Introduction

The ability to construct effective search strategies is fundamental to information retrieval, yet user performance varies significantly depending on expertise, interface design, and query reformulation behaviours. As search systems become more sophisticated, understanding how users adapt their strategies across different interfaces remains an important area of study. This paper investigates how graphical search interfaces influence query construction, reformulation tactics, and overall query quality compared to traditional form-based systems.

Drawing from prior research, it is evident that search behaviour is shaped by factors such as domain knowledge, technical expertise, and interface affordances. Novice users often struggle with advanced search techniques, favouring simpler strategies, while experienced users demonstrate more refined and efficient methods (Liu & Wacholder, 2017; Yoo & Mosa, 2015). Query reformulation frameworks, such as those proposed by Jansen, Booth and Spink (2009), Hu, Lu and Joo (2013), Rha, Shi and Belkin (2017), Tibau et al. (2019), provide valuable insights into how users adjust their queries in response to search challenges.

Despite these advances, the interaction between expertise levels and interface types in shaping reformulation strategies and query outcomes remains underexplored. Most prior studies tend to focus on either user expertise or interface design in isolation, making it difficult to understand how these dimensions interact. For example, while form-based systems may encourage precision among experienced users, graphical interfaces may support broader exploration and engagement—especially for novices. Yet few studies have systematically compared how different user types respond to different interface affordances under controlled conditions. Moreover, although query reformulation frameworks have been widely used to classify search behaviours, their application in interface evaluation—particularly in combination with behavioural metrics such as Boolean use, reformulation frequency, and alignment with expert-crafted search strategies—remains limited.

This study addresses this gap by comparing the search behaviours of two distinct cohorts—search professionals and master's students—across two interfaces: a graphical search interface and a conventional form-based system. By analysing query structure, reformulation tactics, and alignment with expert search strategies, we aim to uncover how interface design and user expertise jointly influence search performance. The findings have implications for the design of search systems and training interventions tailored to diverse user needs.

Literature review

Search behaviour across expertise levels

The current study seeks to uncover if and how the use of a graphical interface improves searching for different user types. Cohorts have been compared in previous studies from different perspectives. Taking an evaluation approach, Osborne and Cox (2015) studied differences in the perception of future OPACs between three groups: librarians, library students and master's students in an interview study. The interviews covered a number of OPAC characteristics, but particularly relevant to the current study are the findings that the graphical appearance of the interface under study received positive feedback from a majority of the two student groups, while almost half of the librarians found room for improvement in the graphical elements. Across the three user groups the authors identified agreements among the participants, but also differing

observations, emphasising that different user groups notice different features and elements when evaluating new interfaces.

Liu & Wacholder (2017) investigated search effectiveness in a comparison of four groups of users with different levels of search expertise and topic knowledge, in particular how they benefit from using the controlled vocabulary MeSH for searching. They found that novice searchers (those with little domain knowledge and search expertise) used controlled terms for searching the least and had the lowest mean precision in their queries. On the other hand, users with extensive domain knowledge and search skills used MeSH terms in about two out of three queries. The highest precision was found among domain experts with little search experience, which leads to the hypothesis that searchers with topic knowledge benefit more from using search tools like controlled vocabularies. The study also found that search novices had difficulties identifying how reformulations could improve search results.

Yoo and Mosa (2015) also did a comparison study, focusing on experienced and inexperienced Pubmed users. The empirical basis of the study was based on an actual Pubmed search log, where queries were divided into sessions to inform the analysis. Being defined as users who use advanced Pubmed functions in their queries, experienced users only accounted for 6% of the queries in the dataset. The study found that experienced users needed fewer queries to locate relevant results, while the number of queries for inexperienced users was higher.

In an earlier study, Elbedweihy, Wrigley and Ciravegna (2012) compared expert and casual users interacting with semantic search. Although the paper does not define if expert refers to search skills or topic knowledge, the study finds both differences and similarities between the two cohorts when testing five different versions of their search tool (e.g., form-based, graph-based, variations of natural language interfaces etc.). Both groups are more efficient with the form-based interface, while considering it to be more boring. This also leads to the assessment that both cohorts have the form-based interface as their first preference, but the experts also rated the graph-based interface. However, the experts differ from casual users in that they are more strategic when planning what to include in queries.

In Okhovati et al. 's (2016) study of medical students, experts and novices were defined according to whether they had previously worked with Scopus or Web of Science. A search test with controlled tasks guided the study. The authors found that both cohorts made the same types of error in the two databases, but inexperienced users made significantly more errors than experienced users. The identification of errors in the search test suggests that specialised query builders could be beneficial for both target groups, and that more training could lead to a reduction in errors for both cohorts.

To sum up, previous research has shown that search expertise does have some impact on query formulation and search behaviour. Fewer errors are made, more advanced queries are composed, and less time is spent. Moreover, across most studies, users seem to prefer form-based search interfaces.

Query reformulation frameworks

Research on query reformulation frameworks offers crucial insights into user behaviours, search strategies, and interaction patterns, contributing to improved information retrieval (IR) systems. Previous studies have applied a variety of different coding frameworks to understand users' query modification behaviours and strategies, emphasising the varied ways searchers refine their queries.

In early work, Jansen et al. (2009) classified query reformulations into several distinct types: “new,” “assistance,” “content change,” “generalisation,” “reformulation,” and “specialisation.” This approach captured nuanced transitions between search stages, revealing how users progress

through exploratory or iterative queries. For instance, generalisation typically involved removing terms, while specialisation added terms. By employing an n-gram modelling approach, the study enabled the prediction of reformulation patterns, offering practical guidance for search system features that anticipate and assist users' next steps.

Building on earlier insights, Hu et al. (2013) focused on the impact of topic familiarity and search skills on query reformulation behaviours specifically in health information searches. They created a framework to categorise both content-related changes—such as specification, generalisation, and parallel movement—and content-unrelated modifications, including synonym use, format adjustments, and error correction. Findings from this study indicate that familiarity and skill level affect reformulation frequency, with more experienced users generalising and specifying terms efficiently. This categorization provides insights for designing health information systems that support diverse user abilities and levels of topic knowledge.

In a further exploration of query reformulation, He, Bron, and de Vries (2013) categorised query reformulations as "new" or "related" to represent direct modifications across task stages. Related to this, Ruotsalo, Jacucci and Kaski (2020) defined a reformulation as a query that shared at least one word with the previous query conducted by the same user. By simplifying reformulation types, He, Bron, and de Vries' (2013) framework enabled the researchers to trace user behaviour across multi-session searches, allowing them to differentiate search stages through patterns of query reformulation. The study suggests that stages of complex search tasks can be tracked independently of specific interfaces, contributing to a more universal understanding of how users modify queries over time in response to evolving information needs.

In a subsequent study, Dempsey and Valenti (2016) analysed keyword and limiter use among students in a discovery service context, coding for specific issues such as misuse of quotation marks, repetitive spelling errors, and lack of keyword variation. Their framework categorises keywording errors on a graded scale, indicating the extent of misuse. For example, quotation marks were coded from 1 (correct use) to 3 (multiple misuses), while keyword variance was rated from 1 (high variance) to 5 (no variance). This framework highlights the need for tailored instruction in information literacy, demonstrating that focused training can address recurring student challenges in query construction.

Most recently, Dahlen and Hanson (2023) provide another perspective with their framework based on "search term modification." They captured specific strategies students used to adjust search terms, including narrowing searches by adding terms, broadening by removing terms, rearranging terms, and using keywords derived from article records. These modifications reflect adaptive behaviours as students navigate searches, often influenced by contextual cues from search results or articles. This study's documentation of organic modifications suggests a responsive design approach, where IR systems could integrate more flexible and context-sensitive support for user-initiated refinements.

Together, these frameworks provide a rich basis for understanding how users adapt their queries in response to system feedback, task complexity, and interface design. However, few studies have applied these models to evaluate how different interface types (e.g. graphical vs. form-based) shape reformulation behaviour, or how this varies with user expertise. In the current study, we build on these frameworks to investigate how searchers of varying experience levels interact with different interface designs. This leads to the following research questions:

1. How does the use of a graphical interface affect query construction among different levels of search expertise?
2. How does the use of a graphical interface affect query reformulation tactics among different levels of search expertise?

3. How does the use of a graphical interface affect query quality among different levels of search expertise?

To answer these questions we investigate the use of a traditional, form-based interface represented by Pubmed (see Figure 1) and an alternative, graphical interface (2Dsearch) (see Figure 2). At the heart of 2Dsearch is a graphical editor which allows the user to formulate search strategies using a visual framework (reference anonymised, 2020). Concepts can be simple keywords or attribute:value pairs representing controlled vocabulary terms (e.g. MeSH terms) or database-specific search operators (e.g. field codes and other commands). Users can combine them using Boolean—and other—operators to form higher-level groups and then iteratively nest them to create complex expressions.

Although visualization of search strategies in this manner offers immediate utility, the true value of the approach is in the interaction design. For example, to edit the expression, the user can move terms from one block to another and create new groups simply by combining terms. They can also cut, copy, delete, and lasso multiple objects. If they want to understand the effect of one block in isolation, they can execute it individually or view the hit counts. Conversely, if they want to remove one element from consideration, they can temporarily disable it. The effects of each change display in real time in the adjacent search results pane, which allows users to rapidly optimize their search queries.

Using form-based query builders to craft syntactically correct search expressions can be an error-prone and tedious process. Line numbers, parentheses, square brackets, punctuation, whitespace characters, and Boolean operators all have the potential for errors. However, a graphical representation can delegate the task of generating syntactically correct expressions to lower-level system functions. In addition, transforming logical structure into graphical structure provides a more direct mapping between the underlying semantics and physical appearance, and offers a more intuitive experience for users wishing to experiment with different approaches. In this way, the graphical approach supports many of the key design principles outlined in Russell-Rose & MacFarlane (2020).

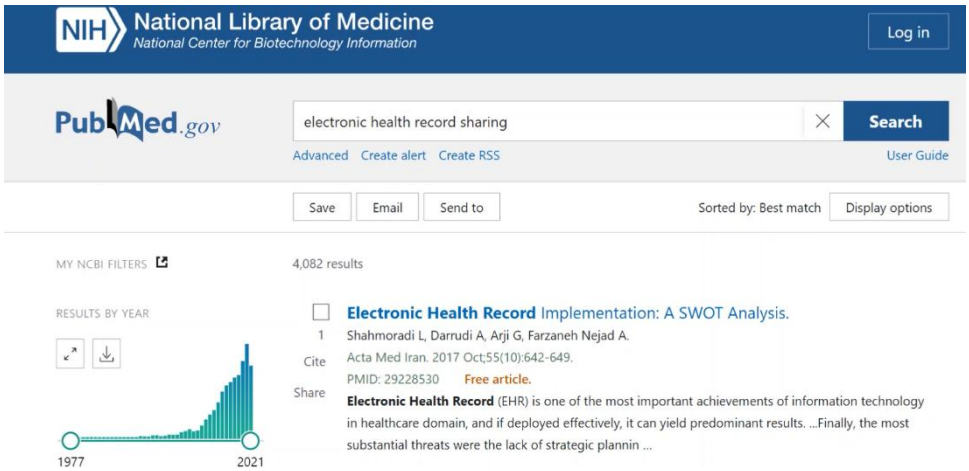

Figure 1. The form-based interface represented by Pubmed.

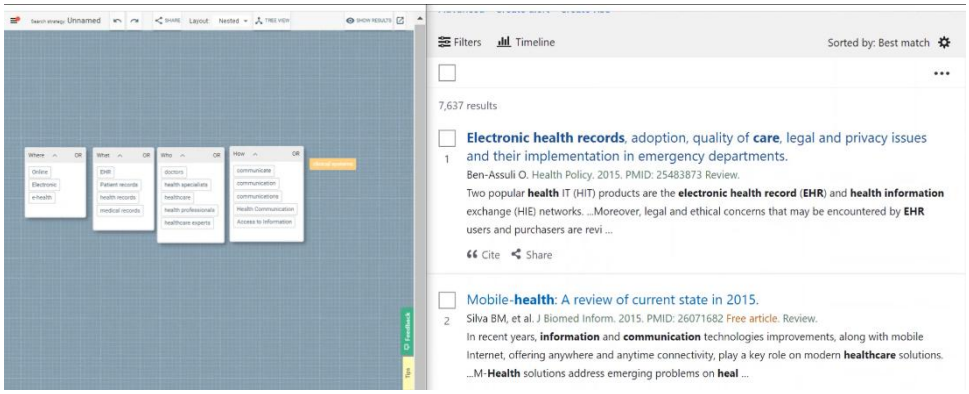

Figure 2. The graphical interface represented by 2Dsearch.

Methods

The aim of this paper is to compare a baseline system (the conventional interface) with an experimental system (the graphical interface). To isolate the effect of the interface as much as possible and minimise the influence of confounding variables, the study was conducted in a controlled lab setting, following the approach outlined by Kelly (2009) and further elaborated in (author anonymised, 2022) and (author anonymised, 2024). As discussed earlier, previous user studies have demonstrated differences in search behaviour across user groups. To reflect this, we recruited participants from two distinct cohorts: search professionals and university students. The professionals were included to represent experienced searchers, while Master's-level students were selected to represent less experienced users—though all had a minimum of three years of academic search experience. 29 participants from the Danish university sector (14 search professionals and 15 master's students of information technology) conducted four controlled search tasks, two using the conventional form-based interface (PubMed) and two using the graphical interface (2Dsearch). The search tasks were designed to elicit exploratory search (Marchionini, 2006) within the digital health domain. Interfaces and tasks were permuted for each test participant to minimise bias or order effects.

Participant interactions were documented in a search log, recording the tasks completed, interfaces used, and the sequence of terms, tokens, and Boolean operators used for each query. For this work the key concepts are defined as:

- Term: a character string delimited by white space
- Token: an instance of a term
- Facet: a conceptual dimension of an information need
- Query: a string of one or more terms submitted to retrieve relevant information
- Query reformulation: a move made to improve the search results from a previous query
- Session: A sequence of queries submitted to complete a controlled search task

The search log provides data for three analyses:

1. A structural analysis of queries and reformulations
2. A taxonomic analysis of the selected reformulation tactics
3. A comparative analysis of participant queries with expert/benchmark queries.

The first analysis consists of a quantitative investigation of the distribution of tokens, terms, facets, and query reformulations for the different combinations of cohorts and interfaces. This analysis used Levenshtein distance ('LD') (Boldi et al., 2011; Wu & Bi, 2017), which is a string metric for measuring the difference between two sequences. LD was calculated using Excel functions.

Category	Definition	Example
Specification (SPE)	To specify the meaning of the previous query by adding more terms or replacing terms with those having more specific meaning	"knowledge sharing" -> (((health professional) AND (education)) AND ("professional development" [Title/Abstract])) AND ("knowledge sharing")
Generalisation (GEN)	To generalize the meaning of the previous query by removing terms or replacing terms with those having more general meaning	app* AND program* -> app*
Parallel movement (PAR)	The previous query and the modified query have partial overlap in meaning, or two queries are dealing with different aspects of a topic	peer to peer OR peer-to-peer -> interpersonal communication
Synonym (SYN)	To replace current terms with those having similar meaning	electronic health record [MeSH Terms] -> ehr[MeSH Terms]
Formatting (FOR)	To change the format of the query without altering the meaning	self?management -> self-management
Inappropriate keywords or structure (INAPP)	Use of inappropriate keywords or structure	(mobile app (mobile AND app))

Table 1. Coding scheme and examples.

The second analysis consists of a taxonomic coding of the reformulation strategies used. The basic unit of analysis is a query reformulation. The coding scheme is an adopted version of the scheme developed by Rieh and Xie (2006) and has been used in previous studies (Hu et al., 2013). Each reformulation was independently coded by two of the authors using the coding scheme in Table 1, and the results were reviewed and revised to reconcile any conflicts. In our analysis we applied the categorisation scheme non-disjunctively, i.e. a given reformulation could be tagged with more than one category. For example, there was one instance where a reformulation was at the same time a generalisation of the previous query (because more synonyms had been added to a facet) and also a specialisation (because an extra facet with terms had been added to the reformulation). This was coded as GEN and SPE, and counted as two reformulation actions. Therefore, the total number of coded reformulations is greater than the number of original query reformulation actions by users. Significant differences between cohorts and interfaces were analysed and identified using Chi-square tests following the precedent of Hu et al. (2013). The analysis was based on contingency tables, and the statistical software SPSS was used for the calculations.

The third analysis compares the participant queries with a 'gold standard' query for each task. Prior to the search test three expert searchers with subject expertise identified relevant search terms and formulated structured queries for all four tasks. The participant queries were then analysed to determine the degree of overlap with the associated 'gold standard' queries. The overlap was calculated in terms of precision (the proportion of participant query terms that match the gold standard) and recall (the proportion of gold standard terms found in the participant query) for each combination of interface and cohort. As many of the gold standard terms consisted of phrases, this

part of the analysis was performed at the term level. The F-measure was used to assess the differences of performance, calculated using Excel functions.

Results

In the following sections present the results of the three analyses: query construction, query reformulation tactics and query quality. Overall, students submitted more queries (new queries and reformulations) in the test to complete the tasks (322 vs. 203, corresponding to an average of 21.5 queries per student and 14.5 per professional). In the analyses below, we focus exclusively on the reformulations.

Search strategy formulation

Table 2 shows the effect of the two interfaces ('form' and 'vis') on the two cohorts ('Professionals' and 'Students') in terms of their use of Boolean operators ('#Bool'), the number of query reformulations ('#reforms'), and the size of those reformulations (measured using Levenshtein distance, 'LD') (Boldi et al., 2011; Wu & Bi, 2017).

Cohorts	#Bool (vis)	#Bool (form)	# reforms (vis)	#reforms (form)	LD (vis)	LD (form)
Professionals	5.07	2.22	2.76	3.36	34.25	50.85
Students	5.61	1.77	5.40	3.17	38.44	28.92
Overall	5.43	1.98	4.20	3.26	37.21	38.71

Table 2. Mean number of Boolean operators, reformulations and edit distance. N=525.

The results show much greater usage of Boolean operators in the graphical interface (5.43 vs 1.98). This effect is particularly pronounced for students (5.61 vs 1.77), although the effect is clearly also present for the professionals group (5.07 vs 2.22). To investigate differences in Boolean operator usage across cohorts and interfaces, we conducted Mann–Whitney U tests. Both professionals and students used significantly more Boolean operators in the graphical interface than in the form-based one ($U = 7885.0$, $p < .001$ for professionals; $U = 20517.0$, $p < .001$ for students). Between cohorts, professionals used significantly more Boolean operators than students in the form-based interface ($U = 7722.0$, $p = .012$), but no significant difference was found in the graphical interface ($U = 8994.5$, $p = .306$). These findings support the interpretation that graphical interfaces encourage richer query construction across user groups, while professionals are more adept at expressing Boolean logic in form-based systems.

Also visible in these results is a clear contrast between the two groups in the number and magnitude of their query reformulations. It can be seen that professionals make a greater number of reformulations (3.36 vs 2.76) with more substantial edits (mean Levenshtein distance of 50.85 vs 34.25) when using the form-based interface. By contrast, the students group does the opposite: they make a greater number of reformulations (5.40 vs 3.17) with more substantial edits (38.44 vs 28.92) when using the graphical interface.

To assess differences in query reformulation frequency, we conducted Mann–Whitney U tests. No significant differences were found within cohorts when comparing the graphical and form-based interfaces ($p = .357$ for professionals, $p = .111$ for students). However, when comparing cohorts, students made significantly more reformulations than professionals in the graphical interface ($U = 255.5$, $p = .043$), while no significant difference was observed in the form-based interface ($U = 440.5$, $p = .753$). These results suggest that the graphical interface encourages more active reformulation behaviour among students in particular.

To examine the magnitude of query reformulations, we analyzed Levenshtein distances using Mann–Whitney U tests. Professionals made significantly larger edits in the form-based interface compared to the graphical interface ($U = 3207.0$, $p = .008$), while students showed the opposite pattern, with significantly larger edits in the graphical interface ($U = 14228.5$, $p = .016$). Comparing across cohorts, there was no significant difference in Levenshtein distance within the graphical interface ($U = 7624.5$, $p = .341$), but professionals made significantly larger edits in the form-based interface than students ($U = 8090.0$, $p < .001$). These findings suggest that professionals engage in fewer but more substantial edits when using structured interfaces, while students tend to iterate more extensively in visual environments.

Reformulation behaviour

In total, there were 115 search sessions (each participant completing one search task is considered one search session), consisting of 57 sessions using the form-based interface and 58 using the graphical interface. Table 3 provides a summary of the coding results by cohort. There were 476 reformulation codes used in total, with 152 observed from the professionals and 324 from the students. The two most commonly observed types were SPE (42.44%) and GEN (31.72%).

Cohorts	SPE	GEN	PAR	SYN	FOR	INAPP	Total
Professionals	77 (50.66%)*	50 (32.89%)*	6 (3.95%)	7 (4.61%)	6 (3.95%)	6 (3.95%)***	152 (100%)
Students	125 (38.58%)*	102 (31.17%)*	14 (4.32%)	24 (7.10%)	13 (4.01%)	45 (13.89%)***	324 (100%)
Total	202 (42.44%)	151 (31.72%)	20 (4.20%)	30 (6.30%)	19 (3.89%)	51 (10.71%)	476 (100%)

Table 3. Query reformulation coding results, by cohort (total count of codes (percentages)). Significance measured by chi square: *<.05; **<.01; ***<.001.

Comparing the two cohorts in Table 3 shows that the professionals have a significantly higher use of SPE (50.66% vs 38.58%) and GEN (32.89% vs 31.17%). Students are marginally higher on PAR (4.32% vs 3.95%), SYN (7.10% vs 4.61%), and FOR (4.01% vs 3.95%), while being significantly higher than the professionals on INAPP (13.89% vs 3.95%).

Table 4 provides a summary of the coding results by interface. There were 476 reformulations in total, with 195 observed using the form-based interface and 281 using the graphical. The graphical interface is associated with significantly higher usage of SPE 43.77% vs 40.51%) and GEN (34.52% vs. 27.69%). The graphical interface is marginally higher for FOR (4.27% vs. 3.59%) and INAPP (11.39% vs. 9.74%), whereas the form-based interface is marginally higher on PAR (6.15% vs. 2.85%) and significantly higher on SYN (11.28% vs. 2.85%).

Cohorts	SPE	GEN	PAR	SYN	FOR	INAPP	Total
Form-based	79 (40.51%)**	55 (27.69%)**	12 (6.15%)	23 (11.28%)**	7 (3.59%)	19 (9.74%)	195 (100%)
Graphical	123 (43.77%)**	97 (34.52%)**	8 (2.85%)	8 (2.85%)**	12 (4.27%)	32 (11.39%)	281 (100%)
Total	202 (42.44%)	151 (31.72%)	20 (4.20%)	30 (6.30%)	19 (3.99%)	51 (10.71%)	476 (100%)

Table 4. Query reformulation coding results, by interface (total count of codes (percentages)). Significance measured by chi square: *<.05; **<.01; ***<.001.

Table 5 lists the frequencies of query reformulation types in a session and the number of sessions with that number of query reformulations. As in other studies (Hu et al., 2013; Jansen, Spink, &

Pedersen, 2005), most sessions were not long. Almost half of the observed sessions had three or fewer reformulation actions (49.52%). The mean number of reformulations per session was around 3.8, which is slightly higher than that of Lu et al. (2017) due to the longer tail in the distribution. This figure is higher for the graphical interface than the form-based interface (4.27 vs 3.32).

Frequency	Form-based	Graphical	Total
0	11 (20.75%)	10 (19.23%)	21 (20.0%)
1	7 (13.21%)	7 (13.46%)	14 (13.33%)
2	9 (16.98%)	8 (15.38%)	17 (16.19%)
3	4 (7.55%)	8 (15.38%)	12 (11.43%)
4	4 (7.55%)	5 (9.62%)	9 (8.57%)
5	4 (7.55%)	2 (3.85%)	6 (5.71%)
6	4 (7.55%)	3 5.77%)	7 (6.57%)
7	6 (11.32%)	4 (7.69%)	10 (9.52%)
8	1 (1.89%)	0 (0%)	1 (0.95%)
9	1 (1.89%)	1 (1.92%)	2 (1.90%)
10	2 (3.77%)	3 (5.77%)	5 (4.76%)
13	0 (0%)	1 (1.92%)	1 (1.90%)
17	0 (0%)	2 (3.85%)	2 (1.9%)
21	0 (0%)	1 (1.92%)	1 (0.95%)
Total	53	52	105

Table 5. Frequencies of query reformulations in search sessions.

Analysing the combinations of a particular cohort with a particular interface gives a further insight into query formulation tactics. Table 6 shows the reformulation tactics broken down by cohort and interface.

Cohorts + interfaces	SPE	GEN	PAR	SYN	FOR	INAPP	Total
Form+pro	35 (47.95%)	27 (36.99%)	2 (2.74%)*	5 (6.85%)*	4 (5.48%)	0 (0.00%)***	73 (100%)
Form+stu	44 (36.07%)	27 (22.13%)	10 (8.20%)*	17 (13.93%)*	3 (2.46%)	19 (15.57%)***	122 (100%)
Vis+pro	42 (53.16%)	23 (29.11%)*	4 (5.06%)	2 (2.53%)	2 (2.53%)	6 (7.59%)	79 (100%)
Vis+stu	81 (40.10%)	74 (36.63%)*	4 (1.98%)	6 (2.97%)	10 (4.95%)	26 (12.87%)	202 (100%)
Total	202 (42.44%)	151 (31.72%)	20 (4.20%)	30 (6.30%)	19 (3.99%)	51 (10.71%)	476 (100%)

Table 6. Frequency of use of query reformulation tactics., by cohorts and interfaces (total count of codes (percentages)). Significance measured by chi square: *<.05; **<.01; ***<.001.

No significant differences were found between the two cohorts for SPE across the two interfaces. Students used significantly more GEN when using the graphical interface, although no significant differences were found for GEN in the form-based interface. In particular, students used the generalisation strategy ito reduce the number of facets (Example 1) or increase the number of synonyms within a facet (Example 2).

Example 1:

"ehr mediated communication" AND **patients AND professions** ->

"ehr mediated communication"

Example 2:

(electronic OR computer OR digital OR electronics OR online OR systems) AND ("health record OR information) AND ("Professional communication" OR "academic communication" OR "specific communication" OR context) ->

(electronic OR computer OR digital OR electronics OR online OR systems) AND ("health record OR information **OR records**) AND ("Professional communication" OR "academic communication" OR "specific communication" OR context **OR professional OR interpersonal**)

Another point that appears from table 5 is INAPP, which is higher for students in both interfaces, with a statistically significant difference for the form-based interface. Example 3 illustrates INAPP in the graphical interface, where different concepts are combined in the same facet, while Example 4 reflects INAPP in the form-based interface, where the structure of the query does not follow Boolean logic.

Example 3:

(diagnosis OR cancer OR online OR "online information") AND **(how OR approach OR information OR seeking)**

Example 4:

Patients (searching) **OR** (online information) AND (Cancer Diagnosis)

Query quality

In this section we evaluate query quality by measuring the alignment with a 'gold standard' set of expert search strategies. As described earlier, the alignment was measured by calculating the overlap at the term level between participant queries and the gold standard. Table 7 shows this overlap, calculated in terms of precision and recall for both interfaces and cohorts.

Cohorts	Precision	Recall	Precision (graphical)	Recall (graphical)	Precision (form-based)	Recall (form-based)
Professionals	0.71	0.10	0.69	0.12	0.73	0.08
Students	0.57	0.10	0.50	0.11	0.68	0.08
Overall	0.62	0.10	0.56	0.12	0.70	0.08

Table 7. Query quality as measured by overlap with the gold standard strategies by cohorts and interfaces. N=525.

Comparing the two cohorts, it can be seen that precision is greater for the professionals than the students (0.71 vs 0.57) but recall is equal in both cases (0.1). This is not unexpected, given the difference in training and expertise. Comparing the two interfaces, we see that precision is higher in the form-based interface (0.70 vs 0.56), but recall is lower (0.08 vs 0.12). This may reflect the greater number of terms entered using the graphical interface, which has the effect of increasing recall at the expense of precision.

Combining precision and recall gives us the F-measure (an overall measure of performance), which is shown in Table 8.

Cohorts	F	F (graphical)	F (form-based)
Professionals	0.18	0.21	0.15
Students	0.17	0.18	0.14
Overall	0.17	0.19	0.14

Table 8. Query quality as measured by F-measure by cohorts and interfaces. N=525.

It can be seen that overall, the graphical interface returns the higher F-measure (0.19 vs 0.14). This effect is present for both cohorts, and somewhat surprisingly the contrast is more apparent in the professional cohort.

Discussion

The results of this study highlight the significant impact that interface design can have on query reformulation behaviors and overall search performance, particularly across cohorts with different expertise levels. Consistent with prior research, our findings reveal that professionals and students approach search tasks differently, with professionals demonstrating more efficiency and precision in query formulation, especially when using form-based interfaces. This aligns with studies such as Liu and Wacholder (2017) and Yoo and Mosa (2015), which found that more experienced users typically perform better in terms of both precision and efficiency, requiring fewer queries and making more effective reformulations.

Overall, students submitted more queries than professionals (302 vs. 203, averaging 21.5 queries per student and 14.5 per professional). This difference can be attributed to at least two factors: (a) professionals, with their advanced search training, may require fewer queries to reach satisfactory results, or (b) students may be more engaged with the interfaces, spending more time refining and iterating on their searches.

The **form-based interface** supported more targeted, precise search behaviors, with professionals demonstrating greater efficiency and fewer reformulations. This is in line with the work of Elbedweihy et al. (2012), who found that experts prefer structured search systems that guide them through the process. The more strategic nature of query construction in the form-based interface likely reflects the professionals' higher level of search expertise, which aligns with previous findings indicating that domain knowledge can significantly enhance search performance (Liu & Wacholder, 2017).

The **graphical interface** led to more frequent use of Boolean operators across both cohorts. However, students, in particular, engaged more with this interface, showing a greater number of reformulations and larger query edits (mean Levenshtein distance of 38.44 vs. 28.92 in the form-based interface). This increased activity in reformulation is consistent with the observations of Marchionini (2006), who argued that graphical interfaces support exploratory search behaviors by allowing users to experiment with multiple facets and terms. While this increased engagement can be beneficial in fostering a more comprehensive exploration of the search topic, it also introduces challenges in the form of **inappropriate keyword/structure (INAPP)**. As seen in the results, students were more prone to inappropriate keyword use in the form-based interface, highlighting the risk that novice users face when using conventional form-based systems. This is consistent with Dempsey and Valenti's (2016) findings, which noted that students often struggle with misusing search terms, underscoring the need for user training in navigating these systems effectively.

In terms of reformulation strategies, both **Specification (SPE)** and **Generalization (GEN)** were the most frequently employed strategies, which is consistent with previous studies (Hu et al., 2013) . These strategies are commonly used when users face difficulties in retrieving relevant information. However, it is notable that students, when using the graphical interface, showed a higher tendency

to generalize their queries (36.63% vs. 22.13% for form-based), which suggests that the graphical interface may encourage a more exploratory, broadening approach to query formulation. By contrast, professionals appeared more strategic, with a greater use of **Specification (SPE)** as they narrowed their focus to improve search relevance. This reflects the results of Osborne and Cox (2015), where novice users tended to broaden their queries more frequently, whereas experts showed a preference for narrowing their queries to enhance precision.

When comparing the **query quality** results (Section 4.3), we observed a typical trade-off between **precision** and **recall**. While **precision** was higher in the form-based interface (0.70 vs. 0.56), suggesting that the structured format supports more accurate searches, **recall** was higher for the graphical interface (0.12 vs. 0.08), reflecting the broader range of terms used. These findings are consistent with those of Jansen et al. (2009), who noted that more complex queries tend to improve recall but often sacrifice precision. This suggests that while the graphical interface facilitates more expansive searches, it also requires a greater effort to maintain focus and precision, especially for novice users.

While the gold-standard analysis involved a simple term-level comparison with expert queries, it does not imply that alternative terms or strategies used by participants were incorrect—only that they diverged from the expert baseline. The relatively low F-measure values observed (ranging from 0.14 to 0.21) should be understood in this context. These values reflect the complexity of the tasks, the strictness of the gold standard, and the natural variability in user strategies, particularly among less experienced searchers. Rather than indicating poor performance, they highlight the diversity of plausible search behaviours and the limitations of using a single "ideal" formulation for evaluation.

Overall, our findings emphasize the **importance of tailoring interfaces** to meet the needs of different user groups. Graphical interfaces can enhance recall and support novice users by encouraging exploration, but they also need to be designed with safeguards against inappropriate keyword use and overly broad queries. Meanwhile, form-based systems, though more restrictive, can provide the precision and structure that experts require to quickly and efficiently retrieve relevant information. The contrast between the two cohorts in terms of query behavior suggests that training and interface customization should be aligned with users' expertise levels.

Further research

Further research should focus on exploring users' intentions during query reformulation and how these intentions may vary across different types of search tasks. Qualitative studies could provide deeper insights into how users conceptualize their search process and how this aligns with or differs from the behaviors observed in this study. Additionally, future work could investigate adaptive interfaces that combine the strengths of both systems, offering users the flexibility of graphical interfaces without sacrificing the precision and structure afforded by form-based systems.

In their 2006 paper, Rieh and Xie suggested several features that IR systems should have to better support user reformulations, such as the ability to efficiently manage multiple queries. It is possible that the two interfaces examined in this study address some of these recommendations. However, Rieh and Xie (2006) also highlight the challenges users face when formulating queries and reformulating insufficient ones. While this study has focused on the characteristics of reformulations, it has not explored users' perceived intentions behind these actions. Future qualitative research should investigate this aspect in greater depth.

Summary and conclusions

This study provides new insights into how graphical and form-based search interfaces influence query reformulation behaviors across user cohorts with varying expertise levels. By analyzing

query construction, reformulation strategies, and alignment with gold standard search strategies, we observed distinct patterns in the search behaviors of professionals and students. Professionals demonstrated greater precision and efficiency in query construction, particularly with the form-based interface, while students benefited more from the affordances of the graphical interface, showcasing higher engagement and reformulation activity.

The findings reveal that the graphical interface prompted greater use of Boolean operators across both cohorts (5.43 on average compared to 1.98 for the form-based interface), suggesting its effectiveness in supporting complex query construction. Students exhibited more frequent and substantial reformulations in the graphical interface (mean Levenshtein distance of 38.44 vs. 28.92 in the form-based interface). Professionals, in contrast, made fewer but more focused adjustments with larger edits when using the form-based interface (50.85 vs. 34.25). These patterns emphasize the role of interface design in shaping user behavior, particularly among novices versus experts.

Reformulation strategies, as analyzed in section 4.2, further illustrate these differences. Specification (SPE) and generalization (GEN) were the most frequently observed strategies across both interfaces and cohorts, reflecting their centrality in refining queries. The graphical interface encouraged a slightly higher use of SPE (43.77% vs. 40.51% for the form-based interface) and GEN (34.52% vs. 27.69%), particularly among students, while professionals displayed more balanced reformulation behaviors across both systems. However, inappropriate keyword use (INAPP), which was most common among students, highlights the ongoing challenges associated with traditional form-based interfaces.

In terms of query quality (section 4.3), measured by precision, recall, and F-measure, the results highlight trade-offs inherent in interface design. Precision was higher overall for the form-based interface (0.70 vs. 0.56 for the graphical interface), reflecting its ability to support more targeted searches. However, recall was higher in the graphical interface (0.12 vs. 0.08), attributed to the broader range of terms entered. Combining these measures, the graphical interface demonstrated a higher F-measure overall (0.19 vs. 0.14), particularly for professionals, indicating its potential to balance precision and recall in complex search scenarios.

The study underscores the need for tailored interface designs and training programs that cater to diverse user needs. While graphical interfaces can support novice users by fostering exploration and recall, they also introduce challenges such as inappropriate keyword use, particularly among students. Form-based systems remain critical for expert searchers requiring precision and efficiency. Future research should explore qualitative aspects of user intentions during query reformulation and extend these findings to other domains and user populations. Additionally, further work could investigate adaptive interfaces that combine the strengths of both systems, offering flexibility for users with varying levels of expertise.