journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Short communication

The publication echo: Effects of retrieving literature in PubMed by year of publication

Cord Spreckelsen, Thomas M. Deserno, Klaus Spitzer*

Institut for Medical Informatics, RWTH Aachen University, Pauwelsstrasse 30, 52057 Aachen, Germany

ARTICLE INFO

Article history:
Received 24 September 2009
Received in revised form
14 January 2010
Accepted 14 January 2010

Keywords:
Information retrieval
Digital libraries
Databases
Bibliographic
PubMed
Chronology as topic

ABSTRACT

Objectives: In PubMed search forms, the publication date refers to both the date of electronic and printed publication. This fact is documented in PubMed, but difficult to anticipate by the users and can provoke misinterpretations of search results. The Technical Note aims at systematically investing the effect (referred to as the publication echo), clarifying onset and extent of the publication echo, and comments on its impact.
Methods: Papers with ambiguous publication dates are systematically retrieved and a trend analysis with seasonal decomposition on monthly publication data is performed.
Results: First doubled search results were found for 1999, their number since then rapidly increasing. Up to 17.6% of all articles of a year are found to be published electronically and in print, which can be before or afterwards. Maximum delay between the two dates is three years, except for one singular publication, where it is five years. Publication trends are exponential and linear when considering echoed and echo-cleaned data, respectively.
Conclusions: As a conclusion, we suggest using a query formulation that unambiguously retrieves literature from PubMed by the date of publication.

1. Introduction

Even if there is a tendency to use general search engines for a rapid access to biomedical information, a systematic research of medical literature relies on PubMed/MEDLINE as an unparalleled source of structured bibliographic information [1]. PubMed is considered to provide "excellent time-resolved data" [2].

Based on publication dates found in PubMed, researchers select articles for producing a review of a period of research. They can also gain an overview of the development of the publication activity in their field, e.g. by using the publication statistics generated by GoPubmed (http://www.gopubmed.com). Such publication data or publication trends also serve as the primary objective of research [3,4].

A PubMed record includes separate data fields for the electronic and the general publication dates. However, the search field labeled "Publication date" of the PubMed Advanced Search and the corresponding [DP]-tag of the query syntax (also named "Publication Date"-tag) do not differentiate between the date of the electronic and the date of the printed publication.

Due to an increasing number of articles being published electronically, a considerable number of publications are equally retrieved by PubMed queries selecting different publication dates.

* *Corresponding author*. Tel.: +49 241 8088870.
E-mail address: CSpreckelsen@mi.rwth-aachen.de (C. Spreckelsen).
1386-5056/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2010.01.007

Filling out the "Publication Date"-field of the search interface is equivalent to directly entering a query into the main search field, that contains the [DP] search tag: a date range is selected by "*startdate:enddate*[DP]", where the variables (*startdate*, *enddate*) are entered as *yyyy/mm/dd* (e.g. 2009/03/02). In order to search literature of a certain year, this can be abbreviated to "yyyy[DP]". A time interval can be excluded by using the NOT operator. For instance, to exclude literature published between 1900 and 2000, one adds "NOT 1900:2000[DP]" to the query.

The online-documentation of PubMed contains the following statement explaining the semantics of the [DP]-tag: "If an article is published electronically and in print on different dates both dates are searchable and may be included on the citation prefaced with an Epub or Print label. The electronic date will not be searchable if it is later than the print date, except when range searching" [5]. This hint is not presented at the search interface, but has to be looked up in the documentation. Furthermore, it is difficult to estimate the resulting effect on the search results: in general, each article "published electronically and in print" will be retrieved twice by entering different dates (using the exact day). Furthermore, if the date of the publication in print precedes the date of the availability of an electronic version, the latter will be ignored in the [DP]-search. As a further complication, this exception does not refer to date range searches, where articles will be retrieved twice, even if the print date precedes the electronic date (e.g. the number of articles retrieved by "2004:2004[DP]" exceeds the number of articles retrieved by "2004[DP]" by about 550—while PubMed users are likely to assume, that they are searching for papers "published in the year 2004" in both cases).

2. Objectives

The purpose of this Technical Note is to clarify the effects of using the "Publication Date" (as the term is used by PubMed) as a selection criterion in PubMed queries.

This article aims at systematically investing the number of papers retrieved twice by selecting different dates, especially the onset, development, and extent of this effect, which we refer to as "publication echo". Consequences for the interpretation of search results and of publication trends are to be discussed. Furthermore, a way to avoid the publication echo while searching PubMed will be derived and presented.

3. Methods

3.1. Onset of the publication echo

Let Q denote a PubMed query, and $n{:=}|\mathrm{res}(Q)|$ the number of papers retrieved by Q (i.e. contained in the result set). Two different queries determine the number of publications within a certain year y:

$$Q_y^{(s)}{:=}y[\mathrm{DP}] \tag{1}$$

$$Q_y^{(e)}{:=}y[\mathrm{DP}]\ \ \mathrm{NOT}\ \ 1900:(y-1)[\mathrm{DP}]\ \ \mathrm{NOT}\ \ (y+1):2100[\mathrm{DP}]. \tag{2}$$

$Q_y^{(s)}$ is the standard query used by the PubMed Advanced Search, whereas $Q_y^{(e)}$ explicitly excludes all other years of publication. Let $n_y^{(s)}{:=}|\mathrm{res}(Q_y^{(s)})|$, $n_y^{(e)}{:=}|\mathrm{res}(Q_y^{(e)})|$ denote the corresponding response sizes. Then, the absolute annual publication echo

$$\Delta_y{:=}n_y^{(s)} - n_y^{(e)} \tag{3}$$

determines the number of publications with an ambiguous publication date in the respective year y. According to (1) and (2), the difference Δ_y in (3) cannot be negative. We determine the *onset* of the publication echo given by the year y, where for the first time $\Delta_y > 0$

The relative annual publication echo is the rate of ambiguous publication dates relative to the number of publications retrieved by the standard query $Q_y^{(s)}$:

$$r_y{:=}\frac{\Delta_y}{n_y^{(s)}} \tag{4}$$

3.2. Extent of the publication echo

The delay between the date of electronic and printed publication is called publication echo time t. It determines the severity of a possible misinterpretation. Let $n_{yz}{:=}|\mathrm{res}(y[\mathrm{DP}]\ \mathrm{AND}\ z|\mathrm{DP}|)|$ denote the number of publications found in two different years y and z. The maximum echo time t_y of a certain year y is given by the maximum difference between the year y and all years z, where $n_{yz} > 0$:

$$t_y{:=}\hat{z} - y, \quad \text{where } \hat{z} = \max\{z | z \geq y,\ n_{yz} > 0\} \tag{5}$$

The average echo time $\bar{t}_y$ is given by

$$\bar{t}_y{:=}\frac{\sum_{z>y}(n_{yz}(z-y))}{\sum_{z>y}(n_{yz})} \tag{6}$$

According to the definition of n_{yz}, we systematically investigated queries of the form:

$$y[\mathrm{DP}]\ \mathrm{AND}\ z[\mathrm{DP}] \tag{7}$$

for all pairs of different years y, z since the onset of ambiguous publication dates and determined the corresponding numbers n_{yz} of publications retrieved.

Since the PubMed Advanced Search interface and the [DP]-tag do not differentiate between the electronic and the print date, the selection of the "Publication Date" can yield four configurations, each referring to two papers that are both retrieved by the same query (Fig. 1):

(A) The selection criterion is met by the print date ([PPDAT]-tag in PubMed) of one paper and by the electronic date ([EPDAT]-tag in PubMed) of another. In this case, the maximum difference of the publication dates is a difference between a print date and an electronic date.
(B) As PubMed contains papers, where the print date *precedes* the electronic date, the date range can start with the print date and end with the electronic date.

(C) For the same reasons, the maximum difference can also be a difference between two print dates.
(D) Finally, the maximum difference can be a difference between two electronic dates.

We systematically investigated the numbers of papers n_{yz} retrieved by queries of the form:

y[PPDAT] AND z[EPDAT] (8)

for all pairs of different years y, z since the onset of ambiguous publication dates. The contribution of the configurations (B), (C) and (D) to the publication echo was derived from the resulting matrix.

3.3. *The impact of the publication echo on publication trend estimation*

An exponential growth of publication output is frequently stated in literature. Hunter and Cohen, for example, stated an exponential growth of biomedical publication output based on curve fitting of PubMed data [3]. De Shazo et al. described an exponential trend for publications addressing Medical Informatics [4]. However, such publication trends may differ when derived from the uncorrected (i.e. echo-based) or corrected (echo-cleaned) data. To determine the echo's impact on publication trend measures, we retrieve the number of papers published per month since the onset of publication echo, apply seasonal trend decomposition [6], and perform a regression analysis (linear vs. exponential model).

4. Results

4.1. *Onset of publication echo*

Fig. 2 shows the relative publication echo. Its onset dates to the year 1999. Since then, the echo effect has strongly increased, reaching a maximum of 17.6% in the year 2007. The theoretical maximum of 50% would be reached if all publications contributed twice to $n_y^{(s)}$ (by the date of the electronic publication and by the print date).

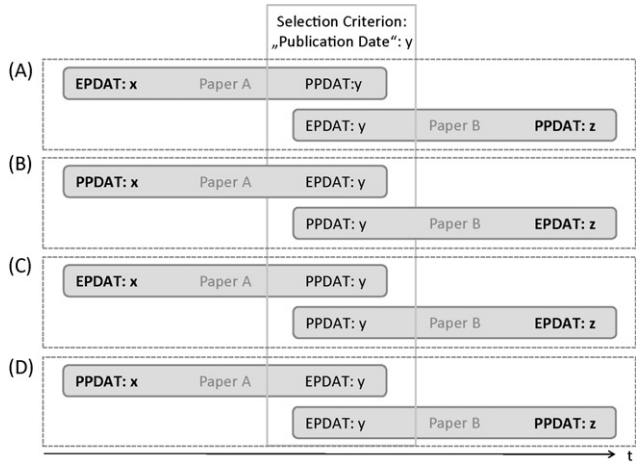

Fig. 1 – Possible configurations contributing to the publication echo ([PPDAT]: print date; [EPDAT]: electronic date; [DP]: "Publication Date" used as selection criterion).

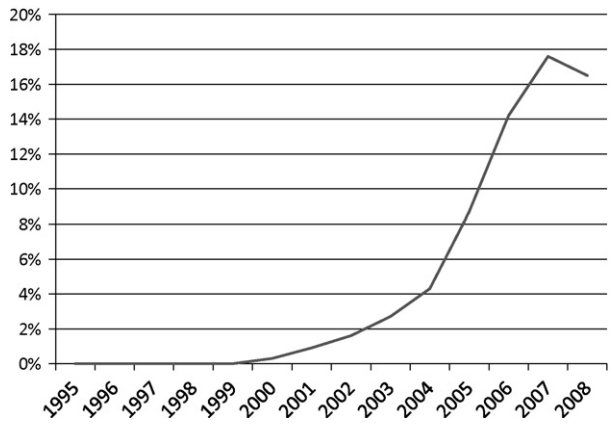

Fig. 2 – Relative publication echo r_y. A further growth is predicted for 2008 when PubMed completes processing of 2008 references.

4.2. *Extent of the publication echo*

Table 1 shows the numbers n_{yz} of papers assigned to two dates of publication. Due to the fact that the "AND"-operator is commutative, the resulting matrix is symmetric.

The maximum echo time was also found increasing: in 1999 and 2000 the maximum delay was one year ($t_{1999} = t_{2000} = 1$), between 2002 and 2005 it was two years ($t_{2002} \ldots t_{2005} = 2$), and since 2005 the publication echo has been three years ($t_y = 3$). There is one exceptional case: paper [7] was published in 2001 electronically and then printed five years later in the year 2006 ($t_{2001} = 5$). The column sum given in Table 1 exactly reproduced the Δ_y found by the first test. The average echo times were all about one year ($\bar{t}_{1999} = \bar{t}_{2000} = 1$, $\bar{t}_{2001} = 1.021$, $\bar{t}_{2002} = 1.003$, $\bar{t}_{2003} = 1.004$, $\bar{t}_{2004} = 1.006$, $\bar{t}_{2005} = 1.009$, $\bar{t}_{2006} = 1.012$, $\bar{t}_{2007} = 1.002$).

Fig. 3 shows the date ranges resulting from the four configurations (A–D) defined in Fig. 1. The date ranges found by the configurations (A) and (B) are given by mixing electronic and print dates. The date ranges resulting from configurations (C) and (D) exclusively refer to electronic and print dates, respectively (e.g. searching papers with a "Publication Date" in the year 2003 by the Advanced Search yields 13 papers printed in 2000 and 37 papers printed in 2005, as shown by (D)). The maximum date range of six years was found in seven cases.

4.3. *Impact of publication echo*

The time series of the numbers of papers published *per month* yielded strong effects of seasonal dates of publication (Fig. 4): each January the number of publications doubles[1]. After sea-

[1] PubMed sets publication dates without a month to January, multiple months (e.g. October–December) are set to the first month, and dates without a day are set to the first day of the month. Dates including a season are set as: winter = January,

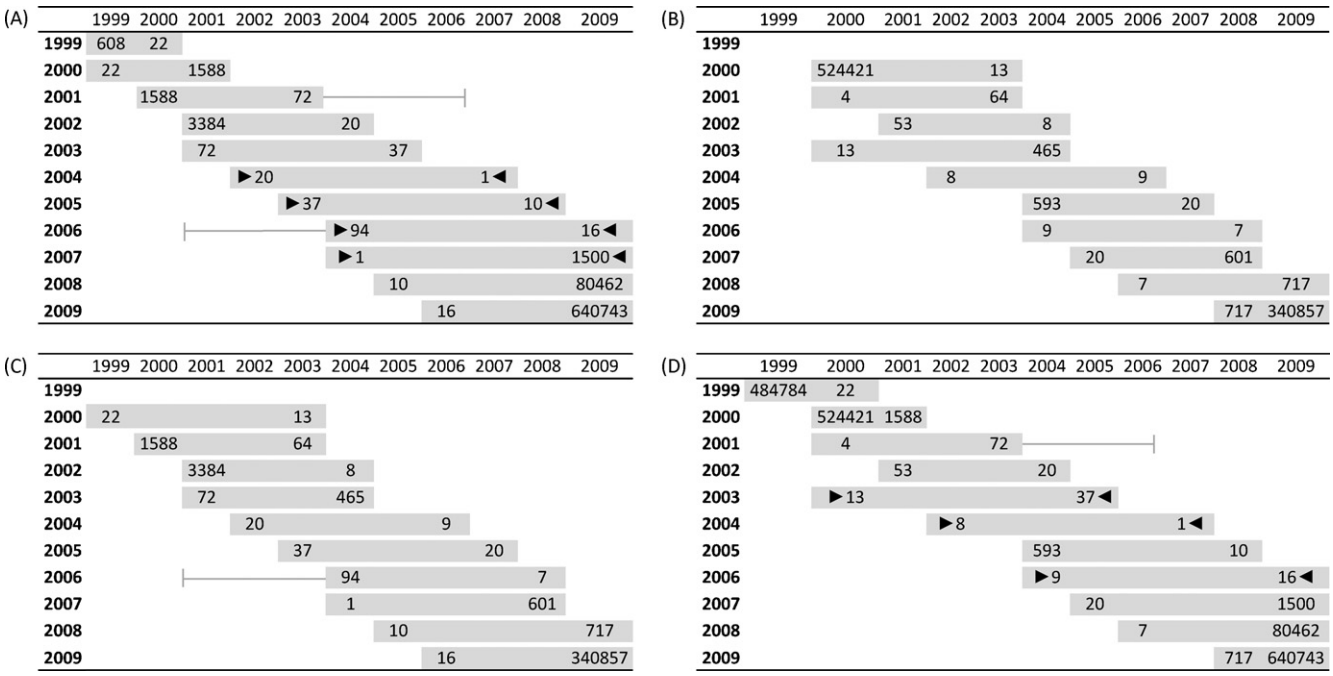

Fig. 3 – Date ranges resulting from the configurations (A–D) defined in Fig. 1. The selection criterion of the respective PubMed query is given by the year on the left side (bold face). The numbers attached to the bars refer to the number of papers with the earliest and latest date respectively (retrieved by the same query). The effect of one paper [7] with an exceptional delay of five years between the print and electronic date is indicated by thin lines, but not considered further. The maximum date range (six years), found in seven cases, is marked by arrows.

Table 1 – Numbers of publications assigned to two different years (*y*, *z*) as retrieved by the PubMed query "y[DP] AND z[DP]". Values below and right of the thick line are expected to increase due to the further processing of articles by PubMed.

y	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
z											
1999		22	0	0	0	0	0	0	0	0	0
2000	22		1588	0	0	0	0	0	0	0	0
2001	0	1588		3384	72	0	0	1	0	0	0
2002	0	0	3384		5743	20	0	0	0	0	0
2003	0	0	72	5743		9838	37	0	0	0	0
2004	0	0	0	20	9838		17324	94	1	0	0
2005	0	0	0	0	37	17324		42461	362	10	0
2006	0	0	1	0	0	94	42461		61095	714	8
2007	0	0	0	0	0	1	362	61095		72612	1114
2008	0	0	0	0	0	0	10	714	72612		59074
2009	0	0	0	0	0	0	0	8	1114	59074	
Sum	22	1610	5045	9147	15690	27277	60194	104373	135184	132410	60196

sonal trend decomposition exponential and linear models were fitted to the trend data. In the case of uncorrected (echo-based) publication data, the exponential model clearly performed better: *R*-square = .97 (exponential) vs. .94 (linear). In contrast, both models fitted almost equally on corrected (echo-cleaned) data: *R*-square = .97 (both, with differences of one order of magnitude below).

spring = April, summer = July, fall = October. This causes the manifest periodic (seasonal) effects.

5. Discussion

Fig. 2 shows a drastic increase of ambiguous results of PubMed queries limiting the publication date. E.g. one out of six publications assigned to the year 2006 has been published electronically or in print before or after that year again. The decrease for the last years can easily be understood when considering Table 1: the number of ambiguous publication dates found for a specific year is given by the sum of the contributions from publications having a second publication date

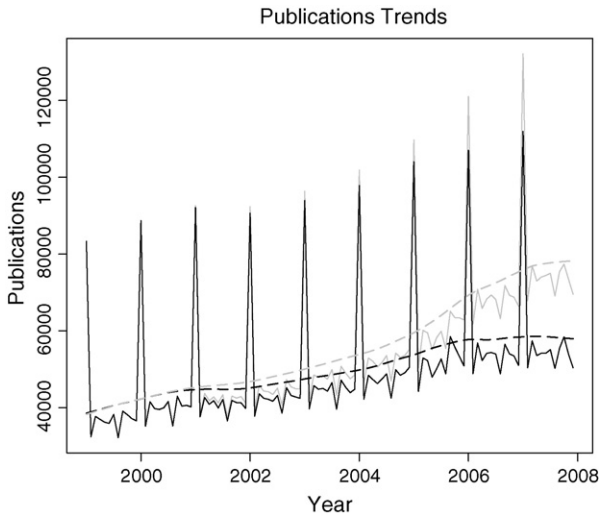

Fig. 4 – Number of new publications per month: echo-based (grey line), echo-cleaned (black line), echo-based trend line (dashed grey line), and echo-cleaned trend line (dashed black line).

prior to the year in question and those published electronically or in print again in the *following* years. For 2008 and 2009 the contributions of the following years are not yet fully available in PubMed. Therefore, the rapid increase of that effect is expected to continue in the next years. The theoretical maximum of the publication echo (at $r = 50\%$ in Fig. 2) would be reached, if each paper had a print date as well as an electronic date. This maximum will not be met in fact, because of the increasing number of papers having only an electronic publication date (due to more eJournals being indexed for Medline). Therefore, the curve in Fig. 2 can be expected to decrease after having reached a maximum of $r < 50\%$.

The cases, where the print date precedes the electronic date are not included in Table 1, because according to the definition of the [DP]-tag, they are not retrieved, when using the [DP]-tag in combination with a single date (or year). In contrast to that, the cases are included by PubMed, when a date *range* is specified, which this the only way to select a publication year in the Advanced Search interface.

The delay between the publication date of the electronic and the printed version increased to three years for a considerable number of publications: therefore, a query limited to a special year of publication yields a set of articles with a broad range of publication dates (Fig. 3). Clearly, the maximum date range in the result set may be caused by configuration (A) of Fig. 1, where the earliest date refers to the electronic publication and the latest date refers to the print publication of some papers from the same result set. But as shown by Fig. 3, the contribution (B), (C), and (D) of Fig. 1 can also contribute to the date range: even the difference between the print dates in the same result set is not limited to three years (maximum delay between the print and electronic dates). Instead it is found to be up to six years. For the majority of articles the electronic date precedes the print date. Thus, one can expect to find more papers with a maximum difference between the dates in this subset than in the smaller one, where PPDAT < EPDAT. Additionally, the delay between the dates was found to increase over the years (Table 1). This explains, why the cases of maximum date range are found in (A) and (D) (Fig. 3), because more recent papers with EPDAT < PPDAT contribute to these configurations (Fig. 1).

The most obvious effect of seasonal dates of publication found in the publication time series per month is due to the fact, that some journals use a season, range of month, or year as the date of publication. Most of these cases add to the corresponding January result set of the year in question (namely, if the publication date is specified only by a year, by a month range starting with January, or by 'winter').

Revisiting the publication trends stated in the literature, our results show that the ambiguous publication dates retrieved with the Advanced Search interface indeed have an effect on the publication trend inferred from the data: in the uncorrected (echo-based) case, the exponential growth stated in the literature is better supported by the data than the alternative linear trend. In the corrected (echo-cleaned) case, there is no such clear preference. For the years investigated, the publication echo makes the difference between observing an exponential and a linear growth of publication output.

However, all the effects are not due to double entries in the PubMed database. Instead, they are caused by the search tag/search interface, which does not distinguish between the electronic and the print date.

Although this search behavior has been documented on the PubMed website, the respective hint is more or less hidden in a special part of the search syntax definition. In particular, users of PubMeds' Advances Search interface are likely to not expect the way their query is answered, because starting from their intuitive concept of a publication date they may not be inclined to read the details in the PubMed documentation.

This can lead to the following mistakes:

- Wrong assumptions referring to the coincidence of ideas or research topics may be induced, because a set of articles seemingly assigned to the same publication period may in fact belong to a much broader range of publication dates.
- The analysis of publication trends may be biased, e.g. a rapidly growing number of publications in a special field caused by exponential growth of numbers of "double" publications, rather than an increased research interest.

As far as we are aware, the established bibliometric indices are not affected by possible ambiguities of the publication date: The Journal Impact Factor [8] is based on Thompson Reuter's proprietary multidisciplinary citation database, which uniquely specifies the publication date by the print date for all printed journals and by the date of the online publication for all eJournals. The same holds for the SCImago Journal Rank [9], which is based on the citation data of the Scopus database, and the Déjà vu project intended to spot double publication and plagiarism [10]. In contrast, systems like GoPubMed (http://www.gopubmed.com), which provide sorted results based on original PubMed queries, reproduce the publication echo in the respective result sets.

To the best of our knowledge, the effect that we refer to as "publication echo" has not yet been addressed in the literature.

Table 2 – PubMed search tags referring to dates.

PubMed date-tag	Use
CRDT	The date the record was added to the PubMed database
DP	The date of electronic or printed publication (i.e. the same article may be selected by two queries using different dates to specify DP)
EDAT	The date the record was added to the PubMed database; EDAT is equal to the publication date for citations that are older than one year when they enter PubMed (policy started in December 2008)
EPDAT	The date of the electronic publication
MHDA	The date the citation was indexed with MeSH terms
PPDAT	The date of the printed publication

Table 3 – PubMed search syntax (examples).

Query	PubMed result
2001/01:2001/06[DP]	. . .published from the 1st of January to the 30th of June 2001 (printed and/or electronic publication) including articles assigned to 2001 without specifying day and month
Intend: to retrieve all articles. . .	PubMed Search Syntax
. . .firstly published in year y either electronically or printed	y[DP] NOT 1900:(y − 1)[DP]
. . .firstly published in 2000 either electronically or printed	2000[DP] NOT 1900:1999[DP]

5.1. *Improved search syntax*

In fact, the PubMed syntax allows to distinct between date of electronic publication and the publication date of the printed version (Table 2). Unfortunately, these additional syntax elements are neither presented at the usual search interface nor listed in the tag overview. Instead, they are mentioned in the documentation text of the [DP]-tag.

Nonetheless, even the use of these search tags does not fully solve the problem: it avoids the ambiguities discussed, but choosing one of these tags will exclude the articles uniquely published in the other medium. So far, the best way to avoid ambiguities is to enter queries of the form:

$$y[\mathrm{DP}]\,\mathrm{NOT}\,y_0 : (y-1)[\mathrm{DP}] \quad (9)$$

where y is the year (resp. date) in question and y_0 is a time before the first publication date available in PubMed (Table 3). This type of query will retrieve a publication only by its first publication date covered in PubMed.

Acknowledgements

The authors wish to thank the reviewers for their statements, which lead to substantial improvements of the paper (especially concerning the difference between "y[DP]" and "y:y[DP]" and the details of the date ranges). They also thank Thees Spreckelsen (Nuffield College, University of Oxford) for suggesting the investigation of seasonal effects.

Authors' contributions: CS had the idea to investigate the effects of the ambiguously interpreted publication dates in PubMed and is the main author of the paper; TD conceptualized the study design and formulated the resulting problems and recommendations; KS contributed to the time series analysis and the discussion.

Summary points

What was already known on the topic:

- The PubMed publication date (DP) search tag and the respective PubMed search forms refer to both the date of electronic and printed publication.
- Bibliometric trend analyses and the appraisal of topicality in the biomedical domain often rely on PubMed search results.

What this study added to our knowledge:

- A substantial and increasing fraction of articles (up to 17.6%) is retrieved by two different years of publication.
- A query selecting a publication year may return a set of articles with a second publication date in a range of up to six years.
- Biomedical researchers should be aware of the ambiguity when considering or comparing publication dates, and should be explicitly instructed by experts in information retrieval.
- Well-defined search results can be achieved by explicitly excluding the time interval preceding the date in question in the respective PubMed query.

REFERENCES

[1] A. Hoogendam, P.F. de Vries Robbé, A.F. Stalenhoef, A.J. Overbeke, Evaluation of PubMed filters used for evidence-based searching: validation using relative recall, J. Med. Libr. Assoc. 97 (3) (2009 Jul) 186–193.
[2] T. Pfeiffer, R. Hoffmann, Temporal patterns of genes in scientific publications, Proc. Natl. Acad. Sci. U.S.A. 104 (2007) 12052–12056.
[3] L. Hunter, K.B. Cohen, Biomedical language processing: what's beyond PubMed? Mol. Cell 21 (2006) 589–594.
[4] J. DeShazo, D. Lavallie, F. Wolf, Publication trends in the medical informatics literature: 20 years of "Medical Informatics" in MeSH, BMC Med. Inform. Decis. Mak. 9 (7) (2009) 1–13.
[5] National Center for Biotechnology Information: NCBI Help Manual [Internet]. Bethesda (MD): NCBI; c2005-2009 [cited 2009 November 24]. Available from: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp#pubmedhelp.Publication_Date_DP.

[6] R.B. Cleveland, W.S. Cleveland, J.E. McRae, I. Terpenning, STL: a seasonal-trend decomposition procedure based on Loess, J. Off. Stat. 6 (1990) 3–73.
[7] J.I. Reddick, A. Goostrey, C.J. Secombes, Cloning of iNOS in the small spotted catshark (*Scyliorhinus canicula*), Dev. Comp. Immunol. 30 (2006) 1009–1022.
[8] E. Garfield, The history and meaning of the journal impact factor, JAMA 295 (2006) 90–93.
[9] D. Butler, Free journal-ranking tool enters citation market, Nature 451 (2008) 6.
[10] M. Errami, J.M. Hicks, W. Fisher, D. Trusty, J.D. Wren, T.C. Long, et al., Déjà vu—a study of duplicate citations in MEDLINE, Bioinformatics 24 (2) (2008 Jan) 243–249.