Is there a consensus when physicians evaluate the relevance of retrieved systematic reviews?

Dympna O'Sullivan[1,2], Szymon Wilk[2,3], Craig Kuziemsky[2], Wojtek Michalowski[2], Ken Farion[2,4,5], Bartosz Kukawka[3]

[1] School of Mathematics, Computer Science and Engineering, City University London
Northampton Square, London, EC1V 0HB, United Kingdom

[2] MET Research Group, Telfer School of Management, University of Ottawa
55 Laurier Ave. E., Ottawa, Ontario, K1N 6N5 Canada.

[3] Institute of Computing Science, Poznan University of Technology
Piotrowo 2, 60-965 Poznan, Poland

[4] Division of Emergency Medicine, Children's Hospital of Eastern Ontario
[5] Departments of Pediatrics and Emergency Medicine, University of Ottawa
401 Smyth Rd., Ottawa, Ontario, K1H 8L1 Canada

Corresponding author:
Szymon Wilk
Institute of Computing Science, Poznan University of Technology
Piotrowo 2, 60-965 Poznan, Poland.
e-mail: szymon.wilk@cs.put.poznan.pl, tel.: +48 61 6652930, fax: +48 61 8771525

Abstract

Background: A significant challenge associated with practicing evidence-based medicine is to provide physicians with relevant clinical information when it is needed. At the same time it appears that the notion of relevance is subjective and its perception is affected by a number of contextual factors.

Objectives: To assess to what extent physicians agree on the relevance of evidence in the form of systematic reviews for a common set of patient cases, and to identify possible contextual factors that influence their perception of relevance.

Methods: A web-based survey was used where pediatric emergency physicians from multiple academic centers across Canada were asked to evaluate the relevance of systematic reviews retrieved automatically for 14 written case vignettes (paper patients). The vignettes were derived from prospective data describing pediatric patients with asthma exacerbations presenting at the emergency department. To limit the cognitive burden on respondents, the number of reviews associated with each vignette was limited to 3.

Results: 22 academic emergency physicians with varying years of clinical practice completed the survey. There was no consensus in their evaluation of relevance of the retrieved reviews and physicians' assessments ranged from very relevant to irrelevant evidence, with the majority of evaluations being somewhere in the middle. This indicates that the study participants did not share a notion of relevance uniformly. Further analysis of commentaries provided by the physicians allowed identifying three possible contextual factors: expected specificity of evidence (acute vs chronic condition), the terminology used in the systematic reviews, and the micro environment of clinical setting.

Conclusion: There is no consensus among physicians with regards to what constitutes relevant clinical evidence for a given patient case. Subsequently, this finding suggests that evidence retrieval systems should allow for deep customization with regards to physician's preferences and contextual factors, including differences in the micro environment of each clinical setting.

Keywords: evidence-based medicine; review, systematic; information storage and retrieval; evaluation study; pediatric asthma

1. Introduction

A large percentage of health professionals' clinical practice relate to information management [1]. A specific aspect of this practice is evidence-based medicine (EBM), described as "an increasingly popular usage model for information within medical informatics" [2] that advocates use of the best evidence to make optimal decisions about patient care [3]. This best evidence usually comes from systematic reviews (e.g. from the Cochrane Library [4]), and is often disseminated in the form of guidelines. Regardless of the source, providing evidence involves identification of high quality scientific publications pertaining to a topic of clinical interest and relevant for a clinical presentation and patient context. Difficulty in assessing how good an information systems is [5], as well as a lack of quick and easy identification, appraisal, and synthesis of best evidence [6] are often pointed out as reasons for low uptake of EBM [7], especially outside academic centers [8].

While the biggest volume of research on EBM is related to its perceptions and use by clinicians (see for example [9]), provision of the right type of information has been identified as a significant challenge for the practice of EBM [10]. This challenge is closely associated with the fact that evidence is often context-independent while the perceived relevance of evidence is influenced by contextual factors [10]. Although it has been reported that various elements modify the relevance of evidence [2], there is less research that has studied how physicians evaluate the relevance of provided evidence for a specific clinical presentation and patient context. To address this void we need a better understanding of whether there is some common ground in evaluating the relevance of retrieved evidence, and subsequently how this common ground (or lack thereof), should contribute to better information management by physicians.

Research described in this paper attempts to assess how physicians perceive the relevance of evidence, in this case defined as systematic reviews from the Cochrane Library, for specific clinical presentations (pediatric asthma exacerbations) and patient contexts. The evidence is provided through a set of written case vignettes (vignettes in short) in defined clinical setting (emergency

department (ED)). The hypothesis behind our research is that *a consensus in evaluating the relevance of retrieved evidence does not exist due to a number of contextual factors affecting how physicians perceive this notion*. To validate this hypothesis we have formulated the following research questions:

- To what extent do physicians agree on the relevance of retrieved systematic reviews for a common set of vignettes?
- What are the contextual factors that may influence the evaluation of relevance of retrieved systematic reviews?

2. Related work

Research on evidence retrieval can be divided into three broad categories. The first category includes research aimed at developing tools (e.g. algorithms, systems) for indexing and retrieval of evidence from various sources (e.g. Medline or the Cochrane Library) [11,12]. The second category is about integrating these tools with point of care clinical workflows [13,14]. The third category of research is concerned with examining factors that affect the acceptance, uptake and use of these tools [15,16].

Our work belongs to the first category. In principle, research from this category is more concerned with the mechanics of indexing, retrieving and using evidence rather than the physicians' perception of the relevance of the evidence for a specific clinical presentation [17]. Subsequently, researchers often assume that relevance is an objective notion shared across the spectrum of physicians [18]. In contrast, in our study we try to establish whether there is a consensus among physicians with regards to their perception of the *relevance of evidence*. If such a consensus exists, then we can say that the expert panels' evaluations represent a "gold standard" for assessing how well indexing and retrieval algorithms perform. Otherwise, any standard used for assessing the quality of indexing and retrieval algorithms is dependent on how this standard was developed and the composition of the physician group who participated in its development.

3. Material and methods

3.1. Study setting and population

The study described in this paper involved pediatric academic emergency physicians of varying clinical experience – measured as years of practice as full time ED clinical staff – working at major university hospitals across Canada. Prospective participants were sent email invitations to participate and enrolled physicians could withdraw their consent at any time during the study.

3.2. Study design

The study design was approved by the Ethics Review Board at the Children's Hospital of Eastern Ontario (CHEO). The study was conducted in three phases: (1) preparation, (2) survey and (3) analysis. It is important to stress that the overall purpose of this study was not to assess how well some retrieval algorithms matched systematic reviews to vignettes, nor to evaluate objective relevance (i.e. not related to any clinical context) of evidence, as in [19]. Instead, the purpose was to examine how physicians subjectively evaluated the relevance of retrieved reviews for the vignettes, and to check if and how these subjective relevance evaluations were consistent or not.

3.2.1. Preparation phase

In this phase we created vignettes that described pediatric patients presenting to the ED with asthma exacerbations of different acuity. First, with a help of a senior ED physician we selected 14 representative patient cases from a set of 82 cases collected prospectively at CHEO [20]. The selected cases were transformed into vignettes where each vignette included a brief textual description of the patient state that highlighted available information, verified diagnosis (e.g. severity of asthma exacerbation) and a therapy formulated according to the Canadian Academy of Emergency Physicians clinical practice guideline.

Subsequently, each vignette was associated with three systematic reviews from the Cochrane Library retrieved by an algorithm described in [11]. Specifically, we selected the three most relevant systematic reviews according to *retrieval ranks* computed by the algorithm – the

number of reviews was limited in order to diminish the cognitive burden imposed on the participating physicians and to reduce time they needed to complete the survey. For brevity, the presentation of each systematic review was limited to Title, Abstract and Plain Language Summary sections.

3.2.2. Survey phase

This phase started with a questionnaire, implemented via a web-based application, where the participating physicians were asked to provide information on basic demographics and about their experience with various clinical information systems. Before starting, all participants were given instructions about how to use the web-based application which included an explanation of the notions of relevance and non-relevance of documents (how well or not a document or set of documents meets the information need of the physician) by providing examples of systematic reviews not used in the study. Upon completing the questionnaire, the participants were requested to evaluate the relevance of systematic reviews associated with the vignettes (participants were blind to the ranks produced by the retrieval algorithm). Specifically, for each retrieved systematic review, physicians were asked to provide an *evaluation rank* to indicate whether and how this review was relevant for the patient case described in the vignette. Possible evaluation ranks were from 1 (most relevant) to 3 (least relevant). There was also a rank *X* to indicate an irrelevant review. The ranks 1—3 had to be unique within a vignette, but more than one review could be marked as irrelevant. At the same time we also asked physicians to give optional commentaries about the reviews associated with each vignette. A sample screenshot from the web-based application is given in Appendix 1.

In order to minimize the influence of the presentation order of systematic reviews on physicians' evaluation ranks, this order was randomized (i.e. the review presented in first position for one physician could be placed last for another), however, all physicians saw the same set of three systematic reviews for a given vignette.

The physicians' evaluations of the reviews were coded as triples following the schema proposed in [21] by comparing the retrieval rank assigned by the retrieval algorithm with the

evaluation rank assigned by a physician. Each element of a triple is one of the *X*, *N* and *Y* codes, where *X* indicates that a retrieved review was deemed irrelevant by a physician, *N* indicates a relevant retrieved review that according to physician was incorrectly ranked by the algorithm, and finally *Y* indicates a relevant retrieved review that was assigned the same rank by a physician and the algorithm. For example, the *YXN* triple means that the first and third retrieved reviews were found relevant by a physician, although she assigned another rank to the latter review than the algorithm, and the second retrieved review was deemed irrelevant.

3.2.3. Analysis phase

In this phase we analyzed the coded responses of the physicians. We used two measures to capture the match between algorithmic retrieval and physicians' evaluation ranks. The first measure is *precision at k* which gives the ratio of relevant documents among the top *k* retrieved results [22], specifically, we used *precision at 3*. The second measure is *group value function* proposed in [21] which is a function obtained by amalgamating preferences of physicians that takes into account not only relevance of retrieved reviews as evaluated by physicians, but also considers position of an evaluation in a triple – see Appendix 2.

Both measures were calculated for each physician-vignette pair, and then averaged by physician over all vignettes. We calculated the Cohen's kappa coefficients [23] to measure agreement between pairs of physicians (i.e. we analyzed coded triples associated with paired physicians) and then used these results to cluster physicians into groups. In particular, we defined the distance between pairs of physicians as (1 – kappa), and we applied a hierarchical clustering algorithm to create a cluster dendrogram [24].

We then used the dendrogram to drive the search for the most appropriate clustering. Specifically, we cut the dendrogram at different levels to obtain groupings with varying numbers of clusters (from 2 to the total number of participating physicians) and selected the clustering with the best overall quality (we used the average silhouette width as the proxy of the clustering quality [24])

and the largest number of clusters. We decided to maximize the number of clusters to obtain finer groupings of physicians assuming that this should help with the interpretation of the results.

4. Results

4.1. Participating physicians and responses

Personalized email invitations were sent to 60 academic emergency physicians, 27 of them consented to participate in the study, and later 1 physician withdrew her consent. 26 physicians finished the survey, however, 4 of them did not complete all the evaluations and their responses were excluded from the analysis. This resulted in 22 physicians included in the study (37% participation rate), which is sufficient for this type of survey – see for example [15]). The participating physicians demonstrated the entire spectrum of professional experience measured as work years as full time staff members. Moreover, the majority of them (20 out of 22) had prior experience with patient tracking systems and clinical decision support systems (including electronic evidence).

Physicians' evaluations of the retrieved reviews resulted in 308 (22 physicians times 14 vignettes) triples. All triples are presented in Appendix 3. Physicians also gave additional commentaries expanding their evaluations. Specifically, we recorded 21 commentaries provided by 6 participants (selected comments are presented in Figure 3).

4.2. Analysis of responses

We started the analysis by computing values of precision at 3 and group value function (see Figure 1 for overview and Appendix 4 for details). There is a group of physicians for whom the precision at 3 is high (0.8 and higher) and this group includes sp14, sp26, sp28 and sp29 – these high values indicate that physicians from this group considered most of the retrieved reviews as relevant for the cases described in the vignettes. It is especially evident for sp26, where precision at 3 is equal to 0.95. On the other hand, there are two physicians – sp5 and sp25 – with very low precision at 3 (0.33 and 0.36 respectively). This indicates that these two physicians evaluated the

majority of the reviews as being irrelevant, revealing a lack of consensus with the previous group. Values of the group value function are lower than those of precision at 3 (indeed this is expected by the definition of the measures where precision at 3 takes into account only what is relevant and what is not, whereas group value function adds ordering to reflect more or less relevance on higher and lower position), however they still confirm the divergent evaluations described above.

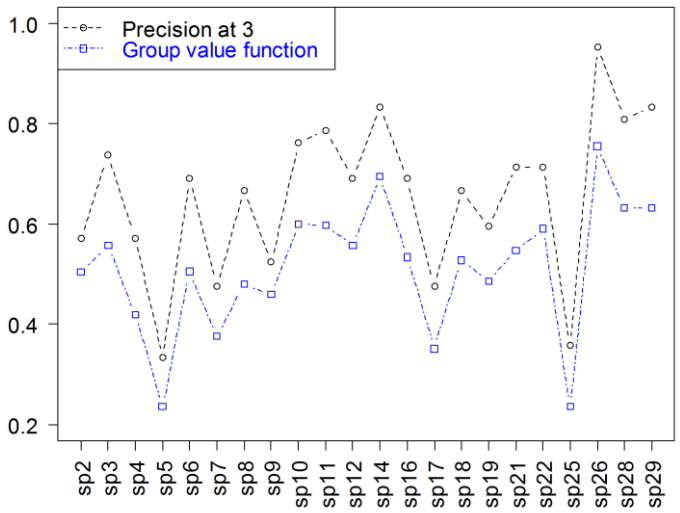


Figure 1. Precision at 3 and group value function averaged over vignettes

In the next step we conducted cluster analysis based on the inter-rater agreement captured by values of Cohen's kappa (see Appendix 5 for detailed results) – the best clustering was achieved by cutting the dendrogram at the level corresponding to 4 clusters. The dendrogram and the resulting clustering are presented in Figure 2.

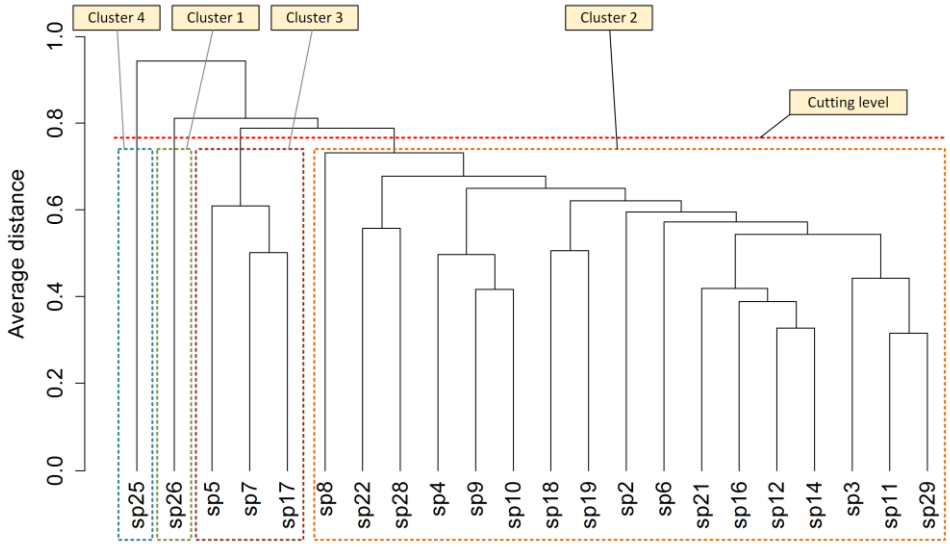


Figure 2. Dendrogram and the associated clustering

The clusters are further expanded in Appendix 6 in the context of precision at 3 and group value function measures. These clusters are as follows:

- Cluster 1: includes a single physician whose evaluation ranks were very similar to the retrieval ranks established by the algorithm (precision at 3 equal to 0.95 and group value function equal to 0.76),
- Cluster 2: includes 17 physicians for whom the match between evaluation and retrieval ranks was moderate to good (precision at 3 ranging from 0.52 to 0.83 and group value function ranging from 0.42 to 0.70),
- Cluster 3: includes 3 physicians for whom the match between evaluation and retrieval ranks was poor (precision at 3 ranging from 0.33 to 0.48 and group value function ranging from 0.24 to 0.38),
- Cluster 4: includes a single physician for whom evaluation ranks poorly matched the retrieval ranks (precision at 3 equal to 0.36 and group value function equal to 0.24).

The separation into clusters 3 and 4 may not be obvious at first glance, but it can be attributed to differences in evaluation ranks provided by physicians from these two clusters (specifically physicians sp5 and sp25) for the same vignettes. Details are given in Appendix 3.

The results of the analysis show large disparities in the evaluations across the physicians, ranging from evaluating systematic reviews as very relevant for the vignettes (sp26) to mostly irrelevant (sp5 and sp25), with the majority of physicians being somewhere in between. These results are also supported by the commentaries. Reading through these commentaries allowed us to identify three contextual factors that might influence how the physicians perceived the relevance of the retrieved systematic reviews:

1. *Expected specificity of evidence*: several physicians made comments that the retrieved evidence while relevant for adult asthma, is less so for a pediatric population (Figure 3a and 3b), or while relevant for chronic asthma, it was less appropriate for acute asthma managed in the ED (Figure 3c and 3d). Interestingly, for a number of physicians, this perceived lack of specificity for a patient case was not an issue.
2. *The terminology used in the reviews*: as illustrated by a comment provided by a physician from cluster 2, some terms used in the reviews (in this case "steroid-naive" – see Figure 3e) were not commonly understood, which prevented evaluating a given review as relevant for a patient case.
3. *The micro environment of clinical setting*: while systematic reviews are population-based, some physicians were looking for evidence better framed for the local context, specifically with respect to treatment options (see commentaries on Figure 3f).

It is interesting to note that according to Spearman's rank correlation (two-tailed test) [25] we did not observe significant impact of professional experience of participating physicians on their evaluations or commentaries. For example, correlation between experience and precision at 3 was equal to -0.1 (p-value = 0.64), while correlation between experience and the number of provided commentaries was equal to 0.29 (p-value = 0.18).

(a)	aminophylline study is useful if you want to remind MD not to bother with it (maybe adult MDs would need this reminder?)… probably more helpful just to include positive studies which tell doc what to do rather than what not to do [**sp17]**
(b)	The first review talks way too much about adults and only has one small line at the bottom about kids, which actually concludes LABAs shouldn't be used. Unless you take time to read the fine print, this review could lead you to change management, which in this case is actually detrimental [**sp7]**
(c)	None of these help me make an acute management decision for my patient. What's more, we don't usually prescribe LABAs in the ED and so studies pertaining to this (especially in the setting where they are potentially not as safe) are not useful to me [**sp7**]
(d)	I would have thought that the retrieved SRs should have been about the ED emergent treatment of asthma [**sp9]**
(e)	I would like to understand the definition of "steroid-naive" inhaled versus systemic before I can decide on the validity of this study with respect to this clinical vignette. [**sp16**]
(f)	More interested in evidence on Magnesium than oral xanthines….would have expected review on addition of anticholinergic to beta agonist to appear here [**sp12]**

Figure 3. Selected comments by participating physicians

Overall, our results indicate that it is difficult to talk about an “objective” notion of the relevance of evidence that is shared by all physicians participating in the study. Clearly, what is relevant or irrelevant from a physician’s perspective is subjective and its perception is influenced by contextual factors. While our analysis points to three of these factors, it is fair to state that the micro environment of the clinical setting to some extent supersedes the factors of expected specificity of evidence and the alignment of the review terminology with the terminology normally used in a given setting.

5. Discussion

5.1. Main results

Analysis of quantitative data gathered during the study showed large differences in the how physicians evaluated the relevance of systematic reviews. Assessments of the same reviews ranged from very relevant to irrelevant, with the majority of evaluations being somewhere in the middle.

Analysis of the commentaries suggests that possible contextual factors include the micro environment of clinical practice, combined in some cases with the need for evidence that is specific for a patient case, and in others with the need for summaries using practice-specific terminology, as opposed to more general terms used in the systematic reviews. All of these factors indicate that physicians seem to ignore the fact that evidence in the form of a systematic review is population-based, comes from a number of (randomized) clinical trials, and therefore is context-agnostic. Thus, we can state that the results of this research support our hypothesis that there is no consensus among physicians evaluating the relevance of evidence. This finding also implies that experts' evaluations that are normally used as a "gold standard" in assessing performance of information retrieval algorithms are of limited reliability because for the same set of retrieved reviews one expert panel may arrive at evaluations that are very different from what other panel might provide.

It is interesting to compare our findings with the research presented in [26], where the authors describe the preparation of the OHSUMED test collection that has become one of the standard benchmark sets for evidence retrieval algorithms. The OHSUMED collection includes queries recording patient data and information needs together with relevant references (titles, abstracts and MeSH keywords) retrieved from MEDLINE. While two reviewing physicians checked only 11% of the included references (the remaining ones were checked by a single reviewer), agreement between their assessments of relevance can be interpreted as moderate according to [27] (associated kappa score was 0.41). In our study the average kappa score (over all pairs of physicians) was 0.30, which is interpreted as fair agreement. This leads us to conclude that physicians do not really agree about what constitutes relevant evidence in the context of specific patient cases. Interestingly, such limited agreement is observed not only when evaluating the relevance of documents. For example, according to results reported in [28] the inter-rater agreement in assessing the quality of diagnostic studies was also fair (mean kappa equal to 0.22). Moreover, the agreement for visual assessment of MRI images observed in the study described in [29] was poor (kappa ranged from 0.19 to 0.39).

5.2. Limitations

The limitations of our research are as follows:

- Physicians recruited for the study came from a single clinical specialty (emergency medicine) that cannot be easily generalized to other practices. On the other hand it resulted in a homogeneous cohort that was less susceptible to responder bias.
- The study asked for the evaluation of pediatric asthma exacerbation cases presented in the ED; such a setting has some unique characteristics that are not easily ported to typical in-patient or out-patient practices.
- The evidence associated with each vignette included only systematic reviews retrieved from the Cochrane Library; while this library contains reviews of the highest quality, it is not used uniformly across different clinical settings and retrieval of the evidence from other repositories might result in more comprehensive assessment.
- The number of reviews per vignette was limited to 3. Increasing the number of reviews (e.g. from 3 to 5) might have resulted in better consistency across participants, however, additional workload might negatively impact the number of completed surveys.

5.3. Implications

There are a number of possible implications of our study for medical informatics. First, our findings suggest that clinical decision support systems with evidence retrieval functionality should allow for customization with regard to physician's preferences and contextual factors, especially the micro environment of the clinical setting. Secondly, broader use of common terms and concepts in describing results of studies on one hand, and in different settings when charting, etc. on the other hand should help with addressing the problem of lack of familiarity with the terminology [30,31]. Finally, the problem of evaluating the relevance of evidence for a specific patient case can be addressed by observing how well physicians perform a clinical task (for example, therapy development) after being presented with evidence of diversified types (coming from large patient population, from specific population, from a population presenting in a given clinical setting, etc.).

Such factors are worthy of further investigation as part of a study to determine pertinent practical features of tools for EBM. A possible approach is also advocated in [32] where the authors postulate to go beyond classical methods of assessing the relevance of retrieved evidence and focus instead on what they call “task-oriented” evaluation. Such evaluation should help discover physician’s preferences with regard to the evidence characteristics, thus facilitating the abovementioned customization of clinical decision support systems with evidence retrieval functionality.

Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (Collaborative Health Research Program), University of Ottawa Research Chairs Program, and the Children’s Hospital of Eastern Ontario Research Institute.

We are gratefully acknowledging the participation of the emergency medicine physicians in the study.

References

1. Shortliffe EH, Cimino JJ, editors. Biomedical Informatics. Computer Applications in Health Care and Biomedicine. 4th ed: Springer; 2014.

2. Cohen AM, Stavri PZ, Hersh WR. A categorization and analysis of the criticisms of Evidence-Based Medicine. Int J Med Inform 2004;73(1):35-43.

3. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996;312(7023):71-2.

4. Wiley InterScience. The Cochrane Library. In: Wiley InterScience; 2014.

5. Ammenwerth E. Evidence-based health informatics: How do we know what we know? Methods Inf Med 2015;54(4):298-307.

6. Montorti VM, Ebbert JO. TRIP database. Evid Based Med 2002;7:104.

7. Kho ME, Brouwers MC. The systematic review and bibliometric network analysis (SeBriNA) is a new method to contextualize evidence. Part 1: description. J Clin Epidemiol 2012;65(9):1010-5.

8. Gonzalez-Gonzalez AI, Dawes M, Sanchez-Mateos J, Riesgo-Fuertes R, Escortell-Mayor E, Sanz-Cuesta T, et al. Information needs and information-seeking behavior of primary care physicians. Ann Fam Med 2007;5(4):345-52.

9. Swennen MH, van der Heijden GJ, Boeije HR, van Rheenen N, Verheul FJ, van der Graaf Y, et al. Doctors' perceptions and use of evidence-based medicine: a systematic review and thematic synthesis of qualitative studies. Acad Med 2013;88(9):1384-96.

10. Pope C. Resisting evidence: the study of evidence-based medicine as a contemporary social movement. Health 2003;7(3):267-282.

11. O'Sullivan D, Wilk S, Michalowski W, Farion K. Automatic indexing and retrieval of encounter-specific evidence for point-of-care support. J Biomed Inform 2010;43(4):623-31.

12. Gatta R, Vallati M, De Bari B, Pasinetti N, Cappelli C, Pirola I, et al. Information retrieval in medicine: an extensive experimental study. In: Bienkiewicz M, Verdier C, Plantier G, Schultz T, Fred ALN, Gamboa H, editors. HEALTHINF 2014 - Proceedings of the International Conference on Health Informatics. Angers, France: SciTePress; 2014. p. 447-452.

13. Timsina P, El-Gayar O, Nawar N. Information technology for evidence based medicine: Status and future direction. In: 20th Americas Conference on Information Systems (AMCIS 2014): Smart Sustainability: The Information Systems Opportunity. Savannah, GA; 2014. p. 1149-1157.

14. Wilk S, Michalowski W, O'Sullivan D, Farion K, Sayyad-Shirabad J, Kuziemsky C, et al. A task-based support architecture for developing point-of-care clinical decision support systems for the emergency department. Methods Inf Med 2013;52(1):18-32.

15. Hung SY, Ku YC, Chien JC. Understanding physicians' acceptance of the Medline system for practicing evidence-based medicine: a decomposed TPB model. Int J Med Inform 2012;81(2):130-42.

16. Ellsworth MA, Homan JM, Cimino JJ, Peters SG, Pickering BW, Herasevich V. Point-of-care knowledge-based resource needs of clinicians. A survey from a large academic medical center. Appl Clin Inform 2015;6(2):305-317.

17. Shepperd S, Adams R, Hill A, Garner S, Dopson S. Challenges to using evidence from systematic reviews to stop ineffective practice: an interview study. J Health Serv Res Policy 2013;18(3):160-6.

18. Oliver D. Evidence based medicine needs to be more pragmatic. BMJ 2014;349:g4453.

19. Haynes RB, Cotoi C, Holland J, Walters L, Wilczynski N, Jedraszewski D, et al. Second-order peer review of the medical literature for clinical practitioners. JAMA 2006;295(15):1801-8.

20. Farion K, Wilk S, Michalowski W, O'Sullivan D, Sayyad-Shirabad J. Comparing predictions made by a prediction model, clinical score, and physicians: pediatric asthma exacerbations in the emergency department. Appl Clin Inform 2013;4(3):376-391.

21. O'Sullivan D, Wilk S, Michalowski W, Slowinski R, Thomas R, Kadzinski M, et al. Learning the preferences of physicians for the organization of result lists of medical evidence articles. Methods Inf Med 2014;53(5):344-356.

22. Manning C, Raghavan P, Schutze H. Introduction to Information Retrieval. New York, NY: Cambridge University Press; 2008.

23. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005;85(3):257-68.

24. Flach P. Machine Learning. The Art and Science of Algorithms that Make Sense of Data: Cambridge University Press; 2012.

25. Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. Radiology 2003;227(3):617-22.

26. Hersh W, Buckley C, Leone TJ, Hickam D. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: Springer; 1994. p. 192-201.

27. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):360-3.

28. Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D, et al. Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A preliminary assessment. Acad Radiol 2006;13(7):803-10.

29. Kapeller P, Barber R, Vermeulen RJ, Ader H, Scheltens P, Freidl W, et al. Visual rating of age-related white matter changes on magnetic resonance imaging: scale comparison, interrater agreement, and correlations with quantitative measurements. Stroke 2003;34(2):441-5.

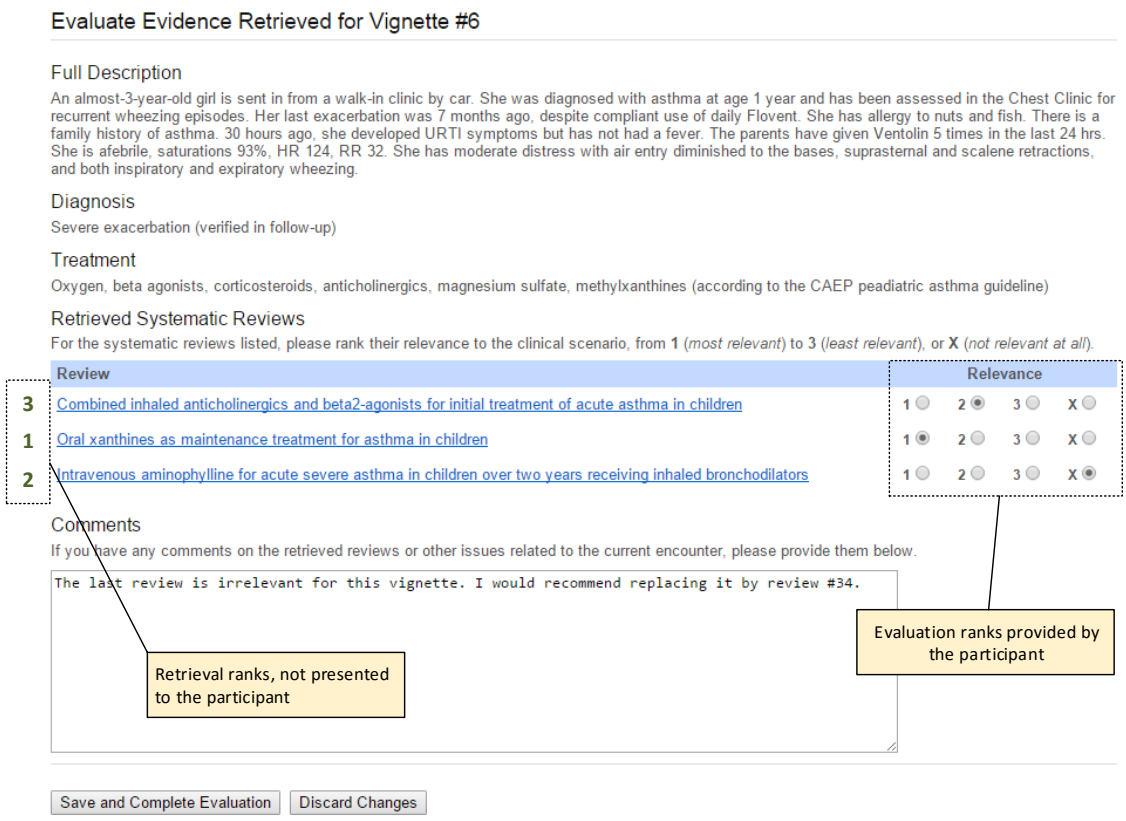
30. Carey TS, Melvin CL, Ranney LM. Extracting key messages from systematic reviews. J Psychiatr Pract 2008;14 Suppl 1:28-34.

31. Berges I, Bermudez J, Illarramendi A. Binding SNOMED CT terms to archetype elements. Establishing a baseline of results. Methods Inf Med 2015;54(1):45-

32. Hersh WR, Crabtree MK, Hickam DH, Sacherek L, Rose L, Friedman CP. Factors associated with successful answering of clinical questions using an information retrieval system. Bull Med Libr Assoc 2000;88(4):323-31.

Is there a consensus when physicians evaluate the relevance of retrieved systematic reviews? Online appendices

Appendix 1. Sample screen from the web-based survey application



Appendix 2. Group value function

	Marginal value		
Code	Position 1	Position 2	Position 3
Y	0.52	0.32	0.16
N	0.26	0.26	0.11
X	0.00	0.00	0.00

Group value function offers a thorough assessment of a triple that is richer than precision at 3. For example, the two coded triples *NXY* and *YXN* have the same value of precision at 3 (it is 2/3 = 0.66), while the group value function gives a better insight into ranking as it indicates that the latter triple is "more preferred" than the former (0.26 + 0.00 + 0.16 = 0.42 for *NXY* vs. 0.52 + 0.00 + 0.11 = 0.63 for *YXN*).

Appendix 3. Coded triples representing relevance evaluations by physicians

	Vignette													
Participant	1	2	3	4	5	6	7	8	9	10	11	12	13	14
sp2	XXX	XXX	YXX	NXN	YNN	NYN	YXX	NYN	XXN	NYN	YXX	NYN	YXX	YYX
sp3	XNN	XXX	XNN	NXN	YNN	NYN	YXX	NYN	XYX	NYN	YNN	NYN	NNY	YXN
sp4	NNY	XXX	NXN	XXN	YXN	XYN	YXX	XYN	YXN	XNX	YXN	XYN	XXN	NNN
sp5	XXX	XXX	XXX	XXX	NXX	XNN	YXX	XNN	NNX	XNN	NXX	XNN	XXX	NXX
sp6	XNX	YXN	XXN	XXN	NXN	NYN	NNN	NYN	XXX	NXN	YNN	NYN	YYY	YXN
sp7	XXN	XXX	XXY	XXX	YXY	NYN	YNX	NNN	XXY	XYN	YXX	XYN	XXY	YXX
sp8	XXX	NXN	XYN	NXN	YNN	XYN	YNN	XYN	NXN	XXN	YNN	XYY	XYN	YXN
sp9	NNY	YXX	XXX	XXN	YXX	XYN	YXX	XYN	XXX	XNN	YXX	XYN	YYY	YNN
sp10	NNY	YXN	NXN	NXN	YNN	XYN	YYY	XYN	XXY	XXN	YYY	XYN	YYY	YNN
sp11	NNY	YXN	NXN	NXN	YNN	NYN	YXX	NYN	XXX	XYN	YNN	NYN	NNN	YNN
sp12	NNN	XXX	NXN	NXN	YXN	NYN	YYX	NYN	YXX	NXN	YYY	XYN	YXN	YXN
sp14	NYN	YXX	NNN	NNN	YNN	NNN	YYX	NYN	YXX	NXN	YYY	NYN	YYY	YXN
sp16	NNN	YXX	XXN	NXN	NXN	NNN	YYX	NYN	YXX	NXN	NNX	NNN	YXN	YXN
sp17	XXN	NXX	XXN	XXX	YXX	XNN	YXX	XNN	XXY	XNN	YXY	XNN	XNN	YXY
sp18	NYN	XXX	NNN	YXN	YNN	NYN	YXX	XYN	YNX	NXN	YXN	XXN	NXN	YXN
sp19	XYN	XXX	YXN	XXX	YYY	NNN	YXX	NYN	YXX	NYN	YXX	XYN	NXN	NXN
sp21	NXN	YXX	XXN	NXN	NXN	NYN	YXX	NYN	XXN	NYN	YYY	NYN	YNN	YXN
sp22	NNN	XNN	NNN	YNN	YXX	NYN	YYX	XYN	YYY	NYN	YXX	XYN	XXN	YXX
sp25	NNX	XXX	NXN	XXX	XXN	NNX	XNX	NNX	XXX	NXX	XXN	XNX	XNX	XXN
sp26	YYY	YNN	NNN	NNN	YNN	NNN	YYY	XYN	YYY	XNN	YNN	NNN	NNY	YNN
sp28	YNN	YNN	NNN	NXN	NXN	NYN	YNX	NYN	YNX	XYN	YXN	XYN	XNN	YNN
sp29	NNY	NXN	NXN	NXN	YNN	NYN	YNX	NYN	NYX	XYN	YYY	NYN	NYN	YXN

There are major differences between evaluations provided by different physicians for the same vignette. For example, physician sp25 evaluated reviews retrieved for vignette 9 as *XXX* (considered all of them to be irrelevant), while physician sp26 evaluated the same reviews as *YYY* (considered all of them not only as relevant but also correctly ranked by the retrieval algorithm).

Appendix 4. Precision at 3 and group value function averaged over vignettes (95% CI)

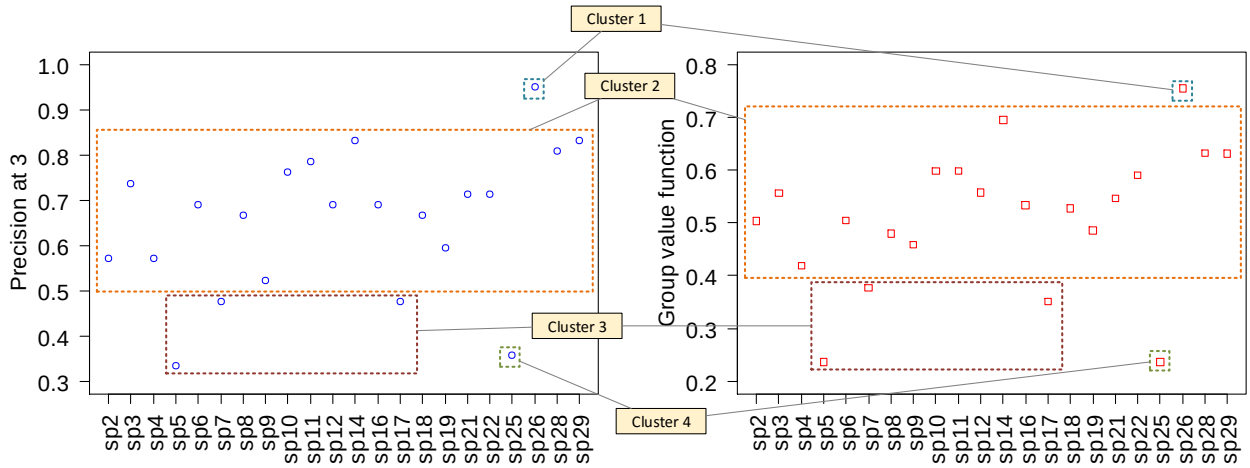
Participant	Precision at 3	Group value function
sp2	0.57 [0.37, 0.77]	0.50 [0.35, 0.65]
sp3	0.74 [0.57, 0.91]	0.56 [0.43, 0.69]
sp4	0.57 [0.43, 0.72]	0.42 [0.30, 0.54]
sp5	0.33 [0.18, 0.49]	0.24 [0.13, 0.34]
sp6	0.69 [0.52, 0.86]	0.51 [0.35, 0.66]
sp7	0.48 [0.31, 0.64]	0.38 [0.23, 0.52]
sp8	0.67 [0.53, 0.80]	0.48 [0.34, 0.62]
sp9	0.52 [0.35, 0.70]	0.46 [0.31, 0.61]
sp10	0.76 [0.64, 0.89]	0.60 [0.43, 0.76]
sp11	0.79 [0.62, 0.95]	0.60 [0.47, 0.73]
sp12	0.69 [0.55, 0.84]	0.56 [0.43, 0.68]
sp14	0.83 [0.70, 0.97]	0.70 [0.60, 0.79]
sp16	0.69 [0.56, 0.82]	0.53 [0.44, 0.63]
sp17	0.48 [0.36, 0.59]	0.35 [0.24, 0.46]
sp18	0.67 [0.51, 0.82]	0.53 [0.40, 0.66]
sp19	0.60 [0.41, 0.78]	0.49 [0.35, 0.62]
sp21	0.71 [0.56, 0.87]	0.55 [0.41, 0.68]
sp22	0.71 [0.56, 0.87]	0.59 [0.47, 0.71]
sp25	0.36 [0.23, 0.48]	0.24 [0.14, 0.34]
sp26	0.95 [0.89, 1.02]	0.76 [0.64, 0.87]
sp28	0.81 [0.72, 0.90]	0.63 [0.52, 0.74]
sp29	0.83 [0.74, 0.92]	0.63 [0.53, 0.73]

Appendix 5. Values of Cohen's kappa for pairs of participants

	Participant																					
Participant	sp2	sp3	sp4	sp5	sp6	sp7	sp8	sp9	sp10	sp11	sp12	sp14	sp16	sp17	sp18	sp19	sp21	sp22	sp25	sp26	sp28	sp29
sp2		0.50	0.27	0.18	0.32	0.44	0.24	0.34	0.27	0.44	0.43	0.34	0.32	0.14	0.32	0.42	0.54	0.26	-0.05	0.08	0.19	0.34
sp3	0.50		0.23	0.03	0.47	0.38	0.30	0.34	0.28	0.61	0.49	0.47	0.38	0.22	0.51	0.43	0.55	0.30	0.09	0.29	0.37	0.50
sp4	0.27	0.23		0.20	0.15	0.26	0.30	0.56	0.44	0.45	0.49	0.19	0.21	0.32	0.45	0.38	0.31	0.35	0.11	0.20	0.38	0.32
sp5	0.18	0.03	0.20		0.00	0.34	0.07	0.29	0.01	-0.05	0.06	-0.01	0.17	0.44	0.10	0.27	0.07	0.12	-0.19	0.02	0.07	0.04
sp6	0.32	0.47	0.15	0.00		0.30	0.41	0.40	0.40	0.43	0.41	0.37	0.46	0.09	0.18	0.21	0.48	0.09	0.17	0.09	0.31	0.36
sp7	0.44	0.38	0.26	0.34	0.30		0.22	0.35	0.21	0.26	0.37	0.20	0.23	0.50	0.22	0.49	0.37	0.37	0.07	0.05	0.30	0.27
sp8	0.24	0.30	0.30	0.07	0.41	0.22		0.26	0.43	0.33	0.27	0.20	0.13	0.27	0.26	0.10	0.27	0.13	-0.09	0.15	0.25	0.41
sp9	0.34	0.34	0.56	0.29	0.40	0.35	0.26		0.58	0.49	0.40	0.32	0.34	0.39	0.29	0.27	0.41	0.26	-0.06	0.22	0.32	0.36
sp10	0.27	0.28	0.44	0.01	0.40	0.21	0.43	0.58		0.57	0.54	0.48	0.29	0.25	0.29	0.09	0.39	0.24	0.04	0.40	0.31	0.52
sp11	0.44	0.61	0.45	-0.05	0.43	0.26	0.33	0.49	0.57		0.48	0.38	0.37	0.19	0.40	0.32	0.55	0.22	0.11	0.38	0.52	0.68
sp12	0.43	0.49	0.49	0.06	0.41	0.37	0.27	0.40	0.54	0.48		0.67	0.67	0.25	0.60	0.50	0.64	0.49	0.24	0.11	0.44	0.56
sp14	0.34	0.47	0.19	-0.01	0.37	0.20	0.20	0.32	0.48	0.38	0.67		0.55	0.16	0.49	0.37	0.52	0.30	0.14	0.34	0.26	0.46
sp16	0.32	0.38	0.21	0.17	0.46	0.23	0.13	0.34	0.29	0.37	0.67	0.55		0.19	0.35	0.38	0.58	0.27	0.15	0.15	0.29	0.31
sp17	0.14	0.22	0.32	0.44	0.09	0.50	0.27	0.39	0.25	0.19	0.25	0.16	0.19		0.19	0.33	0.36	0.22	0.01	0.15	0.16	0.18
sp18	0.32	0.51	0.45	0.10	0.18	0.22	0.26	0.29	0.29	0.40	0.60	0.49	0.35	0.19		0.49	0.33	0.49	0.08	0.25	0.39	0.37
sp19	0.42	0.43	0.38	0.27	0.21	0.49	0.10	0.27	0.09	0.32	0.50	0.37	0.38	0.33	0.49		0.36	0.32	0.12	0.07	0.25	0.30
sp21	0.54	0.55	0.31	0.07	0.48	0.37	0.27	0.41	0.39	0.55	0.64	0.52	0.58	0.36	0.33	0.36		0.24	0.09	0.05	0.39	0.47
sp22	0.26	0.30	0.35	0.12	0.09	0.37	0.13	0.26	0.24	0.22	0.49	0.30	0.27	0.22	0.49	0.32	0.24		-0.03	0.25	0.44	0.30
sp25	-0.05	0.09	0.11	-0.19	0.17	0.07	-0.09	-0.06	0.04	0.11	0.24	0.14	0.15	0.01	0.08	0.12	0.09	-0.03		-0.04	0.12	0.09
sp26	0.08	0.29	0.20	0.02	0.09	0.05	0.15	0.22	0.40	0.38	0.11	0.34	0.15	0.15	0.25	0.07	0.05	0.25	-0.04		0.28	0.21
sp28	0.19	0.37	0.38	0.07	0.31	0.30	0.25	0.32	0.31	0.52	0.44	0.26	0.29	0.16	0.39	0.25	0.39	0.44	0.12	0.28		0.40
sp29	0.34	0.50	0.32	0.04	0.36	0.27	0.41	0.36	0.52	0.68	0.56	0.46	0.31	0.18	0.37	0.30	0.47	0.30	0.09	0.21	0.40	

For each pair of participants kappa value was computed using coded triples obtained for these participants (see Appendix 3). Specifically, for each participant, we concatenated all 14 associated triples (one triple per vignette) into a single vector of relevancy evaluations. Triples were concatenated in the same order (corresponding to the sequence of vignettes) and the resulting vector contained 52 entries corresponding to coded relevancy evaluations of individual systematic reviews. Then, these vectors were exported to the R system, where we calculated kappa values.

Appendix 6. Selected clustering of physicians in the context of precision at 3 and group value function



Physician sp5 from cluster 3 and sp25 from cluster 4 are very similar in terms of these two measures. However, a closer look at their coded triples (Appendix 3) reveals differences in evaluations across vignettes. For example, reviews retrieved for vignette 1 were evaluated as *XXX* by sp5 and as *NNX* by sp25, while for vignette 10 the evaluations were *NNX* for sp5 and *XXX* for sp25. While these differences were compensated after averaging values of both measures over all vignettes, they were captured by the kappa coefficient (that indicated the lack of agreement between sp5 and sp25) and resulted in placing these two physicians in two different clusters.