Learning the Preferences of Physicians for the Organization of Result Lists of Medical Evidence Articles

D. O´Sullivan[1,2]; S. Wilk[2,3]; W. Michalowski[2]; R. Slowinski[3,4]; R. Thomas[5]; M. Kadzinski[3]; K. Farion[6]

[1]School of Informatics, City University London, London, United Kingdom;
[2]Telfer School of Management, University of Ottawa, Ottawa, Canada;
[3]Institute of Computing Science, Poznan University of Technology, Poznan, Poland;
[4]Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland;
[5]Sprott School of Business, Carleton University, Ottawa, Canada;
[6]Departments of Pediatrics and Emergency Medicine, University of Ottawa, Ottawa, Canada

Keywords
Organization of medical evidence, physician preferences, rank-ordered lists, evidence-based medicine, information retrieval, decision support systems, clinical

Summary
Background: Online medical knowledge repositories such as MEDLINE and The Cochrane Library are increasingly used by physicians to retrieve articles to aid with clinical decision making. The prevailing approach for organizing retrieved articles is in the form of a rank-ordered list, with the assumption that the higher an article is presented on a list, the more relevant it is.
Objectives: Despite this common list-based organization, it is seldom studied how physicians perceive the association between the relevance of articles and the order in which articles are presented. In this paper we describe a case study that captured physician preferences for 3-element lists of medical articles in order to learn how to organize medical knowledge for decision-making.
Methods: Comprehensive relevance evaluations were developed to represent 3-element lists of hypothetical articles that may be retrieved from an online medical knowledge source such as MEDLINE or The Cochrane Library. Comprehensive relevance evaluations asses not only an article's relevance for a query, but also whether it has been placed on the correct list position. In other words an article may be relevant and correctly placed on a result list (e.g. the most relevant article appears first in the result list), an article may be relevant for a query but placed on an incorrect list position (e.g. the most relevant article appears second in a result list), or an article may be irrelevant for a query yet still appear in the result list. The relevance evaluations were presented to six senior physicians who were asked to express their preferences for an article's relevance and its position on a list by pairwise comparisons representing different combinations of 3-element lists. The elicited preferences were assessed using a novel GRIP (Generalized Regression with Intensities of Preference) method and represented as an additive value function. Value functions were derived for individual physicians as well as the group of physicians.
Results: The results show that physicians assign significant value to the 1st position on a list and they expect that the most relevant article is presented first. Whilst physicians still prefer obtaining a correctly placed article on position 2, they are also quite satisfied with misplaced relevant article. Low consideration of the 3rd position was uniformly confirmed.
Conclusions: Our findings confirm the importance of placing the most relevant article on the 1st position on a list and the importance paid to position on a list significantly diminishes after the 2nd position. The derived value functions may be used by developers of clinical decision support applications to decide how best to organize medical knowledge for decision making and to create personalized evaluation measures that can augment typical measures used to evaluate information retrieval systems.

Correspondence to:
Dympna O'Sullivan
School of Informatics
City University London
Northampton Square
London EC1V 0HB
United Kingdom
E-mail: Dympna.O'Sullivan.1@city.ac.uk

Methods Inf Med 2014; 53: 344–356
doi: 10.3414/ME13-01-0085
received: July 24, 2013
accepted: February 24, 2014
Epub ahead of print: June 6, 2014

1. Introduction

As part of our research on clinical decision support systems we have been developing methods for automatically retrieving biomedical articles from The Cochrane Library relevant in the current context of a patient-physician encounter, and for presenting these articles to physicians at the point-of-care. These articles are indexed and searched using methods from information retrieval and organized as a rank-ordered list [1], where the rank corresponds to the relevance of an article as computed by the information retrieval system. These automatically computed rankings are only estimates of relevance, the "gold standard" of relevance is provided by real users (who

may also differ in their estimates of relevance) and mistakes by the information retrieval system are also possible. This work prompted us to pose the following research question: "What are physicians' preferences with regards to the organization of medical articles for specific positions on a list?" In other words how do physicians value the correct match between an article's rank and relevance on specific positions on a result list and how do they perceive mistakes on different positions on a rank-ordered list. For example, how do physicians rate the importance of having relevant (or irrelevant) articles on particular list positions (e.g., 1st, 2nd and 3rd)? Or, how do they value articles that are relevant but misplaced on list positions (for example the most relevant article is placed in 2nd instead of 1st position)? We need to underline that in this research we are not concerned with issues of how a given article was placed on a particular position on a list (i.e., how its rank was computed) and we do not examine the relative relevance of retrieved articles. Rather we are interested in discovering physicians' preferences associated with particular list positions (specifically, we consider lists with three positions), where on each of these positions there may be a correctly ranked relevant article, an incorrectly ranked relevant article or an irrelevant article, and subsequently in learning how to better organize medical articles from online medical knowledge repositories for decision-making.

Conventional information retrieval applications (including popular search engines), return lists of ranked articles where article features (e.g. terms and term frequencies) are used to estimate relevance for a given query which is usually in the form of a set of keywords specified by a user. In general, a query does not uniquely identify a single article in the collection. Instead, several articles may match the query, usually with different degrees of relevancy and most retrieval systems compute a numeric score on how well each article matches the query, and ranks the articles according to this value. Generally users do not have time to examine all highly ranked results returned by a query, rather they are most interested in the most relevant articles from a collection, however this maximization of the most relevant results is a difficult task due to the inherent ambiguity of natural language which is characterized by homographs, synonyms and polysemes. This is especially evident if a user provides too few query terms to precisely describe their information need. Furthermore users often do not or cannot articulate the context of their information need or are not familiar with advanced search techniques such as Boolean operators that allow them to choose optimal logical combinations of search terms and therefore the search engine will retrieve many more documents that the user considers relevant to their query. In terms of evaluating information retrieval systems, the established method is to evaluate the relevance of retrieved articles by comparison with a gold standard which is usually provided by an expert. The effectiveness of the automatic application is then measured in terms of precision — the number of relevant articles a query retrieves divided by the total number of articles retrieved, and recall – the number of relevant articles retrieved divided by the total number of relevant articles that should have been retrieved for the query [2–5]. Such metrics are commonly applied to evaluate the performance of clinical information retrieval applications [6–10].

However, these metrics (precision and recall), are set-based measures, therefore they do not take into account the position of an article on a rank-ordered list and how users (in this case, physicians), perceive mistakes with regard to relevant but misplaced (in terms of their position on a list) articles. Some attempts have been made to extend precision and recall to take into consideration rank-ordered results. For example, precision values can be interpolated at standard recall levels (e.g. {0.1, 0.2 ... 1.0}), where the interpolated precision at the j-th standard recall level is the maximum known precision at any recall level between the j-th and (j + 1)-th level [2]. Likewise mean average precision is a measure used to average precision over a number of queries and has the effect of promoting relevant results closer to the top of a result list [2]. Both measures factor in precision at all recall levels but for many applications, including web search, this may not be germane to users who tend to be interested in how many relevant articles there are at the top of returned result lists [11].

This leads to measuring precision at fixed top levels of retrieved results, such as 10, 20 or 30 articles. This is referred to as *precision at k*, for example *precision at 10* [2]. It has the advantage of not requiring any estimate of the size of the set of relevant articles but at the same time it is the least stable of the commonly used evaluation measures and it does not average well [2]. Furthermore it fails to take into account the positions of the relevant articles among the top *k* [2].

A measure that attempts to incorporate user preferences into evaluating retrieval is *binary preference*. It considers the number of articles that were judged as non-relevant that were retrieved with higher rank than relevant articles and it is claimed that the measure is more reliable than mean average precision [12, 13].

However none of the discussed measures have the ability to capture preferences with regard to relevant articles that are out of position on a list. Specifically these measures consider only the retrieval of relevant articles and make the assumption that users equally value all relevant articles. In order to address this issue we introduce the notion of *comprehensive relevance evaluations* that represent the relevance of hypothetical articles, and that specify not only whether an article is relevant for a query, but also whether it has been placed in the correct position on a 3-element list given the gold standard. Consider a sample query where the gold standard indicates that the correct triple of articles retrieved should be (a_1, a_2, a_3), while the information retrieval application retrieved a triple of articles (a_1, a_3, a_5). Comprehensive relevance evaluations representing hypothetical articles from the retrieved list (established by comparing the retrieved list to the gold standard) are the following:

- *Relevant and correctly placed* for articles a_1–a_1 is most relevant and placed in the right position,
- *Relevant but misplaced* for article a_3–a_3 is relevant but placed in the wrong position (2nd instead of 3rd),
- *Irrelevant* for article a_5.

In our analysis we focus on the interplay between the comprehensive relevance evaluation and the position on a list. Specifically we pose the following questions for organizing articles returned by information retrieval applications for clinical decision support:

1. How do physicians value specific positions on a result list?
2. How do physicians perceive the relevance of articles as they move from top to lower positions on a result list?

We posit that an understanding of physician's preferences with regard to retrieved medical articles from online knowledge sources is an important issue as list-based presentation is used to guide and constrain a physician's decision making behaviour [14]. Furthermore physician's decision making is often time-limited and is prone to such cues as position on a list [15–17], amplifying the importance of learning about the interplay between article relevance and its placement on a list.

In order to answer the research questions we designed and conducted a case study described later in the paper. In this study we move away from typical rank-ordered lists of retrieved articles and replace them with lists of comprehensive relevance evaluations (as in the example above) in order to facilitate their analysis. Discussions with physicians during the implementation of a clinical retrieval system revealed their concerns with time pressures experienced during patient encounters and there was a strong consensus for result lists with maximum length of three articles. The physicians' preferences are consistent with results published by Spink et al. on information retrieval on the Web [18] who found that most users do not go beyond the third item in a result list. Thus we focused on lists of three articles in our analysis. In the text hereafter we refer to such lists as *triples of comprehensive relevance evaluations* or *triples* for brevity. We note that although we focus on triples representing 3-element lists of hypothetical articles, longer lists may be easily generated using different combinations of possible comprehensive relevance evaluations – relevant and correctly placed, relevant but misplaced and irrelevant.

The rest of the paper is organized as follows. In the next section we discuss related research on the information seeking behaviour of physicians and list-based organization of search results. In Section 3 we present a novel GRIP (Generalized Regression with Intensities of Preference) method that was used to assess physician's preferences by deriving value functions representing preferences with regard to article positions on a list and describe the design of our case study for eliciting physician preferences, both as individuals and as a group. The results of the analysis and a discussion of these results are presented in Section 4 and the paper concludes with insights and a discussion in Section 5.

2. Related Research

Our research falls into a broad space of work studying how physicians perceive and process organized information (be it in the form of evidence or other information at the point-of-care). The information needs of physicians have been studied for many years since the seminal work of Covell [19]. Studies have focused on the type of information need, for example the extensive "evidence cart" survey in the UK found that 81% of medical evidence sought related to diagnostic and/or treatment decisions [20]; the frequency of the information need, for example [21] reported that doctors asked 5.5 questions per half day; the information sources used by physicians, for example studies by Cullen and Schilling reported that the most commonly used electronic resources were MEDLINE, clinical guideline websites, Internet search engines, and The Cochrane Library [22, 23]; and aspects of search that are necessary to cope with an increasing amount of medical information[24, 25]. These studies concluded that physicians have significant information needs in practice, that electronic resource are an under-utilized information source, and that information literacy training is required so that physicians can effectively perform literature searching [26].

However scant research has focused on the most effective ways to organize retrieved information and in particular the impact that list-based presentation has on physicians valuation of the information (articles) in relation to how much value they place on the position on a list, or the ordering of relevant articles in such a list. Lists are one of the most common ways of managing the presentation of information and have long been used as a means of externalizing cognition and organizing items [27–30]. Several main effects of list-based organization and presentation have been observed in the cognitive psychology literature including the *primacy effect* – items in a list that are first tend to be more memorable, the *recency effect* – items in a list that are last tend to be more memorable than items in the middle, and the *von Restorff effect* – distinctive items are more memorable. List-based presentation also has its critics. For example, it has been claimed that the low precision of search engines coupled with an ordered list presentation style make it hard for users to find the information they are looking for [31]. In spite of such criticisms and subsequent attempts to introduce other mechanisms such as clustering for organizing search results [32], list-based presentation continues to be the dominant method for organizing information presentation.

Researchers have examined user behaviour and interaction with ranked lists to analyse how best to present search results. User actions such as opening articles or the order in which users view entries in result lists have been extensively evaluated. For example, researchers have studied whether users prefer to follow a depth-first strategy, where the user examines each entry in the list in turn starting from the top, and decides immediately whether to open the article in question; or a breadth-first strategy where the user looks ahead at a number of list entries and then revisits the most promising ones [33]. The results were conclusive that a significant majority (85%) of users rely on a depth-first strategy. Other research has demonstrated that subjects change strategies from depth-first to breadth-first patterns of processing as time pressures increase [34].

Eye tracking studies (measuring spatially stable gaze during which visual attention was directed to a specific area of the display), have been employed to estimate

how users process list-based information [35, 36]. Most gaze activity was directed at the first few items with items ranked lower receiving last and least attention. Specifically the results indicated that users tended to view the first and second-ranked entries right away, and then there is a large gap before viewing the third-ranked entry. Another study by Keane et al. [37], confirmed the inclination of users to access items at the beginning of lists. Their study presented search result lists in randomized, counterbalanced, normal and reversed ordering during an experiment involving 30 users. The most relevant item was selected on the first click 70% of the time in the normal (relevance-based) ordering, but only 10% of the time in the reversed ordering. In contrast, the 10th relevance-ranked item was chosen 41% of the time in the reversed condition, and only 2% of the time in the relevance-based order. These findings reflect other analysis of user patterns in web search conducted by [18, 38–40]. For example in a study of 1 million queries [18], 47.6% of users looked at two or less results, and in studies of 154,000 [40] and 51,00 queries [39], the percentage of users looking at one, two or three results were 85.2%, 7.5%, and 3.0%, and 58%, 19% and 9% respectively. The authors conclude that the users have a low tolerance of going in detail through longer lists and that the need for high precision in Web information retrieval algorithms is vital. Considering the potential impact of this inherent user behaviour on search, a school of research is actively devising solutions to overcome the effect of falsely over-promoting web pages by placing them at the top of result lists where they will be selected preferentially by users [41, 42].

Other related studies have examined how memorable search results within a list are. For example in a study involving 245 participants, Teevan found that on average participants followed 1.9 search results and that only about 15% of all results displayed were memorable [43]. The factors affecting how likely a result was to be remembered included among others, where in the result list it was ranked. In particular those results presented first are more memorable compared to later results. Teevan also analysed how result ordering was remembered and found that memory mistakes were less common for highly ranked results and that the first result's rank was remembered correctly 90% of the time [43].

All these findings have strong implications for the organization of medical articles for physicians. Considering the usual cognitive limitations of humans as well as the time pressures experienced by most physicians, it seems fair to state that physicians will most likely consult only a small number of articles from a list. Furthermore, if articles presented close to the top of a list have little relevance or are irrelevant, it is likely the entire list will be discarded. While the above statement is confirmed by the research cited in this section, there is no evidence of how strong physician preferences (either individually or as a group), are with regards to positions on rank-ordered lists and what articles are presented on specific list positions. Specifically, little is known about how much value they place on retrieving relevant articles in the correct order on a list versus how they assess being presented with relevant articles but not necessarily in the right order. The study presented in this paper gives an insight into physician's preferences for specific positions on a list and describes the construction of value functions that can be used to personalize the presentation of relevant articles based on their preferences. Such an approach moves beyond previous work on information presentation in the biomedical domain where the focus has not been on user preferences, rather the order of presentation has been estimated using features of the articles (e.g. specific pre-labelled text segments [44]), as well as metadata about articles (e.g. citation counts and journal impact factors [45]).

3. Methods

3.1 The GRIP Method

In order to answer the research questions formulated in Section 1, we need to elicit preferences of physicians (as individuals and as a group) with regard to comprehensive relevance evaluations at specific positions on a triple, and use this information to construct a preference model in the form of a value function. Such a value function can be then used to assess the overall value of a triple.

In this research we aim to estimate an additive value function, which is constructed as the sum of marginal value functions associated with specific criteria characterizing decision alternatives (or alternatives in short). The additive value function is formally defined as:

$$U(a) = \sum_{i=1}^{n} u_i(g_i(a)),$$

where a is an alternative described by family of n criteria g_1, … g_n, and u_i is a non-decreasing marginal value function associated with criterion *gi*.

In our research, alternatives are triples of comprehensive relevance evaluations characterized by three ordinal criteria – g_1, g_2 and g_3 – corresponding to the 1st, 2nd and 3rd position in a triple respectively. The evaluation scale of each criterion is composed of possible comprehensive relevance evaluations ordered from the least to the most preferred one (irrelevant, relevant but misplaced, relevant).

The additive value function not only provides the overall value of a triple, but through marginal value functions gives a thorough insight into preferences associated with specific positions in a triple. The latter is crucial for answering our research questions. It also differentiates our study from previous work, where value functions have been applied in information retrieval to learn in general about user preferences for article relevance [46, 47].

There are many theoretical approaches to estimating an additive value function, including those that rely on the *ordinal regression paradigm*. In these approaches, preferential information is captured first through pairwise comparisons of a subset of alternatives (so-called *reference alternatives*), and then a value function consistent with this information is built [48, 49]. Such a value function, called *compatible*, represents preferences of a specific decision maker (DM), which means that it compares reference alternatives in the same way as a DM would, and it can be applied to assess other alternatives, not included in the reference set.

In our research we used the GRIP method [50, 51]. Its outline is given in

►Figure 1. GRIP accepts as input preferential information given by a DM in the form of pairwise comparisons of reference alternatives and ordinal intensities of preference among them. This preferential information does not need to be complete, i.e., a DM can provide pairwise comparisons of selected reference alternatives, and intensities of preferences on pairs of some other reference alternatives. Indeed, in the case study described in this paper, preferential information elicited from physicians was limited to pairwise comparisons of reference triples only.

Computations conducted by GRIP are divided to include four steps marked with rectangles in ►Figure 1 that involve solving different mathematical programming models (for detailed descriptions of the steps and models applied in our case study see the online appendix). In step 1 GRIP constructs all possible value functions that are compatible with provided preferential information by solving a model with linear constraints representing this information. If the model has no solution (i.e., no compatible value function exists), it indicates that the preferences of a DM cannot be represented by an additive value function, preferential information is erroneous, or preferential information is contradictory (preferences are unstable or some latent criteria have not been considered).

In such a case GRIP is able to help a DM identify reasons for the incompatibility by proceeding to step 2. In this step GRIP solves a mixed 0–1 linear programming model that identifies a minimal set of constraints that need to be removed from the model constructed in step 1 so at least one compatible value function can be found. The identified constraints point at problematic preferential statements (pairwise comparisons or preference intensities) that need to be assessed by a DM. Once the preferential information has been revised, GRIP returns to step 1.

If step 1 terminates successfully (i.e., at least one compatible value function has been found), GRIP proceeds with step 3 where it constructs necessary and possible preference relations. A necessary preference relation is supported by all compatible value functions obtained in step 1, and a possible preference relation is supported by at least one compatible value function constructed in step 1. Both these relations are established by solving another linear programming model. While possible and necessary relations can be applied to build possible and necessary rankings of alternatives that are presented to a DM as a supplemental result, they are used as additional input to calculations in the last step.

In step 4 GRIP constructs a representative value function on the basis of all compatible value functions from step 1 and preference relations from step 3 [52, 53]. First, it maximizes the minimal difference between values of alternatives, for which the necessary preference holds (this is done by solving the mathematical programming model with constraints from step 1 expanded with additional constraints representing the necessary preference). If there is still more than one such value function, GRIP minimizes the maximal difference between values of alternatives, for which the possible preference holds (the linear programming model used previously is further expanded with constraints representing the possible preference and solved). The resulting value function is presented to a DM as the representative value function. This function is based on the necessary and possible preference relations and, in a sense it gives the most accurate representation of these relations.

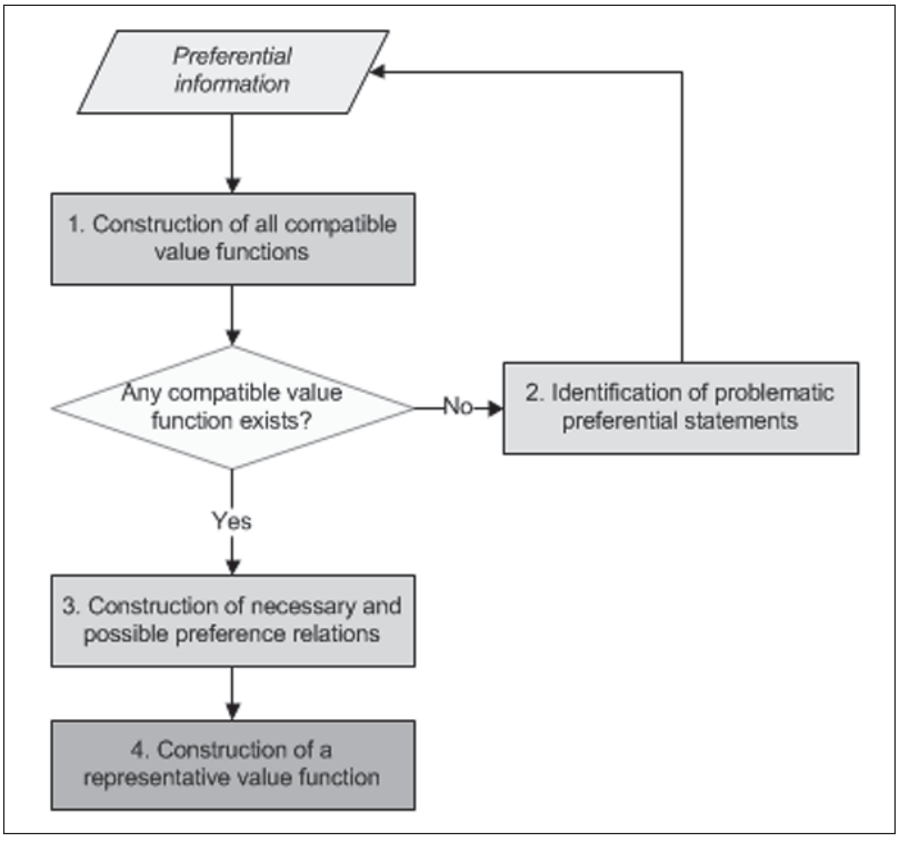

Figure 1 General outline of the GRIP method

3.2 Case Study

We designed a case study to elicit and represent physician preferences with regard to the organization of rank-ordered lists of medical articles using the GRIP method. The case study involved the participation of six physicians, all from teaching hospitals and all with at least 10 years of clinical experience. They represented a range of clinical specialties – emergency medicine, community medicine, internal medicine, intensive care and anaesthesiology. All had prior experience of retrieving medical articles for decision support both off-line and at the point-of-care.

The experimental design of the case study is shown in the form of pseudo-code

```
/*Phase 1                                                          */
all ←set of all feasible triples of rank-ordered articles
ref ← set of pairs of reference triples from all
/*Phase 2                                                          */
for p ∈ set of participants do
   cmp_p ← ask p to compare pairs of triples from ref
   if no value function compatible with cmp_p exists then
      inc_p ← identify inconsistent pairwise comparisons in cmp_p
      cmp_p ← cmp_p \ inc_p
   end
   construct an individual representative value function for p using cmp_p
/* Phase 3                                                          */
cmp_group ← {}
for p ∈ set of participants do cmp_group ← cmp_group ∪ cmp_p
if cmp_group contains contradictory comparisons then
      con_group ← select conflicting responses in cmp_group
      cmp_group ← cmp_group \ con_group
end
construct a group representative value function using cmp_group
```

Figure 2
Pseudo-code representation of experimental design

in ▶Figure 2 and involved three phases, described in detail in the next subsections:

1. Selecting reference triples and combining them into pairs (see Section 3.2.1),
2. Obtaining pairwise comparisons from the physicians and constructing individual representative value functions (see Section 3.2.2),
3. Constructing a group representative value function by amalgamating preferences of all participating physicians (see Section 3.2.3).

3.3.1 Phase 1

This phase started with devising a set of coded triples that represented all feasible combinations of comprehensive relevance evaluations for 3-element lists of retrieved articles. These combinations were established without the need for an auxiliary article retrieval experiment, which significantly streamlined this phase of the study. A comprehensive relevance evaluation in each position in a triple (1st, 2nd, 3rd) was coded as X, N or Y, where X indicates an irrelevant article at a position, N indicates a relevant but misplaced article, and Y indicates a relevant and correctly placed article. Thus, the triple [relevant and correctly placed article, relevant but misplaced article, irrelevant article] mentioned in Section 1 is coded as YNX according to this coding scheme.

The applied coding scheme resulted in 24 feasible triples. This is less than all possible combinations (27), because some triples are not feasible (given the assumption that there are at most three relevant articles), and as a result were not considered in the experiment. For example, a triple YYN is not feasible because it has two articles that are in the correct position and are relevant, thus the third article cannot be misplaced (N) but can be either irrelevant (X) or relevant and correctly placed (Y).

From the set of 24 feasible triples, we selected 12 reference triples and used them to construct 10 pairs of triples for pairwise comparison by physicians. These pairs cor-

Table 1
Pairwise comparisons of triples

PCID	t_1	t_2	P1	P2–3	P4	P5–6
1	NNN	YYX	<	>	>	<
2	NNX	YXY	<	<	<	<
3	NXN	XYY	<	>	~	~
4	NXX	XYX	<	<	<	<
5	XNX	XXY	<	>	~	>
6	XNN	YXX	<	>	<	<
7	NNN	YXY	>	>	<	<
8	NNX	XYY	<	>	~	>
9	XNN	XYX	>	>	<	>
10	NXX	XXY	>	>	<	>

PCID = pairwise comparison ID; t_1 = triple 1; t_2 = triple 2; Y = relevant, correctly placed; N = relevant, misplaced; X = irrelevant; P1–P6 = physicians 1–6; > = t_1 preferred over t_2; < = t_2 preferred over t_1; ~ = t_1 equally preferred to t_2

Table 2 Inconsistent pairwise comparisons excluded from analysis

Physician	Inconsistent comparisons
P2_3	NXN > XYY
P4	NNN > YYX, NXN ~ XYY, NNX ~ XYY
P5_6	NXN ~ XYY

Y = relevant, correctly placed; N = relevant, misplaced; X = irrelevant; > = triple t_1 preferred over triple t_2; < = t_2 preferred over t_1; ~ = t_1 equally preferred to t_2

responded to less obvious evaluations. For example, YYX is intuitively preferred over XYX (retrieving two relevant articles placed correctly on the first two positions is preferred over retrieving one relevant article and placing it correctly on the second position, with two remaining articles being irrelevant); while comparing NNN and YYX is more difficult and subjective (is it preferred that all retrieved articles are relevant but misplaced as opposed to having one irrelevant article and two other relevant and correctly placed articles?). Given the non-trivial nature of the defined pairs, their number was intentionally limited to ensure the lists would be thoroughly assessed by physicians.

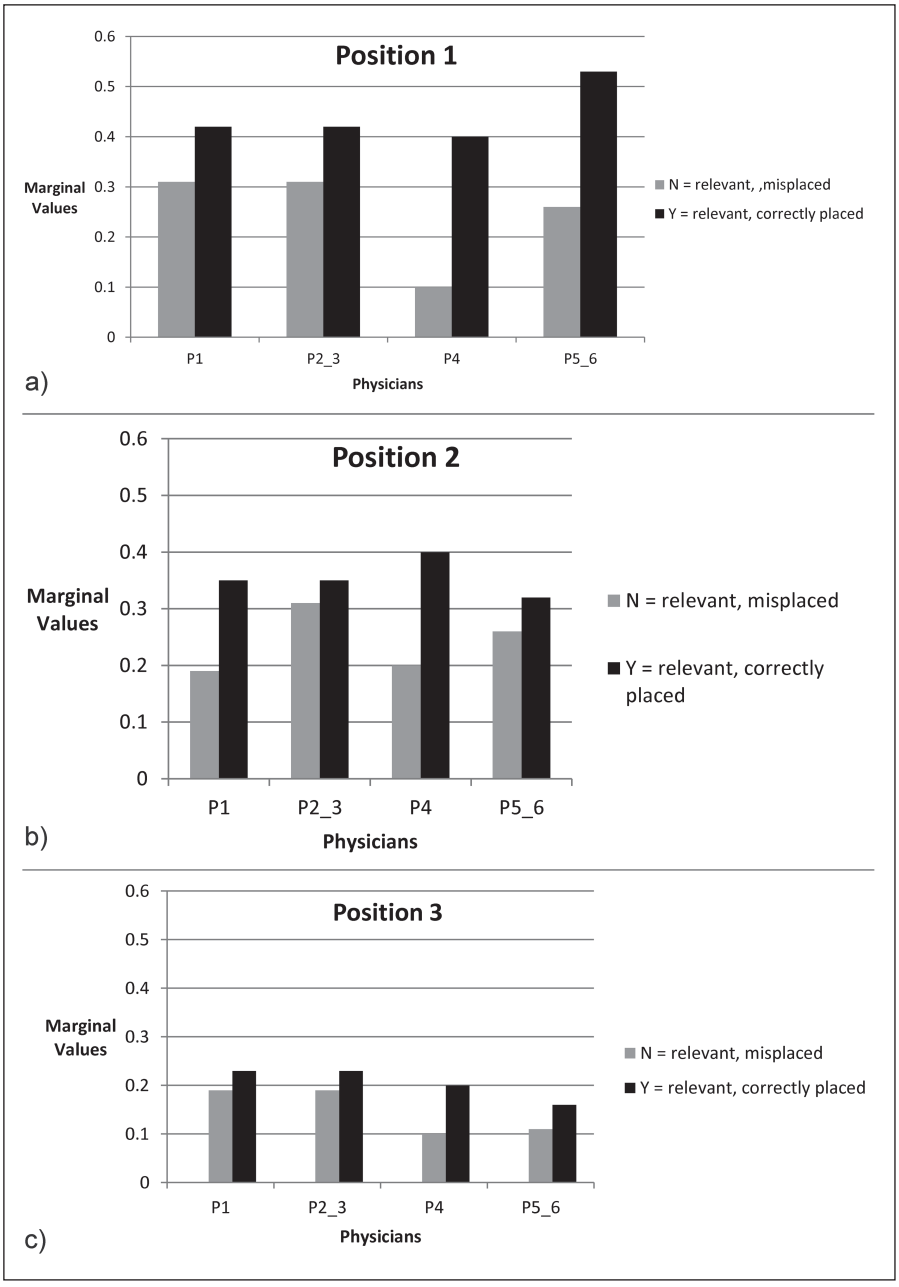

Figure 3 Marginal value functions for individual physicians

3.2.3 Phase 2

At the outset physicians were informed about the purpose of the study and the research questions we were trying to answer. We provided instructions on how they should conduct pairwise evaluations and used triples that were not evaluated in the case study as training examples. After this brief instruction session, physicians were asked to independently assess each pair of reference triples established in the previous phase and to state if one triple was preferred over the other or if they were equally preferred.

When pairwise assessments were completed, GRIP was used to check if individual pairwise comparisons provided by each of the physicians were compatible with any value function. In the case of an incompatibility, inconsistent comparisons by a given physician were identified and removed from the analysis. Finally, GRIP was applied to construct a representative value function for each physician. Specifically, we focused on marginal value functions in order to gain a detailed insight into the preferences of specific physicians for comprehensive relevance evaluations for the particular positions in a triple (1st, 2nd and 3rd). These marginal functions are discussed in detail in Section 4.2.

3.2.3 Phase 3

This phase began by combining individual pairwise comparisons (with inconsistencies removed in Phase 2). The resulting set was checked for conflicting responses between the respondents (contradictory assessments for the same pair of triples provided by different physicians), as in order to construct a group representative value function, conflicts must be resolved. Since there may have been many possible subsets of comparisons to remove, we used a heuristic that removed conflicting comparisons given by the smallest number of physicians. In other words, we kept those comparisons that were supported by the majority of physicians. Having estab-

lished non-contradictory responses, GRIP was applied to construct a group representative value function. We present these results in Section 4.3.

4. Results and Discussion

We need to note that the ability to use statistical techniques in the analysis of the results is limited because of the case study format that focuses on a single group of physicians. Therefore, the analysis presented here is descriptive with the aim of summarizing the data and providing a succinct description of the patterns and relationships that were revealed.

4.1 Pairwise Comparisons of Reference Triples

▶Table 1 presents the results of the pairwise comparisons of reference triples by the physicians. Each of the physicians (denoted as P1, P2 ... P6) was asked to express her/his preferences for one triple (t_1) over another (t_2) as: t_1 preferred over t_2 (symbol ">"), t_2 preferred over t_1 (symbol "<"), and t_1 equally preferred to t_2 (symbol "~"), and where PCID denotes the pairwise comparison ID. Each physician responded individually, independently assessing the pairs of triples and they were not able to see the responses of other participants. It happened that physicians P2 and P3 responded identically and are grouped together as (P2_3), and similarly for physicians P5 and P6 (P5_6). Their responses were amalgamated for the sake of concise representation after the surveys had been completed.

4.2 Individual Representative Value Functions

The responses presented in ▶Table 1 formed an input for the GRIP method. First, GRIP identified inconsistent comparisons to be removed from the analysis. All physicians except P1 made some inconsistent pairwise comparisons and they are listed in ▶Table 2. The largest number of inconsistent comparisons was made by P4. Moreover, the same inconsistent comparison (NXN ~ XYY) was reported by P4 and P5_P6. After removing these responses, we proceeded with constructing representative value functions.

Table 3 Individual marginal values for N and Y on positions 1, 2 and 3

Physician	Position 1		Position 2		Position 3	
	N	Y	N	Y	N	Y
P1	0.31	0.42	0.19	0.35	0.19	0.23
P2_3	0.31	0.42	0.31	0.35	0.19	0.23
P4	0.10	0.40	0.20	0.40	0.10	0.20
P5_6	0.26	0.53	0.26	0.32	0.11	0.16

Y = relevant, correctly placed; N = relevant, misplaced

Marginal value functions derived by GRIP and representing individual physician preferences regarding comprehensive relevance evaluations of articles on a rank-ordered list are shown in ▶Figure 3. ▶Figure 3a shows preferences of physicians regarding the 1st position; ▶Figure 3b shows their preferences regarding the 2nd position; and ▶Figure 3c shows their preferences for the 3rd position. Furthermore, ▶Table 3 gives marginal values for codes Y and N for each physician on the 1st, 2nd, and 3rd positions in a triple. The value of X on any position is equal to 0; therefore it has been excluded from the figures and table.

A quick overview of these results indicates that physicians place the highest value on position 1 on a list, and the least on position 3. This is demonstrated by the marginal values associated with Y on specific positions – highest for position 1, and lowest for position 3 (only for P4 are utilities for Y on position 1 and 2 equal). Moreover, values obtained for P1 and P2_3 on position 1 and also on position 3 are equal, which implies some similarity between these physicians.

- Analysis of results in ▶Figure 3 and ▶Table 3 reveals there are two prevailing patterns of preferences:
- *Less demanding*, where physicians mostly value having a relevant article, and the correct position of the article is of secondary importance (i.e., they are willing to accept a relevant but misplaced article). This pattern is exemplified by much smaller differences of marginal values between N and Y, than between X and N;
- *More demanding*, where physicians value not only the relevance of an article, but also having an article presented in the correct position (i.e., they

Table 4 Pairwise comparisons selected to construct a group representative value function

PCID	t_1	t_2	P1	P2–3	P4	P5–6	Selected
1	NNN	YYX	<	>		<	<(3)
2	NNX	YXY	<	<	<	<	<(6)
3	NXN	XYY	<				<(1)
4	NXX	XYX	<	<	<	<	<(6)
5	XNX	XXY	<	>		>	>(4)
6	XNN	YXX	<	>	<	<	<(4)
7	NNN	YXY	>	>	<	<	<(3)
8	NNX	XYY	<	>		>	>(4)
9ß	XNN	XYX	>	>	<	>	>(5)
10	NXX	XXY	>	>	<	>	>(4)

PCID = pairwise comparison ID; t_1 = triple 1; t_2 = triple 2; Y = relevant, correctly placed; N = relevant, misplaced; X = irrelevant; P1–P6 = physicians 1–6; > = t_1 preferred over t_2; < = t_2 preferred over t_1; "Selected" = comparisons selected for further analysis

accept only relevant and correctly positioned articles). This pattern is characterized by linearly increasing marginal values for X, N and Y.

The *more demanding* pattern of preferences is demonstrated across all positions by P4, while the *less demanding* pattern is consistently presented by P2_3. Preferences of other respondents vary depending on the position. P1 has less demanding preferences for positions 1 and 3, and a more demanding preference for position 2, while P5_6 has a more demanding pattern of preference for position 1 and less demanding for positions 2 and 3.

The overall conclusion from this analysis is that in principle, positioning and relevance is important for physicians when viewing a list of medical articles. In particular a high value is placed on the 1st position on a list. This is coupled with a general reduction in the value of retrieved articles if they are placed on lower positions in a rank-ordered list. Specific preferences vary across physicians and positions – some physicians are more willing to accept relevant but misplaced articles (*less demanding*), while others equally value relevance and correct position of retrieved articles (*more demanding*). These differences in preferences are addressed in the next stage of analysis, where a group representative value function has been developed.

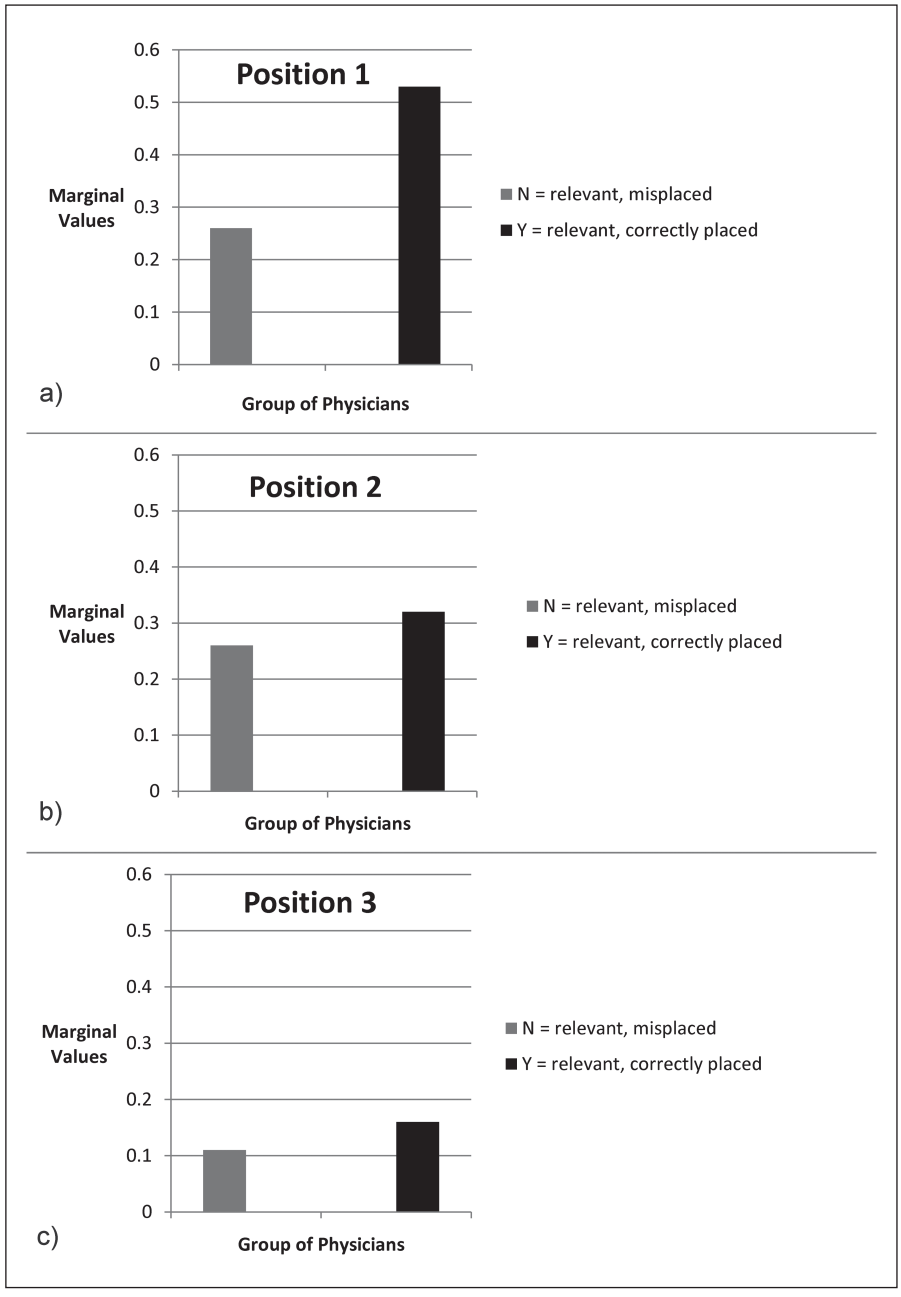

Figure 4 Group marginal value functions

4.3 Group Representative Value Function

▶Table 4 shows the comparisons used to derive individual value functions. Empty cells indicate comparisons that were removed to ensure that representative value functions for individual physicians existed. This table reveals that non-conflicting (consistent) comparisons were given only for 3 out of 10 pairs (PCIDs of 2, 3 and 4). Contradictions for the remaining pairs were addressed before proceeding with the development of a group representative value function. The "Selected" column in ▶Table 4 shows comparisons selected for further analysis – they are augmented with information about the number of physicians (out of 6) who provided them. Moreover, cells in the table corresponding to selected comparisons are marked in grey to better show their prevalence. We note that for pair 7 (NNN and YXY) there were two responses supported by the same number of physicians – NNN > YXY and NNN < YXY, however, the former resulted in no representative value function, so it was discarded.

Marginal value functions derived by GRIP for the group of physicians are presented in ▶Figure 4 and marginal values for codes N and Y are given in ▶Table 5 (as previously, values for X are not shown as they are 0). Interestingly, marginal value functions for the group of physicians are the same as individual marginal value functions for P5_6 (see Section 4.2 for details). This can be easily explained by looking at ▶Table 4. The set of selected consistent comparisons ("Selected") includes all comparisons given by P5_6 – for 9 out of 10 pairs there was at least one other physi-

Table 5 Group marginal values for N and Y on positions 1, 2 and 3

	Position 1		Position 2		Position 3	
	N	Y	N	Y	N	Y
Group	0.26	0.53	0.26	0.32	0.11	0.16

N = relevant, misplaced; Y = relevant, correctly placed

cian who supported the same evaluation, thus these comparisons were most prevalent and ultimately selected.

The pattern of preferences for the group of physicians combines both *more* and *less demanding* patterns as discussed in the previous section. The group is more demanding with regards to the 1st position and require that the most relevant article be presented there, and less demanding with regard to the 2nd and 3rd positions (relevance is more important than the correct positioning). This is intuitively appealing as it highlights the distinctive role of the 1st position on a list, and it is consistent with results of other research e.g. [36]. Moreover, there is a clear difference of importance across positions (as demonstrated by the marginal values for Y) – the 2nd position is twice as important as the 3rd one (0.32 vs. 0.16), and the 1st position is three times as important as the 3rd one (0.53 vs. 0.16).

Finally, ▶Figure 5 presents the ranking of all feasible triples constructed according to the group representative value function (all triples at the same level have the same value of this function). It also indicates reference triples (marked in grey) and shows how they were compared according to the selected consistent responses of the physicians (an arc from t_1 to t_2 means that t_1 was preferred over t_2, the number next to an arc indicates the PCID of a specific pair in Table 4). In other words, the figure shows how preferential information (pairwise comparisons of reference triples) has been embedded into the ranking of all feasible triples by the group representative value function.

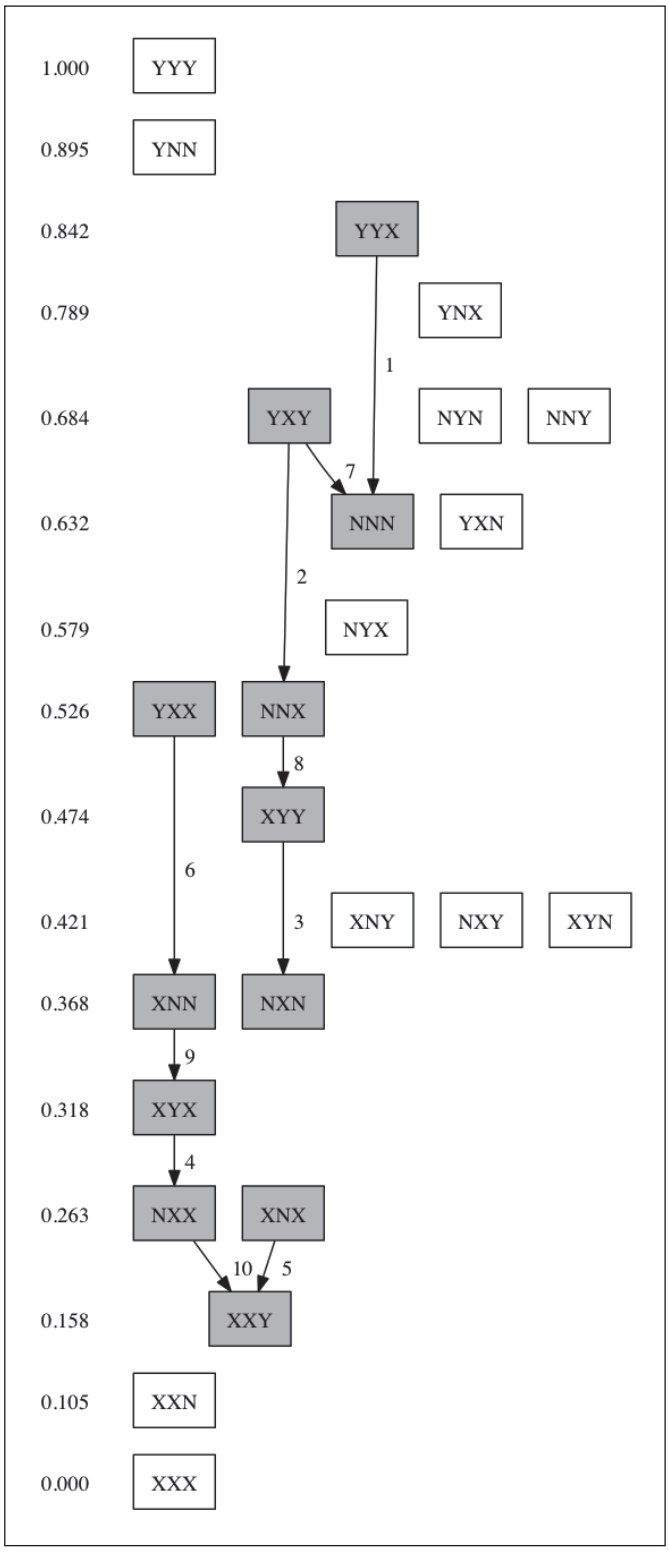

Figure 5 Ranking of all triples by the group representative value function

4.4 Application of the Group Value Function – Sample Scenario

As mentioned in Section 1, we are developing a system for automatically retrieving biomedical articles (i.e., systematic reviews) from The Cochrane Library for presentation to physicians at the point-of-care. Retrieved articles are presented as a rank-ordered list, where the rank corresponds to the relevance of an article for the currently examined patient as computed by the system. In order to evaluate the system we used actual patient cases to derive vignettes combining

Table 6 Evaluation of articles retrieved for vignette 1

Retrieved article ID	Position calculated by retrieval system	Evaluation provided by physician
Article 21	1	Y
Article 3	2	N (should be placed in position 3)
Article 4	3	X

Table 7 Evaluation of articles retrieved for vignette 2

Retrieved article ID	Position calculated by retrieval system	Evaluation provided by physician
Article 60	1	X
Article 48	2	Y
Article 21	3	N (should be placed in position 1)

patient data, confirmed diagnosis and proposed treatment(s). Our system translated these vignettes to queries and then retrieved and ranked relevant articles. Finally, the relevance of retrieved articles was evaluated by physicians of diversified professional experience. The physicians were presented the top three retrieved articles and asked to rate each one using the same codes as in the study described in this paper (Y, N or X). In the case that a participant rated an article N, we asked them to provide what they believed was the correct list position for that article.

We use selected results from the information retrieval study to illustrate application of the representative group value function described in Section 4.3 and to compare it to a typical retrieval metric (precision at 3). Specifically, we focus on articles retrieved for two vignettes and evaluated by a single physician. Results for vignette 1 are given in ▶Table 6 (for brevity we omit the vignette itself and exact titles of the articles). Precision at 3 computed given these evaluations is 2/3 = 0.67, while the group value function is equal to 0.53 + 0.26 + 0.0 = 0.79.

▶Table 7 summarizes results for vignette 2. Here precision at 3 is the same as previously (2/3 or 0.67), while the group value function equals 0.0 + 0.32 + 0.11 = 0.43, thus it is lower than that for vignette 1. This clearly indicates that the retrieval system performed better in the case of vignette 1. However this difference is not captured by the traditional precision at 3, and highlights the need for new measures that take into account preferences of physicians (in this case captured by the group value function).

5. Conclusions

This paper presented the results of a case study that aimed to learn about physician's preferences with regards to the relevance and positioning of articles on rank-ordered lists. To facilitate the analysis we used comprehensive relevance evaluations to characterize articles, thus we considered triples of comprehensive relevance evaluations instead of triples of retrieved articles. Six physicians were asked to provide pairwise comparisons of pre-selected reference triples, and we obtained four unique sets of responses that were analysed using the GRIP method. Representative value functions were developed to model the preferences of individual physicians and the group value function was derived using the four sets of responses.

The study allowed us to answer two research questions regarding the organization of articles in terms of their relevance and position on a list. In terms of specific list positions, the analysis for individual physicians concluded that physicians assign significant value to the 1st position on a list and they expect that the most relevant article is presented first. Regarding other positions on a list, for some *more demanding* physicians, having a correctly positioned relevant article in position 2 on a list is very important, while for *less demanding* physicians, the difference in preferences between relevant articles in the correct position and relevant but misplaced articles is smaller – whilst they still prefer obtaining relevant correctly placed articles in position 2, they are also quite satisfied with misplaced relevant articles.

The analysis for the group of physicians further confirmed these findings about the 1st position on a list and it enforced the pattern of *more* and *less demanding* patterns of preferences. Overall, the group is *more demanding* with regard to the 1st position and finding a relevant article there and *less demanding* about positioning as they move down a list to the 2nd and 3rd positions – for these positions the relevance of a presented article is valued more than correct position. This is combined with an overall pattern of rapidly decreasing importance of articles as physicians move down a result list indicating that physicians perceive articles to be less relevant as they move from higher to lower positions on a list.

Our results correlate with research on general user searches on the Web [34, 36, 37–40], indicating the importance of the 1st position on a result list and in finding a relevant article in this position, and the rapid drop off in importance after that position. However, the results provide further insight into physician search and decision making behaviour by highlighting that it is not sufficient to evaluate only the retrieval of correct articles as is usually done in the information retrieval research community. Rather our analysis shows that physicians clearly value the position on a list where an article is presented as well as relevance.

The findings of our study are beneficial for research into the development of clinical retrieval applications by providing guidance on how articles should be organized and presented for decision making. Rank-ordered lists should be short; both individually and as a group, physicians are most interested in articles positioned at the top of a result list. In particular, the precision of retrieval is essential – it is very important that the article presented in the 1st position is the most relevant article in a given corpus. While some physicians prefer the second most relevant article from the corpus in position 2, many will settle on just retrieving a relevant article (not necessarily the second most relevant article in a corpus). As a group, perceived importance of rank diminishes rapidly the further one moves down a result list. The findings may be more widely applied beyond physicians to any user group that needs to quickly discriminate between articles presented on a result list or to scenarios where

only a limited amount of information can be presented (e.g. on a mobile device).

In future work we intend to use the results of this study in developing a method for more accurately evaluating medical evidence retrieval. As outlined in the sample scenario in subsection 4.4, our findings can be used to supplement traditional evaluation metrics to better take into account the position of an article on a list. In such a way the group representative value function in ►Figure 5 could be used to "calibrate" precision for retrieved articles, i.e., the precision for an article triple represented as YYX should be higher than that for a triple XYY when evaluating the effectiveness of an information retrieval algorithm.

Acknowledgment

The authors would like to thank the physicians who participated in the study. The support of the Natural Sciences and Engineering Research Council of Canada, the Canadian Institutes of Health Research, and the Polish National Science Centre (grant no. NN519 441939) is gratefully acknowledged.

References

1. 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: 623 – 631.
2. Manning C, Raghavan P, Schutze H. Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press; 2008.
3. Dong X, Su L. Search engines on the World Wide Web and information retrieval from the Internet: A review and evaluation. Online Inform Rev 1997; 2: 67 – 82.
4. Harter S, Hert C. Evaluation of information retrieval systems: Approaches, issues, and methods. Online Inform Rev 1997; 32: 3 – 94.
5. Kobayashi M, Takeda K. Information retrieval on the Web. ACM Comput Surv 2000; 32: 144 – 173.
6. Denecke K. An architecture for diversity-aware search for medical web content. Methods Inf Med 2012; 51 (6): 549–556.
7. Westbrook J, Coiera E, Gosling A. Do online information retrieval systems help experienced clinicians answer clinical questions? JAMIA 2005; 12: 315–321.
8. Moskovitch R, Martins SB, Behiri E, Weiss A, Shahar Y. Application of information technology: A comparative evaluation of full-text, concept-based, and context-sensitive search. JAMIA 2007; 14: 164 – 174.
9. D'Avolio L, Nguyen T, Farwell W, Chen Y, Fitzmeyer F, Harris O, Fiore L. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). JAMIA 2010; 17: 375 – 382.
10. García-Remesal M, Maojo V, Billhardt H, Crespo J. Integration of Relational and Textual Biomedical Sources A Pilot Experiment Using a Semi-automated Method for Logical Schema Acquisition. Methods Inf Med 2010; 49 (4): 337 – 348.
11. Büttcher S, Clarke C, Cormack GV. Information Retrieval: Implementing and Evaluating Search Engines. The MIT Press; 2010
12. Allan J. Hard track overview in TREC 2004 (notebook), high accuracy retrieval from documents. Proceedings of the 13th Thirteenth Text Retrieval Conference (TREC 2004) Notebook). pp 226 – 235.
13. Buckley C, Voorhees E. Retrieval evaluation with incomplete information. Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2004). pp 25–32.
14. Zhang J. The nature of external representations in problem solving. Cognitive Sci 1997; 21: 179 – 217.
15. Keinan G, Friedland N, Ben-Porath Y. Decision making under stress: Scanning of alternatives under physical threat, Acta Psychol 1987; 64: 219 – 228.
16. Levin I, Huneke M, Jasper J. Information processing at successive stages of decision making: Need for cognition and inclusion-exclusion effects. Organ Behav Hum Dec 2000; 82: 171 – 193.
17. Starcke K, Brand M. Decision making under stress: A selective review, Neurosci Biobehav R 2009; 9: 1228 – 1248.
18. Spink A, Wolfram D, Jansen B, Saracevic T. Searching the web: The public and their queries. JASIST 2001; 53: 226 – 234.
19. Covell DG, Uman, P. Manning, Information needs in office practice. Are they being met? Ann Intern Med 1985; 108: 596 – 599.
20. Sackett D, Strauss S. Finding and applying evidence during clinical rounds: the "evidence cart". JAMIA 1998; 280: 1336 – 1338.
21. Ely JW, Osheroff J, Chambliss M, Ebell M, Rosenbaum M. Answering physicians' clinical questions: obstacles and potential solutions. JAMIA 2005; 12: 217 – 224.
22. Cullen R. In search of evidence: Family practitioners' use of the Internet for clinical information. J Med Libr Assoc 2002; 90: 370 – 379.
23. Schilling L, Steiner J, Lundahl K, Anderson R. Residents' patient-specific clinical questions: opportunities for evidence-based learning. Acad Med 2005; 80: 51 – 56.
24. Muller H, Hanbury A, Shorbaji NA. Health information search to deal with the exploding amount of health information produced. Methods Inf Med 2012; 51: 516 – 518.
25. Berner ES, McGowan J. Use of diagnostic decision support systems in medical education. Methods Inf Med 2010; 49: 412 – 417.
26. Davies K. The information-seeking behaviour of doctors: a review of the evidence. Health Info Libr J 2007; 24: 78 – 94
27. Murdock B, The serial position effect of free recall. J Exp Psychol 1962;64: 482–488.
28. Hunt R. The subtlety of distinctiveness. Psychonom Bull Rev 1995; 2: 105 – 112.
29. Henson R. Short-term memory for serial order: the start-end model. Cognitive Psychol 1998; 36: 73 – 137.
30. Terry W, Serial position effects in recall of television commercials. J Gen Psychol 2005; 132: 151 – 163.
31. Zamir O, Etzioni O. Grouper: a dynamic clustering interface to web search results. Proceedings of the 8th International Conference on World Wide Web (WWW'09). pp 1361–1374.
32. Senathirajah Y, Bakken S. Visual clustering analysis of CIS logs to inform creation of a user-configurable web CIS interface. Methods Inf Med 2011; 50: 337 – 348
33. Klockner K, Wirschum N, Jameson A. Depth- and breadth-first processing of search result lists. Proceedings of the 22nd SIGCHI Conference on Human Factors in Computing Systems (CHI '04). p 1539.
34. Payne JW, Bettman JR, Johnson E. The Adaptive Decision Maker. New York: Cambridge University Press; 1993.
35. Joachims T, Granka L, Pang B, Hembrooke H, Gay G. Accurately interpreting click-through data as implicit feedback. Proceedings of the 28th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05). pp 154 – 161.
36. Cutrell E, Guan Z. What are you looking for? An eye-tracking study of information usage in web search. Proceedings of Human Factors in Computing Systems (CHI'07). pp 407 – 416.
37. Keane M, O'Brien M, Smyth B. Are people biased in their use of search engines? Commun ACM 2008; 51: 49 – 52.
38. Jansen BJ, Spink A, Bateman J, Saracevic T. Real life information retrieval: A study of user queries on the web. SIGIR Forum 1998; 33: 5 – 17.
39. Jansen BJ, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web. Inform Process Manag 2000: 36: 207 – 227.
40. Silverstein C, Henzinger M, Marais, H, Moricz, M. Analysis of a very large Web search engine query log. ACM SIGIR Forum 1999; 33: 6 – 12.
41. Cho J, Roy S. Impact of search engines on page popularity. Proceedings of the 13th International World Wide Web Conference (WWW'04). pp 20–29.
42. Pandey S, Roy S, Olston C, Cho J, Chakrabarti S. Shuffling a stacked deck: the case for partially randomized ranking of search engine results. Proceedings of the 31st international Conference on Very Large Data Bases (VLDB '05). pp 781 – 792.
43. Teevan J. How people recall search result lists. Proceedings of the 24th SIGCHI Conference on Human Factors in Computing Systems (CHI '06). pp 1415–1420.
44. Moskovitch R, Shahar Y. Vaidurya: A multiple-ontology, concept-based, context-sensitive clinical-guideline search engine. J Biomed Inform 2009; 42: 11 – 21
45. Lin Y, Li W, Chen K, Liu Y. A Document Clustering and Ranking System for Exploring MEDLINE. J Am Med Inform Assoc 2007; 14: 651 – 661.
46. Malo P, Sinha A, Wallenius J, Korhonen P. Concept-based document classification using Wikipedia and value function. JASIST 2011; 62: 2496 – 2511.

47. Roy A, Mackin P, Wallenius J, Corner J, Keith M, Schmick G, Arora H. An interactive search method based on user preferences. Dec Anal 2009; 5: 203–229.
48. Siskos Y, Grigoroudis V, Matsatsinis N. UTA methods. In: J. Figueira, S. Greco, M. Ehrgott (eds.). Multiple criteria decision analysis: state of the art surveys. New York: Springer Science + Business Media Inc.; 2005. pp 297–343.
49. Greco S, Słowiński R, Figueira J, Mousseau V. Robust ordinal regression. In: M. Ehrgott, J. Figueira, S. Greco (eds.). Trends in multiple criteria decision analysis. New York: Springer Science + Business Media Inc.; 2010. pp 241–283.
50. Figueira J, Greco S, Słowiński R. Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. Eur J Oper Res 2009; 195: 460– 486.
51. Greco S, Mousseau V, Słowiński R. Ordinal regression revisited: multiple criteria ranking with a set of additive value functions. Eur J Oper Res 2008; 191: 415– 435.
52. Kadziński M, Greco S, Słowiński R. Selection of a representative value function in robust multiple criteria ranking and choice. Eur J Oper Res 2012; 217: 541– 553.
53. Kadziński M, Greco S, Słowiński R. Selection of a representative value function for robust ordinal regression in group decision making. Group Decis Negot 2013; 22 (3): 429– 462