Is There a Consensus when Physicians Evaluate the Relevance of Retrieved Systematic Reviews?*

Dympna O’Sullivan1,2; Szymon Wilk3,2; Craig Kuziemsky2; Wojtek Michalowski2; Ken Farion4,5; Bartosz Kukawka3

1School of Mathematics, Computer Science and Engineering, City University London, London, United Kingdom; 2MET Research Group, Telfer School of Management, University of Ottawa, Ottawa, Ontario, Canada; 3Institute of Computing Science, Poznan University of Technology, Poznan, Poland; 4Division of Emergency Medicine, Children’s Hospital of Eastern Ontario, Ottawa, Ontario, Canada; 5Departments of Pediatrics and Emergency Medicine, University of Ottawa, Ottawa, Ontario, Canada

Keywords
Evidence-based medicine, systematic review, information storage and retrieval, evaluation study, pediatric asthma

Summary

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 three.

Results: Twenty-two 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.

Conclusions: 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.

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 system 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 setting.

* Supplementary material published on our website http://dx.doi.org/10.3414/ME15-01-0131


http://dx.doi.org/10.3414/ME15-01-0131
received: October 9, 2015
accepted: February 7, 2016
epub ahead of print: March 4, 2016

Correspondence to:
Szymon Wilk
Institute of Computing Science
Poznan University of Technology
Piotrovo 2
60-965 Poznan
Poland
E-mail: szymon.wilk@cs.put.poznan.pl

© Schattauer 2016
To what extent do physicians agree on the relevance of retrieved systematic reviews? 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 emergency physicians from 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 the 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

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 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.
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 – 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 one physician withdrew her consent. Twenty-six physicians finished the survey, however, four 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).

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

© Schattauer 2016

Methods Inf Med 3/2016
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 six 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 (▶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 positions), however, they still confirm the divergent evaluations described above.

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

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 three 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 Figure 3b), or while relevant for chronic asthma, it was less appropriate for acute asthma managed in the ED (Figure 3c and Figure 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” – 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).

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 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.

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 the 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 (e.g., therapy development) after being presented with evidence of diversified types (coming from a large patient population, from a 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 physicians’ preferences with regard to the evidence characteristics, thus facilitating the abovementioned customization of clinical decision support systems with evidence retrieval functionality.

Acknowledgment

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

298

D. O’Sullivan et al.: Is There a Consensus when Physicians Evaluate the Relevance of Retrieved Systematic Reviews?


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


