Lost in PubMed. Factors influencing the success of medical information retrieval.

Klaar Vanopstal[a,b], Joost Buysschaert[a], Godelieve Laureys[c], Robert Vander Stichele[d]

[a]Faculty of Applied Linguistics, University College Ghent
Groot-Brittanniëlaan 45
9000 Ghent
Belgium
0032 9 224 97 23
klaar.vanopstal@hogent.be (corresponding author)
joost.buysschaert@hogent.be

[b]Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9)
9000 Ghent
Belgium
klaar.vanopstal@ugent.be

[c]Faculty of Arts and Philosophy, Ghent University
Rozier 44
9000 Ghent
Belgium
0032 9 264 38 01

godelieve.laureys@ugent.be

[d]Faculty of Medicine and Health Sciences, Ghent University
De Pintelaan 185
9000 Ghent
Belgium
0032 9 240 33 47

robert.vanderstichele@ugent.be

Abstract

With the explosion of information available on the Web, finding specific medical information in an efficient way has become a considerable challenge. PubMed/MEDLINE offers an alternative to free-text searching on the web, allowing searchers to do a keyword-based search using Medical Subject Headings. However, finding relevant information within a limited time frame remains a difficult task. The current study is based on an error analysis of data from a retrieval experiment conducted at the nursing departments of two Belgian universities and a British university. We identified the main difficulties in query formulation and relevance judgment and compared the profiles of the best and worst performers in the test.

For the analysis, a query collection was built from the queries submitted by our test participants. The queries in this collection are all aimed at finding the same specific information in PubMed, which allowed us to identify what exactly went wrong in the query formulation step. Another crucial aspect for efficient information retrieval is relevance judgment. Differences between potential and actual recall of each query offered indications of the extent to which participants overlooked relevant citations.

The test participants were divided into “worst”, “average” and “best” performers based on the number of relevant citations they selected: zero, one or two and three or more, respectively. We tried to find out what the differences in background and in search behavior were between these three groups.

Keywords: medical information retrieval, Medical Subject Headings, bibliographic instruction, nursing education, information seeking behavior

1. Introduction

Several studies have been devoted to possible causes for search failure in information retrieval (Hofstede, Proper, & van der Weide, 1996; McCray & Tse, 2003; Sutcliffe, 2000), trying to find out why some information searches do not yield satisfactory results. The aim of the present study is to contribute to the understanding of the reasons for failure in bibliographic searches executed by - relatively – untrained PubMed users. This should help us to formulate educational objectives in bibliographic instruction and to draw a profile of the better-performing searchers and compare it to that of the worst-performing searchers. As Sutcliffe (2000) claims, training the searchers is sometimes the only remedial action.

The present study focuses on the use of PubMed, an online system to access journal citations and abstracts in MEDLINE. PubMed was developed by the National Center for Biotechnology Information (NCBI) and daily provides hundreds of thousands of users with bibliographic information from the life sciences. It is a global resource of US origin; nevertheless many of its users are non-native speakers of English, which makes efficient retrieval an even more challenging task. Although the recommendation that only MeSH terms should be used is a matter of discussion (Jenuwine & Floyd, 2004), the use of these terms can enhance PubMed searches considerably (Richter & Austin) – provided that the user understands how search terms map to MeSH terms and how PubMed’s search engine works in general. Poor understanding of MeSH is an issue that exceeds the problem of the language barrier: native speakers of English may also experience difficulties in formulating a good query with MeSH terms. Controlled vocabularies can therefore enhance information retrieval, but they can also be a barrier to finding relevant information in a time- and cost-efficient way.

In this study, we want to do an error analysis of the queries that were submitted by our test participants, focusing mainly on quality in terms of the MeSH terms they contain, and the differences between their potential and actual recall. Based on an error analysis, we try to formulate advice on how to address retrieval problems. Some searchers succeed in finding relevant results more easily than others. We also draw a profile of efficient searchers versus those who have more difficulty in finding relevant citations by comparing their characteristics and search strategies.

We will discuss the methods used in this study in part two. The results section of this paper consists of two main parts: query error analysis, and secondly, a comparison of the best, average and worst performers. In the third part we will discuss some of our main findings, and finally, we will present our conclusions and future work in parts four and five.

2. Methods

2.1. Recruitment and test setup

We conducted a test at the nursing departments of two Flemish universities and one British university. A total of 100 respondents with different educational and linguistic backgrounds participated in the test: 31 Dutch-speaking and 8 native English bachelor's students, 40 Dutch-speaking and 21 native English master's students.

Prior to the actual retrieval test, the participants completed a pre-test questionnaire, which allowed us to capture the participants' search experience and - for respondents with non-native English - their self-reported English language skills.

After a short introduction into searching PubMed with the use of MeSH terms, they conducted a literature search for a given subject. The participants in our test were stimulated to use MeSH terms, so their query formulation process consisted of several steps: first, they had to find relevant MeSH terms for each of the components of the search question (falls, elderly, long-term care and prevention). In order to find these MeSH terms, they had to go to the MeSH module in PubMed and enter a free-text search term. Subsequently, PubMed made one or more suggestions for MeSH terms, from which the participants had to select the relevant ones and send them to the search box. This action was repeated until a satisfactory query was obtained. For example, most test participants entered the search term "fall" or "falls" in the MeSH module and then selected the MeSH term "Accidental Falls". Once they had found the right MeSH terms for the other components of the search question and submitted their queries to PubMed, a list of citations was returned by the search engine. From this list, they had to select only those citations that were relevant to all aspects of the search question. The students were given 15 minutes to complete the search. All individual sessions were recorded with Morae software, enabling us to time the subtasks and to reconstruct the queries.

After the experimental task, the participants completed a post-test questionnaire which measured their satisfaction with the search results and with the search system. Additionally, all participants completed an English language test, which enabled us to measure their language skills.

2.2. Query collection and error analysis

We collected all the queries submitted during the literature search task. This resulted in a total of 309 queries, issued by 98 participants – two participants did not submit any queries. The number of queries per participant ranges between 1 and 10 per participant, with a median of three.

For each of the queries in our collection, we determined which errors they contained; this allowed us to make a classification of different error types. Queries that contained no errors and covered the information need were labeled as "good queries".

On the basis of these findings, we will try to make suggestions for the improvement of bibliographic instruction.

2.3. Performance

We developed a gold standard, consisting of 62 to 66 citations, depending on the moment of the test session (for more information see (Vanopstal, Vander Stichele, Laureys, & Buysschaert, 2012)). The students' selections were compared against this gold standard in order to calculate recall.

We are especially interested in the students' search strategies and in their relevance judgment, which is reflected in the selection of citations they considered as relevant. We will not report on the typical performance metrics in information retrieval, i.e. proportional recall scores expressed in percentages, but instead we will discuss performance in terms of absolute recall (R_{abs}), i.e. the number of relevant citations selected by the test participants as relevant to the information need.

We consider three relevant citations a good threshold to designate a search as successful, especially in the limited time frame of this test. Three relevant citations is a good starting point for exploratory work using the "related citations" function of PubMed, and it should provide the searcher with a relevant introduction to the research field. Based on this absolute recall, we will subdivide our test group into a "worst" (no citations), "average" (one or two citations) and "best" (three or more citations) performer group (see 2.4).

Next to absolute recall, we also will calculate the number of missed citations per query and per participant. Missed citations are relevant citations that were returned by the queries, but were not selected as being relevant. Using NLM's E-Utilities (http://www.ncbi.nlm.nih.gov/books/NBK25500/), we simulated the students' searches to obtain their resulting lists of citations. Per search, we registered the number of result pages that were viewed. Each page contained 20 citations, so a participant who looked at two result pages, is considered to have viewed 40 citations.

We compared each result list, i.e. only the pages that were actually viewed, to the gold standard. This allowed us to calculate - absolute - potential recall (R_{pot}), the recall the participants would have obtained had they not overlooked any relevant citations. Potential recall is the "raw" recall of the query itself, without any intervention or selection by the searcher.

$$R_{pot} = \text{\# relevant but missed citations} + R_{abs}$$

For instance, if the participant only looked at the first page (with 20 results per page), and there were two relevant citations in that page, potential recall was two.

2.4. Comparison of the performer types

We will analyze the differences between the worst, average and best performers in our test. This categorization is based on absolute recall. Participants in the worst performer group did submit one or more queries, but did not select any relevant citations. The “average performers” selected one or two relevant citations, and the “best performers” selected three or more.

All comparisons between the performer types were tested using the ANOVA test for variables with normal distribution. The other variables were tested using the nonparametric Kruskal-Wallis test with pairwise comparison and Bonferroni correction. All statistical tests were performed with IBM SPSS Statistics 20.

2.4.1. Search process

We consider the number of queries as an indication of the fluency of the search process. Participants who submitted ten different queries obviously had more problems finding the information they needed than those who submitted only one or two queries.

Other indicators for the fluency of the search process are querying and relevance judgment times. As described in Vanopstal et al. (2012), the querying step is “an alternation of search term formulation and MeSH term selection”. It results in the construction of a query and ends when the user submits the query to the search engine. The total querying time is the sum of the querying times that precede each submission of a query.

Total relevance judgment time is the time spent on assessing the lists of citations returned by PubMed after the submission of each query.

2.4.2. Quality-based assessment of queries

In this part of the study we try to find out whether any of the performer groups makes a higher or lower number of errors of a specific type. We will analyze three error types: incorrect MeSH term, underspecification, and the incorrect use of Boolean operators.

2.4.3. Outcome-based query analysis

Queries can be labeled as “good” or “bad” based on the number of errors they contain, but another way to classify them is based on their potential recall (see figure 1: “adequate” versus “inadequate” queries). In this categorization, good or adequate queries yield at least one relevant citation, whereas bad or inadequate queries either lead to an empty result set, or to a list of citations that are not relevant to the information need. Besides the ability to formulate an adequate query, the participants therefore needed the ability to distinguish relevant from irrelevant citations. We can subdivide the category of adequate queries into queries that led to the selection of relevant citations and queries that did not (see figure 1: “good relevance judgment” versus “relevance judgment errors”).

2.4.4. Query reformulation

Another angle from which we can study queries, next to analyzing the errors they contain, is the reformulation strategies used. As mentioned above, the participants had 15 minutes to complete the literature search task. In an ideal situation, they would have entered one comprehensive query, which covered all the components of the information need. However, as many of these students were not familiar with the search system, and as even more of them were not familiar with the subject of the search, most participants had to iterate the process of finding MeSH terms and combining them into a query. We identify different types of strategies and analyze their use by the different performer types.

3. Results

3.1. Sample description

3.1.1. Respondents

A total of 100 respondents participated in the test, 2 of whom did not formulate any queries and are therefore excluded from the analyses. Although the participants come from different linguistic (English versus Dutch-speaking) and educational (bachelor's versus master's level) backgrounds, a Kruskal-Wallis test indicated that there are no significant differences in recall between these groups, so we can safely concatenate them and use another categorization for the purpose of this study, i.e. best, average and worst performers.

3.1.2. Background

With regard to PubMed experience, our test group was rather heterogeneous: 44% had had an elaborate introduction into the use of the search engine, whereas others had only had a brief introduction (46%). Some (10%) claimed to have had no introduction into PubMed at all, although this was part of their curriculum.

About 97% use a computer several times a week to daily, but only 18% consult PubMed with the same frequency. About 40% of our test participants rarely or never use PubMed to search for medical information.

As for English language skills, 74.4% of the - British and Belgian - students achieved a B2 level in reading and 88.8% achieved a B2 level in vocabulary, indicating that they are “independent users” of the English language, and that they should be able to read and understand complex technical texts and “produce detailed text on a wide range of subjects” (for more information about CEFR levels, see http://www.coe.int/t/dg4/Linguistic/Source/Framework_EN.pdf).

3.2. Query analysis

3.2.1. Quality-based query analysis

We analyzed the queries in our collection (n=309) and distinguish 8 types of errors. Table 1 gives an overview:

Table 1: Error types and their frequencies

Error Type	Description	Example	n
1. Irrelevant MeSH term	Query contains at least one incorrect MeSH term.	(("**Multifactorial Inheritance**"[Mesh] AND "Accidental Falls"[Mesh]) AND "Frail Elderly"[Mesh]) AND "Nursing Homes"[Mesh]	89
2. Overspecification	Query is too narrow and therefore yields few or no results.	"Pharmaceutical Preparations"[Mesh] AND "Aged"[Mesh] AND "Risk Factors"[Mesh] AND "Accidental Falls"[Mesh] AND "Nursing Homes"[Mesh]) AND "Nursing"[Mesh]	36
3. Underspecification	Query is too broad; contains only 1 or 2 concepts and yields a long list of citations.	"Accidental Falls" [Mesh]	125
4. Incorrect non-MeSH term	Query contains incorrect free-text search term. The corrective effect of the MeSH terms is lost, and spelling and translation errors corrupt the queries.	multifactorial **programm** and **faling**	42
5. Spelling error	A misspelled and therefore incorrect non-MeSH term	study for **fallprevention**	7
6. Incorrect translation	Query contains an incorrect translation. This can be an incorrect free-text search term, or a MeSH term which is believed to have another meaning than intended.	("Accidental Falls"[Mesh] AND "**Disabled Persons**"[Mesh]) AND "Nursing homes"[Mesh]	7
7. Incorrect operator	The excessive use of AND can lead to overspecification, whereas the exclusive use of OR leads to underspecification.	• ((("Aged"[Mesh] **AND** "Accidental Falls"[Mesh]) **AND** "Residential Facilities"[Mesh]) **AND** "Nursing Homes"[Mesh]) **AND** "Homes for the Aged"[Mesh] • (("Aged"[Mesh]) OR "Residential Facilities"[Mesh]) OR "Accidental Falls"[Mesh]	27

8. Syntax error	query contains unmatched brackets or quotes, or truncated words	• Accidental Falls"[Mesh]) AND**""**Frail Elderly"[Mesh]) AND "Nursing Homes"[Mesh] • "kine* AND ((("Aged"[MeSH] OR "Frail Elderly"[MeSH])) AND "Accidental Falls"[MeSH] AND "Residential Facilities"[MeSH]	17

These error types are not mutually exclusive, i.e. one query can contain several errors, causing overlap between the error categories. Moreover, some errors induce other errors, e.g. “incorrect operator”, and more specifically the excessive use of “AND”, automatically leads to overspecification. The fourth column in the table shows the number of times each error occurs in our query collection.
A total of 60 queries did not contain any errors and covered all components of the information need.

3.2.2. Impact of query quality on potential recall

We analyzed the impact of the eight different error types on search performance, and noticed that three of those error categories had a significant impact on actual and potential recall: incorrect MeSH terms, underspecification, and the incorrect use of Boolean operators (see table 2).

Table 2: Impact of query quality on potential recall

	n	**R_{pot} = 0**	**Mean R_{pot}**
Good queries	60	0	4.05
Queries with incorrect MeSH term	42	73%	0.78
Underspecified queries	125	77%	0.41
Queries with incorrect Boolean operator	27	81%	0.85

Incorrect MeSH terms This error was made in almost 1 out of 3 queries (29%). A total of 73% of the queries containing an incorrect MeSH term had zero potential recall, either because of empty result sets (33%), or because the results were irrelevant to the search question (40%). In the remaining 27%, the search did yield some relevant results, despite the use of a MeSH term that was not entirely relevant for this search. Queries containing an incorrect MeSH term yielded less than one (0.78) relevant citation on average.

Underspecification The error of underspecification, i.e. when queries consist of only one or two terms and are therefore too broad, was made in 125 queries (40%). About 77% of the underspecified queries had zero potential recall. Underspecified queries yielded 0.41 relevant citations on average.

Incorrect use of Boolean operators In 27 queries (8%), one or more Boolean operators were used incorrectly. This manifests itself mainly in the excessive use of AND (67%) and OR (33%). This error led to zero potential recall in 81%, yielding result sets in 37% of the cases, and yielding only citations irrelevant to the search question in 44%.

Good queries A total of 60 queries (19%) were formulated correctly, with an average potential recall of just above 4 citations. This means that the participants who submitted these queries could have selected an average of four relevant citations, whereas they selected less than two.

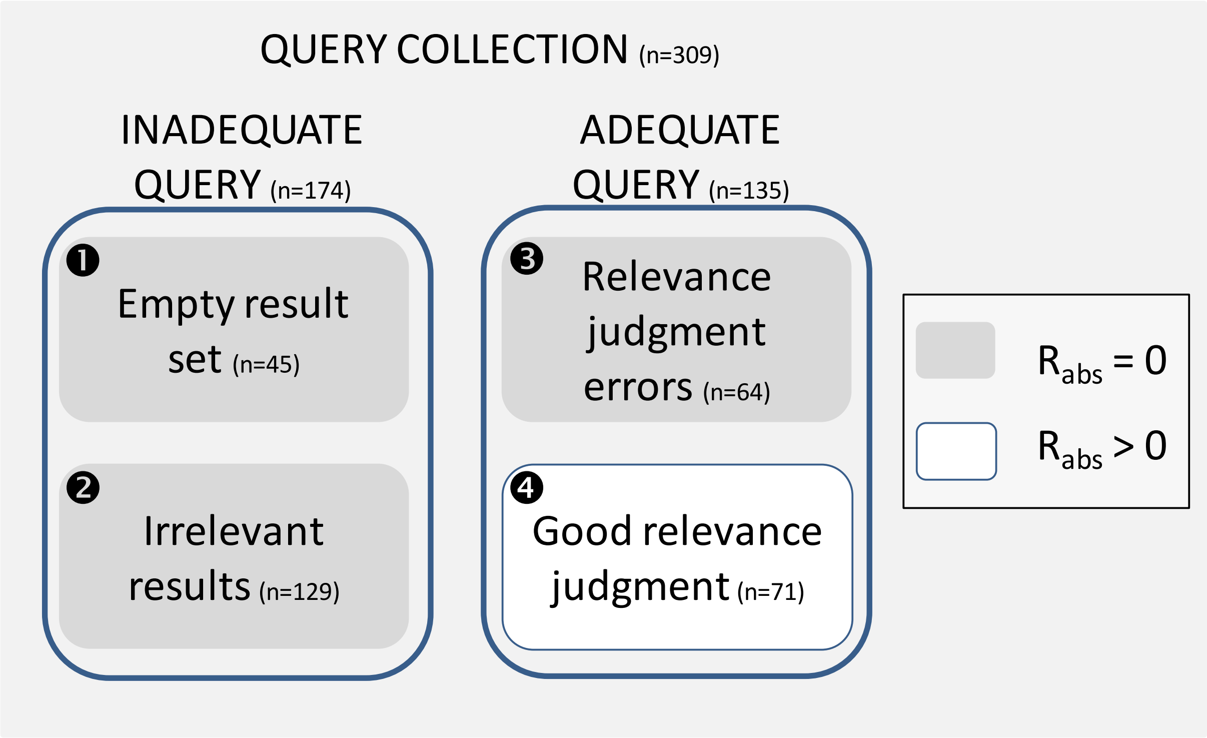
3.2.3. Outcome-based query analysis

Next to the quality of the queries in terms of the number and types of errors they contain, we also assembled data on the potential and actual recall for each query. Potential recall data allow us to determine the direct influence of each error (type) on recall (see 3.2.2), whereas differences between potential and actual recall indicate relevance judgment errors.

We can subdivide our query collection on the basis of their actual and potential recall. Inadequate queries did not yield any relevant results, either because the result set was empty (figure 1 box 1), or because it contained only irrelevant citations (figure 1 box 2). Adequate queries, on the other hand, were well-constructed and covered the information need. However, in some cases relevance judgment errors prevented the searcher from selecting relevant citations (figure 1 box 3). This means that well-formulated queries do not guarantee high recall in the context of our study.

A total of 71 queries (22.9%, figure 1 box 3) were well-formulated, and led to the selection of at least one relevant citation.

Figure 1: Outcome-based classification of queries


QUERY COLLECTION (n=309)
INADEQUATE QUERY (n=174)
ADEQUATE QUERY (n=135)
❶ Empty result set (n=45)
❷ Irrelevant results (n=129)
❸ Relevance judgment errors (n=64)
❹ Good relevance judgment (n=71)
$R_{abs} = 0$
$R_{abs} > 0$

A total of 45 queries returned empty result sets, and another 129 queries had zero potential recall. This means that 56% of the queries in our collection contained errors and did not cover the information need.

A total of 135 queries (44%) were adequate, i.e. they yielded at least one relevant citation. In almost half of those cases (48%), the query itself was acceptable and – although it may contain one or more (minor) errors - had positive potential recall, but the issuer lacked in relevance judgment skills. The remaining 71 (52%) queries had positive potential recall, and their issuers selected at least one relevant citation from the lists of results.

3.3. Performance

During the search task, our test participants selected six citations on average, two of which were relevant (average R_{abs}=2). The potential recall of their searches was four, which means that their search results contained four relevant citations on average, two of which were overlooked by our test participants.

3.4. Comparison of the performer types

3.4.1. Division into performer types

As mentioned in section 2.3, we divided our test group into three performer groups, based on the number of relevant citations they selected. A total of 38 participants are labeled as "worst performers", 28 as "average performers" and 32 as "best performers".

A chi-square test did not reveal any significant differences in the distribution of the student types over the types of performers (see table 3). However, there are more Dutch-speaking master's students in the best performer group than we would statistically expect (observed: 17, expected: 12.8; 53% of the best performers are Belgian master's students).

Table 3: Distribution of participants over 3 performer types

		worst performers (n=38)		average performers (n=28)		best performers (n=32)	
		%	n	%	n	%	n
Dutch	bachelor	27%	10	39%	11	28%	9
	master	39%	15	29%	8	53%	17
English	bachelor	13%	5	3%	1	6%	2
	master	21%	8	29%	8	13%	4

3.4.2. Background of the performer types

There are no significant differences in language skills between the performer types: the average level in all three groups (including the native speakers of English) is B2 for both reading and vocabulary.

A Kruskal-Wallis test showed no significant differences between the performer types in prior experience with PubMed, general computer skills, or in general usage of the Internet to search for information. Although the difference is not significant, we do see that more than half of the participants in the best performer group (56%) are students who had received an elaborate introduction into the use of PubMed.

In the post-test questionnaire, we asked the students whether they were satisfied with their search results and their search process. A one-way ANOVA test (F(2, 97)=28.917; p<.001) showed that the worst performers were significantly less satisfied with their search results than the average and best performers (Bonferroni correction; p<.001 for both groups). The worst performers also experienced their search process as less fluent than the other two groups (F(2, 97)=22.796; p<.001; Bonferroni correction: p<.001) and one in three of the worst performers find PubMed not so user-friendly, as opposed to less than one in five in the average and best performer groups.

3.4.3. Search process

On average, all three performer types submitted three queries during the search. However, we do see that the number of participants who needed only one query to conduct their search task is higher in the best performer group than in the other groups. This means that their searches are more focused from the beginning, whereas the other participants needed more queries to find what they were looking for.

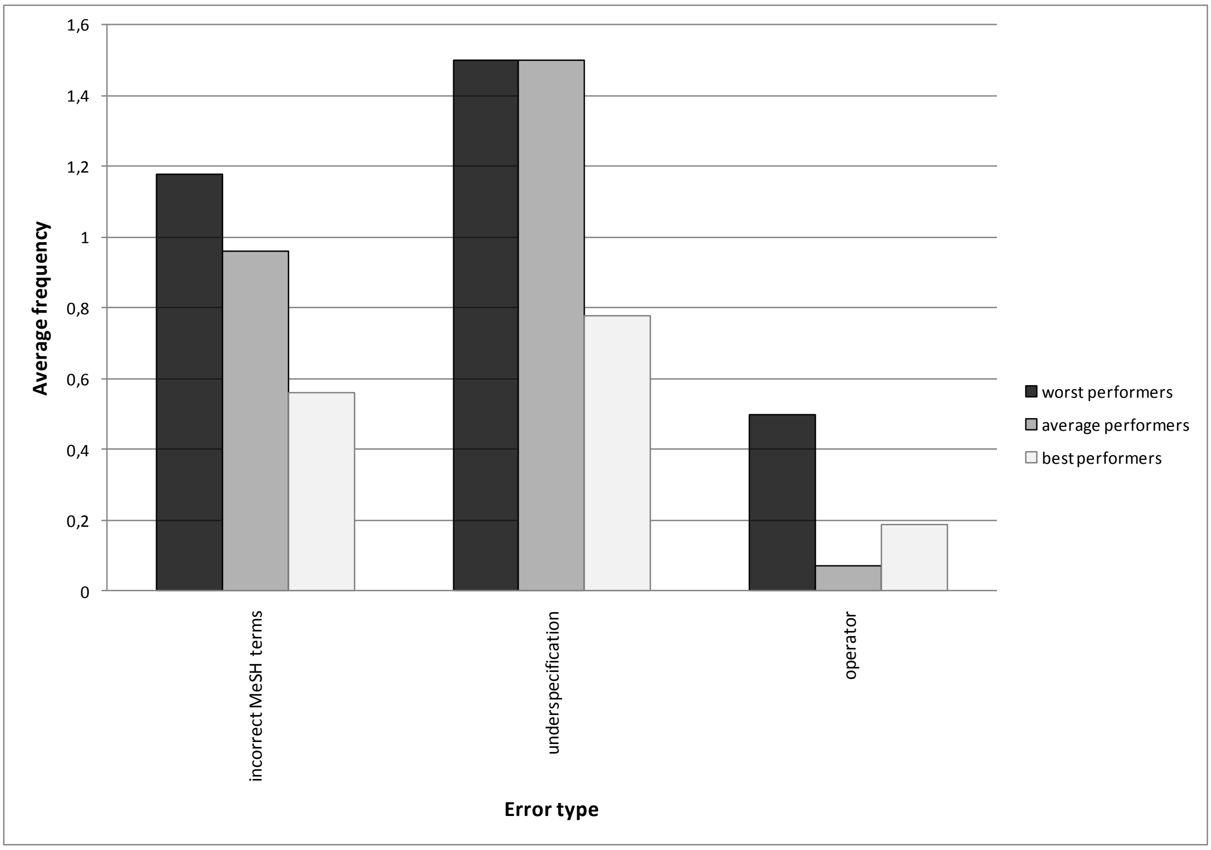
We measured the time spent on querying, i.e. the time spent on searching for MeSH terms and combining them into a query. As we explained in our previous study (Vanopstal, et al., 2012), longer querying times can indicate hesitation. A one-way ANOVA test showed that there were significant differences in querying times between the performer types (F(2, 95)= 11.896, p<.001). Bonferroni post-hoc comparisons of the three groups indicated that the worst performers needed significantly more time to formulate their queries than the average (p=.001) and best (p<.001) performers.

Total evaluation time is the time spent on skimming the result list(s) for relevant results. As the total evaluation times were not distributed normally, we used a Kruskal-Wallis test to find any differences between the three performer types (H= 18.18, p<.001). Post-hoc tests for pairwise comparison showed us that the average and best performers spent significantly more time on the evaluation of the search results (p= .003 and p<.001, respectively).

3.4.4. Quality-based query analysis per performer type

Figure 2 shows a summary of the errors that will be discussed in this section. Although we also see some clear differences in the number of bad MeSH terms used by the performer types, and we have already shown the impact of incorrect MeSH terms on recall (see 3.2.2), we only found significant differences in the number of underspecification errors and in the incorrect use of Boolean operators. We refer to the error analysis for an analysis of the direct impact of different types of errors on recall.

Figure 2: Summary of errors per performer type



• MeSH terms

As described above (see 2.1), our test participants were instructed to use MeSH terms. In previous research (Vanopstal, et al., 2012), we have shown that MeSH terms have a corrective effect; they compensate for possible errors in the free-text search terms that were entered in the MeSH module. Although these free-text search terms have no direct effect on recall, they may have an impact on the fluency of the search process. The worst performers formulated significantly more search terms than the other two groups (H=9.95, p=.007), indicating that they struggled to find the right MeSH terms for their search.

The best performers selected a smaller number of incorrect MeSH terms, which enabled them to construct better queries. Although there is a clear trend in the number of badly chosen MeSH terms, the differences between the performer types is not significant.

• Underspecification

Both worst and average performers made a high number of underspecification errors: 1.5 times on average during the search. A Kruskal-Wallis test showed a significant difference in occurrence of this error between the performer types (H= 8.030; p= .018), more specifically between worst and best performers (Bonferroni correction; p= .028).

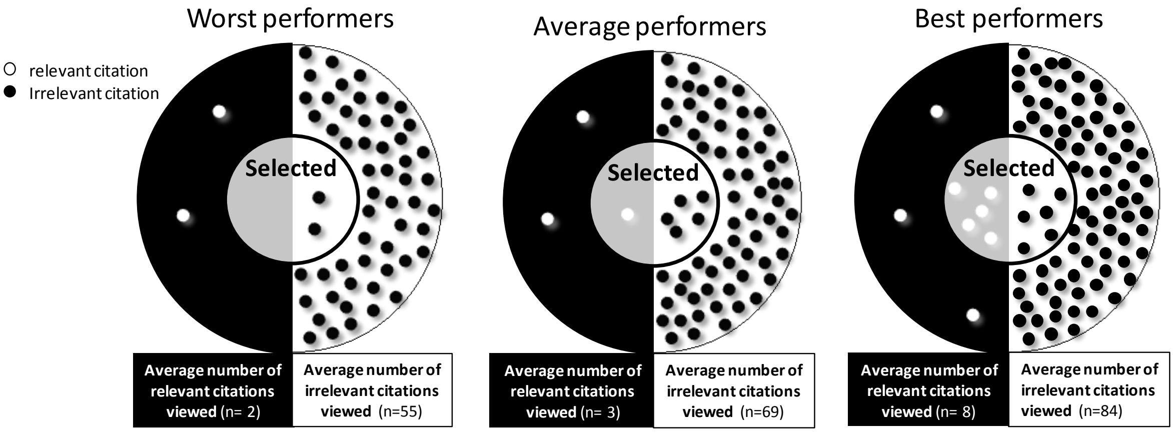
• Boolean operators

A Kruskal-Wallis test revealed a significant difference between the performer types in the use of Boolean operators (H= 8.037, p= .018), usually the excessive use of AND or OR, especially between the worst and average performers (Bonferroni correction; p= .014).

3.4.5. Differences between actual and potential recall as an indication of relevance judgment quality

Figure 3 below gives an overview of the number of citations viewed by each performer group and the proportions of relevant and irrelevant citations. For each PubMed search, we registered how many – titles of - citations in the result list were viewed. When a participant performed more than one search, we added up this number from the several searches. On average, 67 citations were viewed. The worst performers viewed 57 citations on average, 55 (96%) of which were irrelevant. Although the remaining two (4%) were relevant, this group failed to distinguish them from the relevant ones. The average performers viewed 72 citations on average, 69 (96%) of which were irrelevant. They missed some citations, but succeeded in identifying some too. On the other hand, his group also selected more irrelevant citations than the worst performers. Finally, the best performers viewed 92 citations on average, 10% of which were relevant, indicating that their queries were better constructed than those in the other two groups. They were also better at identifying the relevant citations, as they only missed 38% of the relevant ones in the results lists. However, they also selected a relatively high number of irrelevant citations.

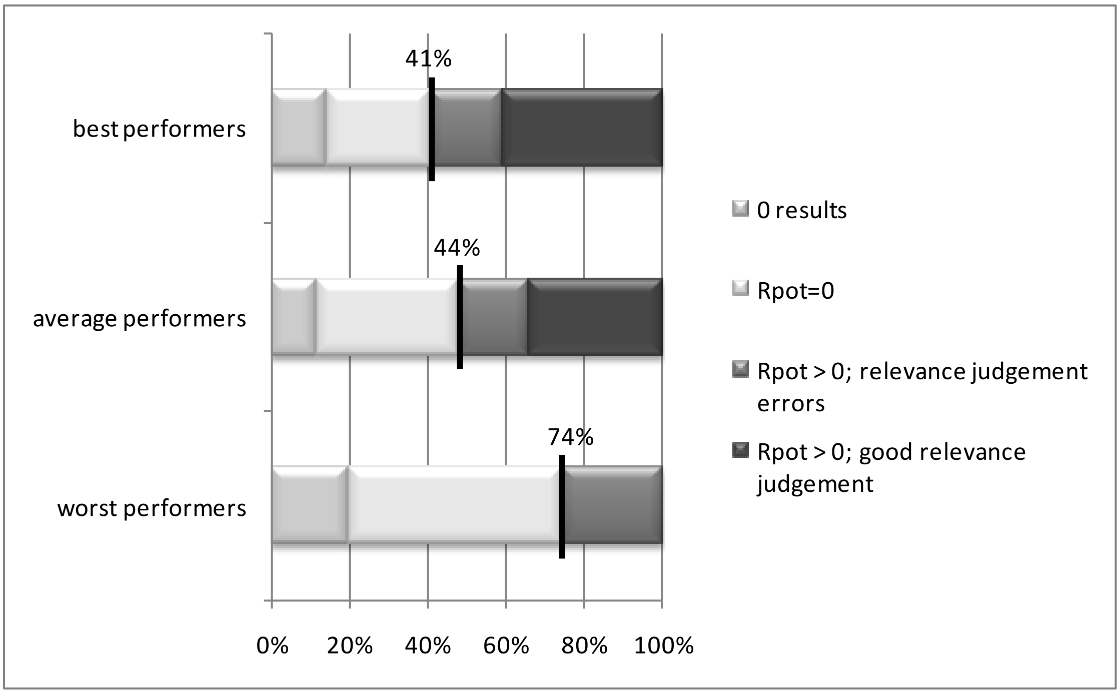
Figure 3: Relevant versus irrelevant citations selected by the performer types


Worst performers
Average performers
Best performers
○ relevant citation
● Irrelevant citation
Selected
Selected
Selected
Average number of relevant citations viewed (n= 2)
Average number of irrelevant citations viewed (n=55)
Average number of relevant citations viewed (n= 3)
Average number of irrelevant citations viewed (n=69)
Average number of relevant citations viewed (n= 8)
Average number of irrelevant citations viewed (n=84)

3.4.6. Outcome-based comparison

We already stated above (see 3.2.3) that ill-formulated queries and bad relevance judgment can cause low recall. In figure 4, this information is linked to the performer types.

Figure 4: Percentage of zero and positive potential recall queries per performer type



Ill-formulated queries can lead to empty result sets, or to zero potential recall. About 74% of the queries issued by the worst performers were ill-formulated, which is almost double of the erroneous queries in the group of average (44%) and best (41%) performers.

About 60% of the queries submitted by the best performers were adequate, i.e. they yielded at least one relevant citation (potential recall > 0). In the group of average performers, this was 56%, whereas no more than 26% of the queries in the worst performer group yielded relevant results.

Due to bad relevance judgment, the worst performers failed to identify any of the relevant citations yielded by those 26% of good queries. The best and average performers failed to identify any of the relevant citations yielded by their adequate queries in 18% of the cases.

3.4.7. Query reformulation

The formulation of a good query requires a conceptual analysis of the information need, and a thorough understanding of the syntax used by the search engine. When a query does not yield satisfactory results, a searcher may have problems finding alternative ways to formulate it. It takes some insight to see what exactly went wrong in a query for a searcher to be able to correct that error.

We distinguish six different types of reformulation: narrowing, broadening, substitution, repetition, trial and error, and a last category which we call “one”.

- **Narrowing**: a more general query is made more specific by adding one or more MeSH terms
  e.g. Query 1= *“Housing for the Elderly [MeSH] AND Accidental Falls [MeSH]”*;
  Query 2= *“(Housing for the Elderly [MeSH] AND Accidental Falls [MeSH]) AND Accident Prevention [MeSH]”*)
- **Broadening**: a query that is too specific - and therefore often yields an empty results set – is made more general by omitting one or more terms from the query

e.g. Query 1= “(Housing for the elderly [MeSH] AND Accident Prevention [MeSH]) AND Nursing Homes [MeSH]”
Query 2= “Accident Prevention [MeSH]) AND Nursing Homes [MeSH]”

- **Substitution:** one MeSH term is substituted for another

  e.g. Query 1= “Accidental Falls [MeSH] AND Frail Elderly [MeSH]”
  Query 2= “Accidental Falls [MeSH] AND Elderly [MeSH]”
  Query 3= “Accidental Falls [MeSH] AND Residential Treatment [MeSH]”
  Query 4= “Accidental Falls [MeSH] AND Combined Modality therapy [MeSH]”
- **Repetition:** re-use of a previous query
- **Trial and error:** formulation of a completely different query, as the previous one did not appear to yield any satisfying results.

  e.g. Query 1= “Critical pathways [MeSH]”
  Query 2= “Accident Prevention [MeSH]”
  Query 3= “(Aged [MeSH] OR Frail Elderly [MeSH] OR Housing for the elderly [MeSH])”
- **One:** only one query was submitted.

In general, there are no significant differences in the use of one specific reformulation strategy between the three performer types, except for the trial and error strategy (Kruskal-Wallis H= 9.010; p=.011). The worst and average performers use this strategy significantly more often than the best performers (Bonferroni correction, p= .046 and p= .018, respectively). This may be another indication that their searches are less fluent.

As pointed out above (see 3.4.4), the best performers used a lower number of incorrect MeSH terms in their queries than the worst performers did. There are three ways in which this error can be corrected: by removing the incorrect MeSH term, which is a way of broadening the query, by replacing the incorrect MeSH term (substitution), or by formulating a completely new query (trial and error). The errors that were made in the best performer group were corrected in 60% of the cases, as opposed to 48% in the worst performer group.

We already showed that there were no significant differences in the incorrect use of Boolean operators between the worst and best performers. The difference between the two groups lies more in their reaction to the - usually poor - results of these searches. In only 26% of the cases did the worst performers succeed in correcting the erroneous query. The other queries either repeated the error, or they were replaced by another erroneous query. This indicates that the searchers did not know exactly what went wrong. The best performers, on the contrary, corrected 83% of the queries containing an error of this type. Correction is done by either replacing the operator (substitution), removing a component of an overspecified query (broadening), or by formulating a completely new query (trial and error).

The best way to correct an underspecified query is to narrow it down to a more specific one. About 60% of the underspecified queries were corrected this way in the best performer group, as opposed to 34% and 31% in the worst and average performer groups, respectively.

An overspecified query should be corrected by broadening. This reformulation strategy was used in 15%, 25% and 33% of the queries in the worst, average and best performer groups, respectively.

Incorrect free-text terms seem to be very difficult to correct, as the searcher mostly does not realize that there is an error in the query. These free-text terms were replaced in nine (out of 43) queries, but only in two of those queries did the searcher (best performer type) replace the incorrect free-text term with a correct one.

4. Discussion

4.1. Main findings

When we look at the separate queries, there are three error types which have a direct impact on potential recall, i.e. which cause the query to yield few or no relevant citations: incorrect MeSH terms, underspecification and incorrect Boolean operators. Between 73% and 81% of the queries containing these error types had zero potential recall.

Good queries do not guarantee high recall: in almost half of the queries with positive potential recall, students failed to identify the relevant citations. This indicates that the participants experienced some problems during the relevance assessment step.

None of the four student types (Dutch-speaking bachelor’s and master’s students, native English bachelor’s and master’s students) outperformed the others, whereas we had expected the English (master’s) students to be the better-performing ones. The Dutch-speaking master’s students are better represented in the best-performing group. This group had had the most elaborate introduction into PubMed during their training. This may indicate that language skills – although obviously important - do not compensate for the lack of facility with the search engine.

There are no significant differences between the performer types in the scores on the language tests, educational background or computer skills. The worst performers did not select any relevant citations, and they are well aware of their poor performance. One in three of these participants assessed the PubMed search system as “not so user-friendly”.

The worst performers struggled to find the correct MeSH terms for their searches and generally needed more time to formulate their queries. On the other hand, they spent less time on the evaluation of the search results, a crucial step in information retrieval.

Making errors may be one indication of poor research skills. However, the correction of an error in the next query demonstrates a certain level of understanding of the system. This study showed that the ability to correct one's own errors distinguishes better performing searchers from the less successful ones.

4.2. Strengths and limitations

One of the limitations of this analysis is the small number of queries available for research. It is difficult to find significant results for such a small dataset. However, we do believe that the fact that these queries were all meant to fulfill the same information need – as opposed to queries from logs, where the information need is unknown – adds to the validity of our conclusions.

4.3. Critical remarks on main findings

4.3.1. Impact of query quality

As argued by Dogan et al. (2009), the quality of a query depends on 3 factors: the searcher's understanding of the information need, his searching skills, and system design on the search engine's side. The retrieval experiment described in this paper was set up to enable us to formulate advice for the improvement of bibliographic instruction. In an earlier paper, we concluded that the nonidentification of concepts in the information need was the main cause for noncoverage. The first factor, i.e. understanding of the information need, is therefore a problem that should be tackled in bibliographic instruction. The second factor, searching skills, should be addressed in bibliographic instruction as well, focusing on three error types: incorrect use of MeSH terms and of Boolean operators, and the formulation of underspecified queries.

Most of the queries that contained an *incorrect MeSH term* did not lead to the selection of any relevant citations, either because of empty result sets, or because the query only yielded irrelevant results, or because relevant citations were overlooked.

Underspecification, also referred to as "the million hits syndrome"(Mulligen, et al., 2004), leads to very long lists of results, which discourage the searcher from skimming the results. In almost two out of three of the underspecified queries, test participants considered cost-effectiveness too low and constructed a new query. Only 12% made the effort of going through the results, and succeeded in identifying at least one relevant citation. Underspecification in itself therefore does not render a query completely useless; however, it makes the relevance judgment step much more labor-intensive and causes people to give up.

The danger of using *incorrect operators* lies especially in overspecification, which usually results in queries with zero potential recall.

Medical students should learn how to construct comprehensive queries that cover the information need, without overspecifying. They need to gain more insight into the use and structure of MeSH, practice combining the terms to a good query, and learn to interpret the MeSH terms assigned to the citations that were retrieved. In this respect, the incorporation of MeSH translations into the search engine may be useful for non-native speakers of English. An understanding of the indexing and relevance sorting algorithms may also help to formulate better queries.(Aula, 2003)

The absence of errors in queries, however, does not guarantee positive recall: bad relevance judgment may cause searchers to overlook relevant citations, as it did in about 25% of the queries. More experience in reading scientific articles, and more familiarity with the display settings in PubMed may facilitate relevance assessment of citations based on their abstract.

4.3.2. Performer profiles

There are no significant differences in the distribution of the two student levels in the groups of performers (see table 3), although the Belgian master students are better represented in the best performer group. We assumed that native speakers of English would do better on a literature search task in PubMed, and therefore that a larger proportion of the native English participants would be in the best performer group. However, their language skills do not seem to compensate for the lack of searching skills.

Although there are no significant differences between the performer types with regard to PubMed familiarity or frequency of use, we do see that more than half of the best performers were Belgian master students – the most experienced PubMed users in our test group. Searching skills therefore definitely play a role in search efficiency.

We did not find any significant differences in language skills between the performer types. However, when we only look at the non-native speakers of English, a Kruskal-Wallis test shows that the best performers scored better on the reading test than the average and worst performers (H= 3.968; p=.047): 81 percent of the best performers achieved a B2 level or higher, as opposed to 60 and 44 percent in the worst and average performer groups, respectively. The differences in scores on the vocabulary test are less obvious, as the scores are relatively high in all three groups. This means that English – reading – skills do play a role in information retrieval, more specifically in non-native speakers of English.

4.3.3. Errors made by the different performer types

Long citation lists resulting from underspecified queries discourage most searchers from scrolling through them. Participants of the worst performer type who made this error failed to select any relevant citations, whereas many of the average and best performers did. This means that the latter are either more perseverant, or their relevance judgment skills compensate for a low-quality query. Underspecification therefore especially has an impact on recall in those searchers who lack in relevance judgment skills.

The incorrect use of Boolean operators was especially found in queries submitted by the worst and best performers, whereas only three average performers committed this error. Differences in system experience may partly explain this difference between worst and average performers, whereas the differences between average and best performers may be caused by the length of the queries. Query length in the average performer group was 4.1, in the best performer group 5.8. Longer queries automatically contain more operators, which makes them more error-prone.

We consider citations that do not contain the crucial components *falls* and *fall prevention* as completely irrelevant to the search question: citations in which these two components are not represented contain too little information to answer the information need. Surprisingly, we see that the best performers selected a significantly larger number of citations without the components *falls* and *prevention* than the worst performers. They selected more relevant, but also more completely irrelevant citations. This illustrates the classical trade-off between precision and recall: the students' selections contain an increasing number of irrelevant citations with increasing performance (r_s=.344, p=.000, n=98). In other words, the higher the recall, the more "noise" we see in the students' selections.

The main difference between bad and average or good performers lies in the query formulation step. The worst performers failed to construct a comprehensive query with relevant MeSH terms and no syntax errors. This issue should clearly be addressed in bibliographic instruction. The difference between average and good performers is subtler, and also mainly originates in the query formulation step. This is illustrated by the average potential recall scores in each of the performer types: average recall in the worst performer group was 0.5, and 1 and 3 in the average and best performer groups, respectively. Although their queries were still rather unsuccessful, the average performers did succeed in identifying most of the relevant citations their queries yielded. The best performers' queries were better-constructed and yielded more relevant results, which, in turn, made it easier for the participants to identify them. The best performers spent more time on relevance judgment, probably because they made strategic decisions in allocating enough time to this crucial last phase.

4.3.4. Query reformulation

Incorrect free-text terms are rarely (twice in our query set) corrected by our test participants, rather they are repeated, or replaced by another incorrect free-text term. This corroborates our previous finding that the extra step of selecting MeSH terms can be very useful to prevent errors from percolating to the final query (Vanopstal, et al., 2012).

Another error that seems very difficult to correct, is the error of overspecification. About one in three of these errors were corrected. This error therefore also deserves some extra attention in bibliographic instruction.

The incorrect use of MeSH terms, and underspecification and overspecification errors are problems that need extra attention, especially in the instruction of novice searchers. They seem to have more difficulty in correcting these errors than the better-performing searchers.

5. Conclusions

We conducted a retrieval experiment in a group of nursing students with mixed linguistic and education level backgrounds: Dutch-speaking master's and bachelor's nursing students, and native English master's and bachelor's nursing students. The aim of this study was twofold: to formulate advice for the improvement of bibliographic information retrieval instruction, and to draw a profile of the best, average and worst performers in the test.

An analysis of the queries submitted by our test participants allowed us to identify the errors with a direct impact on recall, and to determine a focus for bibliographic information retrieval instruction. Although broad queries can be good for a searcher's orientation within a specific domain, exercises on the translation of an information need into a good query should prevent the students from formulating broad or underspecified queries (only). The skills required for this include a thorough analysis of the components of the information need, the translation of these components into free-text search terms and subsequently into MeSH terms. Students may benefit from some practice in the use of these MeSH terms, which can enhance a search considerably, provided the terms are used correctly. We agree with Aula's assertion that an understanding of the indexing and relevance sorting algorithms may also help to formulate better queries (Aula, 2003). Combining MeSH terms using Boolean operators to obtain a comprehensive query is a difficult task which should also be addressed in bibliographic retrieval instruction.

Another problem in information retrieval using PubMed is the relevance judgment step. Relevant citations are often overlooked, even by native English speaking searchers. Skimming exercises may help the students to detect the structure and contents of abstracts more easily. General familiarity with scientific texts may also facilitate the relevance judgment step.

We tried to draw a profile of the "efficient searchers" in our test group and analyzed what they did differently from the less efficient searchers. In non-native speakers of English, the level of English language skills plays an important role in retrieval, as the best performers are those with the highest scores on the English language tests.
More than half of the best performers proved to be Belgian master's students, the group who had received an elaborate introduction into the use of PubMed in their master's training.

The best performers generally formulated better queries, were better at detecting and correcting the errors in their queries and had less difficulty in identifying the relevant citations in the result sets. The correction of one's own errors in queries requires insight into the search system and a critical analysis of the queries. The best performers are better at correcting errors pertaining to incorrect MeSH terms, Boolean operators and underspecification. They do, however, also have problems detecting and correcting the apparently more complex errors of overspecification and incorrect free-text terms.

6. Future work

We would like to experiment with some techniques that facilitate both query formulation and relevance judgment for non-native English searchers. A translated version of the Medical Subject Headings can help them to formulate a good query. This translation can also be integrated for relevance judgment: listing the translated MeSH terms that are assigned to each citation can be helpful do decide whether an article is relevant to the information need or not. We would also like to experiment with simplified abstracts using automatic paraphrasing techniques, and with wikification (He, Rijke, & Sevenster), which may also make the selection of relevant abstracts easier. Applying comprehensibility assessment techniques like OCSLA (Liu & Lu, 2009) to the abstracts in PubMed may provide some insight into the reasons why some texts are more easily understood – and selected – than others.

References

Aula, A. (2003). Query Formulation in Web Information Search. In P. Isaias & N. Karmakar (Eds.), *IADIS International Conference WWW/Internet* (Vol. I, pp. 403-310). Algarve, Portugal: IADIS Press.

He, J., Rijke, d. M., & Sevenster, M. Generating links to background knowledge for medical content. In *None*: ACM.

Hofstede, A. H. M., Proper, H. A., & van der Weide, T. P. (1996). *Query formulation as an information retrieval problem* (Vol. 39). Oxford, ROYAUME-UNI: Oxford University Press.

Islamaj Dogan, R., Murray, G. C., Neveol, A., & Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. *Database (Oxford), 2009*, bap018.

Jenuwine, E. S., & Floyd, J. A. (2004). Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. *J Med Libr Assoc, 92*, 349-353.

Liu, R.-L., & Lu, Y.-L. (2009). Online assessment of content skill levels for medical texts. *Expert Systems with Applications, 36*, 12272-12280.

McCray, A. T., & Tse, T. (2003). Understanding search failures in consumer health information systems. *AMIA Annu Symp Proc*, 430-434.

Mulligen, E. v., Diwersy, M., Schijvenaars, B., Weeber, M., van der Eijk, C., Jelier, R., Schuemie, M., Kors, J., & Mons, B. (2004). Contextual annotation of web pages for interactive browsing. *Medinfo, 11*, 94-98.

Richter, R. R., & Austin, T. M. Using MeSH (Medical Subject Headings) to Enhance PubMed Search Strategies for Evidence-Based Practice in Physical Therapy. *Physical Therapy, 92*, 124-132.

Sutcliffe, A. M. A. M. (2000). Model mismatch analysis: towards a deeper explanation of users' usability problems. In *Behaviour & Information Technology* (Vol. 19, pp. 43-55): Taylor & Francis Ltd.

Vanopstal, K., Vander Stichele, R., Laureys, G., & Buysschaert, J. (2012). PubMed Searches by Dutch-Speaking Nursing Students: The Impact of Language and System Experience. *Journal of the American Society for Information Science and Technology, 63*, 1538-1552.