Identifying evidence for public health guidance: a comparison of citation searching with Web of Science and Google Scholar

Paul Levay, Nicola Ainsworth, Rachel Kettle and Antony Morgan

Aim: To examine how effectively forwards citation searching with Web of Science (WOS) or Google Scholar (GS) identified evidence to support public health guidance published by the National Institute for Health and Care Excellence.

Method: Forwards citation searching was performed using GS on a base set of 46 publications and replicated using WOS.

Outcomes: WOS and GS were compared in terms of recall; precision; number needed to read (NNR); administrative time and costs; and screening time and costs. Outcomes for all publications were compared with those for a subset of highly important publications.

Results: The searches identified 43 relevant publications. The WOS process had 86.05% recall and 1.58% precision. The GS process had 90.7% recall and 1.62% precision. The NNR to identify one relevant publication was 63.3 with WOS and 61.72 with GS. There were nine highly important publications. WOS had 100% recall, 0.38% precision and NNR of 260.22. GS had 88.89% recall, 0.33% precision and NNR of 300.88. Administering the WOS results took 4 h and cost £88–£136, compared with 75 h and £1650–£2550 with GS.

Conclusion: WOS is recommended over GS, as citation searching was more effective, while the administrative and screening times and costs were lower. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: information retrieval; literature searching; public health guidance; citation searching; web of science; google scholar

1. Introduction

Evidence-based guidance is often supported by systematic reviews that use rigorous and transparent methods to reduce the risk of bias and improve external validity (Higgins and Green, 2011). The search process is essential to developing a systematic review so that it does not miss relevant studies, introduce bias or reach invalid conclusions. It is, however, unrealistic to expect a systematic review to locate all relevant evidence given that the time and resources required to do this would be excessive (McGowan and Sampson, 2005). This study examines the role of forwards citation searching in a systematic search and considers how effectively Web of Science (WOS) and Google Scholar (GS) can identify the relevant literature.

The case study examines how the National Institute for Health and Care Excellence (NICE) located evidence for its guidance on assessing body mass index (BMI) and waist circumference thresholds for intervening to prevent ill health among black, Asian and other minority groups (NICE, 2013). It was challenging to complete the review by the deadline because of the practical difficulties encountered when trying to use GS in a systematic way. This
study was motivated by needing to know whether the practical limitations of GS outweighed the fact that it was free to access. The additional work investigated whether using WOS instead of GS would have addressed some of these issues. The value of this study is that the comparison of WOS with GS contributes to the debate on doing reviews efficiently so that they are pragmatic and yet still systematic.

2. Background

2.1. Searching at NICE

NICE is an independent public body in the UK responsible for producing public health guidance to promote good health and to prevent ill health. The guidance is based on a “rigorous assessment of the evidence base” and is developed “using a transparent process and methods” (NICE, 2012). The guidance is usually informed by several systematic reviews that explicitly address the research questions by drawing on a wide range of study methodologies and sources. Once the evidence has been identified, appraised and synthesised, it is presented to an independent committee of experts who use it to formulate their recommendations. The draft guidance is subject to consultation with stakeholders before final publication. NICE has an ongoing commitment to methodological innovations to support guidance development (Kelly et al., 2010), and the manual used in this study (NICE, 2012) has recently been updated (NICE, 2014).

NICE acknowledges that the search will not find every single study that might be relevant and balances this potential risk against the fact that a variety of resources and techniques will be used to capture high-quality evidence (NICE, 2014). Increasing the number of search results to ensure nothing important is missed will mean more screening time is required to identify the publications that are actually relevant. This issue is particularly likely to affect searches for complex problems, where a range of evidence needs to be located across a number of subject areas (Alpi, 2005). A comprehensive search could also increase the ratio of non-relevant to relevant results in such a proportion that the reviewers will miss some important publications as they try to finish the screening in time, especially as titles and abstracts are not always helpful for judging relevance. The risk is that a high volume of results cannot be screened within the time and resources available, leading to a review that is not timely, relevant or up to date (Beller et al., 2013). The aim of a search, therefore, is to locate the best available evidence without creating an unmanageable set of results to screen (NICE, 2014).

2.2. Berrypicking

Systematic review methods that can take over a year to complete may not be appropriate for government departments and other policy developers to make evidence-informed decisions (Thomas et al., 2013). A number of methods have been used to produce reviews rapidly, and those targeted at the searching stage need to be assessed carefully to ensure they do not introduce bias (Ganann et al., 2010). There is a risk that being both comprehensive and timely will inadvertently affect the results if decisions are not based on a clear rationale, for example if an arbitrary date limit is set to reduce the volume of results.

Recent work has questioned whether “comprehensiveness” actually designates a quality search (Booth, 2010). The implication is that efforts should be focussed on the resources that are most likely to affect the outcomes (Ganann et al., 2010), instead of always prioritising databases (Booth, 2010). The use of focussed, and yet robust, searches draws on the berrypicking model of information retrieval. The model states that the query is satisfied not by a single final retrieved set, but by a series of selections of individual references and bits of information at each stage of the ever-modifying search (Bates, 1989).

Searches following this approach can develop iteratively, instead of trawling a standard list of databases with a pre-determined search strategy (Booth et al., 2013). The berrypicking model suggests it is important to investigate other search techniques as well as databases.

Methodological guides for systematic reviews (such as Lefebvre et al., 2011) and for guidance development (such as NICE, 2014) recommend using a range of techniques. Searchers can choose from various methods appropriate to the review (Papaioannou et al., 2010), including citation searching, reference harvesting, contact with experts (McManus et al., 1998), handsearching (Armstrong et al., 2005), pearl growing (Booth, 2008) and related-item searching (Waffenschmidt et al., 2013). The common theme with these methods is that they seek to exploit maximum benefit from a set of relevant publications, in line with the berrypicking model.

2.3. Citation searching

Citation searching can be performed in a backwards or forwards direction. Backwards citation searching, also known as reference harvesting, involves reviewing the bibliography of a relevant publication, thus providing links to earlier research (Horsley et al., 2011). Forwards citation searching means checking whether a specific publication, already known to be relevant, has been cited by any later publications, which might also be of interest (Papaioannou et al., 2010).
The NICE team chose GS for citation searching because it was a free service, was easy to use, required little training and it could be accessed immediately without a password (http://scholar.google.com). GS includes peer-reviewed articles, theses, patents, reports and other grey literature, which it draws from open access materials, pre-prints, university repositories and other sources.

Google does not fully explain how it compiles GS, and it has not published a list of the journals it covers (Mikki, 2009). The automated processes for gathering content for GS make it vulnerable to error and, for example, page numbers could be mistaken for publication years (Jacso, 2010). The results are ordered with the “most relevant” ones first, and this sorting is determined by a proprietary algorithm. Google has not described the exact ranking method, and this is challenging when using GS for a transparent and replicable systematic review. GS only makes the first 1000 results available to view, however many times an article has been cited, and these top results are likely to be the older citations (Bramer et al., 2013). The long-term viability of GS is questionable (Gans, 2013), now that Google does not seem to be prioritising its development and the link is difficult to find from the home page (Schwartz, 2011).

WOS is a subscription-based citation index (http://thomsonreuters.com/web-of-science) that covers several, separate sources, including Science Citation Index (SCI) and Social Science Citation Index (SSCI). The publisher describes WOS as a “curated” source, meaning there is some manual verification of the data (Thomson Reuters, 2014) that helps to reduce the errors that can occur with automated systems (e.g. conflating different authors with the same name). A major strength of WOS is its in-depth indexing and tagging of articles, coupled with a sophisticated search interface, which makes for a powerful retrieval tool (Mikki, 2009). WOS has been criticised for its bias towards North American and English language publications because of the way that citations are harvested and verified (Mikki, 2009).

A number of studies comparing the two resources are already available, as summarised by Mikki (2009) and Bramer et al. (2013). Tables assessing the key features of WOS and GS are available (Falagas et al., 2008; Li et al., 2010). The comparative studies undertaken to date cover unrelated topics (Beckmann and von Wehrden, 2012), analyse clinical topics (Kulkarni et al., 2009; Nourbakhsh et al., 2012), consider the social sciences in general (Levine-Clark and Gil, 2009) or attempt to measure the growth in coverage (De Winter et al., 2014).

Citation searching is already an established part of NICE methods (NICE, 2014), but it was important to examine the implications of using it as the primary evidence retrieval tool. The first study to examine citation searching in a public health review concluded that it was of limited value when used as a supplement to database searches (Wright et al., 2014). It is important to move away from seeing citation searching and other methods as having a supplementary role, as though database searching must always happen first. This study of the NICE process differs from Wright et al. (2014) by testing citation searching without making it secondary to extensive database searching. GS might have been free to access but was it efficient for a systematic review or did it have knock-on effects that were disadvantageous later in the process?

3. Aim and objectives

The aim of this study was to compare how effectively forwards citation searching with WOS and GS identified relevant evidence for a systematic review supporting public health guidance. The purpose was to test citation searching in NICE guidance production and not to compare WOS and GS in isolation.

The objectives were to assess:

- whether citation searching with WOS retrieved more publications relevant to the systematic review than GS.
  - Test 1: Did WOS find the same publications as GS?
  - Test 2: Did WOS find additional publications that GS had missed?

- whether citation searching with WOS retrieved more publications of high importance to the systematic review than GS.
  - Test 3: Did WOS find the same highly important publications as GS?
  - Test 4: Did WOS find additional highly important publications that GS had missed?

- the ease of searching, downloading and screening records from WOS and GS.
- the time and costs of using WOS and GS for citation searching.

4. Methods

4.1. Research questions

Before describing the methods for this WOS and GS comparison, it is first necessary to explain how the evidence was retrieved for the public health guidance (NICE, 2013). The purpose of the search (Figure 1) was to answer four review questions:
Question 1: How accurate are BMI and waist circumference in predicting the future risk of type 2 diabetes, fatal/non-fatal myocardial infarction or stroke and overall mortality among adults from black, Asian and other minority ethnic groups living in the UK compared to the white or general UK population?

Question 2: What are the BMI and waist circumference cut-off points indicating a healthy range for these measures among adults from different black, Asian and other minority ethnic groups living in the UK?

Question 3: What are the BMI and waist circumference cut-off points that indicate an increased risk of type 2 diabetes, fatal/non-fatal myocardial infarction and stroke and the need for preventative action among adults from different black, Asian and other minority ethnic groups living in the UK?

Question 4: What are the cut-off points for BMI and waist circumference among adults from black, Asian and other minority ethnic groups living in the UK that are ‘risk equivalent’ to the current thresholds set for white European populations?

4.2. Search approach

NICE originally intended the searches to follow the traditional method of a database strategy, translated across several sources, supplemented with other techniques where necessary. The initial testing suggested that this would not be an effective use of time, as a MEDLINE strategy answering the four questions retrieved 26,385 hits. The screening would have been much higher once other resources, such as Embase and the Cochrane Library, were searched with this strategy.

There were two difficulties with the database search strategy. Firstly, it was difficult to isolate articles on appropriate cut-off points without also retrieving studies where changes in BMI were the main outcome being measured. This was compounded when additional search terms were added to describe waist-to-hip ratio, waist circumference and adiposity markers. Secondly, the focus was on ethnic minorities in a UK context. It was not possible to restrict the search to studies conducted in a UK setting because, for example, an Indian study on an Indian population would be relevant. On the other hand, searching for BMI cut-off points for “ethnic minorities” would not be appropriate (Pakistanis are clearly not a minority in Pakistan). It was equally difficult to list the ethnicities that were relevant in a strategy because the search term “Chinese” would retrieve articles about the country China and these would not necessarily be relevant (such as a weight management intervention measuring BMI improvements in a mixed population in Beijing).

A new approach was required because of the high volume of results and the difficulty of focussing the database strategy. NICE permits novel approaches where the subject in question and the type of evidence required suggest an alternative approach would be beneficial and avoid producing an unmanageable volume of results (NICE, 2012; NICE, 2014). This topic met both criteria. It was appropriate to follow an approach based on the berry-picking model (Bates, 1989) that would not require a comprehensive and pre-determined search strategy.

4.3. Guidance development methods

The first stage was to identify a core set of publications to be used as the base set for citation searching. Contacting experts has been an effective method for identifying publications when time is limited (Ganann et al., 2010), and this was deemed an appropriate initial step. NICE canvassed a small group of topic experts for key evidence, and they recommended six relevant publications. The bibliographies from the recommendations
were checked (reference harvesting) to produce a further 40 publications. A list of the 46 publications in the base set is available in the Supplementary Information. NICE issued a public call for evidence on its website for 6 weeks while the citation searching was undertaken to ensure a wider spread of evidence and to gather any unpublished studies.

The base set comprised 46 publications, and citation searching was undertaken in March 2012 using the 'cited by' function in GS. The first 100 citations for each of the 46 publications were copied into Microsoft Word for preliminary screening. The decision to screen only the first 100 citations for each publication was taken to ensure the screening could be completed in time. The reviewers double-screened a random sample of 10% to check consistency and then they screened half of the GS results each in Word. The items in the base set, the call for evidence submissions and the GS results selected in Word were entered into a Reference Manager database (Thomson Reuters: version 12.0.3), and duplicates were removed. One of the reviewers examined the full text of the records in Reference Manager to produce a long list of potentially relevant publications for further screening. A third reviewer undertook a full-text screen, using a refined set of criteria, and the selected publications were included in the first draft of the systematic review (Bazian, 2012). The draft review was presented to the NICE committee.

4.4. Supplementary study methods

The supplementary study was conducted by substituting WOS for GS in the guidance development search process (Figure 2). Citation searching was conducted on the same base set of 46 publications using the SSCI and SCI components of WOS in June 2012. All of the results from WOS were downloaded into Reference Manager, along with the base set and the call for evidence submissions, and checked for the publications identified by GS.

A set of results unique to WOS and not retrieved from GS was identified and screened separately for this study. The publications unique to WOS were screened by the same reviewers, following the GS process as far as possible. A random sample of 10% was double screened, followed by one reviewer examining the titles and abstracts to produce a long list, two reviewers reading the full-text publications and a third reviewer making the final selection.

4.5. Outcomes

The searches have been tested for recall and precision in order to explain the trade-off between finding relevant publications and avoiding unmanageable volumes of results (Lefebvre et al., 2011). Recall is calculated by dividing the number of relevant publications found by the number of relevant publications known to exist. Precision is calculated by dividing the number of relevant publications found during the search by the total number of results. A third measure used in this study is the number needed to read (NNR), which measures the number of publications that needs to be screened in order to identify one relevant publication. NNR is calculated as $1/\text{precision} \times 100$ (Bachmann et al., 2002).

Recall, precision and NNR were calculated using the whole retrieval process (base set, call for evidence and searching), as the purpose of the study was to replicate guidance production and not to compare WOS and GS
in isolation. The publications from the base set and the call for evidence would have been available to the reviewers regardless of what other search techniques were used. The relevant publications derived from the base set and the call for evidence were, therefore, counted as retrieved by the WOS or GS processes. NNR was calculated according to how many titles and abstracts would have to be read to show the overall burden of the screening process.

A subset of the results was identified as “highly important”, and the methods required to identify these publications were explored in more detail. The NICE committee decided that question 4 was its highest priority for making recommendations in the guidance. This means that the publications answering question 4 made the most impact on the guidance, and they have been defined as “highly important”.

5. Results

5.1. Did WOS retrieve more relevant publications than GS?

Google Scholar citation searching led to 2407 unique items being screened in Word, and these were combined with the base set and the call for evidence submissions to produce a Reference Manager database containing 872 unique records (Figure 1). The screening processes led to the third reviewer identifying 39 publications relevant to the first draft of the systematic review (Bazian, 2012). The process was replicated in WOS (Figure 2), and this resulted in 2342 unique results, of which 909 had not been retrieved from GS.

5.1.1. Test 1: did WOS find the same publications as GS?. The 39 publications relevant to the four review questions comprised 24 unique to GS, two unique to the base set, eight unique to the call for evidence and five found in both GS and the base set. The WOS process found 33 and missed six of the 39 publications that the GS process had retrieved. WOS itself found 25 of the publications, with an additional eight coming from the call for evidence.

WOS did not miss any publications that helped the committee to make its recommendations. Three of the publications missed by WOS (Lee et al., 2002; Gupta et al., 2005; Gupta et al., 2007) were not included in the final draft of the review (Bazian, 2012), one did not contribute to the recommendations (Shah et al., 2009) and one was relevant only to question three (Sarrafzadegan et al., 2010). The sixth publication was a PhD thesis (Nyamdorj, 2010) relevant to question 4, as discussed in the following.

5.1.2. Test 2: did WOS find additional publications that GS had missed?. Four of the 909 publications unique to WOS met the criteria for inclusion in the systematic review (Kaur et al., 2008; Boggs et al., 2011; Hwang et al., 2011; Jenum et al., 2012). Combining the WOS and GS processes together gives a total of 43 publications relevant to the four review questions (Figure 2). A list of the 43 relevant publications is available in the Supplementary Information.

The WOS process had 86.05% recall and 1.58% precision for the 43 publications, compared with 90.7% and 1.62%, respectively, with the GS process (Table 1). These GS figures relate to the decision to screen only the first 100 results for each publication, which led to 2407 publications being reviewed (Figure 1). GS actually stated that the base set had been cited 15,068 times, and screening all of these records would have seen 100% recall and 0.29% precision. In practice, Google prevents access to anything over 1000 records (Bramer et al., 2013), and so 7692 of the 15,068 (51%) citations were actually accessible. Screening the 7692 results would have found the four publications from the 43 that GS missed, taking recall to 100% and precision to 0.56%. The practicalities of dealing with GS had a demonstrable impact on the review, as the NNR to find a relevant publication would have nearly tripled from 61.72 with the cut off at 100 to 178.88 with the cut off at 1000, although recall would have increased from 90.7% to 100% (Table 1). The NNR to find one relevant publication in the WOS process was 63.3 (Table 1).

<table>
<thead>
<tr>
<th>Method</th>
<th>Relevant publications available</th>
<th>Relevant publications retrieved</th>
<th>No. of results</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
<th>Number needed to read</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS: first 100</td>
<td>43</td>
<td>39</td>
<td>2407</td>
<td>90.70</td>
<td>1.62</td>
<td>61.72</td>
</tr>
<tr>
<td>GS: first 1000</td>
<td>43</td>
<td>43</td>
<td>7692</td>
<td>100</td>
<td>0.56</td>
<td>178.88</td>
</tr>
<tr>
<td>GS: all citations</td>
<td>43</td>
<td>43</td>
<td>15,068</td>
<td>100</td>
<td>0.29</td>
<td>350.42</td>
</tr>
<tr>
<td>WOS: all citations</td>
<td>43</td>
<td>37</td>
<td>2342</td>
<td>86.05</td>
<td>1.58</td>
<td>63.30</td>
</tr>
</tbody>
</table>

GS, Google Scholar; WOS, Web of Science.
5.2. Did WOS retrieve more highly important publications than GS?

Nine of the 43 publications were relevant to question 4 and categorised as highly important (Stevens et al., 2002; Pan et al., 2004; Stevens et al., 2008; Nyamdorj et al., 2009; Cameron et al., 2010; Stommel and Schoenborn, 2010; Taylor et al., 2010; Chiu et al., 2011; Jenum et al., 2012).

5.2.1. Test 3: did WOS find the same highly important publications as GS? All nine of the highly important publications were derived from citation searching, and none of them originated from the base set or the call for evidence. The WOS process did not miss any highly important data.

One of the six publications retrieved by GS and missed by WOS was a PhD thesis (Nyamdorj, 2010). The thesis was not defined as highly important because it did not contribute any unique data to the systematic review. The reviewers identified during guidance development that Nyamdorj had cited the data from the thesis in a peer-reviewed journal article (Nyamdorj et al., 2009). The graphs and tables relevant to question 4 were contained in both publications, and so the reviewers decided to extract data using only the journal article. No highly important data were missed because WOS retrieved the journal article.

5.2.2. Test 4: did WOS find additional highly important publications that GS had missed? WOS outperformed GS in terms of retrieving the highly important publications (Table 2). Three of the four publications found by WOS and missed by GS addressed question 2 (Kaur et al., 2008; Boggs et al., 2011; Hwang et al., 2011), and the fourth was highly important as it was relevant to question 4 (Jenum et al., 2012). This publication would have been retrieved by GS had the results not been cut at 100 and all of the 7692 available citations had been screened.

The WOS process had 100% recall and 0.38% precision, in terms of the nine highly important publications, while the GS process had 88.89% recall and 0.33% precision (Table 2). This means that screening the titles and abstracts from 260.22 WOS results produced a highly important publication, compared with 300.88 with GS. Screening all of the 7692 results would have increased GS recall to 100% at the expense of precision falling to 0.12% and NNR increasing to 854.67 (Table 2).

6. Discussion

A number of themes emerged from the comparison, and they are important considerations when choosing whether to use WOS or GS. The citation searching produced similar results in terms of the volumes to be screened and the recall (Table 1). It is not sufficient to draw conclusions from these figures without considering the practical issues involved in achieving them.

6.1. Ease of searching

The challenge with trying to conduct searches more rapidly than in a traditional systematic review is that they need to maintain the same transparent and robust standards (Ganann et al., 2010). The GS search was not transparent and it is probably not reproducible. It is difficult to maintain transparency and reproducibility if Google only displays the first 1000 results and then a limit of 100 is introduced to keep the volumes manageable. It is impossible to know what would have been retrieved without cutting the results at 100 and, as the results are ordered by Google’s algorithm, the same 1000 hits would not be retrieved when repeating the search. It is likely that the GS results were biased towards older publications, as newer ones will have had less opportunity to be cited and have fewer links to them.

There were no practical problems with WOS, and all of the publications citing the base set were downloaded for screening, which had a manageable number of results with an NNR of 63.3. The searcher can be confident that the WOS editorial process will have identified genuine citations, and there is no need to decide how many of the results to download. It is much easier to defend a search as robust where all of the citations are screened, rather than selecting a manageable number.

<table>
<thead>
<tr>
<th>Method</th>
<th>Highly important publications available</th>
<th>Highly important publications retrieved</th>
<th>No. of results</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
<th>Number needed to read</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS: first 100</td>
<td>9</td>
<td>8</td>
<td>2407</td>
<td>88.89</td>
<td>0.33</td>
<td>300.88</td>
</tr>
<tr>
<td>GS: first 1000</td>
<td>9</td>
<td>9</td>
<td>7692</td>
<td>100</td>
<td>0.12</td>
<td>854.67</td>
</tr>
<tr>
<td>GS: all citations</td>
<td>9</td>
<td>9</td>
<td>15068</td>
<td>100</td>
<td>0.06</td>
<td>1674.22</td>
</tr>
<tr>
<td>WOS: all citations</td>
<td>9</td>
<td>9</td>
<td>2342</td>
<td>100</td>
<td>0.38</td>
<td>260.22</td>
</tr>
</tbody>
</table>

GS, Google Scholar; WOS, Web of Science.
than one where a cut-off point was imposed to help manage the time. WOS helps to guard against the accusation that a more rapid search will lead to a less systematic approach.

6.2. Ease of downloading results

It is important to keep detailed records of how search results have been handled to make systematic reviews transparent and replicable (Moher et al., 2009). It is just as important to record the citation searching process as it is a database strategy, even though it might be more challenging (Rader et al., 2013). Handling the GS data placed a considerable administrative burden on this review as Google has deliberately blocked users from downloading the bibliographic details in batches (Google, 2013). Manually entering 737 GS results into Reference Manager not only had a major time impact but also increased the potential for data-input errors. It would have been more efficient to handle all of the GS results in a single Reference Manager file, instead of having to do a preliminary screen in Microsoft Word to save time.

It is possible to import a whole page of GS results at once into Zotero, an alternative reference management application, although this is less useful now that Google has reduced a page from 100 to 20 results (Bramer et al., 2013). This confirms the findings in Wright et al. (2014), where downloading 1680 individual records “significantly increased” the administration time required. WOS allows users to download results in batches of 500, and it was a simple process to download seven files and upload them into Reference Manager before removing duplicates from the 3211 results.

Once the bibliographic details had been entered into Reference Manager, it was then necessary to find the abstracts as GS did not export these automatically. It was time-consuming to locate the abstracts for 737 publications, even if GS provided links to them. The WOS export function provided both the bibliographic details and the available abstracts in the same operation.

6.3. Costs of administering results

A full cost appraisal of the searching components has not been attempted, as this would involve counting the hours spent in meetings, developing the search and handling results, as well as accounting for overheads and equipment (McGowan and Sampson, 2005). Two cost components have been considered: administering the citation searching results and the time to screen them.

It took approximately 4 h to use WOS, including 46 citation searches, downloading seven batches of results, adding records to Reference Manager and removing duplicates. It took 10 working days (75 h) to process the GS results, including using the ‘cited by’ function on the base set, copying around 315 pages of results into Word, identifying duplicates, entering the bibliographic details of 737 publications into Reference Manager and locating abstracts.

The administrative time required with GS had cost implications, as well as delaying the systematic review. The full on-costs to NICE in 2012 for the relevant staff grades ranged from £22 to £34 per hour. This means that 4 h of WOS administration cost £88–£136 and 75 h processing the GS results cost £1650–£2550 (Table 3). The results from the free-to-access GS cost between £1562 and £2414 more to administer than those from the subscription-based WOS. The WOS subscription would not just be used for one review, and its annual cost should be divided across the range of NICE products developed that year. Organisations would need to make an individual decision on a cost-effective price for them, as database subscriptions usually relate to several factors, such as the type of institution and the number of users.

6.4. Costs of screening results

The second cost element considered is the impact of citation searching on screening costs. Records kept by the reviewers suggest that an average of 62 publications could be screened per hour at the title and abstract stage. The WOS NNR of 63.3 and the GS NNR of 61.72 (Table 1) suggest the reviewers would find nearly one relevant publication every hour. The WOS and GS processes would have taken similar amounts of time to screen, with 38 h (costing £836–£1292) and 39 h (£858–£1326), respectively (Table 3).

There is a clearer difference in the screening resources required to find the highly important publications. The NNR of 260.22 with the WOS process (Table 2) means that one highly important paper would be located about every 4.2 h when screening the results. The NNR of 300.88 with the GS process equates to one highly important paper for every 4.85 h of screening to achieve 88.89% recall (Table 2). Increasing the recall with GS to 100% would have tripled the screening time to 124 h to process 7692 results, and the costs would have risen to £2728–£4216 just to find one more highly important publication. The screening rate of 62 publications per hour over a 7.5 h working day means that any NNR above 465 is equivalent to adding at least a day’s screening to find one relevant publication. The GS process had an NNR of 854.67 (Table 2) to achieve 100% recall, and this would require 13.8 h of screening for every highly important paper.

The results in this study contrast with Wright et al. (2014), which found WOS had 20% recall in a review of multiple-behaviour interventions. The conclusion in Wright et al. (2014) that citation searching “may not be the best use of valuable time and resources” relates to their objective to test “whether citation searching identified
any relevant records in addition to those retrieved by the original database searches”. The key difference is that Wright et al. (2014) treated citation searching as a supplementary method that could only be undertaken after database searching, whereas this study has investigated citation searching as the primary method.

This study has sought to help move searching away from a reliance on comprehensiveness towards using methods that will locate the best evidence more quickly when it is appropriate to do so. Citation searching would be particularly useful for policymakers requiring a rapid evidence review where reliable evidence needs to be found quickly. It is not appropriate to recommend one optimal approach that will be suitable for all reviews, and the extensive database search still has its role when it would be the most effective retrieval method. The choice of methods to make reviewing more efficient needs to account for the time and resources available, the purpose of the review and the policymaker’s requirements (Thomas et al., 2013).

7. Limitations and further research

The comparison is limited by what might be called linking lag. GS was searched in March 2012, and more citations might have been linked to the base set by the time WOS was searched in June 2012. The WOS records do not show the date on which a citation link has been added, and so it is impossible to replicate which citations it would have retrieved in March 2012. One of the disadvantages of citation searching is that it takes time for more recent publications to be cited and for these links to be picked up in WOS or GS. It would be worth investigating the extent to which citation searching would benefit from being supplemented by some focussed searches of PubMed to ensure very recent publications are included.

It is plausible that screening higher yields would have identified further relevant studies. Google does not provide access to the full set of 15 068 results, but additional, currently unknown, publications could have been located in the available set of 7692. This study began with the realisation that it was not feasible to screen the 26 385 MEDLINE results and that it is not known whether they contain other relevant publications. The resources were not available to screen the additional GS and MEDLINE results either for the systematic review or this supplementary study.

The functionality and scope of WOS made it the most appropriate comparator with GS for the NICE review. Other resources are available for citation searching, and Scopus (http://www.scopus.com/) has been compared with WOS and GS in other studies (Falagas et al., 2008; Kulkarni et al., 2009; Li et al., 2010; Wright et al., 2014). The resources were not available to conduct a wider study, but a direct comparison between WOS, GS and Scopus would have merit.

This case study has analysed the effectiveness and efficiency of using citation searching in a single review. A comparison of how citation searching performs in other topic areas would be beneficial to assess whether these conclusions have more general applicability, for example across the other 50 areas of NICE public health guidance. It would also be worth pursuing these methods in other disciplines that deploy complex interventions, have a disparate evidence base or lack controlled trials. It would be particularly useful to examine whether citation searching can help to overcome some of the difficulties experienced in finding evidence on social care topics (Clapton, 2010) since NICE gained additional responsibilities in this area in April 2013.

The costs of developing the base set were not calculated for this study because they were the same for WOS and GS. It would be interesting to compare the cost of developing a base set for citation searching, reference harvesting or other methods with the time taken to develop, test and implement a database search strategy. Further research is required to explore how the timing of citation searching and its relationship to database searching affect its value, given the differences between this study and Wright et al. (2014).

<table>
<thead>
<tr>
<th>Activity</th>
<th>Method</th>
<th>No. of results</th>
<th>Highly important publications retrieved</th>
<th>Hours</th>
<th>Minimum cost (£)</th>
<th>Maximum cost (£)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1</td>
<td>22</td>
<td>34</td>
</tr>
<tr>
<td>Admin</td>
<td>GS: first 100</td>
<td>3176</td>
<td>8</td>
<td>75</td>
<td>1650</td>
<td>2550</td>
</tr>
<tr>
<td></td>
<td>WOS: all citations</td>
<td>3211</td>
<td>9</td>
<td>4</td>
<td>88</td>
<td>136</td>
</tr>
<tr>
<td>Screening</td>
<td>GS: first 100</td>
<td>2407</td>
<td>8</td>
<td>39</td>
<td>858</td>
<td>1326</td>
</tr>
<tr>
<td></td>
<td>GS: first 1000</td>
<td>7692</td>
<td>9</td>
<td>124</td>
<td>2728</td>
<td>4216</td>
</tr>
<tr>
<td></td>
<td>WOS: all citations</td>
<td>2342</td>
<td>9</td>
<td>38</td>
<td>836</td>
<td>1292</td>
</tr>
</tbody>
</table>

GS, Google Scholar; WOS, Web of Science.
8. Conclusions

The choice of which resource to adopt for citation searching depends on the practicalities of using WOS and GS. WOS was more reliable than GS to search, downloading the results was quicker and the costs of administering the results were lower. The practical difficulties experienced with GS meant that the full search results could not be screened, and this led to one highly important publication being missed. The WOS process performed better than the GS process in retrieving 100% of the highly important publications. Administering results from WOS took 4 h and cost £88–£136, compared with 75 h costing £1650–£2550 with GS. These performance and practical issues have to be offset against the cost of a WOS subscription. The results are a reminder to budget holders that resources that are free at the point of access do not necessarily represent better overall value for money than subscription services.

This retrospective analysis found that forwards citation searching was useful for finding the best available evidence without creating an unmanageable volume of results. Citation searches can be developed in a series of focussed steps that avoid unnecessary amounts of results, while maintaining a systematic approach. The new NICE methods for developing guidance (NICE, 2014) place greater emphasis on using a range of search techniques, including citation searching. NICE recommends using these techniques where the topic of the review or the type of evidence required suggest they are “reasonably likely” to be more productive than databases (NICE, 2014). The searches for BMI in this study will encourage NICE to utilise forwards citation searching and to use WOS for this purpose when developing public health guidance.

9. Abbreviations

BMI  Body Mass Index
GS   Google Scholar
NICE National Institute for Health and Care Excellence
NNR Number Needed to Read
SCI  Science Citation Index
SSCI Social Science Citation Index
WOS  Web of Science

10. Declaration

The authors confirm that they have no competing financial or other interests.

References


Supporting information

Additional supporting information may be found in the online version of this article at the publisher’s web site.