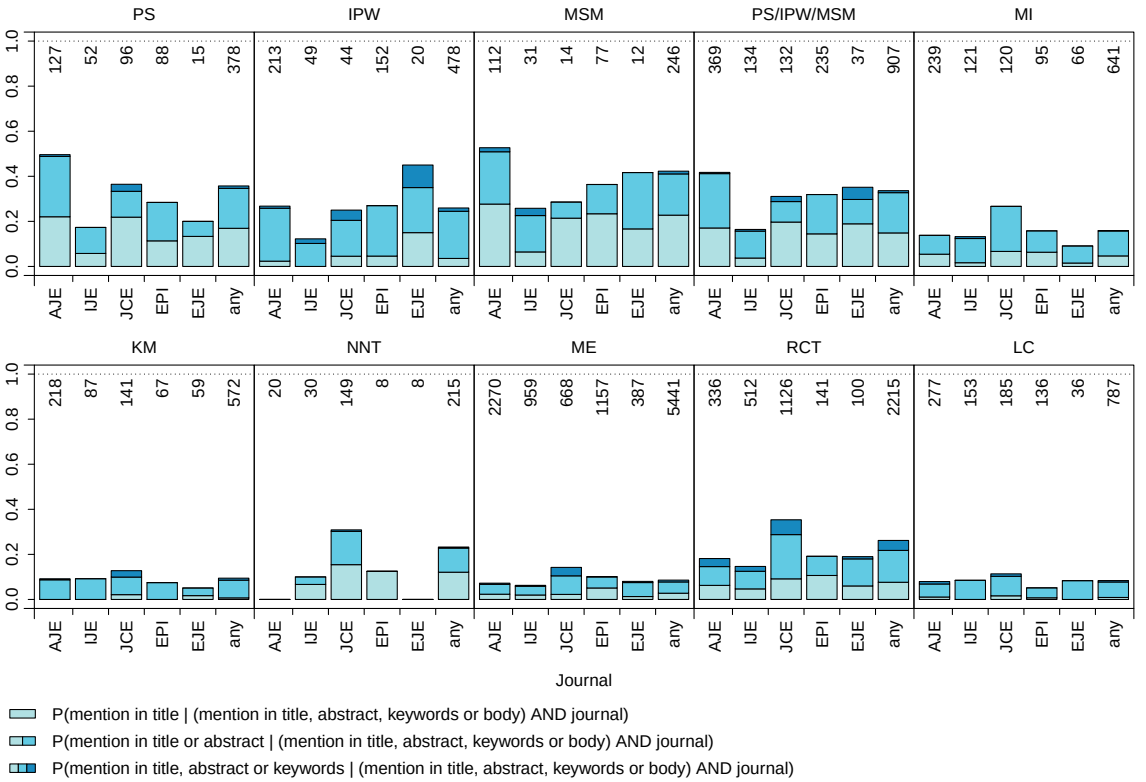

Figure 2.1: Sensitivities of topic mentioning in various text fields stratified by journal. Colors relate to text fields as follows: light blue areas give the proportion of articles with a topic mention in the title among all articles published in the indicated journal with a mention in the title, abstract, 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.

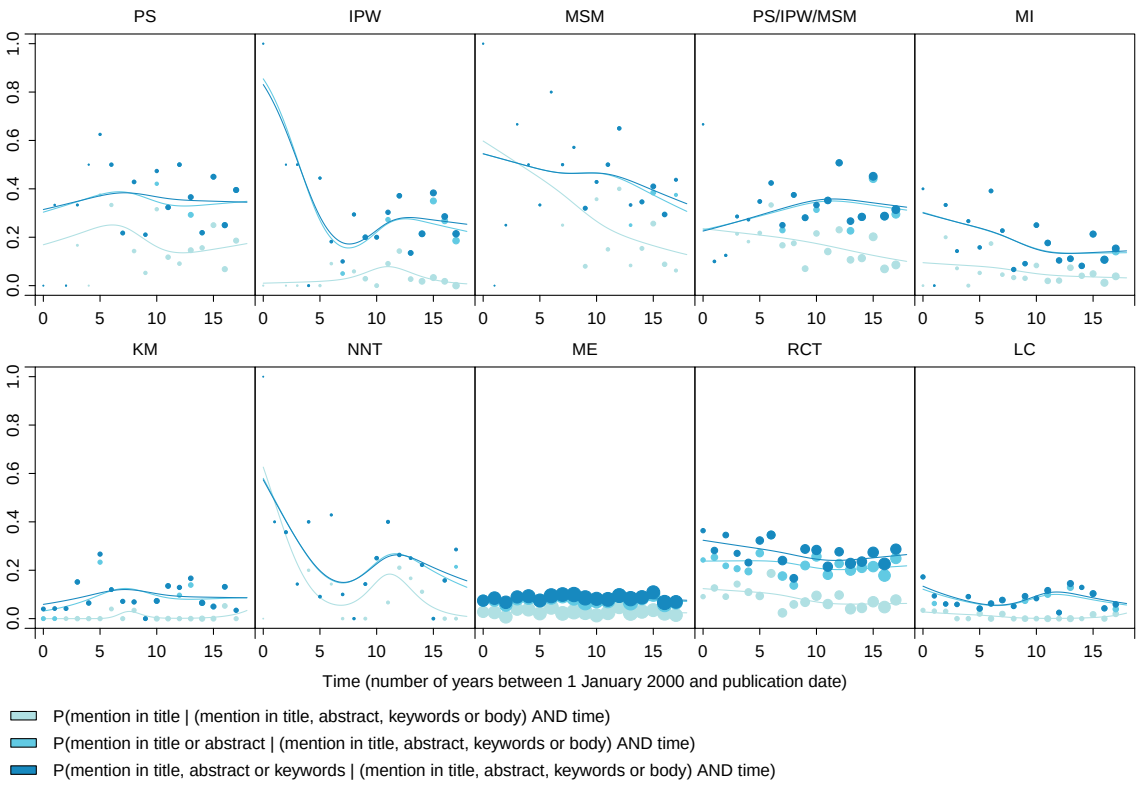

Figure 2.2: Sensitivities of topic mentioning in various text fields over time. Bullets give year-specific sensitivities with bullet size being proportional to number of publications in the given year with a mention of the topic in any text field (title, abstract, keywords, or body). Solid lines reflect 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 t o treat; ME, measurement error; RCT, randomized controlled trial; LC, latent class.

2.3 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 the 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 S2.1 and S2.2).

Figures 2.1 and 2.2 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 2.3 and 2.4 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 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 out 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%).

2.4 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., systematics reviews of the effects of a medical treatment). Our study illustrates that in systematic reviews of research practice and reporting, searches that rely only on these tools may lead

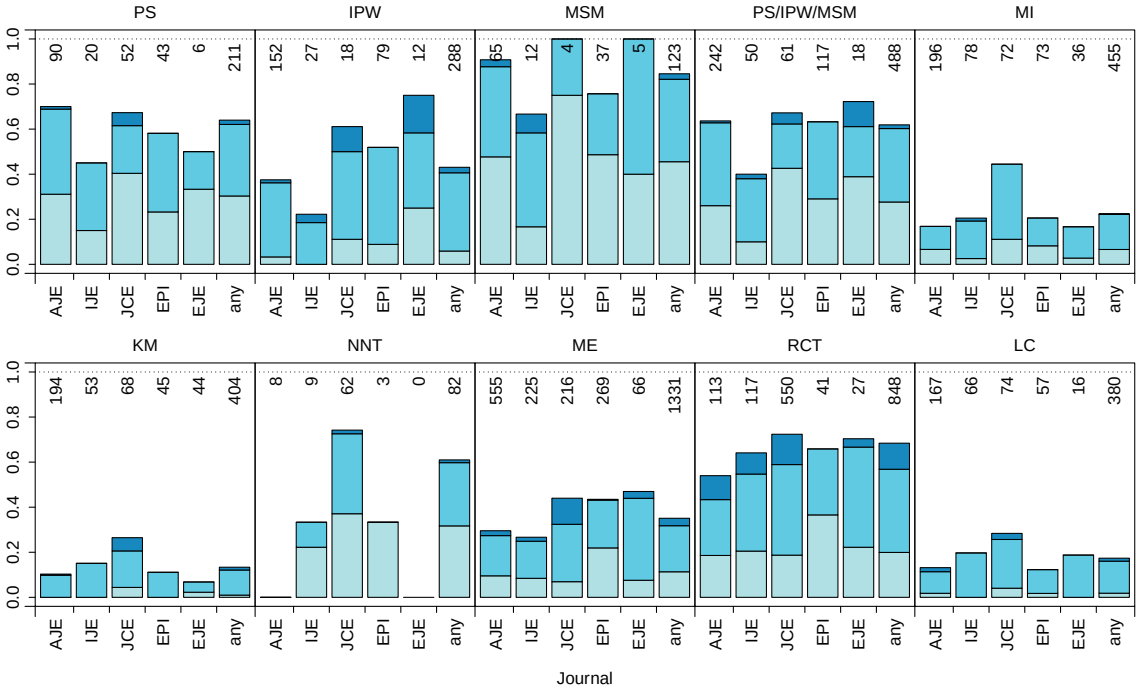

Figure 2.3: Sensitivities of topic mentioning in various text fields stratified by journal, according to post hoc analysis. Colors relate to text fields as follows: light blue areas give the proportion of articles with a topic mention in the title among all articles published in the indicated journal 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.

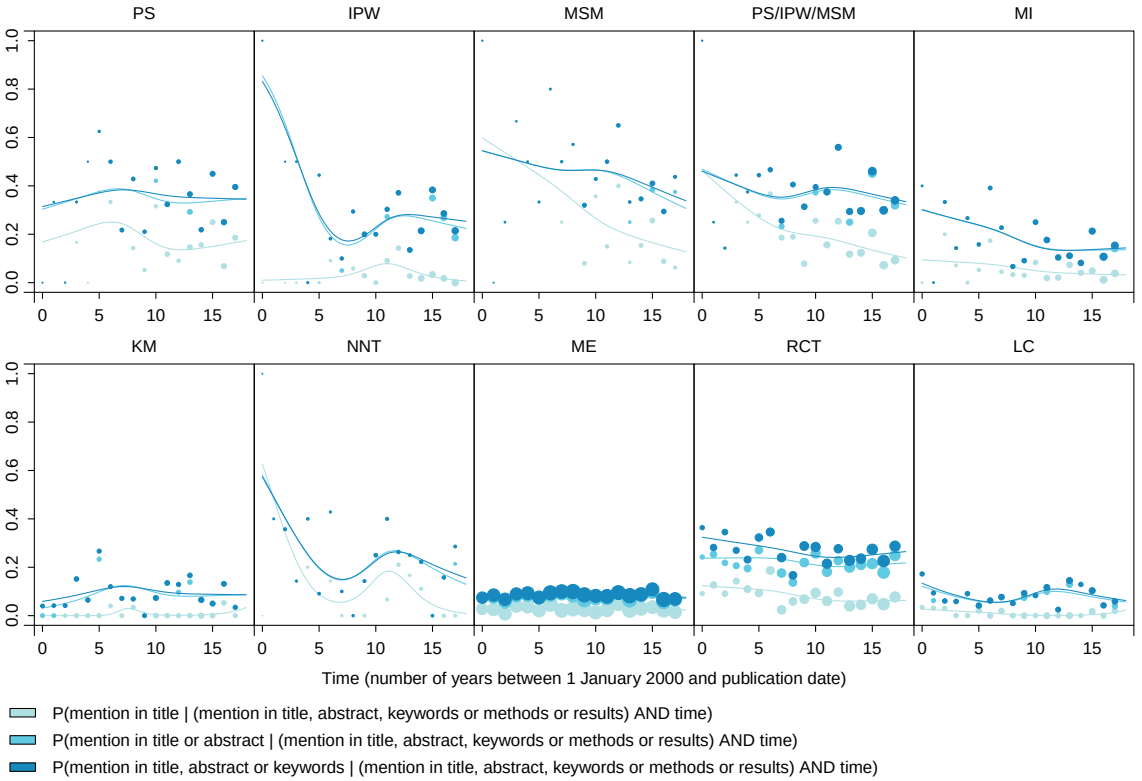

Figure 2.4: Sensitivities of topic mentioning in various text fields over time, according to post hoc analysis. Bullets give year-specific sensitivities for a mention in the title, abstract, keywords, methods, or results text fields, with bullet size being proportional to number of publications in the 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.

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 title, abstract and keywords. 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., 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 in order 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, e.g., where diverse or non-specific 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.

While 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, i.e., studies that may not be relevant to the reviewer (depending on the review goals) (see Figure 2.3). 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 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 to, 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, since 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 (Greenhalgh and Peacock, 2005; Ramer, 2005).

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.

References

S. D. Alfian, I. S. Pradipta, E. Hak, and P. Denig (2019): "A systematic review finds inconsistency in the measures used to estimate adherence and persistence to multiple cardiometabolic medications," *Journal of Clinical Epidemiology*, 108, 44–53.

Ali, M. S., R. H. Groenwold, S. V. Belitser, W. R. Pestman, A. W. Hoes, K. C. Roes, A. de Boer, and O. H. Klungel (2015): "Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review," *Journal of Clinical Epidemiology*, 68, 122–131.

Brakenhoff, T. B., M. Mitroiu, R. H. Keogh, K. G. Moons, R. H. Groenwold, and M. van Smeden (2018): "Measurement error is often neglected in medical literature: a systematic review," *Journal of Clinical Epidemiology*, 98, 89–97.

V. S. Conn, S.-A. Isaramalai, S. Rath, P. Jantarakupt, R. Wadhawan, and Y. Dash (2003): "Beyond MEDLINE for literature searches," *Journal of Nursing Scholarship*, 35, 177–182.

Copsey, B., J. Thompson, K. Vadher, U. Ali, S. Dutton, R. Fitzpatrick, S.E. Lamb, and J.A. Cook (2018): "Sample size calculations are poorly conducted and reported in many randomised trials of hip and knee osteoarthritis: results of a systematic review," *Journal of Clinical Epidemiology*, 105, 52–61.

Crawley, M. J. (2013): "2.12: Text characters strings and pattern matching." In: M. J. Crawley [editor], *The R Book*, Chichester: John Wiley & Sons.

Egger, M., and G. D. Smith (1998): "Meta-analysis bias in location and selection of studies," *BMJ*, 326, 61–66.

Greenhalgh, T., and R. Peacock (2005): "Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources," *BMJ*, 331, 1064–1065.

Lefebvre, C., E. Manheimer, and J. Glanville (2018): "Chapter 6: Searching for studies." In: J. P. T. Higgins, S. Green [editors], *Cochrane Handbook for Systematic Reviews of Interventions: Version 5.1.0*, Oxford: The Cochrane Collaboration.

Mendes, D., Alves, C., and F. Batel-Marques (2017): "Number needed to treat (NNT) in clinical literature: an appraisal," *BMC Medicine*, 15, 112.

O'Mara-Eves, A., J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou (2015): "Using text mining for study identification in systematic reviews: a systematic review of current approaches," *Systematic Reviews*, 4, 5.

M. J. Page, L. Shamseer, D. G. Altman, J. Tetzlaff, M. Sampson, A. C. Tricco, F. Catalá-López, L. Li, E. K. Reid, R. Sarkis-Onofre, et al. (2016): "Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study," *PLoS medicine*, 13, e1002028.

Ramer, S. L. (2005): "Site-ation pearl growing: methods and librarianship history and theory," *Journal of the Medical Library Association*, 93, 397.

R Core Team (2018): *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`.

Supplementary Material

Methodological topic	Search terms
Methodological topic	Search terms
Propensity score methods	propensity score; propensity scoring
Inverse probability weighting	"inverse-probability-weight; inverse probability-weight; inverse-probability weight; inverse probability weight; inverse weight; inverse-weight; inverse-probability; inverse probability"
Marginal structural modelling	marginal structural
Multiple imputation	multiple imputation; multiply imputed
Kaplan-Meier estimation	kaplan-meier; kaplan meier
Number needed to treat	number needed to; number-needed-to
Measurement Error	misclassification; measurement error
Randomised controlled trial	randomized controlled clinical trial; randomised controlled clinical trial; randomized controlled trial; randomised controlled trial
Latent class analysis	latent variable; finite mixture

Table S2.1: Overview of search terms (strings) for each of the methodological topics. Distinct strings are separated by semicolons. Articles with a mention of at least one of the specified search strings were regarded as eligible for review (i.e., as articles potentially referring to the topic of interest). A mention of a term/string was established using case insensitive approximate string matching with unit edit costs; an approximate match was said to exist if and only if the Levenshtein distance was no greater than 10% of the number of search term characters (i.e., based on the default settings of the R/3.5.0 function base::agrep).

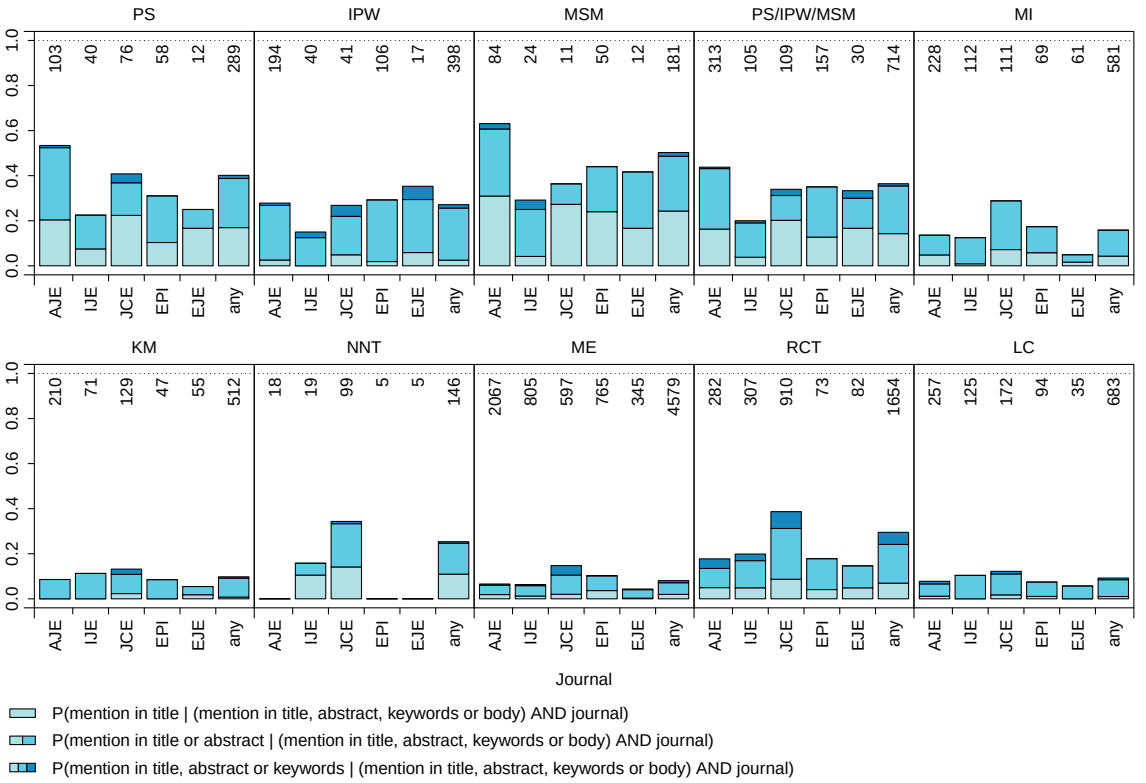

Figure S2.1: Sensitivities of topic mentioning in various text fields stratified by journal among articles of at least 2500 words.

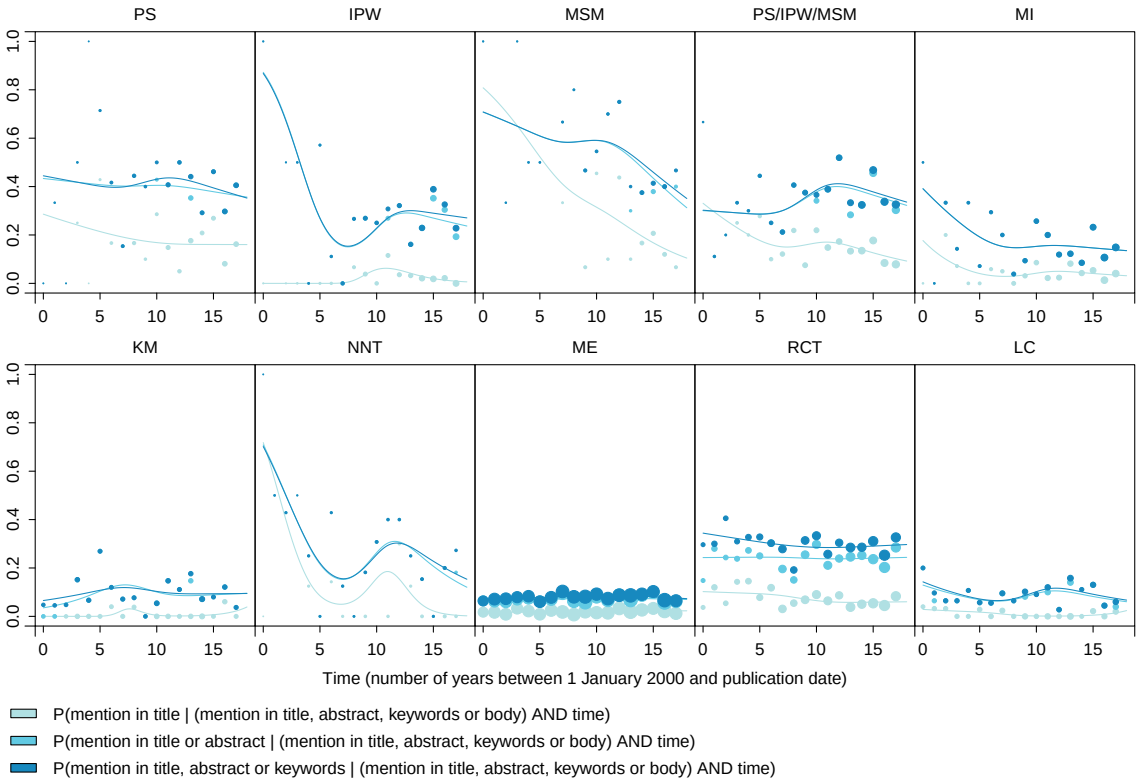

Figure S2.2: Sensitivities of topic mentioning in various text fields over time among articles of at least 2500 words. Bullets give year-specific sensitivities with bullet size being proportional to number of publications of at least 2500 words in the given year with a mention of the topic in any text 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].

Supplementary R Code

```
# The R code below was compiled to illustrate how the page source of
#   the following article may be downloaded, how relevant parts may
#   be extracted or modified, and how term mentions may be
#   identified.

# Mi, X., B. G. Hammill, L. H. Curtis, M. A. Greiner, S. and
#   Setoguchi, 2013. Impact of immortal person-time and time scale
#   in comparative effectiveness research for medical devices: a case
#   for implantable cardioverter-defibrillators, Journal of clinical
#   epidemiology, 66(8), pp.S138-S144.

# ================================================================
# Downloading file
# ================================================================

# NB: The method used below retrieves the original/unrendered page
#  source. What is returned may not contain all elements that are
#  displayed by the web browser. Publisher APIs or headless browsers
#  may be helpful when this occurs. For this example, the original
#  page source is sufficient.

con <- url("https://www.jclinepi.com/article/S0895-4356(13)00163-7/
  fulltext",method="libcurl")
x <- suppressWarnings(paste0(readLines(con),collapse=" \n "))

# ================================================================
# Extracting relevant parts
# ================================================================

# Page sources from the same journal typically have the same
#   structure. Inspection of some source files should help with
#   locating and, in turn, extracting the relevant parts. Below we
#   make use of regular expressions.
?regexpr # But see also
# Crawley, M. J. (2013). 2.12: Text characters strings and pattern
#   matching. In M. J. Crawley [editor]. The R Book, Chichester: John
#   Wiley & Sons.

# Title
m <- regexpr('<h1 class=\"articleTitle\">(.*?)</h1>',x)
title <- regmatches(x,m)

# Abstract
# To isolate the abstract we would like to extract everything
#   between '<h2 class="<section class="abstract'
#   and the matching '</section>'. However, '</section>'
#   occurs multiple times in this target string, so greedily
#   searching for '</section>' after
#   '<h2 class="<section class="abstract' is not effective here.
#   The following functions handle this problem.

fMatchOpenClose <- function(open='<div',close='</div>',text){
```

```
  opn <- gregexpr(open,text)[[1L]]
  cls <- gregexpr(close,text)[[1L]]
  n <- length(opn)
  if(n!=length(cls))
    stop(paste('discrepancy between number of opening and',
      'that of closing strings.')
  wh <- integer(n)
  available <- !wh
  for(i in seq_len(n)){
    sgn <- c(rep(1L,n-i+1),rep(-1,sum(available)))
    loc <- c(opn[i:n],cls[available])
    ord <- order(loc)
    sgn <- sgn[ord]
    loc <- loc[ord]
    clr <- loc[!cumsum(sgn)]
    if(!length(clr)) stop('invalid entry.')
    wh[i] <- clr[1L]
    available[which(cls==clr[1L]&available)[1L]] <- FALSE
  }
  m <- as.integer(opn)
  attr(m,'match.length') <- as.integer(wh+nchar(close)-opn)
  m <- list(m)
  return(m)
}
fMatchOpenCloseStartStop <- function(
  start='<div id="fulltext-body">',
  open='<div',close='</div>',stop='</div>',text){
  m <- fMatchOpenClose(open,close,text)[[1L]]
  str <- as.integer(regexpr(start,text))
  stp <- as.integer(gregexpr(stop,text)[[1L]]+nchar(stop))-1L
  m0 <- as.integer(m)
  m1 <- as.integer(m+attr(m,'match.length')-1L)
  wh <- m0>=str
  m0 <- m0[wh]
  m1 <- m1[wh]
  w <- rep(seq_len(length(m0)),2L)
  s <- c(rep(1L,length(m0)),rep(-1L,length(m1)))
  m <- c(m0,m1)
  o <- order(m)
  w <- w[o][which(!cumsum(s[o]))[1L]]
  stp <- stp[which(stp>=m1[w])[1L]]
  attr(str,'match.length') <- stp-str+1L
  return(str)
}
m <- fMatchOpenCloseStartStop(start='<section class=\"abstract',
  open='<section',close='</section>',stop='</section>',text=x)
abstract <- regmatches(x,m)

# Keywords
m <- regexpr('<div class=\"keywords\">(.*?)</div>',x)
keywords <- regmatches(x,m)

# Body
m <- fMatchOpenCloseStartStop(
  start='<div class=\"content\"><section id="sec1"',open='<div',
```

```
  close='</div>',stop='</div>',text=x)
body <- regmatches(x,m)
# NB: the reference list is not included in this text field.

# ====================================================================
# Modifying/cleaning extracted parts
# ====================================================================

title <- gsub("<(.*?)>","",title)
abstract <- trimws(gsub("^(.*?)Abstract","",gsub("<(.*?)>"," ",abstract)))
keywords <- gsub("<(.*?)>","",keywords)
keywords <- trimws(gsub("Keywords: \n","",keywords))
unlist(strsplit(keywords,", ")) # note that "Defibrillators" and
# "implantable" now appear as distinct keywords because of the crude
#   approach.
body <- gsub(
  paste0('\\[<span class="bibRef\"(.*?)',
     'See all References</a></span></span>\\]'),"",body)
  # This also removes in-text references.
body <- gsub("<(.*?)>"," ",body)

# ====================================================================
# Partial string matching
# ====================================================================

agrepl("Inverse probability weighting",title,ignore.case=TRUE) #F
agrepl("Inverse probability weighting",abstract,ignore.case=TRUE) #F
agrepl("Inverse probability weighting",keywords,ignore.case=TRUE) #F
agrepl("Inverse probability weighting",body,ignore.case=TRUE) #T

# The following function may be used to extract parts where a partial
#   string match is found. A measure of the location of the match is
#   also given.
getExcerpts <- function(term,before='. ',after='. ',count=1L,
  fixed=TRUE,text,min_dist=25L,trim_start=before,trim_end=NULL){
  if(count<1L||!is.integer(count)) stop("count must be positive integer.")
  if(min_dist<1L||!is.integer(min_dist))
    stop("else_nchar must be positive integer.")
  x <- gregexpr(term,text,ignore.case=TRUE)[[1L]]
  y <- as.integer(x)+attr(x,"match.length")-1L
  x <- as.integer(x)
  a <- gregexpr(before,text,fixed=fixed)[[1L]]
  b <- as.integer(a)+attr(a,"match.length")-1L
  c <- gregexpr(after,text,fixed=fixed)[[1L]]
  d <- as.integer(c)+attr(c,"match.length")-1L
  n <- nchar(text)
  fn <- function(i){
    v <- rev(a[b<x[i]])
    m <- min(c(length(v),count))
    p <- if(!is.na(x[i])&&m>0&&x[i]-v[m]>=min_dist) v[m] else
      max(c(1,x[i]-min_dist))
    w <- d[c>y[i]]
    m <- min(c(length(w),count))
    q <- if(!is.na(x[i])&&m>0&&w[m]-x[i]>=min_dist) w[m] else
      min(c(n,x[i]+min_dist))
```

```
    z <- p
    attr(z,"match.length") <- q-p+1L
    out <- regmatches(text,z)
    if(!is.null(trim_start))
      out <- gsub(paste0('^',trim_start),'',out)
    if(!is.null(trim_end))
      out <- gsub(paste0(trim_end,'$'),'',out)
    return(list(excerpt=trimws(out),location=x[i]/n))
  }
  l <- length(x)
  out <- if(l) lapply(seq_len(l),fn) else NA
  out <- list(excerpt=unlist(lapply(out,function(x)x$excerpt)),
    location=unlist(lapply(out,function(x)x$location)))
  return(out)
}
getExcerpts("inverse probability",text=body)
```