An experimental search strategy retrieves more precise results than PubMed and Google for questions about medical interventions

Robert G Badgett, Daniel P Dylla, Susan D Megison, E Glynn Harmon

Objective: We compared the precision of a search strategy designed specifically to retrieve randomized controlled trials (RCTs) and systematic reviews of RCTs with search strategies designed for broader purposes. **Methods**: We designed an experimental search strategy that automatically revised searches up to five times by using increasingly restrictive queries as long at least 50 citations were retrieved. We compared the ability of the experimental and alternative strategies to retrieve studies relevant to 312 test questions. The primary outcome, search precision, was defined for each strategy as the proportion of relevant, high quality citations among the first 50 citations retrieved. **Results**: The experimental strategy had the highest median precision (5.5%; interquartile range [IQR]: 0% - 12%) followed by the narrow strategy of the PubMed Clinical Queries (4.0%; IQR: 0% - 10%). The experimental strategy found the most high quality citations (median 2; IQR: 0 - 6) and was the strategy most likely to find at least one high quality citation (73% of searches; 95% confidence interval 68% - 78%). All comparisons were statistically significant. **Conclusions**: The experimental strategy performed the best in all outcomes although all strategies had low precision.

2 **An experimental search strategy retrieves more precise results than PubMed**

3 **and Google for questions about medical interventions**

5 Robert G. Badgett, MD

6 Department of Internal Medicine

7 Kansas University Medical Center - Wichita

8 1010 N. Kansas

9 Wichita, KS 67214-3199

10 rbadgett@kumc.edu

12 Authors' Present Affiliations:

13 Daniel P. Dylla, MSIS

14 Katy Campus Library

15 Houston Community College NW

16 Houston, TX 77084

17 Daniel.Dylla@hccs.edu

19 Susan D. Megison, MSIS

20 No institutional affiliation

21 Birmingham, AL

22 susan.megison@gmail.com

24 E. Glynn Harmon, Ph.D. (deceased)

25 School of Information, SZB 562D

26 University of Texas at Austin

27 Austin, TX 78712-0390

29 **Objective**: We compared the precision of a search strategy designed specifically to retrieve randomized

30 controlled trials (RCTs) and systematic reviews of RCTs with search strategies designed for broader purposes.

31 **Methods**: We designed an experimental search strategy that automatically revised searches up to five times by

32 using increasingly restrictive queries as long at least 50 citations were retrieved. We compared the ability of the

33 experimental and alternative strategies to retrieve studies relevant to 312 test questions. The primary outcome,

34 search precision, was defined for each strategy as the proportion of relevant, high quality citations among the

35 first 50 citations retrieved.

36 **Results**: The experimental strategy had the highest median precision (5.5%; interquartile range [IQR]: 0% -

37 12%) followed by the narrow strategy of the PubMed Clinical Queries (4.0%; IQR: 0% - 10%). The

38 experimental strategy found the most high quality citations (median 2; IQR: 0 - 6) and was the strategy most

39 likely to find at least one high quality citation (73% of searches; 95% confidence interval 68% - 78%). All

40 comparisons were statistically significant.

41 **Conclusions**: The experimental strategy performed the best in all outcomes although all strategies had low

42 precision.

43
44
45

Introduction

Health care providers are encouraged to answer clinical questions by first consulting evidence-based

summaries.(DiCenso, Bayley, & Haynes, 2009) Summaries are defined as evidence-based practice guidelines

and evidence-based textbooks. (DiCenso, Bayley, & Haynes, 2009) Accordingly, physicians commonly use

online resources such as UpToDate.(Anonymous, 2014; Edson et al., 2010; Duran-Nelson et al., 2013)

Unfortunately, summaries may not always suffice. The evidence-based summaries UpToDate, Dynamed

(Anonymous, 2014b), FirstConsult (Anonymous, 2014), and ACP Smart Medicine (Anonymous, 2014a) have

less than 5% overlap in the studies cited, which implies no resource is comprehensive(Ketchum, Saleh, & Jeong,

2011). Similarly, studies report that UpToDate and the National Guidelines Clearinghouse addressed less than

80% of questions by primary care physicians(Fenton & Badgett, 2007) and hospital-based physicians.(Lucas et

al., 2004)

At times health care providers must search for original studies due to the deficiencies of secondary resources

such as those discussed above; however, practicing physicians tend to have difficulty answering clinical

questions by using electronic databases. This difficulty places physicians in the position of "knowing less than

has been proved." (Mulrow, 1994). In a recent study, only 13% of searches by physicians led to changing

provisional answers to correct while 11% of searches led to changing provisional answers to incorrect.

(McKibbon & Fridsma, 2006) Lucas found that 14% of inpatients were judged to have their care improved after

physicians received unsolicited search results provided as part of a research study.(Lucas et al., 2004)

The best search method for supplementing evidence-based summaries is controversial and difficult to identify

due to the absence of a direct comparison of commonly used methods. The use of PubMed is encouraged by

medical leaders;(Anonymous, 2014b; AAMC-HHMI Scientific Foundation for Future Physicians Committee,

2009) however, physicians prefer the speed and simplicity of Google.(Anonymous, 2014a; Sim, Khong, & Jiwa,

72   2008; Thiele et al., 2010) These methods fundamentally differ in the bibliographic data that are searched and in

73   the sorting of search results. PubMed by default sorts results by date which may obscure a seminal article with

74   more recent results from minor journals. On the other hand, Google, which sorts articles by a mix of estimated

75   importance and relevance, ignores the dates of publication or revision of sources. Thus, Google may not

76   accurately represent critical timing of search results that contain an article from a major journal that was later

77   contradicted in a less impactful journal. (Ioannidis, 2005)   The implications of these differences are not fully

78   known. Google launched Scholar in 2004 in order to improve access to academic publications. As compared to

79   PubMed, Scholar indexes the full text of many journals rather than just the citation and abstract, but does not use

80   MEDLINE's metadata such as the National Library of Medicine's Medical Subject Headings (MeSH) terms and

81   publication types. Like Google web search, Google Scholar by default sorts citations by a mix of estimated

82   importance and relevance. The retrieval algorithms and heuristics deployed by Google Scholar are propriety, not

83   described on the Scholar website, and not clearly discernible.(Anonymous, 2014)

84

85   In 1998, one of the authors (RGB) launched the experimental search engine SUMSearch, which includes

86   PubMed searches and is specifically designed for use in clinical medicine to supplement evidence-based

87   textbooks and practice guidelines. (Badgett, 1999) The current version of SUMSearch is available at

88   http://sumsearch.org. SUMSearch preserves the date sorting feature used by PubMed, but allows automated

89   revisions of searches in order to make older sentinel articles visible. Automatic revisions of searches may

90   address barriers health care providers experience while searching, such as designing search strategies and

91   "uncertainty about narrowing the search...when faced with an overwhelming body of knowledge."(Ely et al.,

92   2002)

93

94   Our objective is to quantify and compare the ability of a search designed specifically for clinical medicine with

95   alternative strategies that are designed for broader purposes. We hypothesized that an experimental search

96   strategy designed specifically for clinical topics would outperform other strategies for retrieving articles about

97   medical interventions.

98   **Materials & Methods**

99   We compared five search strategies taken from four search engines for their ability to answer a collection of

100   clinical questions. In previous comparisons, SUMSearch and PubMed have performed better than Scholar;

101   (Haase et al., 2007; Freeman et al., 2009; Anders& Evans, 2010) however, the current study is the first to

102   compare SUMSearch and PubMed to each other and to Google. While Google and Google Scholar were not

103   designed for clinical purposes, the frequency of their use by health care providers mandates assessment of their

104   ability.

105   **Source of clinical questions**

106   We used questions about from the Clinical Questions Collection of the National Library of Medicine.

107   (Anonymous, 2004; Ely et al., 1999; Ely et al., 2005) The complete collection consists of 4654 questions

108   collected from physicians in Iowa. For each question, personnel at the National Library of Medicine assigned

109   keywords that were almost always taken from the MeSH database.

110

111   From the collection we included questions about treatment of non-pregnant adults. This exclusion allowed us to

112   better monitor development of the project as our clinical expertise is internal medicine. We excluded questions

113   that also had a keyword assigned for diagnosis in order to ensure that the questions focused on treatment and so

114   were best answered with randomized controlled trials and meta-analyses of trials. We excluded questions whose

115   keywords duplicated the keywords of other questions. We included 312 questions after the above exclusions

116   (Figure 1).

117

118   **Query expansion**

119   Each question in the Questions Collection contains a median of 2 keywords, usually based on Medical Subject

120   Heading (MeSH) terms. We linked these keywords with "AND". In addition, we replaced the word "neoplasms"

121   with "cancer" and inverted all keywords that contained commas. For example, "anemia, sickle cell" was inverted

122   to "sickle cell anemia." This inversion allows the search term to also perform well as a text word. The resulting

123   search terms were submitted to the search engines without designation of a search field so that at PubMed's

124   Clinical Queries the terms were searched as both MeSH terms and text words. All searches were performed

125   between June and December of 2009.

126   **Search strategies**

127   The experimental search strategy was based on the PubMed component of a prior version of SUMSearch

128   federated search engine and could perform up to five iterations for each question. Details and examples of the

129   iterations used by the experimental strategy are included in Table 2. This strategy sought randomized controlled

130   trials and systematic reviews of trials. Each iteration was progressively more restrictive. The composition and

131   sequencing of the iterations was based on experience with SUMSearch. The strategy returned the results of the

132   last iteration that retrieved 50 or more citations. The rationale for restricting the numbers of citations is to

133   reproduce the behavior observed in searchers to typically scan a limited number of citations. (Blair D, 1980;

134   Islamaj Dogan et al., 2009) This limit has been called the futility point and occurs when searchers regard

135   reviewing additional citations as being beyond their respective time and manageability constraints. The

136   experimental search strategy imitated PubMed searching by querying Entrez's eSearch utility.(Sayers, 2013)

137   This utility has no user interface and is designed by the National Library of Medicine for external search engines

138   and other automated tools to efficiently query PubMed.

139   We included two strategies from PubMed's Clinical Queries that are publicly available

140   (http://www.ncbi.nlm.nih.gov/pubmed/clinical). We used the current Narrow and Broad strategies for therapies.

141   These strategies were initially developed by Haynes in 1994 and revised by Haynes in 2005. (Haynes et al.,

142   1994; Haynes et al., 2005)

143   We studied two strategies by Google. We used the main Google Web search engine and labeled this strategy as

144   "Google." We used the Google Scholar search engine and labeled this strategy as "Scholar". For both of these

145   strategies, we assessed methods to improve upon simply constructing search queries by using clinical terms.

146   Using test cases, we informally assessed the benefit of adding the following candidate search terms to the search

147   query: "PMID", "DOI", ~random, ~trial, site:.org, site:.edu, and site:.gov. The terms PMID and DOI are

148   abbreviations for "PubMed identifier" and "digital object identifier" and are common numeric identifiers in the

149   Internet addresses and on the Internet pages for articles in health care journals. These identifiers are indexed by

150   Google like any other content on an Internet page or in its Internet address. In addition, formal citations to health

151   care articles, such as in wikis, frequently include these numbers and the abbreviations that indicate the type of

152   number. The final strategy chosen for both Google and Scholar appended the strings "PMID", "~random", and

153   "~trial" to the search terms. The "~" character was required at the time of our study for Google to seek

154   synonyms for an adjacent search term.(Schwartz, 2013) We appended "num=50" to the urls submitted for both

155   strategies in order to retrieve 50 hits per search. Searches were performed on a dedicated server that had Google

156   cookies removed in order to prevent Google from any customization of search results such as prioritizing results

157   based on geographic location.

158   **Outcome ascertainment**

159   All search results were parsed for PMIDs and DOIs. For search results from Google, we also parsed the text in

160   the Internet addresses of hyperlinks. All identifiers found were then submitted to Entrez's efetch utility in order

161   to retrieve full citations including PMIDs, MeSH terms and lists of all articles that commented on the retrieved

162   articles.

163   **Reference standard**

164   The reference standard required articles to be relevant and high quality. An article was considered relevant to the

165   clinical question if the article contained all of the keywords assigned by the Clinical Questions Collection to the

166   clinical question. The keywords could be either MeSH terms or MeSH entry terms, and the keywords could be

167   located in title, abstract, or MeSH terms of the article.

168   An article was considered high quality if it had high quality methodology or was considered important by an

169   expert in the domain of the article. Articles having high methodological standards were considered those that

170   were reviewed by an evidence-based synoptic journal as previously done by

171   Aphinyanaphongs.(Aphinyanaphongs et al., 2005) These journals were ACP Journal Club, InfoPoems, Evidence

172   Based Dentistry, Evidence Based Medicine, Evidence Based Nursing, and Evidence Based Mental Health.

173   Articles considered important by a domain expert were those that were published with an accompanying

174   editorial.

175   To avoid incorporation bias that would artificially inflate our estimated of the accuracy of the searches, all

176   strategies were designed without incorporating search terms that contribute to the definition of the reference

177   standard. For example, one component of our reference standard is abstraction of the article by the publication

178   ACP Journal Club. Some websites, such as PubMed, indicate which citations have been reviewed by ACP

179   Journal Club. Thus, we could have added "ACP Journal Club" to our search strategy to improve its precision.

180   However, we did not add this term, as it would create incorporation bias and limit the ability to generalize the

181   results of our study to topics not covered by ACP Journal Club. An example question from the Clinical

182   Questions Collection and the resulting search strategies is in Table 1

183   **Statistical analysis**

184   The primary outcome was the median average precision of the searches for retrieving studies meeting criteria for

185   the reference standard. We limited the number of search results examined to 50 to control for the varying

186   number of results retrieved by each search engine. For example, searches for medical interventions may retrieve

187   hundreds of thousands of results using the Google strategy while retrieving a much smaller number with the

188   other search strategies. We specified 50 search results because searchers, outside of those performing meta-

189   analysis, are unlikely to review a large number of citations.(Blair D, 1980; Islamaj Dogan et al., 2009) In

190   addition, this limit allows comparison of searches that may retrieve substantially different number of citations.

191   (Herskovic, Iyengar, & Bernstam, 2007) For example, Google may retrieve more citations of high quality than

192   the other strategies due to retrieving many-fold more total citations. However, the Google search is not clearly

193   better because the user had to sift through more citations to find the high quality citations.

194   The precision was calculated as the proportion of the first 50 search results identified by each strategy that were

195   deemed to be relevant, high quality studies according the criteria in the preceding section, "Reference standard."

196   If no qualifying studies were retrieved, the precision was set to 0.

197   The number need to read (NNR) for each strategy is the number of citations that would have to be assessed to

198   yield one qualifying article. The NNR was calculated as the inverse of the precision. (Toth, Gray, & Brice, 2005)

199   Calculations were made with R statistical software package, version 2.11.1.(R Development Core Team, 2012)

200   Pairwise comparisons between individual medians were assessed using a post hoc analysis for Friedman's Test.

201   Rates of dichotomous outcomes were compared with the chi-square test. Chi-square is a conservative choice as

202   it does not consider pairing of data in calculation.

203   **Results**

204   The most common clinical concepts in the 312 questions about treating non-pregnant adults were hyperlipidemia

205   (15 questions), hypertension (10 questions), and urinary tract infections (10 questions).

206

207   The principal outcome, search precision, and all other outcomes were not normally distributed (Lilliefors

208   normality test $p<0.001$), so the median precision became the principal outcome. Using Google as an example to

209   illustrate the results, when the first 50 hits in a Google search were examined, a mean of 23 PubMed citations

210   were retrieved by parsing PMIDs or DOIs from the Google results (not shown in table). Of these 23 PubMed

211   citations, an average of 3.3 were deemed high quality because the citation was abstracted by an evidence-based

212   synoptic journal or published with an accompanying editorial. Of the 3.3 citations, an average of 1.3 was

213   relevant to the original search terms. While this suggests the mean precision for Google was 1.3 divided by 23,

214   or 5.6%, the actual mean precision was lower at 4.6%. The discrepancy is because the average of a series of

215   fractions is not equivalent to the average of the numerators divided by the average of the denominators. Lastly,

216   54% searches performed by Google retrieved no high quality, relevant citations thus the *median* precision for

217   Google was 0% (Table 3). The corresponding values for the numbers needed to read are: Experimental 18,

218   PubMed narrow Clinical Query 25, and PubMed's broad Clinical Query 50. The numbers needed to read cannot

219   be calculated for the Google strategies due to their median precision of 0%.

220   The median precision was significantly different among the strategies by Friedman's rank sum test (Table 3).

221   The experimental strategy and the narrow strategy of the PubMed Clinical Queries had the highest median

222   precision (5.5% and 4.0%, respectively). The experimental strategy had the highest ranked and mean values of

223   precision (Table 3; $p < 0.001$ for both analyses). The experimental strategy was the most likely method to find at

224   least one high quality citation (73% of questions) with $p < 0.001$. The median number of high quality articles

225   retrieved per search was two for both the experimental strategy and the PubMed narrow, while the means were

226   5.0 and 2.6, respectively ($p < 0.001$).

227   In an unplanned analysis, we examined the precision of experimental search strategies based on the number of

228   iterations the experimental strategy required (Figure 2). Searches that required one or two iterations had low

229   precision, whereas searches requiring more iterations had higher precision.

230   For all outcomes, Google and Google Scholar performed worse than the other strategies. This was in part

231   because Google itself sometimes found high quality citations that were not relevant. For example, in a search for

232   bronchiectasis and drug therapy, Google retrieved the Wikipedia pages on acetylcysteine and pulmonary

233   embolism. The acetylcysteine page was retrieved because Wikipedia listed bronchiectasis as treatable with

234   acetylcysteine while the pulmonary embolism page was retrieved only because the page listed bronchiectasis in

235   the page's navigational menu of pulmonary diseases. Unfortunately, the high quality citations that were on these

236   pages were not relevant to bronchiectasis.

237   **Discussion**

238   The experimental search was significantly better than the other strategies in all outcomes. Google and Google

239   Scholar strategies did not perform as well. We believe this is the first comparative study to identify a search

240   strategy that may be comparable to or better than the 2004 version of the PubMed Clinical Queries for common

241 clinical questions. The experimental search is available at http://sumsearch.org/ by changing the default settings

242 so that "Focus" is Intervention, "Target # of original studies" is 50, and "Require abstracts" is not selected.

243

244 Our results support Battelle's hypothesis that domain-specific search strategies should perform better than

245 general strategies.(Battelle, 2005) Google and Google Scholar's poor performance was consistent with prior

246 comparisons with PubMed or SUMSearch. (Haase et al., 2007; Freeman et al., 2009; Anders& Evans, 2010) Our

247 study should be compared to three studies that suggest benefit from using Google Scholar. Gehanno notes

248 perfect coverage by Scholar of trials in a set of Cochrane reviews. (Gehanno, Rollin, & Darmoni, 2013)

249 However, coverage simply relates to the presence of trials in the Scholar database, which is different from our

250 study of how well those trials can be retrieved by search strategies. Two smaller studies, by Nourbakhsh and

251 Shariff suggests that Scholar retrieves more citations that are relevant than PubMed retrieves. (Nourbakhsh et

252 al., 2012; Shariff et al., 2013) Several reasons may underlie the conflicting results. The reference standard used

253 by Nourbakhsh only considered relevance and not study design or quality of citations. Our precision is likely

254 underestimated due to the certain existence of qualifying articles that were missed due to not being abstracted by

255 an evidence based synoptic journal. Also, the PubMed searches used by Nourbakhsh relied exclusively on

256 MeSH terms. (Nourbakhsh et al., 2012) For example, Nourbakhsh used "hypertension, pulmonary [MeSH]"

257 whereas we would have changed this term to "pulmonary hypertension[all fields]". The 'all fields' tag submits

258 the term as *both* a MeSH term and a text word. In addition, the Nourbakhsh study was limited to four questions

259 the researchers were familiar with and the differences did not reach statistical significance. Shariff did not

260 provide details on how the nephrologists used PubMed other than stating that the searches were not revised

261 based on the number of results retrieved.(Shariff et al., 2013) The findings of similar precision of results yet

262 fewer relevant citations among the first 40 citations retrieved by PubMed compared to Scholar indicates that in

263 many cases the PubMed searches retrieved fewer than 40 citations. The conflict between our results and those of

264 Shariff may be due to our use of iterative searching or to the nature of primary versus specialty care questions.

265 Iterative searching may be more important in broad topics that retrieve more citations.

266

267   The domain-specific search strategies that we studied, PubMed and SUMSearch, may perform better for two

268   reasons that have not changed since our study was completed. First, these strategies, unlike Google and Scholar,

269   take advantage of the hierarchal Medical Subject Headings (MeSH) terms that the National Library of Medicine

270   assigns to citations. Second, our results raise the question of whether a Boolean search model should be

271   preferred for the task we studied. Most contemporary research of searching MEDLINE examines search models

272   other than Boolean. Boolean models connect search terms with logical connectors such as 'and' and 'or' are

273   considered weaker than other search models. (Baeza-Yates & Ribeiro-Neto , 2011) A paradoxical advantage of

274   Boolean models is that because they do not rank documents by any grading scale, search results can be sorted by

275   date of publication. Sorting by date can be critical in medicine because of the surprising frequency that research

276   results are contradicted by subsequent authors.(Ioannidis, 2005)

277

278   In addition to providing a comparison of the performances of commonly used search strategies, our results

279   reinforce the difficulty of retrieving clinical studies from MEDLINE. The experimental strategy was most

280   precise but barely achieved a precision of 5%. Our study reported substantially lower precision than a previous

281   comprehensive review by McKibbon.(McKibbon et al., 2009) Common to our study and the review was analysis

282   of the PubMed Clinical Queries narrow filter. McKibbon reported a precision of 55% whereas we found the

283   same filter to have a precision of 2%. We believe our study reflects the precision that health care providers will

284   encounter and is lower than the report of McKibbon for two reasons. First, we measured the precision in

285   answering actual clinical questions. Second, we measured the precision among all journals of PubMed rather

286   than limiting to the 161 journals that publish the highest rate of high-quality studies. Since we executed our

287   study, Shariff reported that nephrologists were able to search MEDLINE with a *mean* precision higher than our

288   report of *median* precision. (Shariff et al., 2013) We reported median rather than mean values for precision due

289   to concern that means will overstate performance. To directly compare studies, the *mean* precision of 10.2% we

290   report for our experimental strategy is higher than found by Shariff.

291

292   **Possible limitations**

293 First, we standardized the design of all search strategies to eliminate variability in the search skill of actual users.

294 Both the precision and number of relevant citations retrieved by human searchers may be less than we report. It

295 is possible that in our study Google's performance was diminished because Google may have found citations

296 that were not counted because they were not accompanied by PMIDs or DOIs. However, in addition to parsing

297 the results displayed by Google, we also parsed the links provided by Google. Any functional link to an article at

298 PubMed will have a PMID embedded and be found by our methodology. Similarly, high quality studies may

299 have been missed by all strategies due to our removing "diagnosis" as a key word. This may have selectively

300 harmed the experimental and PubMed strategies as these incorporated MeSH terms. However, it is unknown

301 whether this affected precision as the total number of studies retrieved is also lower.

302

303 We recognize that our definition of the reference standard might be debatable for three reasons. First, we limited

304 our study to retrieving randomized controlled trials and systematic reviews of randomized controlled trials

305 because treatment questions are important and the standards for the conduct and assessment of these studies are

306 better developed than for other resources. While this information need may be infrequent for many health care

307 providers, we believe the ability to locate randomized controlled trials is very important for peer leaders who

308 may be writing or teaching clinical topics. Second, our definition of high quality articles is imperfect. We

309 believe, however, our definition has the advantage of being determined by experts who determined that an

310 editorial or synopsis was justified and who were not involved in the evaluation of the search strategies. In

311 addition, we believe the results that our definition yields are likely to move in parallel with other definitions of

312 high quality. Third, the use of precision (the proportion of relevant documents retrieved in the search) as a

313 metric instead of sensitivity (the ability of the system to detect all relevant documents) is debatable. However,

314 our goal is to create searches with precision for clinicians rather than comprehensive searches for meta-analysts.

315 For example, high sensitivity may be more useful for meta-analyses that require comprehensive results. High

316 precision may be more useful for time-sensitive tasks that require relevant documents quickly. Regardless, we

317 do provide the numbers of high quality citations retrieved which should correlate with the sensitivity of a

318 strategy for a given question.

319

320   Our results should not be generalized beyond searching for studies of interventions. The randomized controlled

321   trial index term used by the National Library of Medicine's Medical Subject Headings (MeSH) is unusually

322   accurate whereas MeSH terms for other study designs may be less accurate. (Haynes et al., 2005) None of the

323   strategies we tested may be appropriate for the conduct of meta-analysis when very high recall or sensitivity of

324   searches is required. Lastly, our questions all had carefully assigned MeSH terms. Searchers not facile with

325   MeSH terms may have lesser results.

326

327   **Future research**

328   Future research could address the strategies that were studied and compare them to search strategies based on

329   alternative search models. Aside from the search strategies developed by Haynes for PubMed's Clinical Queries,

330   the strategies were not formally developed. For example, we appended Google and Scholar strategies with "

331   PMID ~random ~trial" based on several use cases, but perhaps other restrictions would have performed better.

332   However, Google's performance was so low that substantial improvement from revising search terms seems

333   unlikely. Google frequently revises its search algorithms. (Anonymous, 2011) Until Google makes a major

334   change, such as recognizing MeSH terms and the hierarchical relationship among them, the impact of lesser

335   revisions on searching for medical research is not known. Continual research of Google is warranted. Regarding

336   the experimental strategy, perhaps other iterations, sequences of iteration, and number of iterations would

337   improve the search results. In addition, Wilczynski recently described how to improve the precision of the

338   Haynes strategies by adding "not" terms to searches of MEDLINE (Wilczynski, McKibbon, & Haynes, 2011)

339   Future research could compare our strategy to strategies based on machine learning or citation analysis. Lastly,

340   we hope that search engines in the future will provide more than a list of citations and will add indicators of

341   credibility to citations and display the conclusions in a way to allow users to quickly assess the concordance

342   among conclusions. The former is currently done by SUMSearch by indicating which citations are accompanied

343   by editorials and reviews by synoptic publications. The latter is being developed by AskHermes.(Yu, 2014)

344   **Conclusion**

345   Our results suggest that when health care providers need to supplement evidence-based summaries by searching

346   for high quality randomized controlled trials and systematic reviews of randomized controlled trials, an

347   experimental strategy designed specifically for clinical care may be more appropriate than the more general

348   strategies deployed by Google and PubMed Clinical Queries.

References

AAMC-HHMI Scientific Foundation for Future Physicians Committee. 2009. *Scientific Foundations for Future Physicians*.

Anders, M. E. and D. P. Evans. 2010. "Comparison of PubMed and Google Scholar Literature Searches." *Respiratory Care* 55 (5): 578-583.

Anonymous. "About Google Scholar " Google, http://scholar.google.com/intl/en/scholar/about.html(accessed May 5, 2014).

———. "About the IOWA Questions Collection(s)." National Library of Medicine, last modified September, 2004, http://clinques.nlm.nih.gov/About.jsp(accessed May 5, 2014).

———. "ACP Smart Medicine." American College of Physicians, http://smartmedicine.acponline.org(accessed May 5, 2014).

———. "DynaMed™." EBSCO Publishing, http://dynamed.ebscohost.com/(accessed May 5, 2014).

———. "Facts about Google and Competition ", http://www.google.com/competition/howgooglesearchworks.html(accessed 12/30/2013, 2013).

———. "FirstCONSULT." Elsevier B.V., http://www.firstconsult.com/(accessed May 5, 2014).

———. "UpToDate®." Wolters Kluwer Health, http://uptodate.com(accessed May 5, 2014).

———. 2011. "Friedman Rank Sum Test." Chap. June 3, 2011, In *R: A Language and Environment for Statistical Computing*. Vol. 2010, 1321. Vienna, Austria: R Foundation for Statistical Computing. http://cran.r-project.org/doc/manuals/fullrefman.pdf.

———. "Google." Google, http://www.google.com/(accessed May 5, 2014).

———. "PubMed." National Center for Biotechnology Information, http://pubmed.gov(accessed May 5, 2014).

Aphinyanaphongs, Y., I. Tsamardinos, A. Statnikov, D. Hardin, and C. F. Aliferis. 2005. "Text Categorization Models for High-Quality Article Retrieval in Internal Medicine." *Journal of the American Medical Informatics Association : JAMIA* 12 (2): 207-216.

Badgett, R. G. 1999. "How to Search for and Evaluate Medical Evidence." *Seminars in Medical Practice* 2 (3): 8-14. http://www.turner-white.com/smp/abstractSMP.php?PubCode=smp_oct99_medevid.

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto. 2011. "Modeling." In *Modern Information Retrieval: The Concepts and Technology Behind Search*. 2nd ed., 57-130. Boston, MA: Addison-Wesley Professional.

Battelle, John. 2005. "Perfect Search." In *The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed our Culture*. 1st ed., 274-276. New York: Portfolio.

380   Blair D. 1980. "Searching Biases in Large Interactive Document Retrieval Systems." *J Am Soc Inf Sci* 31: 271-77.

381   DiCenso, A., L. Bayley, and R. B. Haynes. 2009. "ACP Journal Club. Editorial: Accessing Preappraised Evidence:
382   Fine-Tuning the 5S Model into a 6S Model." *Annals of Internal Medicine* 151 (6): JC3-2, JC3-3.
383   http://www.annals.org/content/151/6/JC3-2.citation.

384   Duran-Nelson, Alisa, Sophia Gladding, Jim Beattie, and L. James Nixon. 2013. "Should we Google it? Resource
385   use by Internal Medicine Residents for Point-of-Care Clinical Decision Making." *Acad Med* Publish Ahead
386   of Print (6). doi:10.1097/ACM.0b013e31828ffdb7.

387   Edson, R. S., T. J. Beckman, C. P. West, P. B. Aronowitz, R. G. Badgett, D. A. Feldstein, M. C. Henderson, J. C.
388   Kolars, and F. S. McDonald. 2010. "A Multi-Institutional Survey of Internal Medicine Residents' Learning
389   Habits." *Medical Teacher* 32 (9): 773-775.

390   Ely, J. W., J. A. Osheroff, M. L. Chambliss, M. H. Ebell, and M. E. Rosenbaum. 2005. "Answering Physicians'
391   Clinical Questions: Obstacles and Potential Solutions." *Journal of the American Medical Informatics
392   Association : JAMIA* 12 (2): 217-224. doi:M1608 [pii]; 10.1197/jamia.M1608 [doi].

393   Ely, J. W., J. A. Osheroff, M. H. Ebell, G. R. Bergus, B. T. Levy, M. L. Chambliss, and E. R. Evans. 1999. "Analysis of
394   Questions Asked by Family Doctors regarding Patient Care." *BMJ* 319 (7206): 358-61.

395   Ely, J. W., J. A. Osheroff, M. H. Ebell, M. L. Chambliss, D. C. Vinson, J. J. Stevermer, and E. A. Pifer. 2002.
396   "Obstacles to Answering Doctors' Questions about Patient Care with Evidence: Qualitative Study." *BMJ*
397   324 (7339): 710.

398   Fenton, S. H. and R. G. Badgett. 2007. "A Comparison of Primary Care Information Content in UpToDate and
399   the National Guideline Clearinghouse " *Journal of the Medical Library Association : JMLA* 95 (3): 255-259.

400   Freeman, M. K., S. A. Lauderdale, M. G. Kendrach, and T. W. Woolley. 2009. "Google Scholar Versus PubMed in
401   Locating Primary Literature to Answer Drug-Related Questions." *The Annals of Pharmacotherapy* 43 (3):
402   478-484. doi:10.1345/aph.1L223.

403   Gehanno, J. F., L. Rollin, and S. Darmoni. 2013. "Is the Coverage of Google Scholar enough to be used Alone for
404   Systematic Reviews?" *BMC Medical Informatics and Decision Making* 13: 7-11. doi:10.1186/1472-6947-
405   13-7; 10.1186/1472-6947-13-7.

406   Haase, A., M. Follmann, G. Skipka, and H. Kirchner. 2007. "Developing Search Strategies for Clinical Practice
407   Guidelines in SUMSearch and Google Scholar and Assessing their Retrieval Performance." *BMC Medical
408   Research Methodology* 7: 28.

409   Haynes, R. B., K. A. McKibbon, N. L. Wilczynski, S. D. Walter, S. R. Werre, and Hedges Team. 2005. "Optimal
410   Search Strategies for Retrieving Scientifically Strong Studies of Treatment from Medline: Analytical
411   Survey." *BMJ (Clinical Research Ed.)* 330 (7501): 1179. doi:10.1136/bmj.38446.498542.8F.

412   Haynes, R. B., N. Wilczynski, K. A. McKibbon, C. J. Walker, and J. C. Sinclair. 1994. "Developing Optimal Search
413   Strategies for Detecting Clinically Sound Studies in MEDLINE." *J Am Med Inform Assoc* 1 (6): 447-58.

414 Herskovic, J. R., M. S. Iyengar, and E. V. Bernstam. 2007. "Using Hit Curves to Compare Search Algorithm
415   Performance." *Journal of Biomedical Informatics* 40 (2): 93-99. doi:10.1016/j.jbi.2005.12.007.

416 Ioannidis, J. P. 2005. "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research." *JAMA* 294
417   (1538-3598): 218-228. doi:10.1001/jama.294.2.218.

418 Islamaj Dogan, R., G. C. Murray, A. Neveol, and Z. Lu. 2009. "Understanding PubMed User Search Behavior
419   through Log Analysis " *Database : The Journal of Biological Databases and Curation* 2009: bap018.
420   doi:10.1093/database/bap018.

421 Ketchum, A. M., A. A. Saleh, and K. Jeong. 2011. "Type of Evidence Behind Point-of-Care Clinical Information
422   Products: A Bibliometric Analysis." *Journal of Medical Internet Research* 13 (1): e21.

423 Lucas, Brian P., Arthur T. Evans, Brendan M. Reilly, Yuri V. Khodakov, Kalyani Perumal, Louis G. Rohr, Joseph A.
424   Akamah, Tunji M. Alausa, Christopher A. Smith, and Jeremy P. Smith. 2004. "The Impact of Evidence on
425   Physicians' Inpatient Treatment Decisions." *Journal of General Internal Medicine* 19 (5p1): 402-409.
426   doi:10.1111/j.1525-1497.2004.30306.x.

427 McKibbon, K. A. and D. B. Fridsma. 2006. "Effectiveness of Clinician-Selected Electronic Information Resources
428   for Answering Primary Care Physicians' Information Needs." *Journal of the American Medical Informatics
429   Association : JAMIA* 13 (6): 653-659. doi:10.1197/jamia.M2087.

430 McKibbon, K. A., N. L. Wilczynski, R. B. Haynes, and Hedges Team. 2009. "Retrieving Randomized Controlled
431   Trials from Medline: A Comparison of 38 Published Search Filters." *Health Information and Libraries
432   Journal* 26 (3): 187-202. doi:10.1111/j.1471-1842.2008.00827.x.

433 Mulrow, C. D. 1994. "Rationale for Systematic Reviews." *BMJ (Clinical Research Ed.)* 309 (6954): 597-599.

434 Nourbakhsh, E., R. Nugent, H. Wang, C. Cevik, and K. Nugent. 2012. "Medical Literature Searches: A
435   Comparison of PubMed and Google Scholar " *Health Information and Libraries Journal* 29 (3): 214-222.
436   doi:10.1111/j.1471-1842.2012.00992.x; 10.1111/j.1471-1842.2012.00992.x.

437 R Development Core Team. 2012. *R: A Language and Environment for Statistical Computing*. Vol. 2.15.1.
438   Vienna, Austria: R Foundation for Statistical Computing.

439 Sayers, E. "E-Utilities Quick Start." National Center for Biotechnology Information, last modified December 14,
440   2011, http://www.ncbi.nlm.nih.gov/books/NBK25500/(accessed May 5, 2014).

441 Schwartz, B. "Google Drops another Search Operator: Tilde for Synonyms ",
442   http://searchengineland.com/google-drops-another-search-operator-tilde-for-synonyms-
443   164403(accessed 5/6/2014, 2014).

444 Shariff, S. Z., S. A. Bejaimal, J. M. Sontrop, A. V. Iansavichus, R. B. Haynes, M. A. Weir, and A. X. Garg. 2013.
445   "Retrieving Clinical Evidence: A Comparison of PubMed and Google Scholar for Quick Clinical Searches."
446   *Journal of Medical Internet Research* 15 (8): e164. doi:10.2196/jmir.2624; 10.2196/jmir.2624.

447 Sim, M. G., E. Khong, and M. Jiwa. 2008. "Does General Practice Google? " *Australian Family Physician* 37 (6):
448   471-474.

449 Thiele, R. H., N. C. Poiro, D. C. Scalzo, and E. C. Nemergut. 2010. "Speed, Accuracy, and Confidence in Google,
450   Ovid, PubMed, and UpToDate: Results of a Randomised Trial " *Postgraduate Medical Journal* 86 (1018):
451   459-465. doi:10.1136/pgmj.2010.098053.

452 Toth, B., J. A. Gray, and A. Brice. 2005. "The Number Needed to Read-a New Measure of Journal Value " *Health
453   Information and Libraries Journal* 22 (2): 81-82. doi:10.1111/j.1471-1842.2005.00568.x.

454 Wilczynski, N. L., K. A. McKibbon, and R. B. Haynes. 2011. "Search Filter Precision can be Improved by NOTing
455   Out Irrelevant Content " *AMIA ...Annual Symposium Proceedings / AMIA Symposium.AMIA Symposium*
456   2011: 1506-1513.

457 Yu, H. "AskHermes - the Clinical Question Answering System " University of Wisconsin-Milwaukee,
458   http://www.askhermes.org/(accessed May 5, 2014).

459
460

Table 1. Example clinical question and resulting search strategy

* For users to reproduce the strategies with the current version of Google, settings are configured for "Google Instant Predictions" to be off and Results per page to be 50. The tilde signs are no longer required by Google as Google currently searches for synonyms by default. Since execution of our study, Google has revised Scholar to allow a maximum of 20 results per page.

Table 1. Example clinical question and resulting search strategy (Clinical Questions Collection #

NQ001384).

Original question by primary care physician	"If someone had x-rays for acne treatment, how should they be followed-up regarding thyroid cancer risk?"
Keywords originally assigned by the Clinical Questions Collection	Thyroid neoplasms Radiation Injuries
Search submitted to PubMed's Clinical Queries (Therapy category) and to Experimental search	Thyroid cancer AND Radiation Injuries
Search submitted to Google and Scholar*	Thyroid cancer Radiation Injuries PMID ~random ~trial

* For users to reproduce the strategies with the current version of Google, settings are configured for "Google Instant Predictions" to be off and Results per page to be 50. The tilde signs are no longer required by Google as Google currently searches for synonyms by default. Since execution of our study, Google has revised Scholar to allow a maximum of 20 results per page.

1

Selection of questions

4654 questions

2493 questions about nonpregnant adults

367 questions about treatment

312 questions
· unique keywords
· did not include a diagnostic component

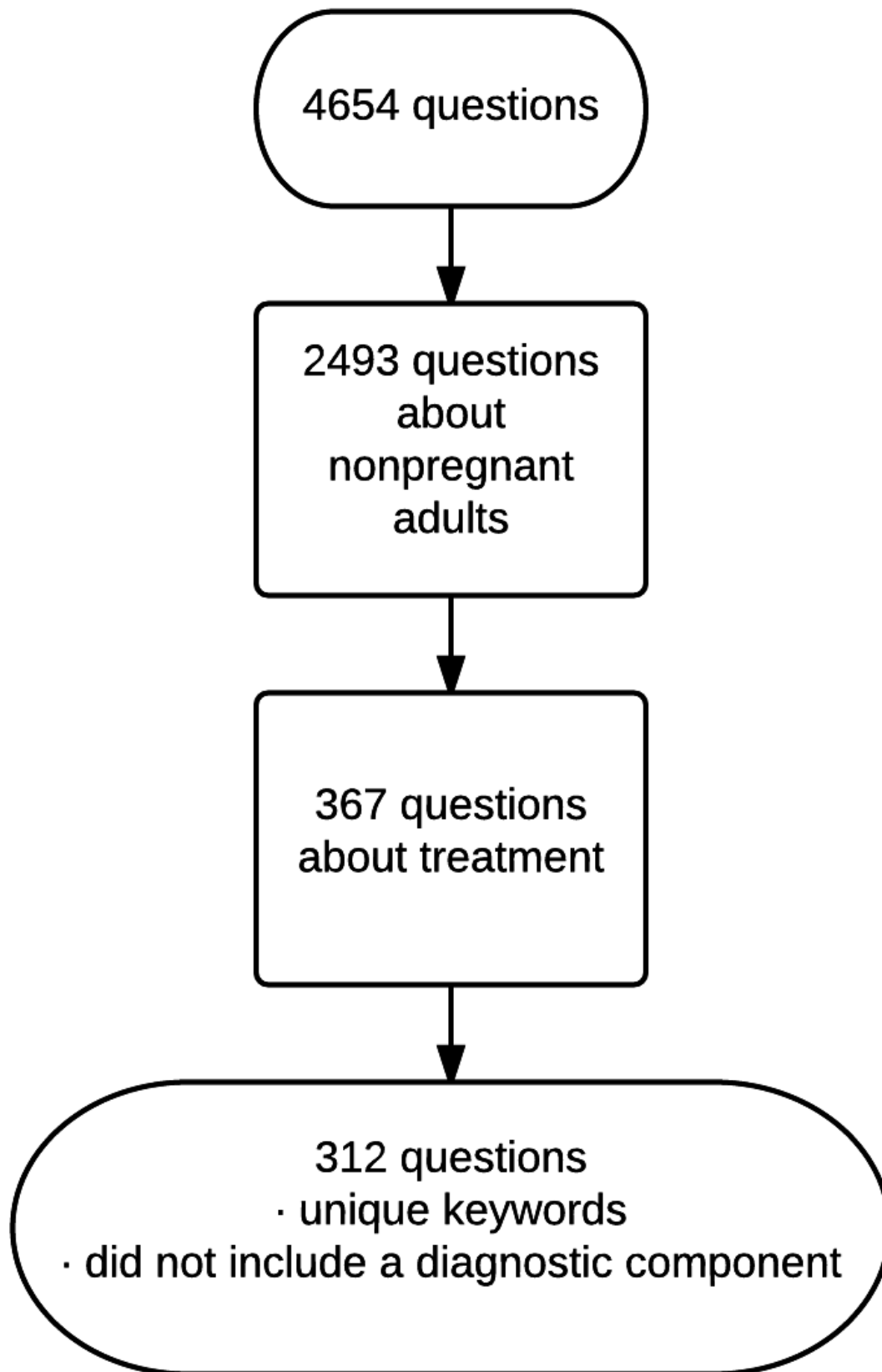



Table 2. MEDLINE iterations of the experimental search strategy

* Filters are detailed in the original study by Haynes(Haynes et al. , 2005) and at http://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx. † Journals are listed at http://hiru.mcmaster.ca/hiru/hedges/MedlineJournalsRead.pdf. ‡ Medical Subject Headings terms assigned by the National Library of Medicine.

Table 2. MEDLINE iterations of the experimental search strategy.

| Iteration | Options to increase specificity of search | | |
	Quality filters	**Publication types**	**Additional query expansion**
1	No filter	None	None
2	*Haynes 2005 sensitive filter* or systematic review subset*	*Excluded publication type of review, letter, editorial*	*Required abstract*
3	*Switched to Haynes 2005 specific filter or systematic review **	No change	No change
4	*Added restriction to 106 journals in McMaster list as of 02/2008 †*	No change	No change
5	No change	No change	*Added restriction of search terms to MeSH‡ major field*

* Filters are detailed in the original study by Haynes(Haynes et al. , 2005) and at

http://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx.

† Journals are listed at http://hiru.mcmaster.ca/hiru/hedges/MedlineJournalsRead.pdf.

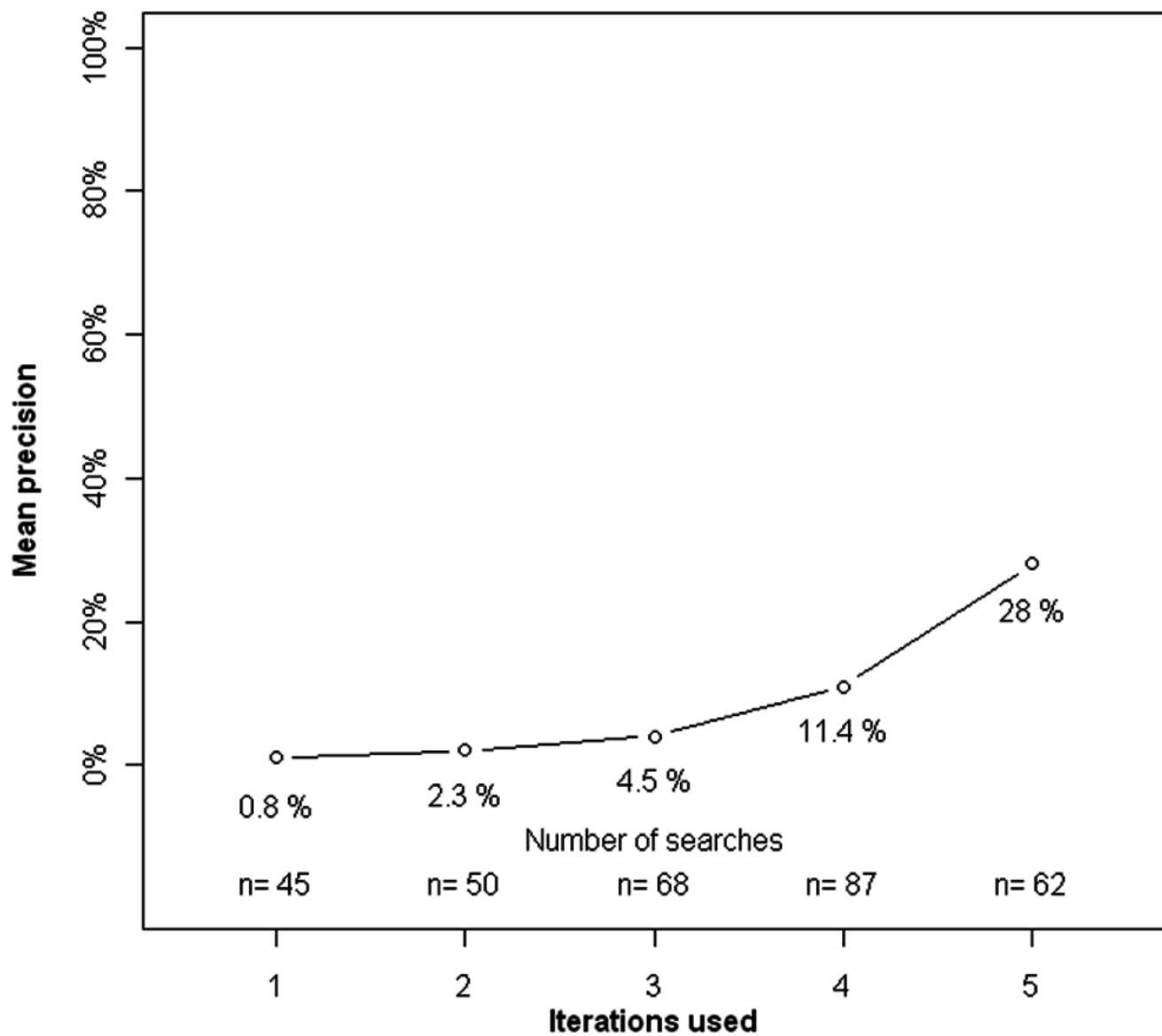
‡ Medical Subject Headings terms assigned by the National Library of Medicine.

Precision by number of interations used by the experimental search engine

Table 3. Comparison of search strategies for retrieving high quality*, relevant PubMed citations

High quality citations were those reviewed by an evidence-based synoptic journal or accompanied by an editorial

Table 3. Comparison of search strategies for retrieving high quality*, relevant PubMed citations.

Experimental	PubMed Clinical Queries for therapies		Google	Google Scholar
	narrow	broad		
Precision of searches, median (interquartile range)†‡				
5.5% §	4.0%§	2.0%	0%§	0%§
(0% to 12%)	(0% to 10%)	(0% to 8%)	(0% to 7%)	(0% to 0%)
Number of citations retrieved, median (interquartile range) †‡				
2§	2	1	0§	0§
(0 to 6)	(0 to 4)	(0 to 3)	(0 to 2)	(0 to 0)
Proportion of searches that retrieved at least one citation (95% confidence intervals) †				
73%§	63%	65%	46%§	20%§
(68% to 78%)	(58% to 68%)	(59% to 70%)	(41% to 52%)	(15% to 24%)

* High quality citations were those reviewed by an evidence-based synoptic journal or accompanied by an editorial.

† P < 0.001 for differences among groups.

‡ Note that rank sums can differ significantly although medians are the same.

§ P < 0.05 for difference compared to other groups.

Search results were limited to a maximum of 50 per search.