Collaborative search in electronic health records

Kai Zheng,[1,2] Qiaozhu Mei,[2,3] David A Hanauer[4,5]

► Additional materials are published online only. To view these files please visit the journal online (www.jamia.org).

[1]School of Public Health Department of Health Management and Policy, The University of Michigan, Ann Arbor, Michigan, USA
[2]School of Information, The University of Michigan, Ann Arbor, Michigan, USA
[3]Department of Electronic Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan, USA
[4]Department of Pediatrics, The University of Michigan, Ann Arbor, Michigan, USA
[5]Comprehensive Cancer Center, The University of Michigan, Ann Arbor, Michigan, USA

Correspondence to
Dr Kai Zheng, Information Systems and Health Informatics, School of Public Health Department of Health Management and Policy, School of Information, The University of Michigan, M3531 SPH II, 109 South Observatory Street, Ann Arbor, MI 48109-2029, USA; kzheng@umich.edu

ABSTRACT

Objective A full-text search engine can be a useful tool for augmenting the reuse value of unstructured narrative data stored in electronic health records (EHR). A prominent barrier to the effective utilization of such tools originates from users' lack of search expertise and/or medical-domain knowledge. To mitigate the issue, the authors experimented with a 'collaborative search' feature through a homegrown EHR search engine that allows users to preserve their search knowledge and share it with others. This feature was inspired by the success of many social information-foraging techniques used on the web that leverage users' collective wisdom to improve the quality and efficiency of information retrieval.

Design The authors conducted an empirical evaluation study over a 4-year period. The user sample consisted of 451 academic researchers, medical practitioners, and hospital administrators. The data were analyzed using a social-network analysis to delineate the structure of the user collaboration networks that mediated the diffusion of knowledge of search.

Results The users embraced the concept with considerable enthusiasm. About half of the EHR searches processed by the system (0.44 million) were based on stored search knowledge; 0.16 million utilized shared knowledge made available by other users. The social-network analysis results also suggest that the user-collaboration networks engendered by the collaborative search feature played an instrumental role in enabling the transfer of search knowledge across people and domains.

Conclusion Applying collaborative search, a social information-foraging technique popularly used on the web, may provide the potential to improve the quality and efficiency of information retrieval in healthcare.

INTRODUCTION

Clinicians' day-to-day interactions with electronic health records (EHR) generate rich, patient-level data that can be utilized for secondary-use purposes such as population health management, epidemic surveillance, and clinical, translational, and health services research.[1 2] As advocated by numerous experts and professional organizations, the value proposition of widespread adoption and meaningful use of EHRs resides not only in improving the quality of care and controlling costs, but also in creating a 'rapid learning' healthcare system that can advance our knowledge in a wide spectrum of clinical and policy domains.[3–6] This anticipated value will not be realized, however, if the data stored in EHRs cannot be effectively searched and retrieved.

While codified data entered through structured templates are generally more desirable, a significant amount of clinical documentation continues to exist in an unstructured, narrative format.[7] Despite the obvious weaknesses,[8] narrative documents offer many advantages that structured data inherently lack, including, for example, expressiveness that allows clinicians' thought processes to develop and rich patient stories to build.[9 10] Thus, it is not surprising that unstructured narrative documents remain pervasive even in highly wired healthcare facilities, and they may be increasingly used in adapted forms such as free-text instructions, comments, and memos accompanying codified data.[7 11–15]

Identifying effective ways to retrieve information from unstructured narrative documents is therefore imperative.[16] Unfortunately, making use of narrative data generated in day-to-day clinical settings is extraordinarily challenging, and costly, onerous, and error-prone manual chart review processes are often inevitable.[17] Recent research advances in natural language processing and other novel approaches such as 'structured narratives' have provided great promise for automatically extracting concepts from narrative documents or directly embedding computer-recognizable terms into them.[18–22] Nonetheless, before these new technologies become widely available and versatile enough to handle assorted, oftentimes vaguely defined information-retrieval needs, a convenient and cost-effective solution continues to be in great demand.

Similar to how Google has changed the way people search for information on the web, a full-text EHR search engine, supporting basic functions such as string matching to advanced functions such as regular expressions, can be an invaluable tool to help practitioners and researchers navigate through large quantities of narrative data stored in EHRs.[23] A full-text search engine does not solve all information-retrieval problems, however: its performance is critically dependent upon the quality of search queries that users are able to construct. Unfortunately, average users often do not have adequate knowledge to construct effective and inclusive search queries especially when the topic of interest is novel, or the subject domain is complex, such as healthcare.[24–26]

To address the issue, information-retrieval tools on the web are increasingly adopting a 'social information-foraging' concept which encourages users to collaboratively refine the quality of search queries as well as the quality of information resources, for example, del.icio.us, which allows users to collectively bookmark web documents that then can be tagged, annotated, and searched (how many times a webpage has been bookmarked by different users in itself serves as an indication of the quality and 'interestingness' of the page), Yahoo! Search Pad that allows users to develop search queries cooperatively, and Google's search-term

recommendation service that suggests alternative search terms based on how the relevant topic has been searched in the past by other users.[27–29] Leveraging the population's collective wisdom, these social information-foraging tools have not only improved search quality and efficiency but also made finding information on the web an engaging and rewarding social experience.[29]

Can this concept be applied to facilitate information retrieval in EHRs? Through a homegrown full-text EHR search engine, we experimented with a 'collaborative search' feature that allows users to preserve their search knowledge and share it with others. The objective of this feature was to nurture a cooperative and participatory culture in the user community so that search queries could be socially formulated and refined, and search expertise could be preserved and diffused across people and domains. Based on the computer-recorded usage data collected over a 4-year period, we conducted an in situ evaluation to assess whether this design objective had been achieved. We also applied a social-network analysis to delineate the structure of the user-collaboration networks engendered by the feature so as to quantify its utility in enabling the diffusion of knowledge of search.

BACKGROUND
EMERSE: the EHR full-text search engine
Because full-text search functionality is largely missing from commercially sold EHR systems, we built one, and successfully integrated it with our institutional EHR environment and the Computerized Patient Record System at the Ann Arbor Veterans Affairs hospital.[30–32] At both institutions, the search engine, named the Electronic Medical Record Search Engine (EMERSE), has been routinely utilized by clinicians, medical coding personnel, quality officers, and researchers to support their chart abstraction tasks that would be otherwise difficult or even impossible.[33–38]

EMERSE provides a full-text search capability analogous to that of Google, in addition to features specifically designed to handle the challenges unique to retrieving information from unstructured medical data. For example, it provides an 'alternative search query recommendation function' based on customized medical dictionaries and open-source phonetic matching algorithms.[39] This function detects and suggests correction of common forms of spelling mistakes or non-standard use of medical terminologies, acronyms, and abbreviations—either contained in the search queries that the user submits or appearing in the narrative documents that are being searched.

Prior evaluation studies have demonstrated that the use of EMERSE, as compared with manual reviews, can help achieve significantly improved sensitivity, specificity, and efficiency in various types of chart abstraction tasks.[31 32] Nonetheless, our inspection of the log data recorded in the system indicated several issues severely undermining its performance, including (1) variable quality of search queries submitted by the users and (2) a considerable amount of redundant effort in repeatedly constructing the same or substantially similar queries by different users.[40] These deficiencies learned from the field prompted us to look for novel approaches to further improve the system's usefulness and usability. The collaborative search feature studied in this paper is one of them.

Collaborative search concept
Sophisticated page-ranking algorithms employed by web search engines such as Google have greatly improved the relevance of documents retrieved. However, what if the search query does not accurately reflect the user's search intention in the first place, due to, for example, the user's lack of search expertise or domain knowledge?

A significant and successful stream of work attempting to mitigate this problem is based on the social information-foraging paradigm, wherein all users collectively contribute to an evolving body of wisdom of search-query construction, results refinement, and knowledge discovery.[28 29] Such machine-mediated, user-driven cooperation may take place (1) passively through automated services such as search-query recommendation (eg, based on an analysis of query logs and clickthrough data),[41–45] (2) or proactively by allowing users to collaborate through tools such as Yahoo! Search Pad,[27] Microsoft SearchTogether,[46] and dedicated social search websites (eg, Yoople! http://www.yoople.net/ and Eurekster Swicki http://www.eurekster.com/). This concept has also been increasingly applied in healthcare; for example, My NCBI developed by the US National Library of Medicine by which users can preserve their PubMed/MEDLINE search queries or results as 'Saved Searches' or 'Collections,'[47] and search 'hedges' or 'filters' developed or enlisted by various initiatives to facilitate information retrieval in biomedical literature.[48 49]

These socially oriented approaches greatly help capture of search expertise, which is usually possessed by only a few experts, and diffuse it widely for the benefit of everybody in the user community. Further, they provide an opportunity to solve complex information-retrieval problems that may be beyond the ability of any individual users (eg, to classify a massive collection of digital image files), for which the community's collective intelligence is often needed.

Implementation of collaborative search in EMERSE
Inspired by the success of the social information-foraging techniques popularly used on the web, we implemented a similar feature in the EMERSE system, referred to as 'collaborative search.' Central to the feature is a concept called 'search-terms bundles,' which are created by end users to hold collections of concepts which may require multiple keywords or complex regular expression formulas for proper identification in unstructured narrative documents stored in EHRs (an illustration is provided in figure A in appendix 1 of the online supplemental data).

'Cancer Staging Terms,' for example, is a popular bundle consisting of 202 distinct search terms. It provides an enumeration of words and phrases commonly used by clinicians in their clinical documentation to describe 'cancer staging,' such as 'gleason,' 'staging workup,' 'restaging,' 'microstaging,' and 'Tmic.' It also contains regular expression logics that allow for blocking out false-positive phrases so that the search query can, for example, highlight the mention of a T2 cancer stage while discarding the mention of T2 MRI weighting. 'Myocardial Infarction,' another commonly used search-terms bundle, is particularly useful at our institution, the University of Michigan. This bundle was created to accommodate the special meaning of 'MI' in our local context; for example, a regular expression contained in the bundle {'myocardial infarction,' -$'MI\s*\d{5},' -$'MI\s*,\s*\d{5},' ~MI} instructs the search engine to look for 'myocardial infarction,' or its abbreviation 'MI,' while ignoring the 'MI, xxxxx' or 'MI xxxxx' combinations.

The search-terms bundles that the users have deposited in the system contain 20 distinct terms on average; the most complex one is composed of 370 distinct terms. Constructing such complex queries requires not only highly adept information-

retrieval skills but also a strong specialty background in the medical domain of interest. To achieve optimal results, it may also require multiple iterations in exploring various combinations of search terms and making sense of the documents returned—as well as those that are not returned. A function built right into the EHR search engine to help preserve this precious user effort and make the resulting knowledge widely available in the user community would therefore be of great value.

In EMERSE, the users have the option to convert a search query into a retrievable and reusable search-terms bundle at any time. The owner of a bundle, referred to hereafter in this paper as 'Creator,' has two options to share it with other people (referred to as 'Consumer'): (1) through a designated list of bundle assignees ('Private Bundles,' see figure B in appendix 1 for an illustration) or (2) listing it in a public bundle registry which then becomes available to all search engine users ('Public Bundles'). When this study was conducted, the bundle creators must choose one sharing mode or the other; simultaneously registering a bundle as both 'private' and 'public' was not possible.

The bundles available to a user are accessible on the search screen next to the search box. They are alphabetically ordered and can be sorted alternatively by modification date or user-name of the bundle creators (see figure C in appendix 1). No formal classification methods were provided to organize the bundles when this study was conducted. After a bundle is selected, the search terms it contains will be appended to the query. The user can then modify it further to meet the need of the particular information retrieval task at hand (see figure D—F in appendix 1, each illustrating the 'Patient Summary' view, 'Notes Summary' view, and 'Document Detail' view during a typical EHR search).

METHODS
Through an empirical study, we studied how users might react to the collaborative search feature provided in EMERSE. In particular, we applied a social-network analysis to delineate the patterns of collaborative construction and shared use of search-terms bundles. The results helped us assess the user acceptance of the feature, which also alluded to the potential value of the feature in facilitating dissemination of search knowledge across people and domains.

Empirical study setting
The empirical study was conducted at the University of Michigan Health System, a 930-bed quaternary academic medical center that has over 40 000 inpatient admissions and 1.5 million ambulatory visits annually. EMERSE was integrated with the University of Michigan Health System's institutional EHR environment that supports both its inpatient services and ambulatory clinics and affiliated health centers. As of December 2009, more than 20 million unstructured or semistructured clinical narratives had been stored in the EHR's core data repository, with approximately 3 million new ones added each year. All of these documents are searchable within EMERSE.

In this paper, we analyzed the computer-recorded usage data collected over a period of 4 years: December 16, 2005 to December 16, 2009. During this period, a total of 451 registered users actively used the system to retrieve EHR data. The majority of them, according to the primary appointment information provided in their user registration, were academic researchers (62.7%) and practicing clinicians (21.6%). The

remainder consisted of medical coding personnel (5.6%), IT staff (6.2%), and QA managers and patient-safety officers (3.9%).

To comprehensively capture research data for study, we built into EMERSE a special logging mechanism that records the user interactions with the system at a very fine level of detail, such as each of the steps during the course of search terms revision/expansion that led to the final query submitted. A considerable portion of this information would not be available to us otherwise. The Medical School Institutional Review Board at the University of Michigan reviewed and approved the research protocol of this study.

Data-analysis methods
First, we examined general information-retrieval behavior of the search-engine users. Then, we applied a social-network analysis to examine the structural properties of the bundle-sharing networks engendered by the collaborative search feature. It was through these networks that the knowledge of EHR search became diffused across individuals and across the boundaries between academic departments ('Department'), medical specialties ('Specialty'), and administrative divisions ('Division').

'Departments' are principal organizational units at our institution. A 'Division' is a medical 'Specialty' group administratively homed within a 'Department.' For example, 'Internal Medicine' and 'Pediatrics' are academic departments, 'Oncology' is a specialty, and 'Internal Medicine/Oncology' and 'Pediatrics/Oncology' are two distinct administrative divisions. Note that some self-contained organizational units such as 'Clinical Trial Office' and 'Health Information Management' are also classified at the academic department level, even though they are technically not academic departments. Appendix 2 of the online supplementary data provides a full list of all organizational entities studied in this paper.

Search-knowledge diffusion networks
Social-network analysis, mathematically underpinned in graph theory, provides an ideal approach for delineating the interactions among the search-engine users (or the organizational entities with which they are affiliated) through their collaborative search activities. Based on the empirical data, we constructed five different bundle-sharing networks, which we call 'search-knowledge diffusion networks' (SKDNs):

► Network 1: Creator—Consumer Network, in which network nodes represent search engine users, and an edge (directed) connects the creator of a bundle to the consumer(s) of the bundle.
► Networks 2, 3, and 4: Organizational Entity Networks at the department, specialty, and division level, respectively. In the Department—Department Network, for example, network nodes represent academic departments, and an edge (directed) connects the department of a bundle's creator to the department(s) of the consumer(s) of the bundle.
► Network 5: Consumer—Consumer Network, a second-order, derived network in which network nodes represent search-engine users, and edges (undirected) join those who had utilized the same search-terms bundle(s) in their EHR search.

The last network conveys an invisible type of relationships that may not be explicitly known to the connected parties. We included this network in our analysis because such relationships disclose information that could be potentially very useful: the overlapped bundle usage of these users may suggest that they have similar EHR search objectives in common; this information could then be utilized, for example, to stimulate offline

collaboration among these users which may lead to unexpected synergic effects in patient-care provision or in research.

For each of the SKDNs, we separately analyzed the network segment based on private bundles and that based on public bundles. This segmentation allowed us to compare the utility of these two different bundle-sharing mechanisms in facilitating knowledge diffusion. Intuitively, privately shared bundles are only available to a limited number of users, but their designated nature warrants higher rates of usage, whereas public bundles, while they may be utilized less frequently, help disseminate knowledge widely to benefit more search-engine users.

Network measures

Table 1 summarizes the network measures assessed in this study, which are key descriptors of the efficacy of an SKDN in mediating information or knowledge transfer. The fraction of singletons measure presented in the first row, for example, suggests the level of user participation in collaborative search: a smaller value indicates that fewer users were left out of the SKDN, so the network had been more effective in spreading out search knowledge to more users in the community.

The last measure, modularity, delineates the partitioning nature of a network, that is, whether there exist distinct subcommunities formed by nodes that have certain characteristics in common (eg, representing users affiliated with the same academic department). Such subcommunities, or 'social cliques,' are characterized as having more intense connections internally than with the rest of the network.[51 52] In this study, we used the modularity measure to assess whether the partitioning of an SKDN may be reflective of the search-engine users' real-world identities, that is, whether or not the knowledge-sharing activities predominantly took place within the boundaries of the academic departments, medical specialties, or administrative divisions.

Analysis of potential gains

We further constructed a hypothetical network that extends the existing Consumer—Consumer Network, by including additional nodes/edges based on similar search queries constructed ad hoc by different users. The objective was to assess the magnitude of redundant search-query construction effort that could have been avoided if these repeatedly appearing and manually entered queries were preserved as shared knowledge. To simplify the analysis, only the queries that contained exactly the same search terms were considered; the search terms, however, may appear in different orders. Therefore, the results of this analysis do not reflect the maximum gains possible, considering there might be many more queries that were not exactly the same but were substantially similar.

All analyses reported in this paper were programmed in Perl using Clairlib v1.08, an open-source library for supporting natural-language processing, information retrieval, and network analyses (http://www.clairlib.org/). The raw data and the Perl scripts can be downloaded from an online analytical processing (OLAP) tool that we developed for this study: http://sitemaker.umich.edu/emerselog/.

RESULTS
General usage

Descriptive statistics of the empirical dataset are reported in table 2. During the 4-year period, the search engine performed nearly a million searches that involved the processing of medical records belonging to over 20 000 distinct patients. (Note that to relieve server load, up to 100 patients can be searched at one time; the same query submitted repetitively to search among different patients were counted separately toward the measure of total number of searches.) About half the searches (444 784) were facilitated by bundled search terms stored in the system. Of these bundle-based searches, 156 971 (35.8%) utilized shared search knowledge made available by others; these included 41 170 searches based on public bundles (26.2%) and 115 801 based on private bundles (73.8%).

Search-terms bundles

As of December 16, 2009, a total of 702 search-terms bundles had been created in the system. More than half (385) were made

Table 1 Summary of the network measures assessed

Measure	Definition	Empirical interpretation	Desirable value*
Fraction of singletons	Proportion of network nodes that are not connected with any other nodes	Larger value indicates more isolated notes, that is, a larger proportion of the search engine users had never participated in collaborative search	⇓
Average degree†	Average number of connections that each network node has	Larger value indicates a more densely connected network, that is, more knowledge sharing activities took place among the users	⇓
Average tie strength	Average number of bundles shared between connected nodes	Larger value indicates that more bundles were being shared among the users who were connected	⇓
Global clustering coefficient[50] †	Degree to which network nodes tend to cluster together, that is, the likelihood that the nodes connected with A's neighboring nodes are also connected with A	Larger value indicates higher network transitivity ('my friends' friends are also my friends'), suggesting that the network had effectively tied together those users who might have shared objectives/interests in common	⇓
Average shortest path length[50] †	Average number of 'hops' needed for a network node to reach another	Smaller value indicates a more tightly connected network, which could help shared search knowledge spread out to more people and more quickly	⇓
Modularity[51 52] †	A descriptor suggesting whether nodes with certain characteristics in common tend to form subcommunities within a network ('social cliques'); larger value indicates denser intracommunity connections and sparser intercommunity connections; see main text for a more elaborate description		Does not apply

*'Desirable value' indicates whether the value of the measure should be larger or smaller in order to achieve better results in mediating search knowledge diffusion.
†Isolated notes were removed from the network before these measures were computed.

Table 2 Descriptive statistics

Measure	Results
Total number of searches conducted	904 215
Distinct patients searched	214 853
Searches utilizing search-terms bundles	438 421
Searches utilizing bundles shared by others	156 971
Based on public bundles	411 70
Based on private bundles	115 801
Total number of active users*	451
Active bundle users†	265
Bundle creators	188
Bundle sharers	91
Bundle leechers‡	77

*Registered users who conducted a least one search using the system.
†Those who participated in creating, sharing, using, or refining the search-terms bundles.
‡Those who used bundles shared by others while not contributing any of their own.

available to other users through either private sharing (241, 34.3% of all bundles) or public sharing (144, 20.5% of all bundles). Table 3 lists the 10 most often used search-terms bundles and their usage statistics. A comprehensive list of all bundles available in the system is provided in the OLAP tool.

The 702 search-terms bundles available in the system were contributed by a total of 188 bundle creators (41.7% of all registered users or 70.9% of the active bundle users). About half of the bundle creators (91) used the collaborative search feature to share their search knowledge with other users.

Nonetheless, in the search-engine user community, 77 users appeared to be 'bundle leechers'—who utilized others' bundles while not contributing any of their own. Because this free riding behavior could be detrimental to the health of the community, which is entirely based upon voluntary contributions, we specifically evaluated the magnitude of this behavior. The results are reported as two scatter plots (figure 1).

Figure 1A compares the number of bundles own by a user (x axis) with the number of bundles shared by the user (y axis). This comparison tells roughly the bundle creators' willingness to share the knowledge they had created. In figure 1B, we plot the number of others' bundles consumed by a user (x axis) in contrast to the number of bundles that the user created and made available to others (y axis), that is, the consumption/ contribution ratio. Dashed (red) lines in both graphs represent linear regression lines.

The information conveyed in figure 1 suggests that despite a considerable number of users who benefited more from the community than what they contributed, there existed a few enthusiastic users who created and shared many bundles to help

to sustain the community. Note that this bundle-sharing willingness measure should be interpreted relatively within the context of the empirical study, that is, there could be many legitimate reasons for a bundle creator to choose not to share a bundle; for example, a bundle that was being constructed and not yet ready to distribute for production use.

Patterns of bundle creation, sharing, and use among the organizational entities

Table 4 reports collaborative search activities at the three organizational levels. Shown in the table are the five most active organizational entities adjusted for unit size (ie, ranked by bundle creation ratio calculated as the number of bundles created by an organizational entity divided by the total number of active users affiliated with the entity); organizational units that had fewer than 10 active users registered in the system were not included.

As shown in the 'Department' portion in table 4, Pediatrics, Clinical Trials Office, Internal Medicine, Psychiatry, and Quality Control were among the most active organizational entities in bundle creation. Together, they contributed nearly 90% of the search-terms bundles available in the EMERSE system. At the other two levels, General Pediatrics, particularly the General Pediatrics Division in the Department of Pediatrics, ranked top in bundle creation. A complete list reporting the participation levels of all departments, specialties, and divisions is provided in appendix 3 of the online supplemental data.

Network plots

In this paper, we use two network plots to illustrate the SKDNs engendered by the collaborative search feature (figure 2A,B). Both graphs were produced using GUESS v0.5-α, an open-source graph-exploration system (http://graphexploration.cond.org). Raw data and scripts for generating the visual representation of the other SKDNs can be found in the OLAP tool.

Figure 2A depicts the Department—Department Network where the network nodes represent academic departments (circles) or search-terms bundles (dots). A gray edge connects a bundle to the department of its creator, and red edges connect a bundle to the department(s) of its consumer(s). The bundle-sharing activities internal to a department were not depicted in the graph. Further, the departments that did not participate in a collaborative search were not plotted. As shown in figure 2A, the transdepartmental bundle-sharing activities served as an important means for relaying the knowledge of search from one department to another. However, the potential of this feature was far from being fully realized, as indicated by a larger number of bundles that were used only internally.

Table 3 Top 10 most frequently used search-terms bundles

Bundle name	Description and number of search terms contained	No of searches used in	Searches based on bundles shared by others	
			Public	Private
Psych Combined	Combined inclusion and exclusion criteria (91)	82 070	0	80 306
Staged Ulcer	Search for wound assessments with staged ulcers (2)	55 252	0	0
Deep Tissue Injuries	Search for wound assessments with deep tissue injuries (2)	40 986	0	0
Surgical site infection	Infection control and epidemiology bundle for surveillance (54)	181 60	192	0
Deep Tissue or Staged Ulcers	Deep tissue or staged ulcers (3)	14 304	0	0
Hemorrhagic Lesion	Search for wound assessments with hemorrhagic lesions (2)	14 131	0	0
Pressure Ulcer	Search for wound assessments with pressure ulcers (2)	12 469	0	0
Height and Weight	A comprehensive list of terms to identify height and weight parameters (152)	11 361	11 329	0
Down Syndrome	Down's syndrome (6)	9 336	0	0
Portal Vein Thrombosis 2	To search for patients with portal vein thrombosis (8)	5685	5685	0

Figure 1 User participation in bundle creation, sharing, and using: (A) number of bundles created versus number of bundles shared; and (B) number of others' bundles used versus number of bundles shared.
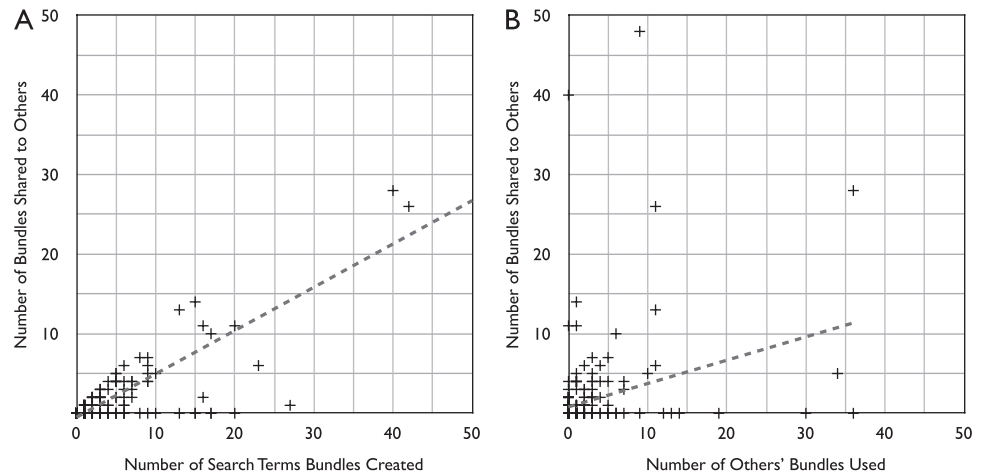

Figure 2B depicts the Consumer—Consumer Network. The network nodes represent bundle consumers, which were distributed in the graph using the Kamada—Kawai layout algorithm to allow each of the distinct network components to stand out.[53] In figure 2B, red edges connect user groups where in all members had used at least two identical bundles, and the width of an edge is proportional to the number of bundles used in common by the two users connected. A salient pattern can be observed based on figure 2B: there exist numerous subcommunities of search engine users who are closely bound together because of their overlapped bundle usage. Such overlapped usage may suggest that these users have similar EHR search objectives, a possible indication of their shared patient care provision or research interests (eg, to identify patient cohorts of similar characteristics). Such relationships, however, are usually not explicitly known to the connected parties and, if made known, could stimulate synergic effects in real-world collaboration beyond the scope of searching for information in EHRs.

Structural property assessments

Network measures assessing the structural properties of each of the SKDNs are reported in tables 5—7. The numbers shown in parentheses are expected values obtained from randomly constructed networks. Each randomly constructed network consists of the same number of nodes and edges as the network of interest, but the edges are randomly placed using the Erdös—Rényi algorithm.[54] The purpose was to derive a baseline in order to determine whether the user collaboration engendered by the empirical stimuli (ie, the collaborative search feature provided through the EHR search engine) demonstrates distinctive attributes significantly different from those of networks formed at random.

The first data row in table 5, the fraction of singletons measure, suggests the overall level of user participation in collaborative search. About a quarter of the search-engine user population participated in knowledge-sharing activities based on private bundles (24.8%) or based on pubic bundles (26.8%). With the two sharing modes combined, 41.9% of the user population participated. Note that this ratio is lower than the ratio of active bundle users as a proportion of all active users of EMERSE (58.7%): a few bundle creators never shared their knowledge with others; they were therefore not participating members of the knowledge-sharing networks.

As shown in tables 5—7, the fraction of singletons of the networks based on private bundles is consistently larger than or

Table 4 Participation in collaborative search at the three organizational levels*

	Entity name	Registered users	Bundles created	Bundles shared internally	Bundles shared privately with other organizational entities	Bundles shared publicly	Searches based on internally shared bundles	Searches based on private bundles shared by other organizational entities	Searches based on public bundles shared by other organizational entities
Department	Pediatrics	53	206	97	53	56	8857	51	2501
	Clinical Trials Office	57	153	120	18	15	8834	1097	1529
	Internal Medicine	122	270	211	26	33	66 097	1724	17 771
	Psychiatry	16	35	33	0	2	102 005	8	6
	Quality Control	42	87	68	14	5	171 908	1763	284
Specialty	General Pediatrics	12	159	57	54	48	2504	60	10
	Quality Improvement (QI)	11	40	31	8	1	5526	9354	23
	Gastroenterology	16	54	37	10	7	8737	1	10 264
	General Internal Medicine	23	73	55	4	14	20 988	1475	19
	Rheumatology	17	48	44	2	2	7554	2	0
Division	Pediatrics/General Pediatrics	12	159	57	54	48	2504	60	10
	Quality Control/QI	11	40	31	8	1	5526	9354	23
	Internal Medicine/General	23	73	55	4	14	20 988	1475	19
	Internal Medicine/Oncology	26	76	50	20	6	18 532	32	2227
	Internal Medicine/Rheumatology	17	48	44	2	2	7554	2	0

*Organizational entities are ranked by their bundle-creation ratio, that is, the number of bundles created by an entity divided by the number of active users affiliated with the entity.
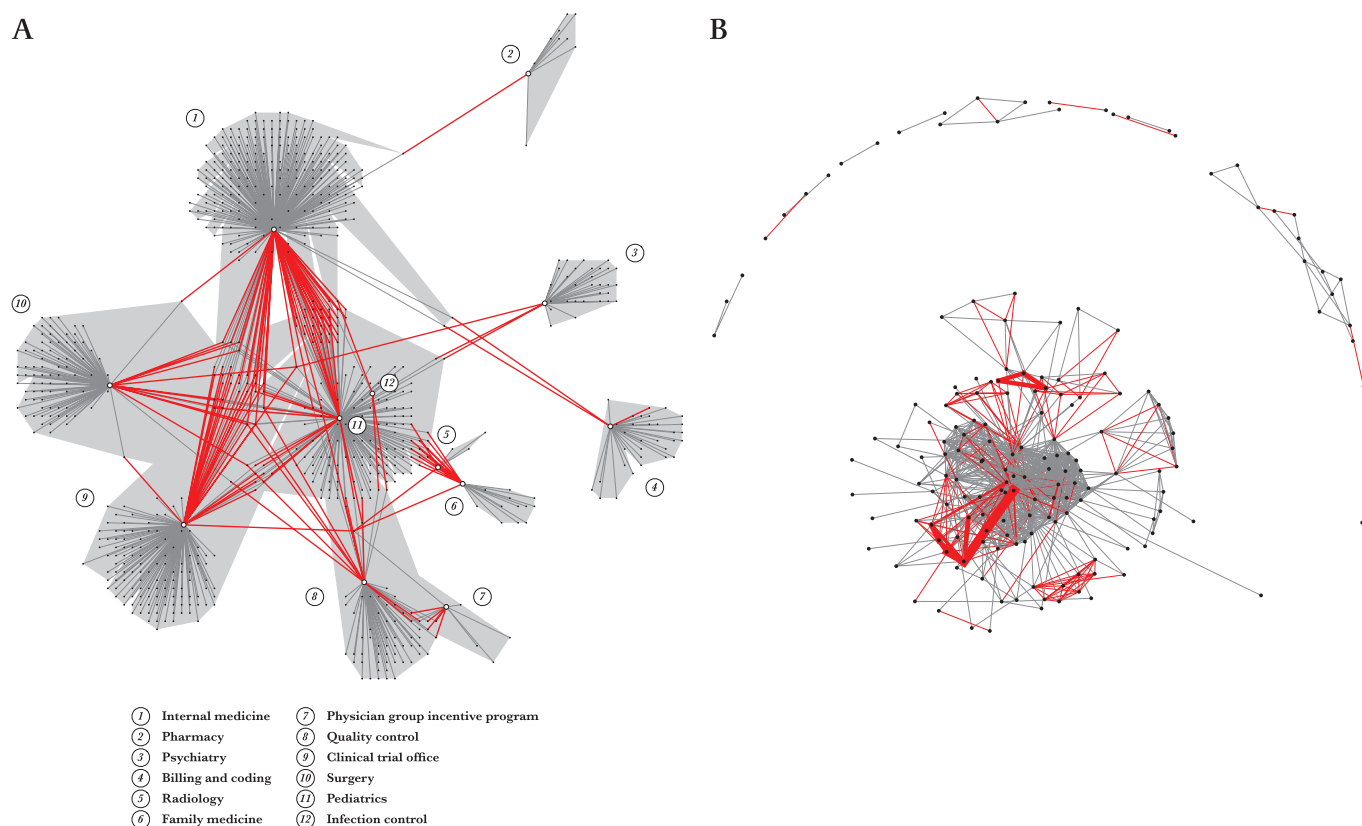
Figure 2   Network plots. (A) Department—Department Network. Circles, academic departments; dots, search-terms bundles; gray areas, zones encompassing all bundles created by users from the same department; edges (gray), connecting a bundle to the department of its creator; edges (red), connecting a bundle to the department(s) with which its consumer(s) are affiliated. (B) Consumer—Consumer Network. Dots, search engine users; Edges (red), connecting user groups wherein all members had used at least two same bundles in common. The width of an edge is proportional to the number of bundles used in common by the two users connected.

equal to that of the corresponding public-bundle-sharing networks. Similarly, the average degree of the private-bundle-sharing networks is consistently lower than that of the networks based on public bundles. For the average tie strength measure, the results of this private versus public comparison are consistently reversed. These findings confirm the hypothesis that private-bundle sharing warrants higher rates of usage at the cost of smaller numbers of beneficiaries, and while publicly available bundles help disseminate the knowledge of search more widely, they are utilized less often.

Global clustering coefficients of the SKDNs are much larger than those derived from randomly generated networks; the average shortest path length is also consistently shorter or at the same level. These two findings jointly suggest that the empirical knowledge diffusion networks engendered by the collaborative search feature demonstrate the properties of small-world networks. Such properties are important prerequisites for a network to function effectively as the substrate mediating the transfer of information or knowledge.[50 55]

As described in the Methods section, the modularity measure assesses the degree to which knowledge-sharing activities occur within a network segment as compared with reaching out to the rest of the network.[51 52] Such network segments, or social cliques, might be administratively formed or organized around the users' medical specialties. Shown in the last three rows in table 7, the modularity assessments of the private-bundle-sharing networks are consistently larger than those of the networks based on public bundles. This finding further confirms that public-bundle sharing is a more effective method for engendering transdomain search-knowledge diffusion. Further, the modularity of the 'Specialty' network is the highest among the three organizational-level networks. Therefore, medical specialty represents a relatively more natural partitioning

Table 5   Creator—Consumer Network (overall)

Network measure	Private bundles	Public bundles	All bundles
Percentage of singletons	75.2	73.2	58.1
Average degree	2.3	2.53	2.89
Average tie strength	2.90	1.36	2.13
Global clustering coefficient	0.2 (0.006)	0.13 (0.043)	0.21 (0.014)
Average shortest path length	2.03 (5.00)	2.36 (5.98)	2.47 (8.28)

Table 6   Creator—Consumer Network at the three organizational levels

	Network measure	Private bundles	Public bundles	All bundles
Department	Percentage of singletons	47.6	47.6	38.1
	Average degree	4.55	5.82	5.85
	Average tie strength	7.96	3.47	8.16
Specialty	Percentage of singletons	42	40	26
	Average degree	3.10	3.53	4.16
	Average tie strength	4.4	2.11	4.03
Division	Percentage of singletons	50.0	43.8	31.3
	Average degree	3.12	3.72	4.23
	Average tie strength	4.08	1.93	3.58

Table 7 Consumer—consumer network

Network measure	Private bundles	Public bundles	All bundles
Percentage of singletons	78.9	74.5	61.9
Average degree	4.27	12.5	10.22
Average tie strength	3.29	1.20	3.46
Global clustering coefficient	0.54 (0.05)	0.74 (0.11)	0.65 (0.066)
Average shortest path length	2.41 (3.24)	2.33 (2.11)	2.52 (2.46)
Modularity (by department)	0.304	0.112	0.185
Modularity (by specialty)	0.303	0.129	0.191
Modularity (by division)	0.285	0.0810	0.158

criterion for delineating distinct user groups wherein the members demonstrate similar EHR search behavior.

Results of the potential gains analysis

In the hypothetical Consumer—Consumer Network constructed based on repetitive queries manually entered by different users, the fraction of singletons measure reduces from 61.9% to 10.6%, the average degree increases from 10.22 to 47.48, and the average shortest path length decreases from 2.33 to 2.14. These results suggest that while the collaborative search feature had contributed to improved search-knowledge diffusion in the study environment, its potential was far from being fully realized.

DISCUSSION

User participation has transformed today's web from a repository of static information into a dynamic and socially constructed information space. With this paradigm shift, many social information-foraging tools have been created to improve information retrieval on the web by leveraging the collective wisdom of millions of users connected by shared interests and goals.[28 29] While prior research has shown that professionals and researchers also possess a positive attitude toward this concept,[56 57] there has been a paucity of empirical studies demonstrating its viability when deployed in practice.

In healthcare, enabling user collaboration using computerized systems is actually not a new phenomenon. It can be found in a variety of forms such as shared EHR documentation templates and shared ordersets. Through developing a full-text EHR search engine, we had a unique opportunity to investigate whether healthcare practitioners and researchers would embrace the social information-foraging concept embodied in a collaborative search feature. The results show that the provision of the feature engendered a considerable level of user enthusiasm, and their participation and contribution facilitated the preservation of EHR search knowledge and diffusion of the knowledge across people and domains.

Utility of the collaborative search feature in facilitating search-knowledge diffusion

The number of shared search-terms bundles contributed by the EMERSE users was much larger than what we had initially anticipated. We expected that only a handful of bundles would be created, and even fewer would be shared across users. For this reason, we did not implement any classification schemas to help organize the bundle repository; nor did we make the bundle repository itself searchable.

Of nearly a million EHR searches processed by the system, about half (0.44 million) were based on stored search-terms bundles, 35.8% of which utilized shared knowledge made available by others in the user community. This means that if the collaborative search feature were not provided, up to 0.16

million search queries would need to be manually created, which could cause tremendous productivity loss in repeatedly constructing the same or substantially similar queries.

More importantly, because developing effective and inclusive EHR search queries is a complex task that requires sophisticated information-retrieval expertise, the quality of search queries that users submit can be highly variable. A bundle creator's act of preserving a query as stored knowledge and making it available to other users signals their relative confidence in the quality of the work. Hence, it may be reasonable to assume that the search-terms bundles shared in the system are generally superior to those prepared ad hoc. Adopting and making judicious use of such bundles, which convey sophisticated search knowledge, could therefore improve the overall quality of EHR search in addition to search efficiency.

Further, the search-terms bundles also serve a crucial function as boundary objects enabling the transfer of search knowledge across domains.[58] This function is particularly important in healthcare because of the highly specialized nature of medicine. Additionally, these stored search-terms bundles represent a vehicle of organizational memory helping to retain the valuable knowledge of EHR search within the organization; such knowledge could otherwise be lost, for example, due to staff turnover.[59]

Partitioning of the knowledge-diffusion networks

Social-network analysis is a particularly useful approach for delineating the structure of the knowledge-diffusion networks to reveal hidden patterns. For example, the network-modularity assessments suggest that partitioning the user population of the search engine according to their medical specialty would yield best results in identifying distinctive subcommunities. This finding provides a valuable design implication in how the search-terms bundles should be organized on the system's user interface: among many facets by which the bundles could be classified or sorted, the medical specialty of the bundle creators may be the most effective facet for presenting bundles that are of high relevance to the user, and therefore should be provided as the default bundle-classifying/-sorting option.

Limitations and future directions

The collaborative search feature studied in this paper was still in a rudimentary form compared to some sophisticated social information-foraging tools available on the web. The feature lacked several important community functions—for example, a reputation-management mechanism collecting peer feedback of the quality of user-contributed content (search-terms bundles), and the provision of latent quality indicators such as how many times a bundle has been borrowed and applied. Further, as revealed in the social-network analysis, the Consumer—Consumer Network conveys an important yet invisible type of relationships that may tie some users together based on their possible shared objectives and interests in common. Upon user consent, this information could be publicized via a research social-networking tool to help these users get to know each other to stimulate synergic effects both online and offline.

The preliminary success of the collaborative feature tested at our institution prompts opportunities for fostering user collaboration at larger scales across institutions. For example, relatively generic EHR search knowledge (eg, identifying the mention of smoking status in clinical narratives) can be developed, evaluated, formalized, and published by authoritative organizations such as the US National Library of Medicine. The

captured knowledge of search, likely in the forms of bundled search terms or more sophisticated natural-language-processing algorithms, can then be consumed by institutional EHR search engines to help healthcare practitioners and researchers accomplish complex information-retrieval tasks more efficiently.

The collaborative search approach, however, has several notable limitations. While it has been demonstrated that over time, 'bad' information usually gets expelled from an online community through group intelligence,[60] it is not clear if this community-based self-distillation mechanism may work in the context of EHR search, given that (1) the size of EHR search-engine user communities is typically much smaller; and (2) the search-terms bundles can be very complex, and appraising their build quality may exceed the ability of most average users. In addition, providing the collaborative search feature could result in an unintended over-reliance on shared search knowledge, particularly among novice users, that is, a shared search-terms bundle may be mindlessly adopted by inexperienced users without a careful evaluation of its appropriateness for the particular information-retrieval task at hand.

CONCLUSION

Through a homegrown full-text EHR search engine, we implemented and evaluated a 'collaborative search' feature for engendering user participation and collaboration so that the knowledge of EHR search could be preserved, collectively refined, and diffused across people and domains. The empirical study results suggest that the search-engine users embraced this concept with considerable enthusiasm, which contributed to improved diffusion of search knowledge and potentially improved search performance. Therefore, we encourage practitioners and researchers to consider applying this, and possibly other social information-foraging techniques popularly used on the web, to help improve the quality and efficiency of information retrieval in healthcare.

REFERENCES

1. **Safran C,** Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;**14**:1−9.
2. **Docteur E,** Berenson R. How will comparative effectiveness research affect the quality of health care? Robert Wood Johnson Foundation Timely Analysis of Immediate Health Policy Issues. 2010:1−15. http://www.urban.org/publications/412040.html.
3. **Hersh WR.** Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Am J Manag Care 2007;**13**:277−8.
4. **Etheredge LM.** A rapid-learning health system. Health Aff (Millwood) 2007;**26**:w107−8.
5. **PricewaterhouseCoopers.** Transforming Healthcare Through Secondary Use of Health Data, 2009. http://www.pwc.com/us/en/healthcare/publications/secondary-health-data.jhtml (accessed 3 Mar 2010).
6. **Bloomrosen M,** Detmer DE. Informatics, evidence-based care, and research; implications for national policy: a report of an American Medical Informatics Association health policy conference. J Am Med Inform Assoc 2010;**17**:115−23.
7. **Cannon J,** Lucci S. Transcription and EHRs. Benefits of a blended approach. J AHIMA 2010;**81**:36−40.
8. **Dick RS,** Steen EB, Detmer DE. The Computer-Based Patient Record: An Essential Technology for Health Care, Revised Edition. Washington, DC: National Academy Press, 1997.
9. **van Ginneken AM.** The physician's flexible narrative. Methods Inf Med 1996;**35**:98−100.
10. **Lovis C,** Baud RH, Planche P. Power of expression in the electronic patient record: Structured data or narrative text? Int J Med Inform 2000;**58−59**:101−10.
11. **Moss J,** Andison M, Sobko H. An analysis of narrative nursing documentation in an otherwise structured intensive care clinical information system. AMIA Annu Symp Proc 2007:543−7.
12. **Madani S,** Shultz E. An analysis of free text entry within a structured data entry system. AMIA Annu Symp Proc 2009:947.
13. **Turchin A,** Shubina M, Breydo E, et al. Comparison of information content of structured and narrative text data sources on the example of medication intensification. J Am Med Inform Assoc 2009;**16**:362−70.
14. **The Healthcare Information and Management Systems Society (HIMSS).** Clinical Analytics: Can Organizations Maximize Clinical Data? 2010. http://www.himssanalytics.org/docs/Clinical_Analytics.pdf (accessed 25 Jun 2010).
15. **Palchuk MB,** Fang EA, Cygielnik JM, et al. An unintended consequence of electronic prescriptions: prevalence and impact of internal discrepancies. J Am Med Inform Assoc 2010;**17**:472−6.
16. **Tange HJ,** Schouten HC, Kester AD, et al. The granularity of medical narratives and its effect on the speed and completeness of information retrieval. J Am Med Inform Assoc 1998;**5**:571−82.
17. **Weiner M,** Stump TE, Callahan CM, et al. Pursuing integration of performance measures into electronic medical records: beta-adrenergic receptor antagonist medications. Qual Saf Health Care 2005;**14**:99−106.
18. **Zhou L,** Tao Y, Cimino JJ, et al. Terminology model discovery using natural language processing and visualization techniques. J Biomed Inform 2006;**39**:626−36.
19. **Friedman C,** Shagina L, Lussier Y, et al. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;**11**:392−402.
20. **Los RK,** van Ginneken AM, van der Lei J. OpenSDE: a strategy for expressive and flexible structured data entry. Int J Med Inform 2005;**74**:481−90.
21. **Johnson SB,** Bakken S, Dine D, et al. An electronic health record based on structured narrative. J Am Med Inform Assoc 2008;**15**:54−64.
22. **Uzuner Ö,** Goldstein I, Luo Y, et al. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;**15**:14−24.
23. **Natarajan K,** Stein D, Jain S, et al. An analysis of clinical queries in an electronic health record search utility. Int J Med Inform 2010;**79**:515−22.
24. **McKibbon KA,** Haynes RB, Dilks CJ, et al. How good are clinical MEDLINE searches? A comparative study of clinical end-user and librarian searches. Comput Biomed Res 1990;**23**:583−93.
25. **Zeng Q,** Kogan S, Ash N, et al. Characteristics of consumer terminology for health information retrieval. Methods Inf Med 2002;**41**:289−98.
26. **Sneiderman CA,** Demner-Fushman D, Fiszman M, et al. Knowledge-based methods to help clinicians find answers in MEDLINE. J Am Med Inform Assoc 2007;**14**:772−80.
27. **Donato D,** Bonchi F, Chi T, et al. Do you want to take notes? Identifying research missions in Yahoo! Search Pad. Proceedings of the 19th international Conference on World Wide Web (WWW '10). New York: ACM Press, 2010:321−30.
28. **Pirolli P.** An elementary social information foraging model. Proceedings of the 27th international Conference on Human Factors in Computing Systems (CHI '09). New York: ACM Press, 2009:605−14.
29. **Chi EH.** Augmented social cognition: Using social web technology to enhance the ability of groups to remember, think, and reason. Proceedings of the 35th SIGMOD international Conference on Management of Data (SIGMOD '09). New York: ACM Press, 2009:973−84.
30. **Hanauer DA.** EMERSE: The Electronic Medical Record Search Engine. AMIA Annu Symp Proc 2006:941.
31. **Hanauer DA,** Englesbe MJ, Cowan JA Jr, et al. Informatics and the American College of Surgeons National Surgical Quality Improvement Program: automated processes could replace manual record review. J Am Coll Surg 2009;**208**:37−41.
32. **Seyfried L,** Hanauer DA, Nease D, et al. Enhanced identification of eligibility for depression research using an electronic medical record search engine. Int J Med Inform 2009;**78**:e13−8.
33. **Ammori JB,** Pelletier SJ, Mathur A, et al. Financial implications of surgical complications in pediatric liver transplantation. Pediatr Transplant 2008;**12**:174−9.
34. **Choi SW,** Kitko CL, Braun T, et al. Change in plasma tumor necrosis factor receptor 1 levels in the first week after myeloablative allogeneic transplantation correlates with severity and incidence of GVHD and survival. Blood 2008;**112**:1539−42.
35. **Englesbe MJ,** Lynch RJ, Heidt DG, et al. Early urologic complications after pediatric renal transplant: a single-center experience. Transplantation 2008;**86**:1560−4.
36. **Debenedet AT,** Raghunathan TE, Wing JJ, et al. Alcohol use and cigarette smoking as risk factors for post-endoscopic retrograde cholangiopancreatography pancreatitis. Clin Gastroenterol Hepatol 2009;**7**:353−8.
37. **Khaled Y,** Mellacheruvu S, Reddy P, et al. Long-term outcomes following myeloablative allogeneic transplantation for multiple myeloma compared to

autologous transplantation and the impact of graft-versus-myeloma effect. *Bone Marrow Transplant* 2009;**44**:325—6.

38. **Paczesny S,** Krijanovski OI, Braun TM, *et al*. A biomarker panel for acute graft-versus-host disease. *Blood* 2009;**113**:273—8.

39. **White T.** *Can't Beat Jazzy: Introducing the Java Platform's Jazzy New Spell Checker API*. http://www.ibm.com/developerworks/java/library/j-jazzy (accessed 30 Jun 2010).

40. **Hanauer DA,** Zheng K, Mei Q, *et al*. Full-Text Search in Electronic Health Records: challenges and Opportunities. In: Kutais BG, ed. *Internet Policies and Issues*. Hauppauge, NY: Nova Science Publishers, 2010;**7**:1—15.

41. **Jones R,** Rey B, Madani O, *et al*. Generating query substitutions. *Proceedings of the 15th international Conference on World Wide Web (WWW '06)*. New York: ACM Press, 2006:387—96.

42. **Zeng QT,** Crowell J, Plovnick RM, *et al*. Assisting consumer health information retrieval with query recommendations. *J Am Med Inform Assoc* 2006;**13**:80—90.

43. **Radlinski F,** Joachims T. Active exploration for learning rankings from clickthrough data. *Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (KDD '07)*. New York: ACM Press, 2007:570—9.

44. **White RW,** Marchionini G. Examining the effectiveness of real-time query expansion. *Inf Process Manage* 2007;**43**:685—704.

45. **Mei Q,** Zhou D, Church K. Query suggestion using hitting time. *Proceeding of the 17th ACM Conference on information and Knowledge Management (CIKM '08)*. New York: ACM Press, 2008:469—78.

46. **Morris MR,** Horvitz E. SearchTogether: An interface for collaborative web search. *Proceedings of the 20th Annual ACM Symposium on User interface Software and Technology (UIST '07)*. New York: ACM Press, 2007:3—12.

47. National Library of Medicine, My NCBI. http://www.ncbi.nlm.nih.gov/sites/myncbi/ NLM (accessed 5 Aug 2010).

48. *Health Information Research Unit, McMaster University, The Hedges Project*. http://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx (accessed 27 Dec 2010).

49. **Centre for Reviews and Dissemination.** *Search filters of RCTs and other trials*. University of York. http://www.york.ac.uk/inst/crd/intertasc/rct.htm (accessed 27 Dec 2010).

50. **Newman MEJ.** The structure and function of complex networks. *SIAM Rev* 2003;**45**:167—256.

51. **Reichardt J,** Bornholdt S. Statistical mechanics of community detection. *Phys Rev* 2006;**74**:016110.

52. **Newman MEJ.** Modularity and community structure in networks. *PNAS* 2006;**107**:8577—82.

53. **Kamada T,** Kawai S. An algorithm for drawing general undirected graphs. *Inf Process Lett* 1988;**31**:7—15.

54. **Erdős P,** Rényi A. On the evolution of random graphs. *Publ Math Inst Hungar Acad Sci* 1960;**5**:17—61.

55. **Watts DJ,** Strogatz SH. Collective dynamics of 'small-world' networks. *Nature* 1998;**393**:440—2.

56. **Spence PR,** Reddy MC, Hall RA. Survey of collaborative information seeking practices of academic researchers. *Proceedings of the 2005 international ACM SIGGROUP Conference on Supporting Group Work (GROUP '05)*. New York: ACM Press, 2005:85—8.

57. **Morris MR.** A survey of collaborative web search practices. *Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. New York: ACM Press, 2008:1657—60.

58. **Star SL,** Griesemer JR. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's museum of vertebrate zoology, 1907—39. *Soc Stud Sci* 1989;**19**:387—420.

59. **Stein EW,** Zwass V. Actualizing organizational memory with information systems. *Inform Syst Res* 1995;**6**:82—117.

60. **Maloney-Krichmar D,** Preece J. A multilevel analysis of sociability, usability, and community dynamics in an online health community. *ACM Trans Comput Hum Interact* 2005;**12**:201—32.