

# 1    **A comparative analysis reveals irreproducibility in searches of scientific**

## 2    **literature**

### 3    **Short title: Search location and the reproducibility of systematic reviews**

4    Gábor Pozsgai<sup>1,2\*</sup>, Gábor L. Lövei<sup>1,2,3</sup>, Liette Vasseur<sup>1,2,4</sup>, Geoff Gurr<sup>1,2,5</sup>, Péter Batáry<sup>6,7</sup>, János  
5    Korponai<sup>8,9,10</sup>, Nick A. Littlewood<sup>11</sup>, Jian Liu<sup>5</sup>, Arnold Móra<sup>12</sup>, John Obrycki<sup>13</sup>, Olivia Reynolds<sup>2,14,15</sup>,  
6    Jenni A. Stockan<sup>16</sup>, Heather VanVolkenburg<sup>4</sup>, Jie Zhang<sup>1,2</sup>, Wenwu Zhou<sup>17</sup>, and Minsheng You<sup>1,2\*</sup>

7

8    <sup>1</sup>State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, Institute of Applied  
9    Ecology, Fujian Agriculture and Forestry University, Fuzhou 350002, China (GPozsgai: ORCID #0000-  
10   0002-2300-6558, MYou ORCID # 0000-0001-9042-6432

11   <sup>2</sup>Joint International Research Laboratory of Ecological Pest Control, Ministry of Education, Fuzhou  
12   350002, China

13   <sup>3</sup>Department of Agroecology, Flakkebjerg Research Centre, Aarhus University, 4200 Slagelse, Denmark

14   <sup>4</sup>UNESCO Chair on Community Sustainability: From Local to Global, Dept. Biol. Sci., Brock University,  
15   Canada (LVasseur: ORCID #0000-0001-7289-2675)

16   <sup>5</sup>Graham Centre for Agricultural Innovation (Charles Sturt University and NSW Department of Primary  
17   Industries), Orange NSW 2800, Australia

18   <sup>6</sup>Agroecology, University of Goettingen, 37077 Goettingen, Germany

19   <sup>7</sup>“Lendület” Landscape and Conservation Ecology, Institute of Ecology and Botany, MTA Centre for  
20   Ecological Research, 2163 Vácrátót, Hungary

21      <sup>8</sup>Department of Water Supply and Sewerage, Faculty of Water Science, National University of Public  
22      Service, 6500 Baja, Hungary

23      <sup>9</sup>Department of Environmental Sciences, Sapientia Hungarian University of Transylvania, 400193 Cluj-  
24      Napoca, Romania

25      <sup>10</sup>Eötvös Lórand University, 1053 Budapest, Hungary

26      <sup>11</sup>University of Cambridge, Cambridge, UK

27      <sup>12</sup>University of Pécs, 7622 Pécs, Hungary

28      <sup>13</sup>University of Kentucky, Lexington, USA

29      <sup>14</sup>cesar, 293 Royal Parade, Parkville, Victoria 3052, Australia

30      <sup>15</sup>Biosecurity and Food Safety, NSW Department of Primary Industries, Narellan, NSW 2567, Australia

31      <sup>16</sup>Ecological Sciences, The James Hutton Institute, Aberdeen, UK

32      <sup>17</sup>State Key Laboratory of Rice Biology, Key Laboratory of Molecular Biology of Crop Pathogens and  
33      Insects, Ministry of Agriculture, Zhejiang University, Hangzhou, China

34

35      \*Corresponding authors: [msyou@fafu.edu.cn](mailto:msyou@fafu.edu.cn), [pozsgaig@coleoptera.hu](mailto:pozsgaig@coleoptera.hu)

36

37 **Abstract**

38 **Repeatability is the cornerstone of science and it is particularly important for systematic reviews.**  
39 **However, little is known on how database and search engine choices influence replicability. Here,**  
40 **we present a comparative analysis of time-synchronized searches at different locations in the world,**  
41 **revealing a large variation among the hits obtained within each of the several search terms using**  
42 **different search engines. We found that PubMed and Scopus returned geographically consistent**  
43 **results to identical search strings, Google Scholar and Web of Science varied substantially both in**  
44 **the number of returned hits and in the list of individual articles depending on the search location**  
45 **and computing environment. To maintain scientific integrity and consistency, especially in**  
46 **systematic reviews, action is needed from both the scientific community and scientific search**  
47 **platforms to increase search consistency. Researchers are encouraged to report the search location,**  
48 **and database providers should make search algorithms transparent and revise access rules to titles**  
49 **behind paywalls.**

50

51 **Key words: Database, search engine, search location, repeatability**

52

## 53      **Introduction**

54      Since the 17<sup>th</sup> century and Newton's strict approach to scientific inquiry[1], research has increasingly  
55      relied on rigorous methodological constraints. One of the cornerstones of the scientific method is  
56      reproducibility. However, a recent study shows that most scientists believe that a substantial proportion of  
57      methods published in peer-reviewed papers are not reproducible, creating a 'reproducibility crisis'[2].  
58      Following similar arguments, narrative reviews are increasingly being replaced by systematic reviews,  
59      also called "evidence-based synthesis"[3]. Transparency and repeatability are also cornerstones of this  
60      method of knowledge synthesis. However, the repeatability of systematic reviews remains rarely  
61      examined. Though repeatability in such studies is of utmost importance, and detailed protocols are  
62      available[4,5], the technical aspects of these underpinning databases and search engines have not been  
63      systematically tested and, at present, there is no recommendation on these technical aspects.

64      As primary scientific literature is rapidly expanding[6], scientists are unable to keep track of new  
65      discoveries by focusing only on the primary literature[7,8], so systematic reviews have become  
66      increasingly important[9]. Recognized weaknesses of the traditional, narrative reviews include the non-  
67      transparency of the literature selection process, evaluation criteria, and eventual level of detail devoted to  
68      individual studies[10]. With the advent and rapid development of Internet-based databases and search  
69      engines, the role of narrative reviews is now being overtaken by new, quantitative methods of evidence  
70      synthesis[11,12]. A core requirement in these activities, repeatability, crucially depends on reliable  
71      databases[13]. Large scientific databases/search engines, such as PubMed, Web of Science and Scopus,  
72      are essential in this process. They have been primary electronic search engines for scientists since 1997  
73      with the inauguration of PubMed[14]. Today, nearly all scientists working on various forms of evidence-  
74      based synthesis use these databases/search engines to find relevant papers as the basis for further analysis.  
75      An important condition in the whole process is that the evidence base must be solid: a given search string  
76      in a database should generate identical results, independent of search locations, provided the searches are

77 running at the same time. If this assumption were violated, it would have serious consequences for the  
78 reliability and repeatability of the data and papers selected for a specific systematic review. Therefore,  
79 there is a need to know what variables and/or parameters should be included in the methodology of any  
80 search to ensure its repeatability. One of the most crucial steps is to define which database and engine  
81 search is going to be used for obtaining the data to be synthesized.

82 Differences among the most commonly used scientific search engines and databases are well  
83 documented[13,15,16] but knowledge of the consistency within databases in relation to geographical  
84 location where the search is requested from (but see Gusenbauer and Haddaway[13]), software  
85 environment, or computer configuration remain surprisingly limited. Since the search histories of users  
86 may be stored in the browsers' cache, and considered by the scientific search engines, repeated and  
87 identical searches may result in different outcomes. During a recent systematic review in ecology, we  
88 accidentally discovered that a multi-locus search performed on 1 February 2018, using an identical search  
89 string in Web of Science, produced radically different number of hits at different institutions at Hangzhou  
90 and Fuzhou, in China, and in Denmark (2,394, 1,571, and 7,447, respectively).

91 Since there is no known study comparing the consistency of returned papers over successive identical  
92 searches using several databases in one machine, we examined the way search engines deliver results and  
93 decided to systematically explore the inconsistencies found. Our study aimed to evaluate the consistency  
94 of search engines by comparing the outcomes from identical search strings ran on different computers  
95 from a wide range of localities across the world, with various software backgrounds, and using different  
96 search engines.

97 To investigate the repeatability of scientific searches in four of the major databases and search engines,  
98 Web of Science, Scopus, PubMed, and Google Scholar, we generated search strings with two complexity  
99 levels in ecology and medicine and ran standardized searches from various locations in the world, within  
100 a limited timeframe. According to our null hypothesis, every search engine should give the exact same

101 number of results to the same search (after the search term has been adjusted to match the specific  
102 requirements for each of these search engines), and therefore, a metric, showing the proportional deviance  
103 of the search hits, should always be zero. We, therefore, first tested if summarized *average absolute*  
104 *deviation proportions* (AADPs) for each search engine were significantly different from the ideal value  
105 (zero) by using robust non-parametric tests. AADPs of search engines were compared to each other and  
106 factors driving the differences were investigated. Similarly, the publications found by any given search  
107 engine from identical searches should be also identical, thus, the mean similarities between search runs  
108 should be 100%, and the scatter of the ordinated points should be zero. In order to test whether these  
109 requirements were met, Jaccard distances[17] of the first twenty hits were used for within and between  
110 group ordinations and multivariate analysis.

## 111 **Results**

112 Our time-synchronized, cross-institution and multi-location search exercise resulted in a large variation  
113 among the hits obtained using any of the search terms. Google Scholar generally yielded a greater number  
114 of hits than any other databases for all the locations (Table 1). As expected, less complex and medical  
115 search terms tended to result in greater hit numbers than complex ecological ones.

116 **Table 1. Comparison of the mean numbers of hits (SD) resulting from simple vs. complex search strings in the fields of ecology**  
117 **and medicine using different search engines, different browsers and cache handling**

Search Engine	Browser	Cache	Number of hits of search strings in thousands			
			Ecology		Medicine	
			Simple	Complex	Simple	Complex
Google Scholar	Chrome	Full	1157.188± 991.840	2.069± 1.663	1165.170± 1167.252	28.117± 25.262
		Cleaned	871.186± 1065.303	1.595± 1.699	1013.800± 1178.801	22.718± 25.643
	Internet Explorer	Full	1077.496± 1018.818	1.945± 1.685	1263.791± 1154.650	28.140± 25.266
		Cleaned	862.614± 1054.802	1.595± 1.699	1012.371± 1177.043	22.689± 25.608
	Firefox	Full	905.849± 1026.956	1.945± 1.684	1263.791± 1154.650	28.113± 25.266

		<b>Cleaned</b>	985.978± 1036.853	1.816± 1.693	1169.975± 1179.213	26.100± 25.602
<b>PubMed</b>	<b>Chrome</b>	<b>Full</b>	2.881± 0.001	0.006± 0	147.726± 0.030	0.233± 0
		<b>Cleaned</b>	2.881± 0.001	0.006± 0	147.727± 0.030	0.233± 0
	<b>Internet Explorer</b>	<b>Full</b>	2.881± 0.001	0.006± 0	147.729± 0.030	0.233± 0
		<b>Cleaned</b>	2.881± 0.001	0.006± 0	147.734± 0.030	0.233± 0
	<b>Firefox</b>	<b>Full</b>	2.881± 0.001	0.006± 0	147.728± 0.030	0.233± 0
		<b>Cleaned</b>	2.881± 0.001	0.006± 0	147.731± 0.030	0.233± 0
	<b>Scopus</b>	<b>Chrome</b>	19.912± 0	0.078± 0	545.558± 0	0.711± 0
		<b>Cleaned</b>	19.912± 0	0.078± 0	545.558± 0	0.711± 0
		<b>IE</b>	19.912± 0	0.078± 0	545.558± 0	0.711± 0
		<b>Cleaned</b>	19.912± 0	0.078± 0	545.558± 0	0.711± 0
		<b>Mozilla</b>	19.912± 0	0.078± 0	545.558± 0	0.711± 0
		<b>Cleaned</b>	19.912± 0	0.078± 0	545.558± 0	0.711± 0
<b>Web of Science</b>	<b>Chrome</b>	<b>Full</b>	17.295± 1.214	15± 0	190.899± 24.163	0.357± 0.041
		<b>Cleaned</b>	17.561± 0.798	15± 0	195.432± 22.271	0.367± 0.026
	<b>Internet Explorer</b>	<b>Full</b>	17.642± 0.740	15± 0	200.904± 15.646	0.373± 0.018
		<b>Cleaned</b>	17.587± 0.832	15± 0	199.665± 17.580	0.372± 0.020
	<b>Mozilla</b>	<b>Full</b>	17.492± 0.967	14.9± 0.49	192.108± 24.784	0.364± 0.031
		<b>Cleaned</b>	17.370± 0.978	14.8± 0.55	203.694± 38.988	0.36± 0.035

118

119 The AADP (see Materials and Methods) of every search engine and database, except Scopus,  
120 significantly deviated from the desirable zero (Table 2). However, we have noticed that both PubMed and  
121 Web of Science were updated during the search process, at 17:00 GMT and 19:00 GMT, respectively.  
122 When the results from PubMed and Web of Science were split into two groups, before and after the time  
123 of the daily update, none of the AADPs from PubMed searches significantly differed from zero. In  
124 contrast, the results from Web of Science searches consistently showed significant deviation, indicating  
125 inconsistency in the number of returned hits by search location.

126 **Table 2. Mean and standard deviations of recorded average absolute deviation proportions (AADP) for each investigated**  
127 **search engines, separated by search topic and search expression complexity. Values are shown in percentage.**

Topic/Complexity	GScholar	PubMed	Scopus	WoS
Ecology/Complex	85.319±9.426	0.000±0.000	0.000±0.000	0.629±1.964
Ecology/simple	98.107±4.063	0.035±0.000	0.000±0.000	4.009±3.459
Medicine/Complex	90.889±5.966	0.000±0.000	0.000±0.000	5.845±5.757
Medicine/simple	94.609±2.964	0.014±0.000	0.000±0.000	7.818±9.852

128

129 The WelshADF test revealed significant differences in AADPs among groups (92.45% variance  
130 explained), with search engines being the most important explanatory variable (WJ = 69265.22, df = 3, p  
131 < 0.001). Effects of the search topic (WJ = 8.49, df = 1, p = 0.005), keyword complexity (WJ = 71.71, df  
132 = 1, p < 0.001), the interaction of search topic and keyword complexity (WJ = 20.40, df = 1, p < 0.001),  
133 and their combination with search engine (Search engine × Topic: WJ = 11959.03, df = 3, p < 0.001,  
134 Search engine × Keyword complexity: WJ = 61790.69, df = 3, p < 0.001) on the outcome were all  
135 significant. The effect of browsers used was not significant, either alone (WJ = 0.06, df = 2, p = 0.941) or  
136 as a covariant of search engine choice (WJ = 0.29, df = 6, p = 0.943). Cache, whether it was emptied or  
137 not, did not have a significant effect, either in its own or as a covariant (Fig 1, Supplementary Information  
138 1, Supplementary Information 2-3). In spite of not being a significant predictor in the entire dataset, both  
139 browser and cache showed a tendency to influence the outcome of the Google Scholar results. None of  
140 these influenced the search platforms with a background database. There were no differences in search  
141 results when using Web of Science, PubMed and Scopus but different machines at the same location but  
142 Google Scholar sometimes produced different results.

143 **Fig 1. Average absolute deviation proportions (AADP) of hit numbers**

144 AADPs are grouped by searched platforms, and separated by keyword complexity (complex, simple), and  
145 research area (ecology, medicine). Boxes represent interquartile range (IQR), with median AADP values  
146 represented as a thick horizontal band. Whiskers extend from Q1-1.5IQR to Q3+1.5IQR. Abbreviated  
147 search platforms: GScholar – Google Scholar, WoS – Web of Science.

148

149 The multivariate analysis run on the first twenty papers collected from each search revealed significant  
150 differences among the search engines ( $p = 0.01$ ) but did not show a significant influence on browser  
151 choice or cache state. Areas of convex hulls defined by these ‘paper-communities’ (see Methods) of the  
152 first twenty hits were zero for Scopus only, and they were the largest for Google Scholar (Table 3). When  
153 PubMed and Web of Science datasets were split by their update time, hulls for both PubMed subsets  
154 became zero but remained greater than zero for Web of Science. Distance measures showed an analogous  
155 pattern; they were zero for Scopus, indicating no difference between the first twenty papers, and deviated  
156 from zero for all other platforms (Fig 2). After correcting for the database update, only Web of Science  
157 and Google Scholar hulls remained significantly greater than zero.

158 *Table 3 Areas of complex hulls for each search engines, separated by terms of topic and complexity.*

Topic/Complexity	GScholar	PubMed	Scopus	WoS
Ecology/Complex	491.90	0.00	0.00	0.00
Ecology/simple	322.24	490.37	0.00	8.82
Medicine/Complex	476.45	4.99	0.00	0.02
Medicine/simple	625.03	428.56	0.03	41.81

159

160 **Fig 2. Average similarities of the first twenty papers within each search engine-topic-keyword**  
161 **complexity group, for each search platform.**

162 Similarities were calculated based on binary matrices, using Jaccard distances. Median similarities are  
163 indicated with a thick black line on the pirate plots. Abbreviated search platforms: GScholar – Google  
164 Scholar, WoS – Web of Science.

165 **Discussion**

166 In this study, we identified a shortcoming of scientific search platforms that can decrease the transparency  
167 and repeatability of the synthesis of quantitative evidence synthesis relying on database searches. Hence,  
168 the creditability and reliability of the conclusions drawn from these syntheses may be compromised.

169 Our results showed significant differences in search platform consistency in terms of both the number of  
170 hits (the size of the body of available evidence) and its composition when identical search terms were  
171 queried at different geographic locations. We found that PubMed and Scopus had high consistencies,  
172 whilst Google Scholar and Web of Science were not consistent in the number of hits they returned.  
173 Google Scholar provided the greatest number of hits for every search, it also proved to be the least  
174 consistent among different search runs, varying greatly in the number of hits, i.e. the total number of  
175 papers. Contrarily, the composition of the evidence collected, characterized by the first twenty papers it  
176 returned, was relatively consistent. Web of Science, however on a lower magnitude, showed similarly  
177 poor consistency in terms of the number of hits returned from identical searches initiated from different  
178 locations. Both the hit numbers and the returned list of articles from Scopus searches were consistent.  
179 PubMed varied in hit numbers and had great dissimilarities among the returned sets of papers, especially  
180 in those related to more general searches that necessarily had more hits. These dissimilarities were likely  
181 due to a database update that happened during our search exercise. Indeed, data showed that 0, 6, 10, 25  
182 papers (complex ecology, complex medicine, simple ecology, and simple medicine terms, respectively)  
183 were added to the database during the course of this worldwide exercise. Since the papers listed were  
184 ordered according to their time of inclusion in the dataset, the first 20 collected papers would greatly  
185 differ and especially the larger values in the newly added articles can cause a disproportionately large  
186 effect on the similarity of the 20 collected papers. Once the differences before and after database update  
187 were accounted for, PubMed showed no deviation either in the number of returned papers or the list of the  
188 first 20 listed papers. A similar change in the dataset happened with Web of Science during our search,  
189 but differences remained even after correcting for the update. This suggests that discrepancies were

190 caused by other sources, such as geographic locations. Overall, in our tests, Scopus and PubMed proved  
191 to be the most consistent databases, and Web of Science and Google Scholar produced highly inconsistent  
192 results.

193 Although we could not thoroughly decipher the influence of browser or cache on the search results, there  
194 was an indication that these factors only affected Google Scholar outcomes. Google Scholar is known to  
195 optimize search hits according to the search history of its users, thus, even the differences between  
196 browsers are likely to be the results of participants' previous browser use, and therefore different cache  
197 contents in different browsers.

198 While the disadvantages of the inconsistencies in Google Scholar search results have been repeatedly  
199 illustrated[18,19], the similar behavior from Web of Science has only recently been reported[13] but in  
200 neither case was the variability estimated nor were the potential solutions discussed. Given the  
201 widespread use of Web of Science, neglecting this discrepancy can mislead scientists when drawing  
202 conclusions from their evidence synthesis, when the body of evidence was collected by Web of Science  
203 searches alone. The use of only one database is generally discouraged[5], and although some authors  
204 mainly target Google Scholar-based reviews[18,20], it is clear here that relying on Web of Science alone,  
205 or another single source, may lead to missing data or can make data-synthesis studies irreproducible. In  
206 spite of the recommendations of the need to use multiple sources for such studies (see the PRISMA  
207 statement[4]), a rapid scan of 20 recent papers in leading journals showed that recent, potentially highly  
208 cited, ecology-related systematic reviews still used Web of Science as their only search engine  
209 (Supplementary Information 4). In the light of the fact that using inadequate databases/search engines  
210 makes systematic reviews unreliable, our findings are concerning.

211 There are various means of overcoming this issue:

212 a) Researchers conducting systematic reviews should be aware of this potential problem, and be explicit  
213 about the methodology they use to ensure sufficient consistency and replicability. A detailed description

214 should be included on the search engines used (ideally more than one), search dates, the exact search  
215 strings, as well as whether the same search was replicated by more than one person. As our study showed,  
216 the location from which the search was conducted should also be reported, preferably along with the IP  
217 address of the computer and the locality/institution the queries were initiated from. The exact time of the  
218 search or the time window of the query are also essential. The holdings of databases, however, are not  
219 constant, historical records can be added over time, and, therefore, queries even within a clearly limited  
220 time period can deliver different result sets. Thus, reporting the time window of the queries can provide  
221 only a partial solution.

222 b) The use of adequate search engines for a particular task should be an important consideration. All of  
223 the large databases have different strengths; Google Scholar searches grey literature, Web of Science has  
224 the largest (combined) dataset and, as our study confirmed, that Scopus and PubMed are the most  
225 consistent. Moreover, some databases may be more suitable for collecting information on a particular  
226 topic or have a greater historical coverage than others[14].

227 c) Providers of scientific search platforms should consider opening their search code and moving their  
228 paywalls to make reference lists publicly available[21], thus contributing to search consistency, and hence,  
229 scientific repeatability. Particularly Web of Science, as the most commonly used search engine, should  
230 act on making its search hits equally reachable to all users and, rather than *a priori* filtering them  
231 according to the institutions' paywall, restrict access only *after* the primary result set has been provided to  
232 the user.

233 d) Google Scholar, on the other hand, should open its computer code to allow researchers to understand  
234 how hit lists are generated and how results are ordered. Google Scholar has been criticized by the  
235 scientific community for the obscurity of its search algorithms[22]. Although we acknowledge that this  
236 can be against business policies for some companies, we argue that compromises must be made for the  
237 sake of research integrity and scientific rigor.

238 e) Providing well-documented, standard application programming interfaces (APIs) and generating  
239 unique identifiers for searches, combining search term, result list, search time and location, and additional  
240 metadata (e.g. computing environment) is required. Using an API for standardized searches would be  
241 particularly beneficial for searches using Google Scholar that shows a strong dependence on the  
242 computing environment. Although this solution could control for a great deal of variation derived mostly  
243 from computing background and would be able to keep detailed records on the metadata of the searches,  
244 it also brings up novel challenges. Firstly, APIs can be more complex to use than simple web interfaces  
245 that may discourage users to use them. Moreover, collecting detailed data about search locations, or even  
246 computing environment, raises both security and privacy concerns. Finally, storing individual searches  
247 along with the necessary metadata may be resource heavy over a long period of time, which is likely to  
248 increase maintenance costs, and therefore the subscription fees, of these services.

249 Should these steps towards ensuring repeatability not happen, the critical voices to web-based systematic  
250 reviews can claim unreliability of this method[11]. Given that the systematic review methodology was  
251 originally developed to handle contentious issues with various, often conflicting bodies of evidence[5],  
252 this is a critical issue. This matter can only be exacerbated by the appearance of automatic systematic  
253 reviews, relying on artificial intelligence[23].

254 Despite the limited number of institutions that participated in this exercise, and the overrepresentation of  
255 Europe, the lack of contribution from African, South American and other Asian countries, we found, even  
256 within the European countries, variation among the numbers of search hits. This suggests that adding  
257 more countries would have led to even greater variability in the resulting datasets. It may be interesting to  
258 test a wider range of search platforms and subjects to gain further understanding of the level of reliability  
259 of various systems and collect reliable knowledge on their strengths and weaknesses.

260 Since, the original set of raw data input can significantly alter/skew the output of the study and, in the age  
261 of big data, studies on already published results are becoming more common, an unbiased and timely way

262 of data extraction is needed. At present, updating systematic reviews using precisely repeated  
263 methodology is impossible[24]; hence a clear decision map on the advantages and disadvantages of  
264 particular databases and search engines should be drawn to ensure the integrity of publication-based  
265 studies.

266 **Materials and methods**

267 **Queried databases**

268 Three major scientific databases, PubMed, Scopus, and Web of Science, and Google Scholar, as the most  
269 used and largest scientific were used in this study. Although Google Scholar is markedly different from  
270 the other three traditionally used databases, both in business politics and search method[14,18], the  
271 increasing use of this search engine [20] justifies its inclusion in the study. The main differences between  
272 these databases have been catalogued and reviewed by Falagas et al.[14].

273 **PubMed** (<https://www.ncbi.nlm.nih.gov/pubmed>) is a freely available scientific database, focusing  
274 mostly on biomedical literature, which holds ca. 28 million citations covering a variety of aspects of life  
275 sciences (<https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp>. PubMed\_Coverage, accessed  
276 15/08/2018). It was developed and is being maintained by the National Center for Biotechnology  
277 Information.

278 **Scopus**, currently owned by the Elsevier group, contains bibliographic data of over 1.4 billion  
279 publications dating back to 1970. It indexes ca. 70 million items and 22,500 journals from 5,000  
280 publishers (<https://www.elsevier.com/solutions/scopus/how-scopus-works/content>, accessed: 17. August  
281 2018).

282 **Web of Science** (<https://webofknowledge.com>) is the oldest scientific database, owned by the Clarivate  
283 Analytics (previously Thomson Reuters). Web of Science, running under its current name since 1997, is

284 the successor of the first scientific citation database, the Science Citation Index, which was launched in  
285 1964. It currently indexes 34,200 journals, books and proceedings, and, as of the last update, on 26  
286 August 2018, it covers 151 million records altogether and over 71 million in its Core Collection  
287 (<https://clarivate.libguides.com/webofscienceplatform/coverage>). Currently it also includes Zoological  
288 Records, CABI Abstracts, and a number of other, formerly independent databases.

289 **Google Scholar** (<https://scholar.google.com>) is a free online tool, the sub-site of the search mogul Google  
290 Inc., which is particularly designed for scholarly searches. Whilst Google Scholar has been often  
291 criticized for not sharing its search algorithms, for its untraceable way of ordering search hits, and for the  
292 inclusion of material from non-scholarly sources in its research hits[18,19,25], it has been playing an  
293 increasing role in daily lives of scientists since its launch in 2004[20,26]. It is also estimated to include  
294 160 million individual scientific publications in 2014[27] and to be the fastest growing resource for  
295 scientific literature[28]. Its usefulness, however, for systematic reviews and meta-analyses has been  
296 debated[16,18,19]

## 297 **Web searches**

298 In order to investigate the reproducibility of scientific searches in the four major search platforms, we  
299 generated keyword expressions (search strings) with two complexity levels using keywords that focused  
300 on either an ecological or a medical topic and ran standardized searches from various locations in the  
301 world (see below), all within a limited timeframe.

302 Simple search strings contained only one main keyword, whereas complex ones contained both inclusion  
303 and exclusion criteria for additional, related, keywords and key phrases (i.e. two-word expressions within  
304 quotation marks). Wildcards (e.g. asterisks) and logical operators were used in complex search strings.  
305 The main common keyword for ecology was “ecosystem” and “diabetes” was used for the medical topic.  
306 Search language was set to English in every case, and only titles, abstracts and keywords were searched.

307 Since different search engines use slightly different expressions for the same query, exact search terms  
308 were generated for each search (Table 4).

309 **Table 4 Search strings for each keyword complexity and topic, adjusted according to the search engines.**

Search engine	<b>Ecology</b>		<b>Medicine</b>	
	Complex search string	Simple search string	Complex search string	Simple search string
GScholar	"ecosystem service" +"promoting" + "crop" - "livestock"	"ecosystem services"	"diabetes" + "sugar" + "fructose" - "saccharose"	"diabetes mellitus"
PubMed	"ecosystem service" [Title/Abstract] AND "promoting" AND "crop" [Title/Abstract] NOT "livestock" [Title/Abstract] AND "english" [Language]	"ecosystem services" [Title/Abstract] AND "english" [Language]	"diabetes" [Title/Abstract] AND "sugar" AND "fructose" [Title/Abstract] NOT "saccharose" [Title/Abstract] AND "english" [Language]	"diabetes mellitus" [Title/Abstract] AND "english" [Language]
Scopus	TITLE-ABS-KEY ("ecosystem service" AND "promoting" AND "crop" AND NOT "livestock") AND (LIMIT-TO (LANGUAGE, "English"))	TITLE-ABS-KEY ("ecosystem services") AND (LIMIT-TO (LANGUAGE, "English"))	TITLE-ABS-KEY ("diabetes" AND "sugar" AND "fructose" AND NOT "saccharose") AND (LIMIT-TO (LANGUAGE, "English"))	TITLE-ABS-KEY ("diabetes mellitus") AND (LIMIT-TO (LANGUAGE, "English"))
WoS	TS= ("ecosystem service" AND "promoting" AND "crop" NOT "livestock")	TS= ("ecosystem services")	TS= ("diabetes" AND "sugar" AND "fructose" NOT "saccharose")	TS= ("diabetes mellitus")

310  
311 Searches were conducted on one or two machines at 12 institutions in Australia, Canada, China, Denmark,  
312 Germany, Hungary, UK, and the USA (Supplementary Information 5), using the three main browsers  
313 (Mozilla Firefox, Internet Explorer, and Google Chrome). Searches were run manually (i.e. no APIs were  
314 used) according to strict protocols, which allowed to standardize search date, exact search term for every  
315 run, and data recording procedure. Not all databases could have queried from every location: Google was  
316 not available in China, and Scopus was not available at some institutions (Supplementary Information 5).  
317 The original version of the protocol is provided in Supplementary Information 6. The first run was  
318 conducted at 11:00 Australian Eastern Standard Time (01:00 GMT) on 13 April 2018 and the last search  
319 run at 18:16 on 13 April 2018 Eastern Daylight Time (22:16 GMT). After each search the number of

320 resulted hits was recorded and the bibliographic data of the first 20 articles were extracted and saved in a  
321 file format that the website offered (.csv, .txt). Once all search combinations were run and browsers'  
322 cache had been emptied, the process was repeated. At four locations (Flakkebjerg, Denmark; Fuzhou,  
323 China; St. Catharines, Canada; Orange, Australia) the searches were also repeated on two different  
324 computers.

325 Results were collected from each contributor, bibliographic information was stripped out from the saved  
326 files, and was stored in a standardized database, allowing unique publications to be distinguished. If  
327 unique identifiers for individual articles were missing, authors, titles, or the combination of these were  
328 searched for, and uniqueness was double checked across the entire dataset.

329 For the rapid scan, if authors used Web of Science as the main search platform, and if search locations  
330 were reported, we chose the first twenty papers from a Google Scholar search (7 November, 2018) with  
331 the search term “systematic review” and “ecology”. Sites were restricted to sciencemag.org, nature.com,  
332 and wiley.com.

### 333 **Statistical analysis**

334 To investigate how consistent the number of resulting hits from each search string (i.e. the combination of  
335 the search topic and keyword expression complexity) was for each of the search engines, *average*  
336 *absolute deviation* (AAD, i.e. the absolute value of the difference of the actual value and the mean) was  
337 calculated and expressed as a percentage of the mean of each group (‘*average absolute deviation*  
338 *proportion*’, AADP, i.e. search topic, search term complexity, and search engine). AADP was calculated  
339 using the equation:

$$340 AADP = \frac{|e - \hat{e}_{gr}|}{\hat{e}_{gr}}$$

341 where  $e$  was the number of hits from one particular search and  $\hat{e}_{gr}$  was the mean number of hits of pooled  
342 numbers from one topic and search term complexity combination and one search engine (e.g. complex

343 ecological search expression queried using Scopus). This grouping was necessary because the number of  
344 hits substantially differed depending on these three factors. Since the aim of the study was not to compare  
345 the efficiency of different search engines, this grouping did not interfere with our analysis.

346 Normality of the data and homoscedasticity were tested using Kolmogorov-Smirnoff test and the Breusch  
347 Pagan test, respectively. These tests confirmed that neither the distribution of AADPs followed normal  
348 distribution, nor were the variances of residuals within each group homogenous. Indeed, the high number  
349 of zeroes resulted in a zero-inflated, an unbalanced beta distribution, as suggested by the *descdist()*  
350 function in the *fitdistrplus* R package[29], under an R programming environment[30].

351 AADP is expected to be zero in cases when search engines consistently give the same number of hits  
352 within groups, regardless where the search is initiated from, browser used, or whether the cache was  
353 emptied or not. Therefore, one-sided Wilcoxon signed rank tests were performed for the AADP values for  
354 each search engines within each group to test if they were significantly different from zero.

355 To address non-normality, unequal variances and control Type I error, non-parametric, Welch-James's  
356 statistic with Approximate Degrees of Freedom (Welch ADF) was used to investigate the differences  
357 between search engine consistencies and to select the most influential factors driving these differences.  
358 This robust estimator uses trimmed means and winsorized variances to avoid biases derived from  
359 heteroscedasticity. Bootstrapping was used to calculate empirical p-values both for between group and  
360 pairwise comparisons[31], with the help of WelchADF R package[32].

361 Moreover, average similarities of the first twenty papers within each of the search engine-topic-keyword  
362 complexity groups were calculated based on binary matrices, in which rows corresponded to search runs  
363 from various institutions and computers, whilst columns contained individual papers. Due to its suitability  
364 for using binary data, Jaccard distance measures were applied for similarity calculations. Distance-based  
365 redundancy analysis (dbRDA, capscale() function) was used with the same similarity matrices to ordinate  
366 the resultant article collections in each search topic-keyword complexity group. Convex hulls of the

367 points resulted from this ordination were then delimited for each search engine and their areas were  
368 calculated. Since similarities between article collections resulted from searches with a search engine  
369 giving consistently the same hits, regardless of search location, browser used, and cache content, should  
370 always be zero, the ideal size of these hulls would be also zero.

### 371 **Data availability statement**

372 All data and computer code are deposited on the Open Science Framework (OSF) website and will be  
373 openly available for the readers through a stable URL or DOI upon acceptance.

### 374 **Acknowledgements**

375 We are grateful for Dr. Mei Ling Huang (Brock University, St. Catharines, Canada) for her insightful  
376 comments on the statistical analysis. This work is supported by a grant of "111 project" in China. Gabor  
377 Pozsgai is supported by a postdoctoral fellowship by the State Key Laboratory of Ecological Pest Control  
378 for Fujian and Taiwan Crops. Arnold Móra was supported by the grants #20765-3/2018/FEKUTSTRAT  
379 and #TUDFO/47138/2019-ITM.

### 380 **Author contributions**

381 Gábor Pozsgai and Geoff Gurr conceived the project. Gábor Pozsgai designed the experiment, and did the  
382 statistical analysis. Gábor Lövei, Gábor Pozsgai, Jie Zhang, and Wenwu Zhou performed the preliminary  
383 searches. All contributors were involved in running the searches and providing raw data in the given  
384 format. The first drafted version of the manuscript was prepared by Gábor Pozsgai. This draft was first  
385 edited by Gábor Lövei, Liette Vasseur, Geoff Gurr, Olivia Reynolds, and Minsheng You. All authors  
386 were included in editing the subsequent versions of the manuscript. Minsheng You funded the work.

## 387 Competing interests

388 The authors declare no competing interest.

## 389 References

- 390 1. McMullin E. The Impact of Newton's Principia on the Philosophy of Science. *Philos Sci.* 2001;68:279–310. doi:10.1086/392883
- 391 2. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature.* 2016;533: 452–454. doi:10.1038/533452a
- 392 3. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: Synthesis of best evidence for clinical decisions. *Ann Intern Med.* 1997;126: 376–380. doi:10.7326/0003-4819-126-5-199703010-00006
- 393 4. Moher D, Liberati A, Tetzlaff J, Altman DG. Academia and Clinic Annals of Internal Medicine Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annu Intern Med.* 2009;151: 264–269. doi:10.1371/journal.pmed.1000097
- 394 5. Higgins J, Green S. *Cochrane Handbook for Systematic Reviews of Interventions.* Chichester, UK: The Cochrane Collaboration; 2008. Available: <http://www.cochrane-handbook.org/>
- 395 6. Bornmann L, Mutz R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *J Assoc Inf Sci Technol.* 2015;66: 2215–2222. doi:10.1002/asi.23329
- 396 7. Pain E. How to keep up with the scientific literature. *Science Careers.* 30 Nov 2016. doi:10.1126/science.caredit.a1600159
- 397 8. Landhuis E. Scientific literature: Information overload. *Nature.* 2016;535: 457–458. doi:10.1038/nj7612-457a

408 9. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research  
409 synthesis. *Nature*. 2018;555: 175–182. doi:10.1038/nature25753

410 10. Clarke M, Horton R. Bringing it all together: Lancet-Cochrane collaborate on systematic reviews.  
411 *Lancet*. 2001;357: 1728. doi:10.1016/S0140-6736(00)04934-5

412 11. Ioannidis JPA. The Mass Production of Redundant, Misleading, and Conflicted Systematic  
413 Reviews and Meta-analyses. *Milbank Q*. 2016;94: 485–514. doi:10.1111/1468-0009.12210

414 12. Garg AX, Hackam D, Tonelli M. Systematic review and meta-analysis: When one study is just not  
415 enough. *Clin J Am Soc Nephrol*. 2008;3: 253–260. doi:10.2215/CJN.01430307

416 13. Gusenbauer M, Haddaway NR. Which Academic Search Systems are Suitable for Systematic  
417 Reviews or Meta-Analyses? Evaluating Retrieval Qualities of Google Scholar, PubMed and 26  
418 other Resources. *Res Synth Methods*. 2019; jrsm.1378. doi:10.1002/jrsm.1378

419 14. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of  
420 Science, and Google Scholar: strengths and weaknesses. *FASEB J*. 2007;22: 338–342.  
421 doi:10.1096/fj.07-9492LSF

422 15. Gavel Y, Iselid L. Web of Science and Scopus: a journal title overlap study. *Online Inf Rev*.  
423 2008;32: 8–21. doi:10.1108/14684520810865958

424 16. Boeker M, Vach W, Motschall E. Google Scholar as replacement for systematic literature searches:  
425 Good relative recall and precision are not enough. *BMC Med Res Methodol*. 2013;13.  
426 doi:10.1186/1471-2288-13-131

427 17. Jaccard P. The distribution of the flora in the alpine zone. *New Phytol*. 1912;11: 37–50.  
428 doi:10.1111/j.1469-8137.1912.tb05611.x

429 18. Jacsó P. Google Scholar revisited. *Online Inf Rev*. 2008;32: 102–114.

430 doi:10.1108/14684520810866010

431 19. Jacsó P. As we may search – Comparison of major features of the Web of Science, Scopus, and  
432 Google Scholar citation-based and citation-enhanced databases. *Curr Sci.* 2005;89: 1537–1547.  
433 Available: [http://muse.jhu.edu/content/crossref/journals/library\\_trends/v056/56.4.jacso.html](http://muse.jhu.edu/content/crossref/journals/library_trends/v056/56.4.jacso.html)

434 20. Haddaway NR, Collins AM, Coughlin D, Kirk S. The role of google scholar in evidence reviews  
435 and its applicability to grey literature searching. *PLoS One.* 2015;10: 1–17.  
436 doi:10.1371/journal.pone.0138237

437 21. Shotton D. Funders should mandate open citations. *Nature.* 2018. doi:10.1038/d41586-018-00104-  
438 7

439 22. van Dijck J. Search engines and the production of academic knowledge. *Int J Cult Stud.* 2010;13:  
440 574–592. doi:10.1177/1367877910376582

441 23. Beller E, Clark J, Tsafnat G, Adams C, Diehl H, Lund H, et al. Making progress with the  
442 automation of systematic reviews: Principles of the International Collaboration for the Automation  
443 of Systematic Reviews (ICASR). *Syst Rev.* 2018;7: 1–7. doi:10.1186/s13643-018-0740-7

444 24. Garner P, Hopewell S, Chandler J, MacLehose H, Schünemann HJ, Akl EA, et al. When and how  
445 to update systematic reviews: Consensus and checklist. *BMJ.* 2016;354: 1–10.  
446 doi:10.1136/bmj.i3507

447 25. Noruzi A. Google Scholar: The New Generation of Citation Indexes. *Libri.* 2005;55: 170–180.  
448 doi:10.1515/LIBR.2005.170

449 26. Halevi G, Moed H, Bar-Ilan J. Suitability of Google Scholar as a source of scientific information  
450 and as a source of data for scientific evaluation—Review of the Literature. *J Informetr.* 2017;11:  
451 823–834. doi:10.1016/j.joi.2017.06.005

452 27. Orduna-Malea E, Ayllón JM, Martín-Martín A, Delgado López-Cózar E. Methods for estimating  
453 the size of Google Scholar. *Scientometrics*. 2015;104: 931–949. doi:10.1007/s11192-015-1614-6

454 28. Larsen PO, von Ins M. The rate of growth in scientific publication and the decline in coverage  
455 provided by science citation index. *Scientometrics*. 2010;84: 575–603. doi:10.1007/s11192-010-  
456 0202-z

457 29. Delignette-Muller ML, Dutang C. *fitdistrplus: An R Package for Fitting Distributions*. *J Stat Softw*.  
458 2015;64: 1–34. Available: <http://www.jstatsoft.org/v64/i04/>

459 30. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria; 2012.  
460 Available: <http://www.r-project.org/>

461 31. Keselman HJ, Algina J, Lix LM, Wilcox RR, Deering KN. A generally robust approach for testing  
462 hypotheses and setting confidence intervals for effect sizes. *Psychol Methods*. 2008;13: 110–129.  
463 doi:10.1037/1082-989X.13.2.110

464 32. Villacorta PJ. *welchADF: Welch-James Statistic for Robust Hypothesis Testing under  
465 Heterocedasticity and Non-Normality*. 2018. Available: [https://cran.r-  
project.org/package=welchADF](https://cran.r-<br/>466 project.org/package=welchADF)

467 **Supporting Information 1.** The results of the Welch-James's statistic with Approximate Degrees of  
468 Freedom. Significant ( $p < 0.05$ ) relationships are highlighted with bold font.

469 **Supporting Information 2.** Average absolute deviation proportions (AADP) of hit numbers, grouped by  
470 searched platforms, and separated by grouped keyword complexity (complex, simple) – research area  
471 (ecology, medicine) and cache state. Boxes represent interquartile range (IQR), with median AADP  
472 values represented as a thick horizontal band. Whiskers extend from Q1-1.5IQR to Q3+1.5IQR.  
473 Abbreviated search platforms: GScholar – Google Scholar, WoS – Web of Science.

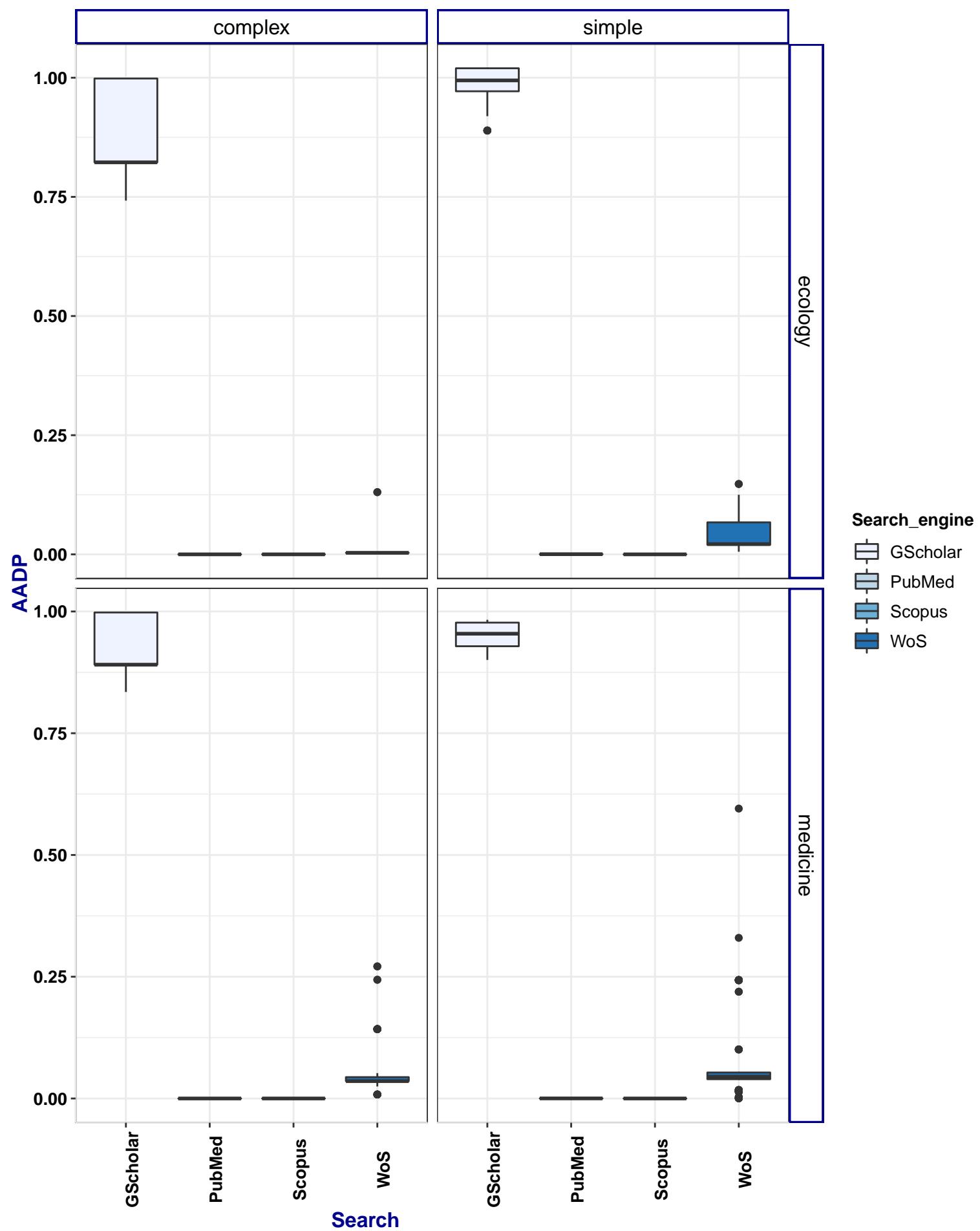
474     **Supporting Information 3.** Average absolute deviation proportions (AADP) of hit numbers, grouped by  
475     searched platforms, and separated by grouped keyword complexity (complex, simple) – research area  
476     (ecology, medicine) and browser type. Boxes represent interquartile range (IQR), with median AADP  
477     values represented as a thick horizontal band. Whiskers extend from Q1-1.5IQR to Q3+1.5IQR.  
478     Abbreviated search platforms and browsers: GScholar – Google Scholar, WoS – Web of Science, Chrome  
479     – Google Chrome, IE – Internet Explorer, Mozilla – Mozilla Firefox.

480     **Supporting Information 4.** The list of papers used in the rapid screen and the results showing how many  
481     different search platforms were used and whether or not the date, search location and browser were  
482     indicated.

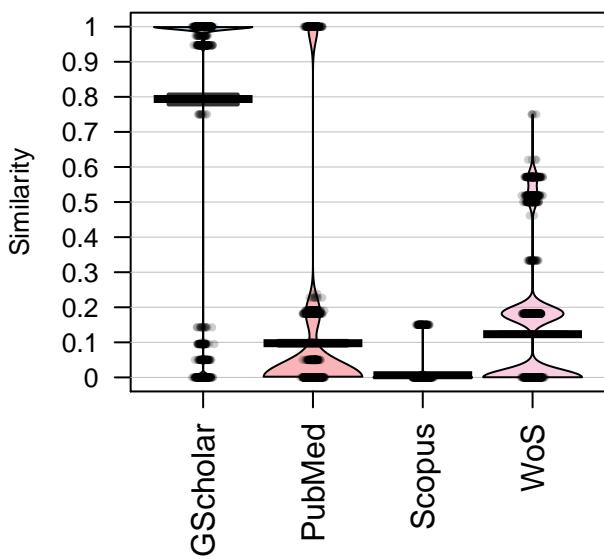
483     **Supporting Information 5.** Names and affiliations of contributors and list of scientific search platforms  
484     accessed during the search exercise.

485     **Supporting Information 6.** The exact protocol which was circulated to contributors, describing how  
486     searches should be performed and how data should be saved.

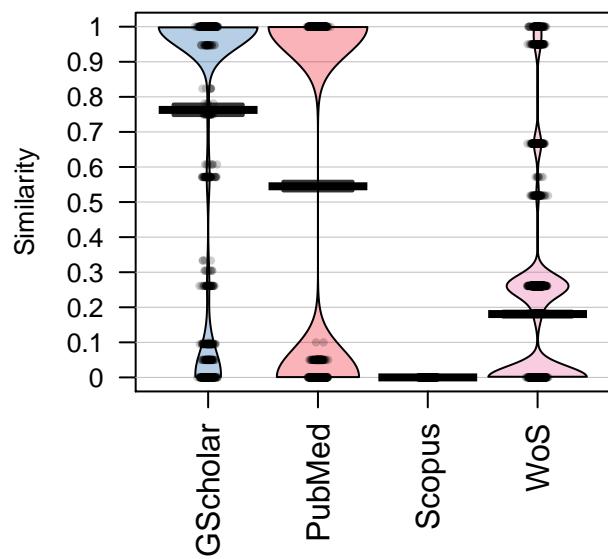
487



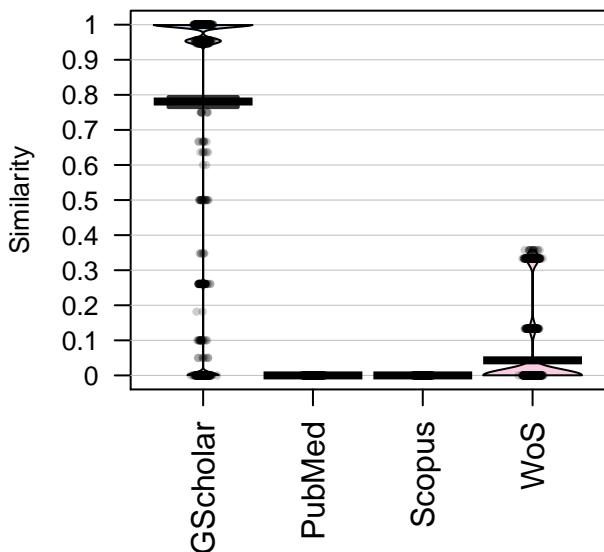
**medicine & complex**



**medicine & simple**



**ecology & complex**



**ecology & simple**

