Open Issues in Intelligent Personal Health Record – An Updated Status Report for 2012

Gang Luo
Department of Biomedical Informatics, University of Utah
HSEB Room 5725B, 26 South 2000 East, Salt Lake City, UT 84112
gang.luo@utah.edu

Abstract To improve the capability and usability of the personal health record (PHR) as a tool to empower consumers in the management of their own health, we have proposed the concept of an intelligent PHR (iPHR) and built a prototype iPHR system with four functions. These four functions use various health knowledge and computer science techniques to automatically provide users with personalized healthcare information to facilitate their well-being. This paper discusses several open issues in iPHR, including two enhancements to an existing function and two potential new functions. The two enhancements are for automatically compiling relevant self-care activities for each health issue and automatically identifying contraindicated self-care activities, respectively. One potential new function is personalized search for individual healthcare providers. Another potential new function is personalized local search for health-related services to help maintain patients in their homes. We include some preliminary thoughts on how to address these open issues with the hope to stimulate future research work on iPHR.

Keywords Personal health record · Self-care activity · Medical information extraction · Automatic contraindication identification · Personalized search for individual healthcare providers · Personalized local search for health-related services

1. Introduction

With the deployment by the Veterans Health Administration [98, 125], Medicare, CIGNA [69], and several major Internet companies including Microsoft [97], WebMD [146], and Office Ally [112] over the past few years, Web-based personal health records (PHRs) are now available to ordinary consumers. These PHR systems enable consumers to actively manage their health records and subsequently their health through a Web interface, but have limited intelligence and can fulfill only a small portion of users' healthcare needs. To improve PHR's capability and usability [105], we have proposed the concept of an intelligent PHR (iPHR) [85, 91] by introducing and extending expert system technology, Web search technology, natural language generation technology [117], database trigger technology, and signal processing technology into the PHR domain.

iPHR serves as a centralized portal for dynamically providing users with comprehensive and personalized healthcare information to facilitate their well-being. It can help many people, particularly those who are limited in their daily activities. For example, chronic conditions alone cause 20% of Americans to become limited in their daily activities in some way [143]. Due to a lack of health knowledge, consumers are often unaware of their healthcare needs and/or unable to identify proper keywords to effectively search healthcare information. To address this problem, iPHR extensively uses health knowledge to (a) anticipate users' needs, (b) guide users to provide the most important information about their health condition, (c) automatically form high-quality queries, and (d) proactively push relevant healthcare information to users whenever their potential need for it is detected.

Our effort on building a prototype iPHR system started in 2006. Previously, we published a paper [91] describing both the experience obtained and the open issues identified during development and operation of iPHR up to 2010. Since then, three years have passed and more progress has been made with iPHR. At present, our iPHR system provides four functions:

(1) guided search for disease information [82],
(2) recommending self-care activities (SCAs) [90],
(3) recommending home health products [92], and
(4) continuous user monitoring [85].

Most of these functions fall into the category of automatically generating personalized healthcare content. Personalization in Web search is traditionally based on the user's search history and browsing history [130], whereas iPHR's personalization is mainly based on the user's health issues.

There are many consumer health information Web sites, such as WebMD.com. Most of the information on them is manually compiled, static, non-comprehensive, non-personalized, and focused on one or a few healthcare application scenarios, such as diagnosis and treatment. Manual compilation is labor intensive and slow. Static information is quickly outdated. Non-comprehensive, non-personalized, and application-specific information is of limited use to most users. To address these limitations, a

key methodology of iPHR is to use health knowledge to automatically form multiple high-quality queries based on the user's needs and personal situation, use these queries simultaneously to retrieve search results, and merge and organize all of the search results to dynamically generate personalized healthcare information that is easy to navigate. This differs from traditional information retrieval, where the user manually forms one query at a time. The use of multiple, personalized, high-quality queries increases the likelihood of obtaining useful, relevant, and relatively comprehensive healthcare information. Dynamic content generation facilitates attaining current information available on the Web. Automatic query and content generation dramatically reduces human effort for both users and system developers. This general methodology is application-independent and reusable across multiple functions of iPHR by plugging in various health knowledge bases. As a result, iPHR can scale to serve as a suite of tools for many healthcare application scenarios.

This paper presents the open issues that we identified in the past three years in developing and operating iPHR, including two enhancements to an existing function for automatically constructing health knowledge bases and two potential new functions that can be helpful to consumers. The two enhancements are for automatically compiling relevant SCAs for each health issue and automatically identifying contraindicated SCAs, respectively. One potential new function is personalized search for individual healthcare providers, which introduces and extends data mining technology [57] and recommender system technology [121] into the PHR domain. Another potential new function is personalized local search for health-related services to help maintain patients in their homes. These open issues represent the most urgent needs in iPHR according to our experiences and interactions with patients and their caregivers. We include some preliminary thoughts on how to address these open issues with the hope to stimulate future research work in the new area of iPHR.

The rest of the paper is organized as follows. Section 2 gives an overview of iPHR's SCA recommendation function. Section 3 discusses automatic compilation of relevant SCAs for each health issue. Section 4 addresses automatic identification of contraindicated SCAs. Section 5 describes the potential new function of personalized search for individual healthcare providers. Section 6 presents the potential new function of personalized local search for health-related services. Section 7 concludes this paper. Appendix A provides a list of acronyms used in this paper. Appendix B includes a list of symbols used in this paper.

2. Overview of iPHR's SCA recommendation function

In this section, we provide an overview of iPHR's SCA recommendation function. The details of this function are described in our previous publications [90, 91].

Consider a consumer with a particular health issue such as asthma. Hundreds of SCAs, such as "Coach in breathing/relaxation techniques," are relevant to the health issue. They can be performed at home or in the community to achieve desirable outcomes and facilitate a wide range of daily activities [32]. However, interaction time during a clinical encounter is usually short, on average no more than 20 minutes [31]. In this limited amount of time, she frequently cannot obtain from healthcare professionals sufficient information about relevant SCAs. With the trend toward more self-care and in-home care, this is problematic.

For people with chronic conditions, more than 80% of their care is performed by themselves and their caregivers [30]. In various countries, one-third to two-thirds of people with a chronic condition are not given a home self-care plan by their healthcare providers [136, 137], although such a plan is a critical component to improving or maintaining their health [11, 106]. Most patients want detailed self-care information covering a wide range of issues [65, 106, 110]. When patients receive pamphlets about their health issues from their healthcare providers, the healthcare information provided on these pamphlets is often overly simplified and lacking in details, thereby limiting the usefulness of these pamphlets [71]. The written home self-care plans provided by their healthcare providers, if any, frequently have the same problem as these pamphlets. In fact, between 20% and 40% of discharge plans are deficient, partly because both the healthcare professionals and patients are often unable to anticipate what the patients' needs will be upon returning home from the hospital [93].

The inability of patients to obtain sufficient, detailed information on self-care from their doctors is a major barrier to successful care transitions and has great impact on healthcare cost [11]. For example, about 20% of fee-for-service Medicare beneficiaries discharged from the hospital are readmitted within 30 days. 75% of these readmissions costing about 12 billion dollars a year are considered potentially preventable, particularly with improved care transitions [11].

To compensate for the insufficient self-care information provided by healthcare professionals, the consumer can search for relevant healthcare information online, starting by entering the health issue name as a keyword query into a major Web search engine. Typically, the search results are incomplete: she retrieves information on a small subset of the relevant SCAs without realizing the existence of others. This is because information on relevant SCAs is currently scattered on numerous Web pages on various Web sites. Different Web pages use varying characteristic phrases indicating the presence of SCA information. Consequently, each keyword query can retrieve information on only a few relevant SCAs, at least in the case of the returned top several search result Web pages. To find a comprehensive set of SCAs relevant to the health issue, multiple searches need to be conducted using in-depth health knowledge to

form appropriate health keyword queries. This exceeds the capability of the average consumer.

Due to the reasons mentioned above, many consumers have unmet needs for self-care information and need help obtaining adequate information. In particular, patients with chronic conditions have an ongoing need for updated information as their conditions evolve and their information needs change [106, 141, 149]. Both they and their caregivers continue to need help obtaining self-care information many months and years after discharge [37, 106]. According to several surveys, about 10% to 25% (depending on the category of SCAs) of patients recently discharged from a hospital [58], 75% of stroke patients at 6 months post discharge [59], 31% of cancer patients [141], and more than 60% of long-term cancer survivors [73] have unmet needs for self-care information.

To fulfill consumers' needs for self-care information, we introduced expert system technology and Web search technology into the PHR domain and developed iPHR's SCA recommendation function that can automatically retrieve SCAs relevant to the user's health issues [90]. Each nontrivial SCA is made clickable for the user to find various, detailed implementation procedures for it on the Web. Consumers can review the information at their own pace when they are ready and as often as needed rather than being suddenly challenged with processing and understanding a lot of information at a stressful time, e.g., upon discharge [7, 149].

Our main idea is to use nursing knowledge to construct a SCA information knowledge base. A set of relevant SCAs is pre-compiled and stored in the knowledge base for each health issue. For each current health issue of the user, iPHR retrieves the corresponding relevant SCAs from the knowledge base and returns them to the user. Moreover, when the user clicks a nontrivial SCA, iPHR automatically submits a corresponding, pre-compiled phrase stored in the knowledge base as a query to a large-scale Web search engine to retrieve various, detailed implementation procedures for the SCA. In this way, we eliminate the challenge for users to independently build appropriate health keyword queries.

This paper focuses on SCAs. Nevertheless, all approaches discussed in this paper for compiling various kinds of information related to SCAs can also be applied to other types of care activities, such as those performed by nurses in the acute care or hospital setting. There is no fundamental difference between the different types of care activities that would prevent the use of our compilation methods. Moreover, much overlap exists between SCAs and other types of care activities, as the same care activity is often used by different people in various settings.

3. Compiling relevant SCAs for each health issue

In this section, we discuss the open issues in iPHR's SCA recommendation function relating to semi-automatically compiling relevant SCAs for each health issue.

iPHR's SCA information knowledge base essentially contains a standardized self-care plan for each health issue. The comprehensiveness of the SCA information in the knowledge base directly affects the accuracy and effectiveness of iPHR's SCA recommendation function. As mentioned in Luo *et al*. [91], no comprehensive SCA information knowledge base currently exists for all health issues. The nursing community has compiled standardized care plans for several hundred, but not all, health issues and published them as nursing textbooks [32]. However, these standardized care plans do not fully serve our purpose, as they typically are tailored to the hospital setting and include some but not all of the relevant SCAs.

Manual compilation of SCA information is labor intensive [32]. There exist thousands of health issues. Typically, hundreds of SCAs are relevant to a single health issue whereas information about these SCAs is scattered in numerous sources, such as health journal articles, health textbooks, and health Web sites. A nurse needs several weeks to conduct literature reviews to compile SCAs relevant to a health issue, but the resulting compilation is often still fairly incomplete.

To facilitate constructing a comprehensive SCA information knowledge base, multiple automatic approaches can potentially be used to extract SCA information from various sources. These approaches are complementary to each other in the sense that each approach may extract additional SCA information missed by the other approaches. Thus, all of the approaches can be used together to construct the SCA information knowledge base.

As is the general case with artificial intelligence, none of these automatic extraction approaches is perfect. Some extracted results will be incorrect. Thus, human experts need to manually review the extracted results to filter out erroneous ones and deposit the remaining ones into the SCA information knowledge base. That is, the construction of the SCA information knowledge base is a semi-automatic process rather than a fully automatic process. Nevertheless, compared to manual compilation of SCA information, these automatic extraction approaches can greatly reduce the amount of human labor needed, making the daunting task of constructing a comprehensive SCA information knowledge base much more manageable.

Following, we describe three potential approaches for automatically extracting SCA information.

3.1 Approach 1: Extracting SCA information from historical, coded care activity data in electronic health records

The first potential approach is to automatically extract SCA information from historical, coded care activity data in electronic health records.

In some healthcare facilities, such as Intermountain Healthcare [61, 64], all care activities applied to a patient are stored as coded data in the facility's electronic health record system. For each health issue, the historical, coded data can be used to determine which SCAs have been applied to patients with the health issue. For each SCA, counting is conducted to obtain the number of patients with the health issue who were provided with the SCA. Intuitively, a SCA is unlikely to be used to address the health issue if this number is too small. A pre-determined number is used as the threshold to identify every SCA that has been applied to at least this number of patients with the health issue. The SCAs meeting this criterion are likely to be relevant to the health issue.

It is easy to implement this approach using the historical, coded care activity data of those patients with a single health issue. However, the situation can become tricky if we would like to also use the historical, coded care activity data of those patients with more than one health issue. In the case where a SCA was applied to a patient with multiple health issues, we may not know which health issue of the patient was addressed by the SCA unless clearly defined by the coded care activity data recorded in the healthcare facility's electronic health record system. If this information is unavailable, statistical hypothesis testing can be conducted to check whether an association exists between a SCA and a health issue. For instance, we can use a method similar to the one described in Chen *et al.* [24], where the χ^2 statistic is adopted to check whether an association exists between a drug and a disease.

At present, iPHR's SCA recommendation function recommends SCAs based on the user's health issues and ranks relevant SCAs according to a set of pre-determined, fixed priority weights [90]. It would be interesting to investigate whether this SCA recommendation function can be made more personalized using both historical, coded care activity data in electronic health records and the concept of patient similarity [151]. For example, given a particular user, we check which SCAs have been applied to other patients similar to him, such as those patients with the same health issues, gender, and age group as him. For each SCA, we compute the number of other patients who are similar to him and were provided with the SCA, weighting each count obtained from a similar patient by the corresponding degree of patient similarity. This number is used in providing more personalized ranking of relevant SCAs.

3.2 Approach 2: Extracting SCA information from historical, textual data in electronic health records

The second potential approach is to automatically extract SCA information from historical, textual data in electronic health records, such as nursing notes. For instance, we can adopt an approach similar to the one described in Chen *et al*. [24], where text mining and statistical techniques are used to analyze clinical documents to identify disease-drug associations. For each health issue, information extraction is performed on historical, textual data in electronic health records to determine which SCAs have been applied to patients with the health issue. Statistical hypothesis testing is then conducted to check whether these SCAs are associated with, or likely to be used to address, the health issue.

3.3 Approach 3: Extracting SCA information from electronic health knowledge resources

The third potential approach is to automatically extract SCA information from electronic health knowledge resources, such as health Web pages and health textbooks. This approach falls into the area of information extraction [8, 22, 99, 108, 124], in which much research has been conducted. Almost all such work focuses on extracting information expressed as one or more phrases, such as entities, relationships between entities, and properties of entities. In our case, due to SCAs' inherent complexity, each SCA is usually expressed as one or more sentences or an entire paragraph. Also, there is no known method that can automatically identify which sections of a document contain SCA information. Consequently, existing information extraction techniques cannot be directly used to extract SCA information from electronic health knowledge resources.

Our following discussion on the third approach focuses on health Web pages. In addition to health Web pages, there are other electronic health knowledge documents, including health textbooks, clinical guidelines, health journal articles, and internal healthcare facility documentation. We can build a search engine for these other documents and then handle them in a way similar to health Web pages.

Our high-level idea is to use existing SCA information that was previously compiled manually as the seed and conduct information retrieval and machine learning to extract additional SCA information from electronic health knowledge resources. To build an effective electronic tool for extracting SCA information, a method needs to be developed to automatically identify the document sections likely to contain SCA information. This requires knowing the characteristic features of these document sections, such as the characteristic phrases that typically appear in these document sections. For this purpose, existing SCA information that was previously compiled manually is used to construct keyword queries. Then the document sections retrieved by these keyword queries are used to learn multiple characteristic phrases. Finally, these characteristic phrases are combined with health issue names to construct more queries to retrieve additional SCA information.

More specifically, we use a bootstrap approach [5] to develop an electronic tool for extracting SCA information from electronic health knowledge resources in the following several steps as shown in Fig. 1.

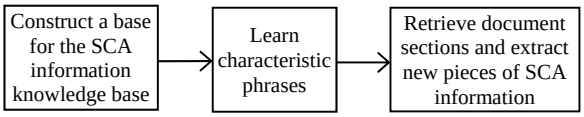

Fig. 1 List of steps to develop an electronic tool for extracting SCA information from electronic health knowledge resources.

3.3.1 Step 1: Constructing a base for the SCA information knowledge base.

We use existing SCA information that was previously compiled manually, such as the standardized care plans in nursing textbooks [32] and the SCA information collected on the health topic Web pages on MedlinePlus [96], to construct a base for the SCA information knowledge base. Each piece of SCA information stored in the knowledge base is a pair of the form (*H*, *A*), where *H* is a health issue and *A* is a SCA relevant to *H*. In the base of the knowledge base, many so-called base pieces of SCA information are readily available, e.g., from a large number of existing standardized care plans. These base pieces cover several hundred, but not all, health issues. Likewise, for a covered health issue, these base pieces include some but not all SCAs relevant to it. In the following steps, these base pieces are used as seeds to help obtain additional pieces of SCA information. In this way, the SCA information knowledge base will be more comprehensive by covering more health issues and an increased number of relevant SCAs for each individual health issue.

3.3.2 Step 2: Learning multiple characteristic phrases that typically appear together with SCA information.

For a base piece of SCA information (*H*, *A*), we combine the name of the health issue *H* and some important phrases manually extracted from the description of the SCA *A* to form a keyword query. By inputting this query into a large-scale Web search engine such as Google, we retrieve some Web pages likely to mention that *A* is relevant to *H*. As explained in Luo *et al*. [91], since the anchor text of links to a Web page usually cannot help identify the relationship between the SCA and the health issue, we would require all keywords of this query to appear in the text of the Web page, e.g., "allintext:" would prefix the query according to the syntax of the Google search engine. For each retrieved Web page document, the same query and passage retrieval techniques [80] are used to find the sections within the document likely to mention that *A* is relevant to *H*. The document sections are then manually reviewed to identify the specific ones mentioning that *A* is relevant to *H*. These document sections contain SCA information.

The above approach uses one base piece of SCA information at a time. In general, we can use several base pieces of SCA information related to the same health issue *H* at a time. In this case, the formed query will include both the name of *H* and important phrases manually extracted from the descriptions of the SCAs.

By performing the above procedure for multiple base pieces of SCA information, many document sections containing SCA information are obtained. We use these document sections as a training set and conduct machine learning to learn multiple characteristic phrases that typically appear together with SCA information. For example, possible characteristic phrases include "treat," "manage," and "prevent" [42]. A healthcare professional would then manually review the characteristic phrases to identify those that intuitively make sense. They will be used in the next step.

3.3.3 Step 3: Retrieving document sections likely to contain SCA information and extracting new pieces of SCA information from them.

We keep a list of health issues, including many not covered in the base of the SCA information knowledge base. For each health issue, its name is combined with one or more characteristic phrases to form a keyword query. By inputting this query into a large-scale Web search engine such as Google, Web pages likely to mention SCAs relevant to the health issue are retrieved. Again, we would require all keywords of the query to appear in the text of the Web page.

In general, the more characteristic phrases the query contains, the greater the chance the retrieved Web pages will mention SCAs relevant to the health issue. However, if the query includes too many characteristic phrases, there is a risk the Web search engine will not return any search result, as most major Web search engines only return Web pages containing all keywords of the query. We can achieve balance and obtain relevant search results by determining an appropriate number of characteristic phrases. For example, we can start from one characteristic phrase and keep adding more characteristic phrases to the query until the number of search result Web pages returned by the Web search engine reaches a certain lower-bound threshold.

For each Web page document retrieved by the query, we use the same query, passage retrieval techniques [80], and section detection techniques [104] to identify the sections in the document likely to mention SCAs relevant to the health issue. By using different combinations of characteristic phrases, we obtain distinct keyword queries and subsequently retrieve various document sections. Negation detection techniques [20] are used to remove those retrieved document sections where SCA information is negated. Finally, a healthcare professional manually reviews the remaining retrieved document sections to extract new pieces of SCA information and deposit them into the SCA information knowledge base.

The above approach starts from a health issue and finds SCAs relevant to the health issue. As is the general case with natural language processing, this approach is

imperfect. It will uncover a lot of SCA information, but possibly not everything. To find more SCA information, we also start from a SCA and find health issues that can be managed by the SCA in a way similar to that of finding indications of a drug [42]. More specifically, we keep a list of known SCAs, such as those mentioned in the Nursing Interventions Classification nursing interventions [12] and in the standardized care plans in nursing textbooks [32]. For each SCA, some keywords extracted from its description are combined with one or more characteristic phrases to form a keyword query. By inputting this query into a large-scale Web search engine such as Google, we retrieve some Web pages likely to mention health issues that can be managed by the SCA. Then we proceed in the same way as mentioned above to identify document sections likely to contain SCA information and extract SCA information from them.

In practice, some retrieved Web pages will come from non-authoritative health Web sites and contain incorrect SCA information. For a particular piece of SCA information extracted from a retrieved Web page, it is possible the healthcare professional will doubt its accuracy and cannot determine its validity using her health knowledge. In this case, she can form one or more keyword queries, such as by extracting keywords from its description, to see whether she can find a high-quality health Web page verifying its accuracy. If no high-quality health Web page is found, by default we conservatively assume that it is incorrect. In general, this approach works for any type of healthcare information extracted from health Web pages, including SCA contraindication information that will be discussed in Section 4.

3.4 Displaying SCA information

iPHR's SCA recommendation function displays the SCAs relevant to the user's health issues. However, this may be insufficient because many SCAs have side effects that the user is unaware of. For a particular SCA, some of its side effects are general, whereas the others can be specific to the concrete health issue being addressed by the SCA. To help the user make informed decisions, it would be beneficial for iPHR to list the side effects of each relevant SCA, if any, next to the SCA. For each listed side effect, its incidence statistics reflecting the likelihood of developing it could be displayed. To compile side effects of SCAs, we can adopt a method similar to those used for compiling drug side/adverse effects from various resources, such as electronic health knowledge resources [27, 68, 89], historical data in electronic health records [148], and medical message board posts [15]. To obtain incidence statistics of side effects of SCAs, we can adopt a method similar to those used for estimating incidence statistics of drug adverse effects from various resources [36, 50, 68].

Users usually prefer receiving comprehensive self-care information tailored to their individual situation [65, 147]. Compared to generic information, tailored information is more likely to be read and remembered, and is more effective in changing health behaviors and improving self-efficacy [60]. For the same health issue, differing users can be interested in different SCAs relevant to it. To help the user quickly find desired SCAs, we can further personalize the displayed SCAs based on his properties (e.g., age, gender, race, level of physical activity) in addition to his health issues. Since a user can be interested in many SCAs and it is difficult to anticipate all possible combinations of properties and associated potential needs beforehand, we can also build one or more hierarchies on the displayed SCAs to facilitate easy navigation.

4. Automatically identifying contraindicated SCAs

In this section, we discuss the open issues in iPHR's SCA recommendation function relating to automatically identifying contraindicated SCAs.

4.1 Background on contraindication identification

Many people are "complex patients," meaning that they have two or more health issues simultaneously [3]. For example, 21% of Americans have multiple chronic conditions [145]. The care of complex patients is often involved. In particular, a SCA suitable for a single health issue can become undesirable in the presence of another health issue. This is called contraindication in healthcare [10]. For instance, the SCA massage is contraindicated for the health issue cancer because massage increases lymphatic circulation and hence may potentiate the spread of cancer through the lymphatic system. In general, the concept of contraindication can refer to any type of health intervention, not just SCAs.

Medical errors related to contraindications are prevalent. For instance, according to the survey Radley *et al*. [119] conducted on patients with dementia, hip/pelvic fracture, or chronic renal failure, about 10% to 30% of patients in multiple different areas have been incorrectly prescribed contraindicated drugs. Safety strategies in healthcare have traditionally focused on hospitals. Nevertheless, according to the survey Schoen *et al*. [136] conducted on adult patients with chronic conditions, about 60% to 80% of medical errors occur outside the hospital. Hence, it is beneficial to develop tools to help consumers prevent medical errors, including those related to contraindications, outside the hospital.

When recommending SCAs, ideally iPHR should automatically identify all contraindicated SCAs so that they will not be inadvertently applied to complex patients [91]. For example, iPHR can present all contraindicated SCAs to the user as a warning message. To build a component in iPHR for automatically identifying contraindicated SCAs, a SCA contraindication information knowledge base is essential. For each SCA, a list of contraindicated health

issues is stored in the knowledge base. The comprehensiveness of the information in the knowledge base would directly affect the accuracy and effectiveness of the automatic contraindication identification component that relies on the knowledge base.

For drugs [48, 77] and medical tests [113], several automatic contraindication identification tools tailored to certain specific cases are available and have been shown to help reduce the number of medical errors [46]. Also, drug contraindication information knowledge bases already exist. For example, First Databank Inc. has a commercial product, FDB MedKnowledge. This product has multiple modules, one of which is the drug-disease contraindications module [17]. Several other similar products are available, e.g., from Medi-Span [95] and Cerner [16]. All of these companies built and continuously maintain their drug contraindication information knowledge bases through a rather time-consuming and labor-intensive process, by hiring lots of human experts to manually conduct extensive literature reviews and compile contraindication information.

At present, no comprehensive contraindication information knowledge base exists for SCAs. Once such a knowledge base becomes available, we can integrate it with our SCA contraindication identification algorithm described in Luo *et al*. [91] to build an automatic contraindication identification component for SCAs in iPHR. The main idea of our SCA contraindication identification algorithm is to check each SCA linked to a health issue of the user with every other health issue of the user to detect any contraindications. This algorithm performs hierarchical propagation using the medical terminology of International Classification of Diseases (ICD-10) [66].

4.2 Challenges for compiling SCA contraindication information

SCA contraindication information is currently scattered in numerous sources, such as health journal articles, health textbooks, and health Web sites, and frequently not indexed in the back of health textbooks [10]. To the best of our knowledge, the only systematic compilation of SCA contraindication information available is Batavia's book [10]. It covers a small subset of all SCAs, about 100 of them in the area of physical rehabilitation. For a SCA not covered in this book, individual sources often contribute somewhat varying contraindications (see part I of [10] for a detailed discussion). Consequently, to obtain a comprehensive set of contraindications for even a single SCA, typically many sources need to be checked.

It is labor intensive to manually compile contraindication information. For example, as mentioned in the Preface and Acknowledgment sections of Batavia's contraindication book [10], the author spent five years compiling this book through extensive and time-consuming literature reviews. To facilitate constructing a comprehensive SCA contraindication information knowledge base, it would be desirable to develop an electronic tool for automatically extracting SCA contraindication information from electronic health knowledge resources, such as health Web pages and health textbooks.

This automatic extraction approach falls into the area of medical information extraction [8, 22, 99, 108, 124], for which much research has been conducted and multiple electronic tools have been developed. Several of these tools, such as SemRep [118], can recognize medical entities and extract their relations and/or properties in a wide spectrum. Several other tools specialize in pharmacogenomics and can extract information related to drugs, such as drug adverse events [38], drug contraindications [120], and drug-drug interactions [135, 140]. Moreover, the medical library science community has conducted much research work on developing search filters for drug adverse events [50]. A search filter is a predefined combination of keywords designed to retrieve information on a particular topic.

To the best of our knowledge, Rubrichi *et al*. [120] is the only published work describing how to automatically extract drug contraindication information from textual documents. This work uses the Summary of Product Characteristics (SPC) documents of drug products as the textual data resource. It first identifies the contraindication section in each SPC document using the property that all SPC documents are written in a fixed format with a dedicated contraindication section. Then it extracts drug contraindication information from this section.

SCA contraindication information has its own unique properties and the text containing it is often written in a specific format. As a result, existing medical information extraction tools cannot be directly used to extract it. For example, no known method can automatically identify which sections of a document contain SCA contraindication information. Thus, the method described in Rubrichi *et al*. [120] cannot be directly used to extract SCA contraindication information.

As a second example, consider the sample text in Fig. 2 describing multiple contraindications for the SCA "massage." These contraindications are displayed as a bullet list: fever, hernia, and osteoporosis, but none of them occurs in the same sentence as the word "massage." Contraindication is a relation between two medical entities. Most of the existing medical information extraction tools are designed to handle the case in which two medical entities and their relation occur in the same sentence and thus cannot handle the scenario in Fig. 2.

… The list of contraindications for massage may be longer than you expect, and it includes some conditions that at first glance don't seem like massage would affect at all. Take a look:

- **Fever:** When you have a fever, your body is trying to isolate and expel an invader of some kind. Massage increases overall circulation and could therefore work against your body's natural defenses …
- **Hernia:** Hernias are protrusions of part of an organ (such as the intestines) through a muscular wall. It's not a good idea to try to push these organs back inside. Surgery works better.
- **Osteoporosis:** Elderly people with a severe stoop to the shoulders often have this condition, in which bones become porous, brittle, and fragile. Massage may be too intense for this condition ...

Fig. 2 A sample description of contraindications for the SCA "massage." This description is extracted from http://www.dummies.com/how-to/content/knowing-when-not-to-massage.html.

To address this problem, one might think about performing anaphora resolution [47] to connect a reference (e.g., a pronoun) to a medical entity with this medical entity mentioned elsewhere in the text. However, although anaphora resolution can help in many cases, it cannot handle this particular scenario as long as the scope of the information extraction algorithm is limited to the case in which two medical entities (or references to them) and their relation occur in the same sentence.

4.3 Proposed solution for compiling SCA contraindication information from electronic health knowledge resources

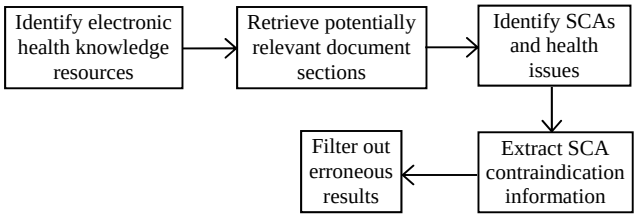

Fig. 3 List of steps to develop an electronic tool for extracting SCA contraindication information from electronic health knowledge resources.

To develop an effective electronic tool for extracting SCA contraindication information from electronic health knowledge resources, we need to consider the unique properties of the language in which contraindications are typically expressed in narrative text. At a high level, our idea is to extend existing algorithms of extracting information related to drugs from electronic health knowledge resources [47], while taking into account these unique properties. We use a bootstrap approach [5] to develop this electronic tool in the following several steps as shown in Fig. 3.

4.3.1 Step 1: Identifying electronic health knowledge sources.

SCA contraindication information is buried in various electronic health knowledge documents, including health textbooks, clinical guidelines, health journal articles, pages of health Web sites, and internal healthcare facility documentation. We identify several major sources of such documents and then obtain documents from these sources, e.g., via Web crawling.

It is common for a health journal article to mention SCA contraindication information in its body but not in its title or abstract. For example, more than half of the studies that documented harm in the body of the article did not report this harm in the abstract [13]. Contraindication is one type of harm. Many health journal databases, including Medline and Embase, limit their native search capability to only titles and abstracts [36]. Hence, using the native search interfaces of these databases to search could result in missed SCA contraindication information. To avoid missing SCA contraindication information in health journal articles, the full text is obtained whenever possible rather than only titles and abstracts.

4.3.2 Step 2: Reducing the amount of text by retrieving potentially relevant document sections.

As is typical with natural language processing, information extraction is computing intensive. To reduce the amount of computing time needed for an information extraction task, information retrieval techniques are often used to quickly filter out most of the irrelevant documents [47]. Information extraction is then performed on the remaining set of documents.

For each SCA, one or more keyword queries are constructed to retrieve documents that may contain contraindication information about it. Each keyword query is a search filter [50] and includes both the name of the SCA and a contraindication-related phrase representing the meaning of contraindication. For example, this phrase can be one of the phrases that Batavia used to compile his contraindication book [10]: adverse effect, adverse event, caution, complication, contraindication, danger, harm, iatrogenic, precaution, risk, and safety. For each retrieved document, both passage retrieval techniques [80] and section detection techniques [104] are used to identify the sections within it that may contain contraindication information about the SCA. Information extraction is performed on only these document sections.

Like the case with adverse effect [50, 89], there are many different ways of expressing the meaning of contraindication, some of which are tricky [131]. Initially, it is difficult to manually construct a comprehensive list of contraindication-related phrases. Consequently, some document sections containing SCA contraindication information may not be retrieved. To address this problem,

a bootstrap method [5, 35] is used to expand the list of contraindication-related phrases.

From Batavia's contraindication book [10], an initial set of SCA contraindication information is obtained as a list of pairs of the form (SCA, health issue contraindicated with this SCA). This set is used to construct multiple keyword queries, each of which includes the name of a SCA and the names of one or more health issues contraindicated with the SCA. Each keyword query is used to retrieve document sections that are likely to contain SCA contraindication information. The frequencies of the keywords appearing in these document sections are then analyzed to identify additional contraindication-related phrases, possibly with manual verification of the appropriateness of these phrases. For instance, some additional potential contraindication-related phrases are shown in Table I.

Table I. Additional potential contraindication-related phrases and their corresponding illustrating examples.

additional potential contraindication-related phrase	illustrating example
warning	Tips & ***Warnings*** Never perform deep tissue massage on swollen tissue or areas that feel warm, as these are symptoms of acute injury. (Extracted from http://www.ehow.com/how_2316841_give-deep-tissue-massage.html.)
avoid	Massage should be ***avoided*** over stents or other prosthetic devices because displacement can occur. (Extracted from http://www.scribd.com/doc/7228870/Acupuntura-Artigo-Medicina-Alternativa-Para-CA.)
should not receive	People with rheumatoid arthritis, goiter (a thyroid disorder characterized by an enlarged thyroid), eczema, and other skin lesions ***should not receive*** massage therapy during flare-ups. (Extracted from http://www.vlad-massage.com/benef.html.)
consult your physician	***Consult your physician*** before scheduling a massage if you've recently had surgery. (Extracted from http://www.livestrong.com/article/136034-deep-tissue-massage-pain/.)
check with your doctor	If you have cancer, ***check with your doctor*** before considering massage because massage can damage tissue that is fragile from chemotherapy or radiation treatments. (Extracted from http://www.vlad-massage.com/benef.html.)

In general, bootstrapping is an iterative process. A set of SCA contraindication information is used to identify more contraindication-related phrases. These phrases are used to retrieve additional, possibly-relevant document sections and extract more SCA contraindication information from these document sections. The process is then repeated. In our case, since the number of contraindication-related phrases is likely to be small, it could be possible that the process can be performed only once to obtain a relatively comprehensive list of contraindication-related phrases.

4.3.3 Step 3: Identifying SCAs and health issues in retrieved document sections.

Before we can start extracting SCA contraindication information from a retrieved document section, named entity recognition needs to be performed first to identify all SCAs and health issues mentioned in the document section. To accomplish this task, we can resort to MetaMap [6], a medical named entity recognition program. The performance of MetaMap relies on the quality of the underlying UMLS Metathesaurus and the associated Specialist Lexicon.

As mentioned in Xu *et al*. [153], medical terminology is highly dynamic. Individual authors may write the same SCA or the same health issue in many different ways. Also, both new SCAs and new health issues keep emerging. As a result, many SCAs and health issues are not recognized by the current MetaMap program. To address this issue and build a more comprehensive contraindication information knowledge base, a method can be developed to automatically construct a dictionary of SCAs and health issues from electronic health knowledge resources. For example, we can extend an existing rule-based or machine-learning-based medical named entity recognition method, such as the one described in Xu *et al*. [153] for automatically constructing a disease dictionary from electronic health knowledge resources.

4.3.4 Step 4: Extracting SCA contraindication information from retrieved document sections.

At this point, for each retrieved document section, all SCAs, health issues, and contraindication-related phrases mentioned in it have been identified. Next, for each such SCA, we need to determine the contraindicated health issues among those identified in the document section. We can start by integrating contraindication-related phrases into SemRep [118], a medical information extraction program designed to extract the relation between two medical entities when they and their relation occur close to each other in the same sentence. Once SemRep can recognize contraindication-related phrases after this extension, we will be able to use it to extract SCA contraindication information.

This approach of extending SemRep could extract some, but not all, of the SCA contraindication information

mentioned in the document section. For example, this approach cannot handle the scenario in Fig. 2, where multiple contraindications are displayed as a bullet list but none of them occurs in the same sentence as the SCA. To address this problem, we can resort to automatic pattern discovery and pattern matching techniques for extracting SCA contraindication information. These techniques fall into the category of wrapper generation in data mining [78].

More specifically, from Batavia's contraindication book [10], an initial set of SCA contraindication information is obtained as training examples. We identify the unique properties of SCAs, design multiple templates for the characteristic patterns according to which SCA contraindication information is written, and use the training examples to learn these characteristic patterns. For instance, some characteristic patterns can be related to lists, tables, bullets, numbering, and/or indentation [23, 38, 75]. The learned characteristic patterns are then used to extract additional SCA contraindication information from each retrieved document section.

As is the general case with bootstrapping [35], the process of learning more characteristic patterns and then extracting additional SCA contraindication information can be performed iteratively. In our case, since our initial set of training examples is reasonably large, it could be possible that the process can be performed only once to obtain a relatively comprehensive list of characteristic patterns. In general, characteristic patterns can be used to help extract from health Web pages any type of healthcare information, including SCA information discussed in Section 3.

4.3.5 Step 5: Filtering out erroneous results and depositing the remaining results into the SCA contraindication information knowledge base.

As is the general case with natural language processing, our information extraction algorithm is imperfect. Some extracted results will not be contraindication information. Thus, human experts need to manually review the contexts of the extracted results to filter out erroneous results and deposit the remaining results into the SCA contraindication information knowledge base. That is, the construction of the knowledge base is a semi-automatic process rather than a fully automatic process. Nevertheless, our automatic extraction tool can greatly reduce the amount of text that human experts need to review, making the daunting task of constructing a comprehensive SCA contraindication information knowledge base much more manageable.

To reduce the amount of text that human experts need to review, we proceed in the following way. The extracted results are aggregated into a list of distinct items, each of which is a pair of the form (SCA, health issue contraindicated with this SCA). For each item, the contexts of all of the corresponding extracted results are displayed to the human expert. The contexts are sorted according to some heuristics so that a higher-ranked context is more likely to show that this item is correct and indeed represents SCA contraindication information. For instance, a machine learning classifier can be trained to predict the probability that a specific context shows that this item is correct. If the human expert knows whether this item is correct according to his healthcare knowledge, he can proceed correspondingly. Alternatively, once he reads any context showing that this item is correct, he can immediately ignore all of the other contexts and deposit this item into the SCA contraindication information knowledge base. If none of these contexts shows that this item is correct, he cannot tell whether this item is correct and hence does not deposit this item into the SCA contraindication information knowledge base. In general, this approach works for any type of healthcare information extracted from health Web pages, including SCA information discussed in Section 3.

4.3.6 Combining text mining with reasoning

Tari *et al*. [140] combined text mining and reasoning together to obtain a more comprehensive set of information on drug-drug interactions. This approach can be extended to obtain a more comprehensive set of SCA contraindication information. For example, via text mining, we discover that a certain property will lead to contraindication for a specific SCA or a specific category of SCAs, and one or more heath issues have this property. Then by reasoning, we can deduce that these health issues contraindicate with the SCA or the category of SCAs.

4.4 Extracting SCA contraindication information from other electronic data resources

Electronic health knowledge resources include mainly SCA contraindication information that has already been published. Similar to adverse drug events, many contraindications related to SCAs are unknown or remain unreported in healthcare literature for various reasons.

Treatments have risks. One type of risk is that a treatment may be contraindicated. Randomized clinical trials (RCTs) are the gold standard for determining the risks and benefits of a treatment. However, many contraindications have not been identified by RCTs. Premarketing RCTs have inherent limitations and cannot detect uncommon (incidence of 1 in 1,000) or long-term (latency of more than 6 months) risks of a treatment [148]. Existing RCTs are typically conducted only on patients with few comorbidities [68] and therefore are not good at discovering contraindications that involve multiple morbidities by definition. Moreover, benefit tends to be reported more frequently than harm in RCTs [13].

After a treatment becomes available on the market, the public can report its risks through several voluntary reporting systems, e.g., at the U.S. Food and Drug Administration and at Health Canada [36]. Due to the voluntary nature of these systems, more than 94% of all adverse drug reactions remain unreported to these systems [62]. We would expect the case with contraindications

related to SCAs to be similar. That is, most contraindications related to SCAs remain unreported to these systems.

In addition to electronic health knowledge resources, there are other electronic data resources. We can use some of them together with multiple potential approaches to discover new contraindications related to SCAs that have not been reported in the healthcare literature. Each approach will extract some candidate, potentially novel SCA contraindication information. Controlled studies then need to be conducted to verify the correctness of the extracted, candidate SCA contraindication information [15]. Following, we describe two potential approaches for automatically extracting potentially novel SCA contraindication information from other electronic data resources.

4.4.1 Approach 1: Using historical data in electronic health records

Wang *et al*. used text mining and statistical techniques to detect potentially novel adverse drug events from historical data in electronic health records [148]. Contraindication is one type of harm, whereas harm is often recorded in electronic health records. Hence, an approach similar to that in Wang *et al*. [148] can be used to extract potentially novel SCA contraindication information from historical data in electronic health records.

Like adverse events [100], contraindication relations are usually uncertain in electronic health record data. They can be suspected, but words indicating the suspicion are not stated. To help identify adverse events from electronic health record data, people have used adverse event monitoring rules [70], such as those related to abnormal laboratory test results, abnormal physiological data, and orders for known antidotes. People have also used keywords related to adverse effects [100], such as "stop" and "change." Similar techniques can be used to facilitate extracting SCA contraindication information from electronic health record data.

4.4.2 Approach 2: Using medical message board posts

As mentioned in Benton *et al*. [15], many medical message boards contain a large number of candid messages and have moderators to remove spam messages, resulting in an inexpensive and relatively trustworthy way to learn more about a given population. Benton *et al*. used text mining and statistical techniques to detect potentially novel adverse drug events from medical message board posts [15]. Contraindication is one type of harm, whereas harm is often discussed on medical message boards. Hence, an approach similar to that in Benton *et al*. [15] can be used to extract potentially novel SCA contraindication information from medical message board posts. For example, consider a particular SCA and a specific health issue. The SCA is used to manage other health issues but not this specific health issue. There is no published report documenting that the SCA contraindicates with the health issue. Suppose we discover that the SCA is mentioned together with an adverse event a number of times on a message board pertaining to the health issue. If the adverse event is not a known side effect of the SCA, we may then suspect that a potentially novel contraindication relationship exists between the SCA and the health issue.

4.5 Providing incidence statistics of the identified contraindications

In addition to automatically identifying contraindicated SCAs, it would also be beneficial for iPHR to provide incidence statistics of the identified contraindications. Some contraindications are relative, meaning that the client is at a higher risk of treatment complications, but the risk may be outweighed by other considerations or mitigated by other protective measures. For example, valve heart disease is a relative contraindication for the SCA of aerobic exercises because stress tolerance is low in severe valve heart disease. For each identified relative contraindication of a particular SCA, iPHR could provide its incidence statistics reflecting the likelihood of developing the corresponding treatment complications. This can help the user assess the efficacy to risk ratio of this SCA and make informed, balanced decisions [50]. For instance, when a user needs to make a choice between two SCAs that are equally efficacious, she can choose the SCA with less severe treatment complications and/or a smaller probability of developing treatment complications [36]. One possible way to obtain incidence statistics of relative contraindications is to conduct text mining and data mining using historical data in electronic health records [70, 100, 148].

4.6 Complex contraindication information

The above discussion on contraindications focuses on the case that each contraindication is related to a single health issue. In general, the concept of contraindication is not limited to the case of a single health issue. A SCA can be contraindicated for the combination of two or more health issues, but not contraindicated for either single health issue. For example, exercise is absolutely contraindicated if a woman is pregnant and also has primary pulmonary hypertension, whereas most moderate exercise is not contraindicated for either pregnancy by itself or primary pulmonary hypertension by itself [1]. To extract complex SCA contraindication information related to the combination of two or more health issues, additional techniques beyond what is discussed above are likely to be needed.

5. Personalized search for individual healthcare providers

In this section, we discuss the potential new function of iPHR for personalized search for individual healthcare providers (IHPs), such as physicians, dentists, and dietitians. The user can use this function to find an IHP who is likely to be good at managing his health issues and serve his needs well.

5.1 The need and challenges for finding satisfactory IHPs

Consumers frequently change their IHPs. As reported in [18], in a 2-year time period, people in almost 50 million U.S. households selected or switched to a new physician. The statistics provided by RAND Health [116] is similar: 38% of Americans changed doctors within the past two years. 36% of people who changed physicians encountered difficulty in finding a new one they like [116]. For people in fair or poor health, this percentage increases to 55%. This phenomenon is undesirable. The degree of satisfaction of a patient with his IHP correlates with his health outcome. A patient who is more satisfied with his IHP is more likely to listen to his IHP's advice and realize various health benefits including medication compliance, adopting a healthy lifestyle, and complying with preventive measures [128, 139].

In general, patients prefer owning the responsibility of selecting new physicians rather than being assigned to new physicians. Those patients who choose new physicians on their own are more satisfied, more trusting, and more likely to adhere to physicians' recommendations [51, 56, 76, 138, 142]. Moreover, consumers are willing to spend time on finding IHPs. For example, before the Web came into existence, 25% of people who were not referred spent more than two weeks searching for an IHP [74].

As the Internet becomes widely available, online search is growing in popularity as a method for finding IHPs. For instance, 20% of visitors to the major health Web search engine Healthline [52] use its DocSearch tool [53, 127]. In addition to this tool, several other online IHP search tools have been developed for consumers, such as UCompareHealthCare [144] and Insider Pages Doctor Finder [67]. Consumers can also check physician information on physician rating Web sites, such as those listed in Luo [86].

At present, consumers have access to only limited information on IHPs. They are interested in obtaining more information that can be used in the IHP selection process. According to several surveys, 32% of people found it rather difficult to find needed information on IHPs [116]. Two thirds of adults wished they could find more comprehensive information on IHPs online [127]. Over 85% of consumers regard information about the quality of care provided by IHPs as essential [133], but this information is often unavailable.

Frequently, a consumer would like to determine which IHP is best for her based on her personal condition [127]. Partly due to a lack of sufficient information on IHPs, existing tools cannot facilitate this task well. In fact, many consumers attribute the finding of a satisfactory IHP to good luck [18]. It would be desirable for iPHR to provide a function for personalized search for IHPs to help users find satisfactory IHPs.

5.2 Overview of our proposed solution for personalized search for IHPs

To illustrate our idea of building iPHR's function for personalized search for IHPs, consider a user with a particular health issue, such as asthma. Intuitively, all other things being equal, an IHP with a lot of patients having this health issue is more likely to be familiar with it and hence more suitable for managing the user's case [64]. If many of the IHP's patients have good outcomes, the likelihood that the IHP can provide satisfactory service to the user becomes even higher. For each IHP, both the number of her patients with this health issue and their health outcomes can be obtained from historical data stored in the provider's electronic medical record system. We use this information to predict both the outcome measure of the user and his degree of satisfaction with the IHP if he is managed by the IHP. All available IHPs are sorted according to a combination of the two predicted values and then presented to the user. The higher-ranked IHPs are more likely to satisfy the user as well as help him achieve a good health outcome.

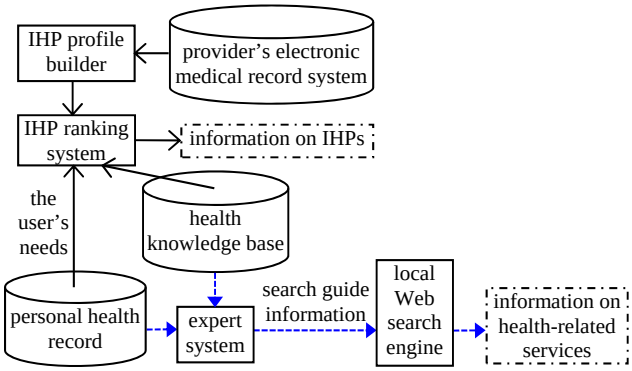

Fig. 4 Architecture of iPHR for the two potential new functions.

Based on the intuition mentioned above, our high-level idea of building this function of iPHR is to use health knowledge and introduce and extend data mining technology [57] and recommender system technology [121] into the PHR domain. To obtain more information on an IHP and expand her profile, the historical data of all of her patients is first extracted from the provider's electronic medical record system and then aggregated as the extended part of her profile. The needs of the user are obtained from the information stored in his PHR, the priorities specified by him, his desired type of IHP, and his inputted preferences. Then using health knowledge stored in the health knowledge

base, all available IHPs of the user's desired type are sorted according to how well their profiles match his needs. The solid arrow paths in Fig. 4 illustrate the workflow of this function. The dashed arrow paths in Fig. 4 illustrate the workflow of iPHR's potential new function for personalized local search for health-related services. The description of that workflow is provided in Section 6.4.2.

Our approach requires extracting data from the provider's electronic medical record system. This is doable, e.g., when the PHR in iPHR is an integrated health record connecting to the provider's electronic medical record system [69, 125]. Several major PHRs are integrated health records. For example, EpicCare of Epic Systems Corporation is the most widely used electronic medical record system in the U.S. Epic's PHR, MyChart [34], connects to EpicCare. As a second example, the Veterans Health Administration is the largest healthcare system in the U.S. Its PHR, My HealtheVet [125], connects to its electronic medical record system VistA. There are multiple ways to extract data from the provider's electronic medical record system. One way is to use its data export function. Another way is to use HL7 messaging.

In the following, we describe the three steps of our proposed algorithm for personalized search for IHPs one by one.

5.3 Step 1: Building a profile for each IHP

The profile of an IHP consists of two parts: a basic part containing standard information about her and an extended part constructed from all of her patients' historical data in the provider's electronic medical record system.

The basic part of the profile of an IHP provides conventional information about her, including some or all of the following items [133]: whether new patients are accepted, age, board certification, disciplinary actions, whether email communication with patients is provided, faculty appointment/academic affiliation, gender, office location(s), honors and awards, insurances accepted, languages spoken, malpractice suits, affiliation with specialists and hospitals, medical school attended, office hours, frequently asked questions, practice philosophy, publications, race/ethnicity, residency training, specialty, and years in practice. These items are available from various sources, such as the public Web site of the health insurance plan and the provider's administration information system.

The extended part of the profile of an IHP is automatically obtained by aggregating all of her patients' historical data in the provider's electronic medical record system. For each health issue, the number of her patients with it, the average outcome measures of these patients, and the average healthcare cost of these patients are included in the extended part. In addition, the extended part encompasses some or all of the following items: number of specific procedures performed, average waiting time for appointment, and average waiting time in the office. Some of these aggregated values appear in the report card on her. Periodically, new data is extracted from the provider's electronic medical record system to re-compute the aggregated values. Incremental maintenance methods from the data stream literature [49] are applied to reduce re-computation overhead and maximally reuse previous computation results.

The idea of using patients' historical data in the provider's electronic medical record system to build profiles for healthcare professionals has been previously adopted in other application scenarios. In the nursing domain, Hall and Thornton [64] used this idea to obtain two kinds of counts for each nurse. The first kind of count is the number of patients who had been managed by her and were of a particular type, such as with a body mass index greater than 35 kg/m^2. The second kind of count is the number of times that she documented the application of a specific care activity.

In the medical domain, Neuvirth *et al*. [111] used this idea to predict the future health condition of each diabetic patient and identify the diabetic patients at high risk. For each physician, two sets of features are computed. In the first set, each feature is the mean of a feature of her diabetic patients with desirable outcomes. In the second set, each feature is the mean of a feature of her diabetic patients with undesirable outcomes. To the best of our knowledge, what is described in this section is the first attempt ever of using this idea to facilitate consumers to perform personalized search for IHPs.

Aggregation is a widely used technique in data mining for reducing the size of a data set [57]. Due to the special properties of our IHP search application scenario, multiple factors need to be considered in performing aggregation to construct the extended part of the profile of each IHP.

Number of an IHP's patients with a particular health issue

As mentioned in Hall and Thornton [64], to take good care of a patient with a particular health issue, ideally the nurse in charge should have recent experience managing other patients with the same health issue. Naturally, this intuition would also apply to IHPs. To factor in this intuition, the time t_l when an IHP last saw a patient with a particular health issue H is used in computing a primitive number of the IHP's patients with H. One way to do this is to multiply the count from the patient by a factor $e^{-\alpha_H(t_c-t_l)}$ that is always between zero and one so that this primitive number is of the form $\sum e^{-\alpha_H(t_c-t_l)}$. Here, α_H is a positive constant controlling the degree of discounting over time for H. t_c is the current time. The longer the time that the IHP last saw the patient, the smaller this factor.

Intuitively, an IHP tends to have a deep impression of a case of a rare health issue and his memory of this case is likely to degrade slowly. Hence, a less prevalent health issue is given a smaller α_H. As described in Section 5.5, in ranking IHPs, classification or regression is performed to

predict the outcome measure of a patient, his degree of satisfaction with an IHP, and his healthcare cost. The value of α_H is learned through conducting machine learning to maximize the classification accuracy or to minimize the regression error.

Differing IHPs see different numbers of patients per week. For example, an IHP who has a faculty appointment with a lot of research responsibilities may see patients for only one day per week. In contrast, a pure clinician may see patients for five days per week. For each IHP and each health issue *H*, the primitive number of her patients with *H* is computed in the way mentioned above. Then a normalized number of her patients with *H* is computed by dividing the primitive number by the average number of patients that she sees per week. Alternatively, a normalized number is computed by dividing the primitive number by the average number of hours for which she sees patients per week. This normalized number is comparable across different IHPs and used in the extended part of her profile.

Average outcome measures of an IHP's patients with a particular health issue

Each health issue has its own set of outcome measures, such as death, complication, functional status, blood pressure level, and blood sugar level (hemoglobin A1c). iPHR's function for personalized search for IHPs uses outcome measures that can be expressed as numbers, possibly after some transformation if needed. For each outcome measure related to a health issue together with each IHP, an average value of this outcome measure is computed using the historical data of all of her patients with the health issue. This average outcome measure may need to be adjusted, e.g., based on the patients' risks, so that it can become comparable across different IHPs [129].

Average healthcare cost of an IHP's patients with a particular health issue

The healthcare cost of a patient is computed based on multiple items in his historical data, such as medications, lab tests, procedures, the amount of time of each of his visits with IHPs, and the number of his visits with IHPs.

5.4 Step 2: Obtaining the user's needs

The needs of the user are reflected by the content of his profile, the desired type of IHP he inputted, and his priorities. His profile consists of two parts: a basic part automatically built from the information stored in his PHR and an extended part describing his preferences.

The basic part of the profile of the user provides fundamental information about him and includes some or all of the following items: health issues of concern by him, age, gender, home address, health insurance, and languages spoken. The health issues of concern by him are based on his current health issues and obtained in a way similar to that in Luo *et al*. [92].

The extended part of the profile of the user encompasses his preferences on various properties of the IHP, such as age range, gender, and medical school attended. All of these properties are included in the basic part of the profile of the IHP. The user is provided with an option to manually input these preferences. If this option is not used, the default settings of these preferences are learned from the preferences of other users similar to him using the concept of patient similarity [151].

As described in Section 5.5, matching between each IHP's profile and the user's needs is based on one or more criteria. Each criterion *c* has an associated priority with a corresponding pre-specified weight w_c. The user is provided with an option to manually specify the priority. If this option is not used, the default level of the priority is learned from the priority levels of other users similar to him.

Various users have different priorities [133]. For example, a healthy user may care more about the convenience of the IHP's office location. In contrast, an ill user may care more about the IHP's experience with a particular health issue and its corresponding outcomes. The priorities of a user can change over time and his past priorities may not be able to predict his current priorities [87]. For instance, this can be the case when his health or financial status changes significantly [74]. Thus, it is important to allow the user to manually specify or change his priorities at any time.

For two criteria that are of the same type, the same level of priority may correspond to different weights. One example of when this holds true is when the two criteria are about different health issues, because differing health issues usually do not have equal importance.

5.5 Step 3: Ranking IHPs by matching their profiles with the user's needs

We sort IHPs according to how well their profiles match the needs of the user and then present the sorted list of IHPs to him. During this ranking process, we consider only those IHPs who accept new patients and are of the type desired by him, such as primary care physician. For each IHP shown to him, multiple attributes in her profile are selected for presentation based on his needs. Some of these attributes, such as an average outcome measure related to one health issue of concern by him, are quality measures of her. Displaying quality measure data is part of the healthcare industry trend. For example, according to the Affordable Care Act, by 2013 the Department of Health and Human Services needs to develop and implement a plan for reporting physician-level quality measure data on the new Physician Compare Web site http://www.medicare.gov/find-a-doctor/provider-search.aspx [11]. An attribute value will not be presented if its disclosure can potentially violate the privacy of another patient [21]. For instance, this can be the case if an IHP has a very small number of patients with a specific health issue,

the number of patients is presented, and the attribute value is computed based on these patients.

The degree of matching between the profile of an IHP and the user's needs is computed using a fixed set F of criteria, such as the outcome measure of the user, the user's degree of satisfaction with the IHP, the user's healthcare cost, and the distance between the IHP's office location and the user's home. Let F' denote the set of criteria in F that are relevant to the user and have non-empty utilities. For each criterion $c \in F'$ whose weight is w_c, we compute a utility u_c reflecting how well the IHP's profile matches the user's needs solely based on c. The degree of matching is computed using multi-criteria recommender system techniques [121], e.g., as a normalized weighted sum of all of these utilities: $\sum_{c \in F'} w_c \cdot u_c / \sum_{c \in F'} w_c$.

Computing utilities of the criteria

The utilities of some criteria are computed by performing machine learning classification or regression using multiple features. Since missing values are prevalent in our case, we use classifiers and regression functions, such as decision trees [156], which are good at handling missing values. For the criterion of the user's degree of satisfaction with the IHP, its utility is the predicted degree of satisfaction if the user is managed by the IHP. The classifier or regression function for its utility is trained using historical patient satisfaction survey data, data of the patients, and data of the IHPs. The case with the criterion of an outcome measure of the user and the case with the criterion of the user's healthcare cost are handled in a similar way.

As is the general case with machine learning, our predicted values are not 100% accurate. Consequently, the top ranked IHPs who are computed based on the predicted values may or may not be the optimal ones for the user. Nevertheless, if our predicted values are reasonably accurate, we would expect that the suitability of the top ranked IHPs for the user will be close to that of the optimal IHPs for the user, or at least it is rather unlikely that the top ranked IHPs will be the least favorable ones for the user. This is similar to the case in database query optimization. The query optimizer usually does not have very accurate cost estimates of different query execution plans, but can choose a reasonably good query execution plan based on the cost estimates [155].

Computing feature values

For some features, the value of each of them is computed based on the value of an attribute in the IHP's profile. One such feature is the logarithm of the normalized number of an IHP's patients with a particular health issue. Here, a logarithm is taken to reduce the difference in the number across different IHPs. For the other features, the value of each of them is computed using both the value of an attribute in the IHP's profile and the value of the same attribute shown in the user's needs. One such feature is related to the distance between the IHP's office location and the user's home. Another such feature is related to whether the IHP and the user are of the same gender [132].

In computing the values of certain features, the problem of insufficient data needs to be addressed. For example, consider a feature that is the average outcome measure of an IHP's patients with a specific health issue H. If H is relatively rare, the IHP may have only one or two patients with H and hence the average outcome measure may not well reflect her capability of managing H [133]. To handle this issue, one of the following two approaches can be used.

In the first approach, if the primitive number of an IHP's patients with the health issue H is smaller than a pre-determined threshold specific to H, the average outcome measure computed from all IHPs' patients with H is used as the average outcome measure of H for the IHP.

In the second approach, the medical terminology of SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) [152] is used to generalize the health issue H to a category of health issues covering H. If the primitive number of an IHP's patients with H is smaller than a pre-determined threshold specific to H, the average outcome measure of H for the IHP is computed on all of her patients with a health issue in the category.

Cole *et al*. [26] have used SNOMED CT before to build a Web search engine for helping consumers find physicians with particular expertise. There, each physician has a Web page describing her expertise. When a user inputs a health issue H into the Web search engine, if no physician's Web page mentions H but some physicians' Web pages mention the category of health issues covering H, these Web pages are returned to the user.

Some advanced issues

When ranking IHPs, the cold start problem needs to be addressed. Consider an IHP, such as a new one in the clinic, who has zero or few patients with a specific health issue H. It would be undesirable to rank her lower among all IHPs just because other IHPs have already treated many patients with both H and good outcomes. Otherwise, this IHP will have little chance to obtain new patients with H and demonstrate her capability of managing H. To take care of this cold start problem, we can borrow some techniques developed for it in the area of recommender systems [121].

In the above discussion on matching IHPs with the user, each feature used is related to at most one health issue. In addition to these features, we can also use other more complex features, each of which is related to two or more health issues. For example, one such feature is the number of an IHP's patients who have both health issues H_1 and H_2. Consider the following scenario with two IHPs I_1 and I_2. The user has both H_1 and H_2. I_1 has two patients, one with H_1 only and the other with H_2 only. I_2 has two patients, one with both H_1 and H_2 and the other has no health issue. Intuitively, all other things being equal, we would expect I_2 to match better with the user than I_1.

6. Personalized local search for health-related services

In this section, we discuss the potential new function of iPHR for personalized local search for health-related services (HRSs) to help maintain patients in their homes [54]. The user can use this function to find HRSs that are relevant to his specific health issues and available in his local community. Some examples of these services are professional home health services, rehabilitation services, social work services, mental health services, the Meals on Wheels program, home grocery & supermarket delivery services, house cleaning and maid services, chore services, personal emergency response system, durable medical equipment suppliers, transportation services, personal care services, home remodeling services for disabled people, legal services for seniors, financial aid for elderly care, the respite program, the adult day care program, senior centers, nursing homes, assisted living facilities, hospice, alternative healthcare services (e.g., acupuncture, massage, meditation, and nutrition consulting), health clubs, healthy living and health education programs (e.g., cardiovascular disease prevention, smoking cessation, weight management, stress management, flu shot, healthy cooking, yoga, and Tai Chi), health-issue-specific helpline, and local/online support/advocacy groups for a particular health issue. In these support/advocacy groups, people exchange miscellaneous information and provide various kinds of (mutual) help related to the particular health issue.

6.1 The need for finding local HRSs

Consider the following exemplary scenario. A senior citizen lives alone whereas his only daughter lives and works at a place thousands of miles away. One day, he develops a serious health issue and is sent to stay in a hospital for a few days. Subsequently, his daughter is notified and attempts to make various arrangements related to the health issue in preparation for his transition from the hospital back to his home. In this case, suddenly, in a short amount of time, and possibly remotely, she needs to figure out what kinds of HRSs are available in his local community and can be used to facilitate his well-being and keep him safe in his own home.

However, since his daughter lives far way, she is unaware of such local information and also has no idea about who to contact for this information. The discharge nurse and hospital social workers may offer some help, but the information they provide on local HRSs is usually limited [29]. The range of HRSs is so vast that each healthcare professional typically knows only a small portion of it and is knowledgeable of only a few local HRS providers. As new HRSs arise at various locations while others are closed, it is a challenge for healthcare professionals to keep up with current information. For these reasons, his daughter needs to work out the search process mainly by herself.

Scenarios similar to the one above are prevalent. 5% of American adults need long-term care services [72]. 18% of American children under age 18 have special health care needs due to chronic conditions and need HRSs [79]. Consumers and caregivers frequently want to know the HRSs that are relevant to a specific health issue and available in a particular local community. For a person with one or more health issues, several different types of HRSs useful for him are usually available in his local community. The more health issues he has and the more severe his health issues are, the more HRSs he is likely to need and use.

The effective use of HRSs is a key component of care after hospital discharge. It can improve patients' quality of life and health outcomes, reduce risk of readmission, delay institutionalization, and alleviate caregiver burden [29]. In fact, the quality of care transitions including follow-up care is so critical that the Centers for Medicare and Medicaid Services is establishing programs, often coupled with financial incentives, to improve care transitions [11].

6.2 Challenges for finding local HRSs

Information on HRSs is currently scattered in numerous sources. Several consumer-oriented health information books, such as the series of books entitled "The Comfort of Home" [103], include incomplete collections of this information. This is also the case with multiple Web sites, such as those of some government agencies, public libraries, and health-issue-specific organizations [150]. For a particular Web site, the information on it is usually tailored to the mission of its owner and hence has a limited scope of coverage. For example, Salt Lake County Aging Services is the largest Area Agency on Aging in Utah and lists its own programs and services at http://aging.slco.org/html/programs_and_service.html. As a second example, the Alzheimer's Association Utah Chapter lists a few resources related to Alzheimer's disease at http://www.alz.org/utah/. At present, no single book or Web site offers a comprehensive collection of information on HRSs [9].

About 40% of American Internet users have searched for HRSs online [123, 134]. However, it is challenging to conduct these searches, e.g., using a large-scale Web search engine such as Google. According to several surveys, 54% of parents of children with chronic conditions [40], 40% of people with disabilities [101], and 22% of the parents of children or youth with special health care needs who called the Parents' Place of Maryland for support [9] required help finding information on essential HRSs. The most important reason for this difficulty is that consumers are unaware of what types of HRSs are available for a health issue [9, 29]. For instance, in a survey conducted on seven commonly-used types of HRSs, the percentage of caregivers for American seniors who were unaware of the specific type of needed HRS varied between 17% and 40% [29].

There are many different types of HRSs. Each health issue has its own unique set of relevant types of HRSs. To find a comprehensive set of HRSs that are both relevant to a specific health issue and locally accessible, one needs to use broad health knowledge to form appropriate and targeted queries and perform multiple searches, e.g., one search per relevant type of HRS. This exceeds the capability of the average consumer.

Moreover, even if a consumer knows a particular type of HRS that he needs, it can be non-trivial for him to find the HRSs that are of this type and available in his local community. Multiple HRSs are of the same type if they are either essentially the same or similar to each other. Frequently, the same type of HRS is referred to by different names in differing places. The average consumer is unlikely to know all of these names and often has difficulty in figuring out the appropriate name for a particular type of HRS within his local community.

Here are some examples:

(1) The home grocery & supermarket delivery service helps a person buy various items from grocery stores and supermarkets, and then delivers these items to his home. This service is called the home shopping service in several locations.
(2) The house cleaning and maid service is referred to as the home cleaning service in some areas.
(3) A health club is often called a fitness center or a gym.
(4) Long-term care facilities are also known as nursing homes.
(5) The Meals on Wheels program delivers meals to individuals at home who are unable to purchase or prepare their own meals. This program is termed the home-delivered meals program in certain areas. There are other places that provide this program for seniors and call it Meal Delivery Service for Seniors. Several other places provide similar programs such as the Senior Citizen Nutrition Program or Gray Gourmet, where seniors can eat almost for free at one or more specified locations. If specifically needed, home delivery service of meals can be arranged.

Consumers can conduct geographically constrained searches using local Web search engines [28, 81, 88, 102], such as Google Maps [44], Yelp [154], and location-based search services for smart phones. However, for the same reasons as mentioned above, it is difficult for a consumer to use an existing local Web search engine to find a comprehensive set of HRSs that are relevant to his specific health issues and available in his local community. This is different from the case of a local search for an IHP. In that case, the consumer usually knows the desired type of IHP [127], such as primary care physician. Each type of IHP has its own name that typically does not vary from one location to another. Hence, the consumer can easily find a list of IHPs that are of this type and available in his local community.

6.3 The MedlinePlus Go Local project

To help consumers find local HRSs, the National Library of Medicine supported the MedlinePlus Go Local project [63] between 2001 and 2010. This project attempted to build directories of HRSs in various locations using a manual approach. One result of this project is the NC Health Info [109] Web site, which collects many HRSs in the State of North Carolina, has more than 1,140 Web sites linking to it, has consistently experienced good Web traffic of about 30,000 page views per month, and keeps operational today [134].

The MedlinePlus Go Local project created a controlled vocabulary of HRSs as well as a set of linkages between health issues and HRSs [134]. Each term in the controlled vocabulary corresponds to a particular type of HRS. Each health issue is linked to zero, one, or more terms in the controlled vocabulary corresponding to the types of HRSs relevant to it. A linkage is formed between it and every such term. Some health issues, such as acoustic neuroma, fainting, and fever, have no linked term because no HRS is specifically relevant to any of them [63].

The MedlinePlus Go Local project constructed a federated database of HRSs at various locations. The user could conduct local search of HRSs stored in the federated database in one of several ways, e.g., based on the health issue, based on the location, or based on the type of HRS.

The federated database of HRSs was created and maintained manually. For each business or organization providing HRSs, a record was first created and then updated once every six months. This record contained a pre-defined set of information about the business or organization, including its service areas, its contact information, and the HRSs provided by it. Typically, a cataloger read Web pages about the business or organization to obtain this set of information.

At the time the record was created, the cataloger classified each HRS provided by the business or organization first into a term in the controlled vocabulary of HRSs and then into one or more pairings between health issues and this term. The HRS is relevant to the health issue in each pairing. This manual classification was challenging [126], but addressed the retrieval difficulty caused by the same type of HRS being called different names in various places.

During record maintenance time, the cataloger reviewed the information in the record and kept it current, e.g., by adding new HRSs or deleting those HRSs that were no longer provided by the business or organization. This was typically done through rereading Web pages about the business or organization.

It was extremely labor intensive to manually create and maintain the federated database of HRSs. On average, a cataloger took 30 minutes to create a new record and 5 to 30 minutes to update a record already in the database once [41]. The larger the federated database, the more maintenance work was needed. Once the federated database

reached a certain size, the catalogers could only afford to maintain the existing records and had almost no time to create new ones [134]. Even at that stage, the federated database included only a small portion of all available HRSs [55]. A similar problem occurs in existing information & referral services, such as 2-1-1 (http://211us.org), that provide information on human services using manually constructed databases of human services.

The manual approach for building directories of HRSs requires excessive overhead and has great shortcomings. To keep current with the rapidly changing world with an affordable overhead, an automatic approach for indexing HRSs is essential. As an analogy, Google's automatic approach for indexing Web contents [14] has significant advantages over Yahoo!'s manual approach and eventually replaced it.

6.4 Overview of our proposed solution for personalized local search for HRSs

It would be desirable for iPHR to provide a function for personalized local search for HRSs. This function is valuable for not only ordinary consumers but also healthcare professionals. In providing Patient-Centered Medical Homes [94], healthcare professionals are expected to refer patients to a variety of HRSs but face substantial barriers to obtaining and maintaining robust knowledge of local HRS providers. Healthcare professionals can use this function to become familiar with local HRSs [71]. They can then use their health knowledge to help patients select appropriate and locally accessible HRSs, e.g., upon hospital discharge. In fact, healthcare professionals are among the most active users of consumer health information tools including local search tools for HRSs [4, 71].

For long-term success, this function should minimize its developers' manual overhead for building and maintaining a database of HRSs. To achieve this goal, we can reuse an existing local Web search engine and implement this function using an automatic approach. The local Web search engine is based on a large database of businesses and organizations. Upon receiving both the name of a HRS and a location, it performs keyword matching for the name and retrieves information about local businesses and organizations that provide the service around the location. The database of businesses and organizations was first created through crowdsourcing [19] and then is maintained continuously and mostly automatically.

6.4.1 Google's Web service for local search

One such local Web search engine is the Web service that Google provides for local search [45]. This Web service was launched in 2011 and is built on a database of over 80 million businesses and organizations around the entire world. Some of these businesses and organizations provide HRSs while the others do not. Google uses crowdsourcing, user moderation, and automatic crawling to minimize the database developers' manual overhead for building and maintaining the database and ensure the quality of the information in the database.

More specifically, for each business or organization, its owner enters its information into the database in the form of a pre-defined set of textual information items. The textual information items include name, address, phone number, email address, service area(s), brief description, business categories, operating hours, uniform resource locator (URL) of its Web site, if any, and additional details about it. Typically, the services provided by the business or organization are mentioned either in the brief description about it or on its homepage. The combination of the textual information items and the content on its homepage serves as the base for its profile stored in the database. Local search in the database is conducted on all such profiles. Moreover, Google users can provide reviews and ratings for a business or organization, which are also stored as part of its profile.

To keep the information stored in the database current while minimizing the database developers' manual maintenance overhead, Google adopts the following three methods. First, the owner of a business or organization can update its textual information items stored in the database at any time. Second, Google periodically crawls the homepage of each business or organization, if any, and automatically downloads its content into the database. Thus, Google can quickly identify any update on the homepage, such as the addition of a new service or the deletion of a service that is no longer provided by the business or organization. Third, if a Google user finds a problem in the information stored in the database, he can report it to Google. The database developers will fix it shortly thereafter.

6.4.2 Our key ideas for implementing personalized local search for HRSs

Our high-level idea of implementing iPHR's function for personalized local search for HRSs is to adopt an approach similar to the one used in iPHR's function for automatically recommending home health products [92]. Based on the user's health issues and location, both expert system technology and health knowledge are used to automatically form zero, one, or more queries corresponding to the various names of the relevant HRSs. These queries are then fed into the local Web search engine to retrieve HRSs that are relevant to the user's health issues and available in the user's local community. The dashed arrow paths in Fig. 4 illustrate the workflow of this function.

Our goal is to retrieve a comprehensive set of relevant HRSs available in the user's local community. To achieve this goal in the MedlinePlus Go Local project, the catalogers manually indexed HRSs by classifying every HRS provided by a business or organization first into a term in the controlled vocabulary of HRSs and then into one or more pairings between health issues and this term. This manual classification is rather time-consuming. Many

businesses and organizations provide HRSs. Also, a single business or organization often offers multiple HRSs.

In iPHR's function for personalized local search for HRSs, the developers are waived of this manual classification work. Instead, to achieve our goal of retrieving a full set of relevant HRSs available in the user's local community, our strategy is to first pre-compile a comprehensive list of the names of HRSs and then use this list to automatically form a complete set of queries corresponding to the various names of the relevant HRSs. Compared to manual classification, this can greatly reduce the amount of manual work needed because the total number of variations on names for all types of HRSs is much smaller than the total number of businesses and organizations providing HRSs. For example, the same type of HRS under the same name is often provided by numerous businesses and organizations in many different places.

6.5 User interface

The user interface of iPHR's function for personalized local search for HRSs is similar to that of iPHR's function for automatically recommending home health products. It consists of a topic-selection input interface [92] and a navigation output interface [84]. In this personalized local search function, multiple queries are used simultaneously to retrieve HRSs that are of different types and/or relevant to differing health issues. The navigation output interface helps the user move through the search results. This is different from the case of traditional local search, where all search results are retrieved by a single query and thus a navigation output interface is usually unnecessary.

6.5.1 Topic-selection input interface

To perform personalized local search for HRSs, iPHR needs to know the health issues of concern by the user, e.g., his current health issues. The topic-selection input interface automatically gathers this information from his PHR in a way similar to that in Luo *et al*. [92]. Recall that iPHR includes his PHR as one of its components. For each health issue of concern by him, he can indicate whether it is highly important to him.

6.5.2 Navigation output interface

The navigation output interface displays the businesses and organizations retrieved by iPHR. Each retrieved business or organization is in the user's local community and likely to provide one or more HRSs relevant to the user's health issues. As shown in Fig. 5, each search result Web page contains three parts. On the left side, a navigation hierarchy based on various categories is provided. In the middle, the retrieved businesses and organizations are listed sequentially. On the right side, these businesses and organizations are shown on a map centered at the user's location. The user can move, zoom in, or zoom out on this map in a way similar to that in Google Maps [44].

Fig. 5 The navigation output interface of iPHR's function for personalized local search for HRSs.

In the navigation hierarchy, the categories without any corresponding search result are omitted. When the user clicks a category, iPHR will display the retrieved businesses and organizations in it.

As shown in Fig. 6, the default navigation hierarchy has two levels: the first level for the health issues of concern by the user and the second level for the relevant types of HRSs. On the first level of the hierarchy, the health issues of concern by the user are sorted in descending order of their importance as specified by the user in the topic-selection input interface, in the same way as that in Luo *et al*. [92]. On the second level of the hierarchy, the types of HRSs relevant to a health issue are sorted alphabetically. When the user moves his mouse over a particular type of HRS, text will be displayed explaining why the type of HRS is relevant to the health issue, unless the reason is obvious to the average consumer.

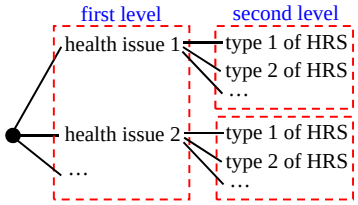

Fig. 6 The default navigation hierarchy with two levels.

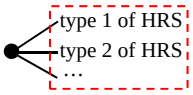

Fig. 7 The alternative navigation hierarchy with a single level.

By clicking a button, the user can switch to an alternative navigation hierarchy shown in Fig. 7. This navigation hierarchy has a single level for the types of HRSs relevant to the user's health issues. The types of HRSs are sorted alphabetically. When the user moves his mouse over a particular type of HRS, text will be displayed. For each

health issue that is of concern by the user and the type of HRS is relevant to, the text presents the health issue and explains the corresponding relevancy if it is non-obvious to the average consumer.

6.6 Our proposed algorithm for personalized local search for HRSs

Our proposed algorithm for personalized local search for HRSs consists of four steps. In step 1, we use an expert system equipped with a health knowledge base to convert the user's health issues into a set of queries. These queries are called "search guide information" and represent the various names of the HRSs relevant to the user's health issues. In step 2, we use each query and the local Web search engine to retrieve businesses and organizations that are in the user's local community and provide HRSs relevant to the user's health issues. In step 3, we combine and rank the businesses and organizations retrieved by the different queries, while taking into account various relevant factors. In step 4, the search results are diversified and returned to the user.

6.6.1 Step 1: Obtaining search guide information.

As mentioned in Section 6.5.1, the topic-selection input interface collects a list L_h of health issues of concern by the user. Both expert system technology and health knowledge are used to provide semantic translation from these health issues to the names of the HRSs relevant to them. The results are the search guide information for the local Web search engine. For each health issue $H \in L_h$, iPHR uses H's search guide information to perform local search for HRSs relevant to H.

Overview of the semantic translation

This semantic translation occurs in two sub-steps. In the first sub-step, for each health issue $H \in L_h$, all types of HRSs relevant to H are obtained. In the second sub-step, the various names of these types of HRSs are found. Both sub-steps use knowledge on HRSs, which is stored in iPHR's health knowledge base K_b as a set of pre-compiled information.

The MedlinePlus Go Local project [63] has already gathered much knowledge on HRSs, such as the controlled vocabulary of HRSs and the set of linkages between health issues and HRSs [134]. This knowledge is incomplete. For instance, it does not cover certain health issues, such as some rare ones. It also does not cover certain types of HRSs, such as home grocery & supermarket delivery services and house cleaning & maid services. Nevertheless, it serves as a good starting point for our knowledge compilation work. Additional knowledge is obtained through checking miscellaneous sources of information on HRSs. These sources include several consumer-oriented health information books, such as the series of books entitled "The Comfort of Home" [103], the Alliance of Information and Referral Systems (AIRS) taxonomy of human services [2], and the Web sites of NC Health Info [109], Seniors Blue Book [122], Medical Home Portal [94], some government agencies, and some health-issue-specific organizations.

Assembling the set of pre-compiled information on HRSs

The set of pre-compiled information that is on HRSs and stored in the health knowledge base K_b is assembled in the following way. Using the controlled vocabulary of HRSs developed in the MedlinePlus Go Local project [134] and the types of HRSs listed in the AIRS taxonomy of human services as the base, a healthcare professional compiles a list of various types of HRSs and stores it in K_b. For each type T of HRS, she compiles a set S_T of phrases and stores S_T in K_b. Each phrase in S_T represents one possible name that T is called. For example, if T is nursing home, both phrases "nursing home" and "long-term care facility" are included in S_T.

As mentioned in Ford and Hannon [39], consumers have difficulty knowing the exact meaning of the names of certain HRSs. To help users understand the meaning of each type of HRS, if needed, iPHR displays in the navigation output interface multiple names for the same type of HRS. More specifically, for each type T of HRS, a healthcare professional selects one or more phrases in the set S_T that are relatively easy for consumers to understand, uses slashes to concatenate these phrases into one single piece of text D_T, and stores D_T in the health knowledge base K_b. D_T will be displayed as the name of T in the navigation hierarchy of the navigation output interface. For instance, if T is the Meals on Wheels program, D_T is "Meals on Wheels / home-delivered meals."

Initially, it is difficult to manually construct a complete set S_T of phrases covering all possible names of a type T of HRS. To make S_T comprehensive, a bootstrap approach [5] can be used. From miscellaneous sources including the NC Health Info Web site [109], we obtain one or more names of T as well as the Web pages of some companies and organizations providing HRSs of type T. Each Web page mentions at least one HRS that is both of type T and provided by the corresponding company or organization. Each name of T is inputted as a query into a large-scale Web search engine such as Google to retrieve some of these Web pages.

We use these Web pages as a training set and conduct machine learning to learn multiple characteristic phrases that typically appear on such Web pages. A healthcare professional manually reviews the characteristic phrases to identify those that intuitively make sense. They are used to construct more queries. Each query combines one or more characteristic phrases and retrieves additional Web pages. The healthcare professional reviews the Web pages in attempt to discover new names of T mentioned on them. Each new name of T serves as a query to retrieve more such Web pages. Then we iterate this process of learning more characteristic phrases, retrieving additional Web pages, and discovering new names of T.

In addition to the bootstrap approach, iPHR can also use a crowdsourcing approach [19] to collect phrases that its users suggest in the form of the names of HRSs not found by iPHR. This can help make up the phrases that should be included in the set S_T but are missed by the bootstrap approach. A similar approach can be used to help make up the types of HRSs that are missed in the health knowledge base K_b.

The knowledge that the MedlinePlus Go Local project [63] gathered on HRSs covers many health issues. For each health issue H, the MedlinePlus Go Local project developed a set of linkages between H and HRSs [134]. A healthcare professional uses this set of linkages as the base to compile a list S_H of types of HRSs relevant to H and stores S_H in the health knowledge base K_b. For each type $T \in S_H$ of HRS, she specifies a weight $w_{T,H}$ indicating how important T is to H. One way to do this is to define multiple levels of importance, each with its own weight. She assigns T to a level of importance and obtains the corresponding weight as $w_{T,H}$. T's normalized importance to H is reflected by a normalized weight $n\_w_{T,H} = w_{T,H} / \sum_{V \in S_H} w_{V,H}$. Unless the reason is obvious to the average consumer, the healthcare professional compiles text $T_{T,H}$ explaining why T is relevant to H and stores $T_{T,H}$ in K_b. $T_{T,H}$ will be displayed when the user moves his mouse over T in the navigation hierarchy of the navigation output interface.

For a health issue not covered by the knowledge that the MedlinePlus Go Local project and other information sources gathered on HRSs, it can initially be difficult to obtain a comprehensive list of types of HRSs relevant to it. To address this issue, we can use similarities among various health issues. For instance, consider a new health issue for which we need to compile a list of types of HRSs relevant to it. Suppose an "old" health issue is similar to the new health issue, such as both health issues belong to the same category of health issues. A list of types of HRSs relevant to the old health issue has been previously compiled. Each type of HRS relevant to the old health issue is likely to be relevant to the new health issue, e.g., when the type of HRS addresses a property shared by both health issues. Hence, we can use the already-compiled list of types of HRSs relevant to the old health issue to facilitate compiling the list of types of HRSs relevant to the new health issue.

Forming search guide phrases

For each type $T \in S_H$ of HRS linked to a health issue H, a healthcare professional uses the set S_T of phrases pre-compiled for T to form one or more search guide phrases $S_{T,H}$ and stores $S_{T,H}$ in the health knowledge base K_b. There are two possible cases.

In the first case, the type T of HRS is context insensitive and not specifically tailored to the health issue H. Two examples of such types of HRSs are nursing home and the Meals on Wheels program. In this case, each phrase in the set S_T directly serves as a search guide phrase. We have $S_{T,H} = S_T$.

In the second case, the type T of HRS is context sensitive. The link between T and the health issue H indicates that T is specifically tailored to H. Two examples of such types of HRSs are support group and education program. In this case, the combination of each phrase in the set S_T and the name of H forms a search guide phrase in the set $S_{T,H}$. For example, if T is support group and H is breast cancer, the corresponding formed search guide phrase is "breast cancer support group." As a second example, if T is education program and H is diabetes, the corresponding formed search guide phrase is "diabetes education program."

In addition to the name of the health issue H, some synonyms of the name of H and some health issues that are hypernyms of (i.e., include) H are also used to form search guide phrases. These synonyms and hypernyms of H are found using the medical terminology of SNOMED CT [152]. The combination of each phrase in the set S_T and each synonym or hypernym of H forms a search guide phrase in the set $S_{T,H}$. For example, consider the case that the type T of HRS is support group and H is myocardial infarction. Heart attack is a synonym of myocardial infarction. Both ischemic heart disease and heart disease are hypernyms of myocardial infarction. In this case, "myocardial infarction support group," "heart attack support group," "ischemic heart disease support group," and "heart disease support group" are all included in $S_{T,H}$.

For each search guide phrase in the set $S_{T,H}$, if needed, one or more search operators [43] are added to it to reduce the likelihood of retrieving irrelevant search results. For example, the quote search operator tells the local Web search engine that matching should be performed for an exact set of words in a specific order.

The complete set of search guide information for all health issues in the list L_h is $C = \bigcup_{H \in L_h, T \in S_H} S_{T,H}$.

6.6.2 Step 2: Finding businesses and organizations that are in the user's local community and provide relevant HRSs.

From the PHR of the user, we automatically extract the location L in which he currently lives. By default L is used to perform local search. The user is provided with an option to change the location based on which local search is conducted.

Both the location L and the complete set C of search guide information are used to search for businesses and organizations that are in the user's local community and provide HRSs relevant to the user's health issues. Each phrase in C provides one way of retrieving relevant HRSs available in the user's local community. The search results for all phrases in C are combined together and returned to the user.

More specifically, the combination of each search guide phrase $G \in C$ and the location L is inputted as a query $Q_{G,L}$

into the local Web search engine to retrieve a set $S_{G,L}$ of businesses and organizations. Each business or organization is located near L and likely to provide one or more HRSs relevant to the user's health issues. These businesses and organizations are merged into a set $R_{all} = \cup_{G \in C} S_{G,L}$ as the search results that will be returned to the user.

For each query $Q_{G,L}$ such that $G \in C$, the local Web search engine returns a map scale s_G for displaying the set $S_{G,L}$ of businesses and organizations on a map in the output interface. We use the minimum one of these map scales, $\min_{G \in C} s_G$, for displaying all businesses and organizations in the set R_{all} on the map in the navigation output interface. This will ensure that all businesses and organizations in R_{all} appear on the map. When the user clicks a category of the navigation hierarchy, a similar approach is used to display the retrieved businesses and organizations in this category on the map.

In general, a business or organization in the set R_{all} can be retrieved by more than one search guide phrase in the set C. For each business or organization in R_{all}, we use the methods described in Luo [85] to obtain its snippet, i.e., some words extracted from its profile stored in the local Web search engine.

6.6.3 Step 3: Ranking businesses and organizations.

To properly rank businesses and organizations retrieved by multiple search guide phrases, we use health knowledge, consider various relevant factors, and fold all of these factors into a single formula in a way similar to that in Luo *et al.* [92]. In Section 6.6.4, this method is enhanced to provide diverse search results. In both this section and Section 6.6.4, our discussion focuses on the case of ranking all businesses and organizations in the set R_{all}. When the user clicks a category in the navigation hierarchy, a similar method is used to rank the retrieved businesses and organizations in the category.

Let N_h denote the user's general need of HRSs regardless of his location. N_h is reflected by the search guide phrases formed for all types of HRSs relevant to the user's health issues. Accordingly, N_h is written into a disjunctive form: $N_h = \vee_{H \in L_h, T \in S_H, G \in S_{T,H}} (H \wedge T \wedge G)$, where G is a search guide phrase in the set $S_{T,H}$. Recall that S_H is the list of types of HRSs relevant to the health issue $H \in L_h$. $S_{T,H}$ is the set of search guide phrases formed for the type $T \in S_H$ of HRS when T is linked to H. The disjunction operator reflects the fact that N_h is satisfied if any search guide phrase formed for a type of HRS relevant to the user's health issues is "hit." We have a conceptual query $Q_c = N_h \wedge L$ that includes the location L of the user and represents his overall need.

For each retrieved business or organization $B \in R_{all}$, a relevance score $score(B, Q_c)$ according to which B is ranked is computed. We start from the following probability computation:

$$\begin{aligned} p(B|Q_c) &= p(B|N_h, L) \\ &= p(N_h, L|B) \cdot p(B)/p(N_h, L) \\ &\propto p(N_h, L|B) \cdot p(B). \end{aligned} \quad (1)$$

As mentioned in Section 6.4.1, each business or organization B has a profile P_B stored in the local Web search engine. It is natural to assume that the location L of the user and his general need N_h of HRSs regardless of his location are conditionally independent of each other, given B or more precisely P_B. That is,

$$p(N_h, L|B) = p(N_h, L|P_B) = p(N_h|P_B) \cdot p(L|P_B). \quad (2)$$

Plugging Formula (2) into Formula (1) and ignoring the second- and higher- order terms, we have

$$\begin{aligned} & p(B|Q_c) \\ &\propto p(N_h|P_B) \cdot p(L|P_B) \cdot p(B) \\ &= p(\vee_{H \in L_h, T \in S_H, G \in S_{T,H}} (H \wedge T \wedge G)\,|P_B) \cdot p(L|P_B) \cdot p(B) \\ &\approx \sum_{H \in L_h, T \in S_H, G \in S_{T,H}} p(H, T, G|P_B) \cdot p(L|P_B) \cdot p(B) \\ &= \sum_{H \in L_h, T \in S_H, G \in S_{T,H}} [p(G|H, T, P_B) \cdot p(T|H, P_B) \cdot p(H|P_B) \cdot \\ &\qquad p(L|P_B) \cdot p(B)]. \end{aligned} \quad (3)$$

We make several additional natural assumptions as follows:

(1) The probability of generating the search guide phrase G depends only on the profile P_B. That is, $p(G|H, T, P_B) = p(G|P_B)$.

(2) The probability of selecting a type $T \in S_H$ of HRS depends only on the linked health issue H and is proportional to T's weight $w_{T,H}$ for H. That is, $p(T|H, P_B) = p(T|H) = w_{T,H} / \sum_{V \in S_H} w_{V,H} = n\_w_{T,H}$.

(3) In the same way as that described in Luo *et al.* [92], each health issue $H \in L_h$ is assigned a weight w_H reflecting its importance. H's normalized importance is reflected by a normalized weight $n\_w_H = w_H / \sum_{U \in L_h} w_U$. The probability of selecting H is independent of the profile P_B and proportional to H's weight. That is,
$$p(H|P_B) = p(H) = w_H / \sum_{U \in L_h} w_U = n\_w_H.$$

Under these assumptions, Formula (3) becomes

$$\begin{aligned} p(B|Q_c) \propto \sum_{H \in L_h, T \in S_H, G \in S_{T,H}} & [p(G|P_B) \cdot p(L|P_B) \cdot p(B) \cdot \\ & n\_w_{T,H} \cdot n\_w_H]. \end{aligned} \quad (4)$$

As the prior probability that the business or organization B is relevant to a query, $p(B)$ is often computed using the customer rating scores of B, the number of customer reviews on B, the click rate of B that is obtained from search logs [88], and link analysis results such as the PageRank score [114] of the Web site or the homepage of B. As the conditional probability of obtaining L given the profile P_B, $p(L|P_B)$ is computed based on the distance between L and B's location [28, 102]. As the conditional probability of producing the search guide phrase G given P_B, $p(G|P_B)$ is computed using both the text in P_B and some retrieval model such as the language modeling approach [92, 115].

We define the ranking score $score(B, Q_{G,L}) \triangleq p(G|P_B) \cdot p(L|P_B) \cdot p(B)$. When the combination of the search guide phrase G and the location L is inputted as a query $Q_{G,L}$, the local Web search engine uses $score(B, Q_{G,L})$ to rank

businesses and organizations retrieved for $Q_{G,L}$. To see this, we use the following probability computation:

$$\begin{aligned} p(B|Q_{G,L}) &= p(B|G,L) \\ &= p(G,L|B)\cdot p(B)/p(G,L) \\ &\propto p(G,L|B)\cdot p(B). \end{aligned} \quad (5)$$

Similar to the way Formula (2) is derived, we have

$$p(G,L|B) = p(G,L|P_B) = p(G|P_B)\cdot p(L|P_B). \quad (6)$$

Plugging Formula (6) into Formula (5), we have

$$p(B|Q_{G,L}) \propto p(G|P_B)\cdot p(L|P_B)\cdot p(B) = score(B,Q_{G,L}).$$

If the source code of the local Web search engine is available to us, we can obtain the ranking score $score(B,Q_{G,L})$ in a way similar to that in Luo *et al.* [92]. In case of too many search results, some top-*k* method can be used to speed up the ranking process. For example, we can use an approach similar to that in Cong *et al.* [25] to obtain the top-*k* search results for each search guide phrase $G \in C$. All of these search results are then merged together to obtain the overall top-*k* search results for the conceptual query Q_c.

In our case of iPHR, the ranking score $score(B,Q_{G,L})$ computed by the local Web search engine is unavailable to us. To overcome this difficulty, an approach similar to the one described in Luo [85] is used. As a rough approximation we assume that this ranking score is inversely proportional to $rank(B,Q_{G,L})$, the rank of the business or organization B among all businesses and organizations that the query $Q_{G,L}$ retrieves from the local Web search engine. That is,

$$score(B,Q_{G,L}) \propto 1/rank(B,Q_{G,L}).$$

This assumption reflects the fact that $score(B,Q_{G,L})$ decreases as $rank(B,Q_{G,L})$ becomes larger. If B is in the set $S_{G,L}$ of businesses and organizations retrieved by $Q_{G,L}$, $rank(B,Q_{G,L})$ is available to us. Otherwise, if $B \notin S_{G,L}$, we take $rank(B,Q_{G,L})$ to be infinite so that $score(B,Q_{G,L}) = 0$. Then Formula (4) becomes

$$p(B|Q_c) \propto \sum_{H\in L_h, T\in S_H, G\in S_{T,H}, B\in S_{G,L}} \frac{n\_w_{T,H}\cdot n\_w_H}{rank(B,Q_{G,L})}. \quad (7)$$

Formula (7) is used to compute the relevance score $score(B,Q_c)$.

6.6.4 Step 4: Diversifying search results.

If we only use the method described in Section 6.6.3 to rank the retrieved businesses and organizations, the top-ranked businesses and organizations can easily concentrate on a few relevant types of HRSs rather than all relevant types of HRSs. For example, this is the case if the businesses and organizations retrieved by one search guide phrase in the set *C* are ranked higher than the businesses and organizations retrieved by the other search guide phrases in *C*.

In the past, studies have shown that searchers usually prefer diverse search results [33, 92]. Ideally in our case, the first few businesses and organizations returned should cover as many health issues in the list L_h and as many relevant types of HRSs as possible.

To provide diverse search results, we enhance the ranking method described in Section 6.6.3 in a way similar to that in Luo *et al*. [92]. The set R_{all} contains $|R_{all}|$ retrieved businesses and organizations sorted in descending order of their relevance scores. We use a constant N=1,000 to control the amount of time spent on search result diversification, re-rank the top $J = \min(N, |R_{all}|)$ businesses and organizations in *J* passes, and generate one result page of ten diverse businesses and organizations at a time. In each pass, we pick a business or organization that strikes a balance between two factors: (1) having a large relevance score and (2) providing a balanced coverage of different health issues in the list L_h, various relevant types of HRSs, and differing search guide phrases in the set *C*. These two factors are combined to re-compute the relevance score. The concrete method is as follows.

We form two sets: $S_{remaining}$ and $S_{returned}$. At any time, $S_{remaining}$ contains the businesses and organizations remaining to be returned to the user, while $S_{returned}$ contains the businesses and organizations already returned to the user. Initially, $S_{remaining}$ contains the top *J* businesses and organizations with the largest relevance scores, whereas $S_{returned}$ is empty.

In the *i*-th ($1 \le i \le J$) pass, the business or organization $B_l \in S_{remaining}$ with the largest relevance score is moved from $S_{remaining}$ to $S_{returned}$ as the *i*-th business or organization returned to the user. For each type *T* of HRS relevant to a health issue $H \in L_h$, appropriate discounts are given to the weights and normalized weights related to *T* if B_l is retrieved by at least one search guide phrase formed for *T* (i.e., $\exists G \in S_{T,H}$ s.t. $B_l \in S_{G,L}$). Specifically, *T*'s normalized weight $n\_w_{T,H}$ is discounted by d_T. *H*'s normalized weight n_w_H is discounted by d_H. d_T and d_H are two constant factors whose default values are both 0.5. According to Formula (7), the relevance scores of the businesses and organizations in $S_{remaining}$ depend on both $n\_w_{T,H}$ and n_w_H, and thus need to be re-computed. As a result, the more businesses and organizations that provide HRSs relevant to the health issue *H* and have been returned to the user, the less likely the next returned business or organization will provide HRSs relevant to *H*. A similar property exists for types of HRSs.

6.6.5 Retrieving HRSs unavailable from the local Web search engine

The local Web search engine misses certain types of HRSs, such as online support/advocacy groups for a particular health issue, because they are not stored in its database of businesses and organizations. To retrieve such types of HRSs in the user's local community, we can form proper keyword queries and input them into a large-scale Web search engine such as Google.

Let L_u denote the location of the user at a fine level of geographical granularity, typically the town in which he lives. For each type *T* of HRS that is both relevant to a

health issue of his and unavailable from the local Web search engine, we combine every search guide phrase formed for T and the name of L_u to construct a keyword query. This query is fed into the large-scale Web search engine to retrieve Web pages on HRSs that are both of type T and in the user's local community. The search results are then merged together in a way similar to that described in Sections 6.6.3 and 6.6.4 and returned to the user in a separate section at the bottom of the navigation output interface.

It is possible that some HRSs that are both of type T and in the user's local community are for coarser levels of geographical granularities but not for L_u. To address this issue, L_u is upgraded to the user's location at a coarser level of geographical granularity, such as county. We repeat the above process to obtain more search results and add them at the end of the previously-obtained search results. Then if needed, more rounds of upgrading of L_u are performed to obtain more search results. The different levels of geographical granularities are pre-compiled beforehand for automatically generating the keyword queries.

7. Conclusions

The intelligent personal health record system iPHR is developed following a modular design approach [83, 85, 91]. Each function of iPHR is implemented as multiple software modules interacting with each other. The inputs and outputs of each module are defined in a way similar to that in the Substitutable Medical Applications, Reusable Technologies (SMART) Platforms project [107]. This increases the chance that with some modification, the source code of many functions of iPHR can be re-used as applications on other vendors' PHR systems, if these systems provide application programming interfaces similar to those in the SMART Platforms project. As a result, more benefits can be obtained from the efforts spent on developing iPHR.

Intelligent personal health record is a new and rapidly moving field. This paper presents the open issues we identified in the past three years in developing and operating iPHR. We outline some preliminary thoughts on how to address these open issues. To fully address these open issues, much more research work is needed. We hope this paper can stimulate future research work on iPHR.

In the future, we will investigate various techniques to address these open issues, implement these techniques, and evaluate their performance. Most of these techniques will fall into the areas of information retrieval, information extraction, data mining, and recommender systems. Hence, we would expect that many standard performance metrics in these areas, such as precision, recall, classification accuracy, and user satisfaction degree, can be used in our performance evaluation.

Acknowledgments

We thank Selena Thomas, Guilherme Del Fiol, Libin Shen, Xiaotong Zhuang, Mollie R. Cummins, Linda S. Edelman, Leslie A. Lenert, Lewis J. Frey, Stéphane M. Meystre, Susan Terry, Qing T. Zeng, Chuck Norlin, John F. Hurdle, and Scott P. Narus for helpful discussions.

References

1. ACOG guidelines for exercise during pregnancy. http://www.completefitness.com.au/articles/prepostnatal/acogguidelines.php, 2012.
2. AIRS taxonomy homepage. http://www.211taxonomy.org, 2012.
3. Agency for Healthcare Research and Quality (AHRQ). Coordinating care for adults with complex care needs in the patient-centered medical home: challenges and solutions. http://pcmh.ahrq.gov/portal/server.pt/gateway/PTARGS_0_11787_956295_0_0_18/Coordinating%20Care%20for%20Adults%20with%20Complex%20Care%20Needs.pdf, 2012.
4. Abrahamson, J.A., Fisher, K.E., and Turner, A.G. *et al*., Lay information mediary behavior uncovered: exploring how nonprofessionals seek health information for themselves and others online. *J Med Libr Assoc*. 96(4): 310-323, 2008.
5. Agichtein, E., and Gravano, L., Snowball: Extracting relations from large plain-text collections. *Proceedings of ACM DL'00*, pp. 85-94, 2000.
6. Aronson, A.R., and Lang, F., An overview of MetaMap: historical perspective and recent advances. *JAMIA* 17(3): 229-236, 2010.
7. Atack, L., Luke, R., and Chien, E., Evaluation of patient satisfaction with tailored online patient education information. *Comput Inform Nurs*. 26(5): 258-264, 2008.
8. Ananiadou, S., and Mcnaught, J., *Text Mining for Biology and Biomedicine.* Artech House, 2005.
9. Background brief executive summary: Community-based services organized for easy use (post summit). http://marylandcoc.com/uploads/User_Friendly_Systems_Outcome_Brief_22311.docx, 2008.
10. Batavia, M., *Contraindications in Physical Rehabilitation: Doing No Harm*. Saunders, 2006.
11. Burton, R., Improving care transitions. *RWJ Health Policy Brief*. September 13, 2012. http://www.healthaffairs.org/healthpolicybriefs/brief.php?brief_id=76.
12. Bulechek, G.M., Butcher, H.K., and Dochterman, J.M. *et al*., *Nursing Interventions Classification (NIC)*, 6th ed. Mosby, 2012.
13. Bernal-Delgado, E., and Fisher, E.S., Abstracts in high profile journals often fail to report harm. *BMC Med Res Methodol*. 8:14, 2008.
14. Brin, S., and Page, L., The Anatomy of a large-scale hypertextual Web search engine. *Computer Networks* 30(1-7): 107-117, 1998.

15. Benton, A., Ungar, L., and Hill, S. *et al*., Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. *JBI* 44(6): 989-996, 2011.
16. Cerner Multum solutions and services (including a drug disease contraindications API). http://www.multum.com/mdoc/Multum_Overview.pdf, 2012.
17. Clinical modules of FDB MedKnowledge (including a drug-disease contraindications module). http://www.fdbhealth.com/fdb-medknowledge-clinical-modules/, 2012.
18. Consumer Information and Education Committee**.** *How We Choose Doctors: What Is and What Could Be.* Midwest Business Group on Health, 2000.
19. Crowdsourcing. http://en.wikipedia.org/wiki/Crowdsourcing, 2012.
20. Chapman, W.W., Bridewell, W., and Hanbury, P. *et al*., A simple algorithm for identifying negated findings and diseases in discharge summaries. *JBI* 34(5): 301–310, 2001.
21. Cooper, T., and Collman, J., Managing information security and privacy in healthcare data mining: state of the art. In [22].
22. Chen, H., Fuller, S.S., and Friedman, C. *et al*., *Medical Informatics: Knowledge Management and Data Mining in Biomedicine.* Springer, 2005.
23. Cohen, W.W., Hurst, M., and Jensen, L.S., A flexible learning system for wrapping tables and lists in HTML documents. *Proceedings of WWW'02*, pp. 232-241, 2002.
24. Chen, E.S., Hripcsak, G., and Xu, H. *et al*., Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *JAMIA* 15(1): 87-98, 2008.
25. Cong, G., Jensen, C.S., and Wu, D., Efficient retrieval of the top-*k* most relevant spatial Web objects. *PVLDB* 2(1): 337-348, 2009.
26. Cole, C.L., Kanter, A.S., and Cummens, M. *et al*., Using a terminology server and consumer search phrases to help patients find physicians with particular expertise. *Stud Health Technol Inform*. 107(Pt 1): 492-496, 2004.
27. Campillos, M., Kuhn, M., and Gavin, A.C. *et al*., Drug target identification using side-effect similarity (supplementary information). *Science* 321(5886): 263-266, 2008.
28. Chen, Y., Suel, T., and Markowetz, A., Efficient query processing in geographic web search engines. *Proceedings of SIGMOD'06*, pp. 277-288, 2006.
29. Casado, B.L., van Vulpen, K.S., and Davis, S.L., Unmet needs for home and community-based services among frail older Americans and their caregivers. *J Aging Health* 23(3): 529-553, 2011.
30. de Silva, D., *Evidence*: *Helping people help themselves*. http://www.health.org.uk/public/cms/75/76/313/2434/Helping%20people%20help%20themselves%20publication.pdf?realName=03JXkw.pdf, 2011.
31. Dugdale, D.C., Epstein, R., and Pantilat, S.Z., Time and the patient-physician relationship. *J. Gen Intern Med*. 14(S1): S34-S40, 1999.
32. Doenges, M., Moorhouse, M., and Murr, A., *Nursing Care Plans: Guidelines for Individualizing Client Care across the Life Span*, 8th ed. F.A. Davis Company, 2009.
33. Drosou, M., and Pitoura, E., Search result diversification. *SIGMOD Record* 39(1): 41-47, 2010.
34. Epic MyChart homepage. http://www.epic.com/software-phr.php, 2012.
35. Etzioni, O., Cafarella, M., and Downey, D. *et al*., Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence* 65(1): 91-134, 2005.
36. Etminan, M., Carleton, B., and Rochon, P.A., Quantifying adverse drug events: are systematic reviews the answer? *Drug Saf*. 27(11): 757-761, 2004.
37. Eames, S., Hoffmann, T., and Worrall, L. *et al*., Stroke patients' and carers' perception of barriers to accessing stroke information. *Top Stroke Rehabil*. 17(2): 69-78, 2010.
38. Friedlin, J., and Duke, J., Applying natural language processing to extract and codify adverse drug reaction in medication labels. http://omop.fnih.org/sites/default/files/omop_white_paper_friedlin_08_26_10.pdf, 2010.
39. Ford, E., and Hannon, T., Oregon Health Go Local: A retrospective look. *J Consum Health Internet* 14(2): 95-108, 2010.
40. Farmer, J.E., Marien, W.E., and Clark, M.J. *et al*., Primary care supports for children with chronic health conditions: identifying and predicting unmet family needs. *J Pediatr Psychol*. 29(5): 355-367, 2004.
41. Fahrmann, M., Pesch, W.T., and Interiano, L.F., The Louisiana Go Local experience. *J Consum Health Internet* 15(3): 277-290, 2011.
42. Fiszman, M., Rindflesch, T.C., and Kilicoglu, H., Summarizing drug information in Medline citations. *AMIA Annu Symp Proc.* 2006: 254-258.
43. GoogleGuide. Search operators. http://www.googleguide.com/advanced_operators.html, 2012.
44. Google Maps homepage. http://maps.google.com/, 2012.
45. Google Places API homepage. https://developers.google.com/places/, 2012.
46. Greenes, R.A., *Clinical Decision Support: The Road Ahead*. Academic Press, 2006.
47. Garten, Y., Coulet, A., and Altman, R.B., Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics* 11(10): 1467-1489, 2010.
48. Galanter, W.L., Didomenico, R.J., and Polikaitis, A., A trial of automated decision support alerts for

contraindicated medications using computerized physician order entry. *JAMIA* 12(3): 269-274, 2005.
49. Gehrke, J., Korn, F., and Srivastava, D., On computing correlated aggregates over continual data streams. *Proceedings of SIGMOD'01*, pp. 13-24, 2001.
50. Golder, S., and Loke, Y., Search strategies to identify information on adverse effects: a systematic review. *J Med Libr Assoc*. 97(2): 84-92, 2009.
51. Grumbach, K., Selby, J.V., and Damberg, C. *et al*., Resolving the gatekeeper conundrum: what patients value in primary care and referrals to specialists. *JAMA* 282: 261-266, 1999.
52. Healthline homepage. http://www.healthline.com, 2012.
53. Healthline's DocSearch. http://www.healthline.com/doctors, 2012.
54. Home care. http://en.wikipedia.org/wiki/Home_care, 2012.
55. http://nnlm.gov/pnr/funding/NLM_GoLocal_Announcement_2-18-10.pdf, 2010.
56. Hart, P.D., *California Consumers Talk about Health Care Quality: Findings from Focus Group Discussions.* California Healthcare Foundation, 1999.
57. Han, J., Kamber, M., and Pei, J., *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.
58. Holland, D.E., Mistiaen, P., and Bowles, K.H., Problems and unmet needs of patients discharged "home to self-care". *Prof Case Manag*. 16(5): 240-250, 2011.
59. Hoffmann, T., McKenna, K., and Herd, C., Written education materials for stroke patients and their carers: perspectives and practices of health professionals. *Top Stroke Rehabil*. 14(1): 88-97, 2007.
60. Hoffmann, T., McKenna, K., and Worrall, L. *et al*., Randomised trial of a computer-generated tailored written education package for patients following stroke. *Age Ageing* 36(3): 280-286, 2007.
61. Hall, E.S., Poynton, M.R., and Narus, S.P. *et al*., Modeling the distribution of nursing effort using structured labor and delivery documentation. *JBI* 41(6): 1001-1008, 2008.
62. Hazell, L., and Shakir, S., Under-reporting of adverse drug reactions: a systematic review. *Drug Saf*. 29(5): 385-396, 2006.
63. Hilligoss, B., and Silbajoris, C., MedlinePlus goes local in NC: The development and implementation of NC health info. *J Consum Health Internet* 8(4): 9-26, 2004.
64. Hall, E.S., and Thornton, S.N., Generating nurse profiles from computerized labor and delivery documentation. *AMIA Annu Symp Proc*. 2008: 268-272.
65. Hafsteinsdóttir, T.B., Vergunst, M., and Lindeman, E. *et al*., Educational needs of patients with a stroke and their caregivers: a systematic review of the literature. *Patient Educ Couns*. 85(1): 14-25, 2011.
66. International Classification of Diseases (ICD-10) homepage. http://www.who.int/classifications/icd/en/, 2012.
67. Insider Pages Doctor Finder. http://www.insiderpages.com/about/doctor_finder.html, 2012.
68. Ioannidis, J.P., Mulrow, C.D., and Goodman, S.N., Adverse events: the more you search, the more you find. *Ann Intern Med*. 144(4): 298-300, 2006.
69. Jones, D.A., Personal health records: selected Webliography. *J Consum Health Internet* 16(3): 307-315, 2012.
70. Jha, A.K., Kuperman, G.J., and Teich, J.M. *et al*., Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *JAMIA* 5(3): 305-314, 1998.
71. Jenkins, C.G., Marshall, J.G., and McDuffee, D., MedlinePlus goes local in NC: context and concept. *J Consum Health Internet* 8(4): 1-8, 2004.
72. Kaiser Commission on Medicaid and the Uninsured. Medicaid and long-term care services and supports. http://www.kff.org/medicaid/upload/2186_06.pdf, 2009.
73. Kent, E.E., Arora, N.K., and Rowland, J.H. *et al*., Health information needs and health-related quality of life in a diverse population of long-term cancer survivors. *Patient Educ Couns*. 89(2): 345-352, 2012.
74. King, K.W., and Haefner, J.E., An investigation of the external physician search process. *J Health Care Mark*. 8(2): 4-13, 1988.
75. Korkontzelos, I., Mu, T., and Ananiadou, S., ASCOT: a text mining-based web-service for efficient search and assisted creation of clinical trials. *BMC Med Inform Decis Mak* 12(Suppl 1): S3, 2012.
76. Krupat, E., Stein, T., and Selby, J.V. *et al*., Choice of a primary care physician and its relationship to adherence among patients with diabetes. *Am J Manag Care.* 8: 777-784, 2002.
77. Kuperman, G.J., Teich, J.M., and Gandhi, T.K. *et al*., Patient safety and computerized medication ordering at Brigham and Women's Hospital. *Jt Comm J Qual Improv*. 27(10): 509-521, 2001.
78. Kushmerick, N., Weld, D.S., and Doorenbos, R.B., Wrapper induction for information extraction. *IJCAI* (1) 1997: 729-737.
79. Krauss, M.W., Wells, N., and Gulley, S. *et al*., Navigating systems of care: results from a national survey of families of children with special health care needs. *Children's Services: Social Policy, Research, and Practice* 4(4): 165-187, 2001.
80. Kaszkiel, M., and Zobel, J., Passage retrieval revisited. *Proceedings of SIGIR'97*, pp. 178-185, 1997.
81. Local search (Internet). http://en.wikipedia.org/wiki/Local_search_(Internet), 2012.

82. Luo, G., Design and evaluation of the iMed intelligent medical search engine. *Proceedings of ICDE'09*, pp. 1379-1390, 2009.
83. Luo, G., Lessons learned from building the iMed intelligent medical search engine. *Proceedings of EMBC'09*, pp. 5138-5142, 2009.
84. Luo, G., Navigation interface for recommending home medical products. *JMS* 36(2): 699-705, 2012.
85. Luo, G., Triggers and monitoring in intelligent personal health record. *JMS* 36(5): 2993-3009, 2012.
86. Luo, J.S., Physician ratings Websites. *Primary Psychiatry* 14(12): 26-30, 2007.
87. Laiteerapong, N., Huang, E.S., and Chin, M.H., Prioritization of care in adults with diabetes and comorbidity. *Ann N Y Acad Sci.* 1243: 69-87, 2011.
88. Lv, Y., Lymberopoulos, D., and Wu, Q., An exploration of ranking heuristics in mobile local search. *Proceedings of SIGIR'12*, pp. 295-304, 2012.
89. Loke, Y.K., Price, D., and Herxheimer, A., Systematic reviews of adverse effects: framework for a structured approach. *BMC Med Res Methodol*. 7:32, 2007.
90. Luo, G., and Tang, C., Automatic home nursing activity recommendation. *AMIA Annu Symp Proc.* 2009: 401-405.
91. Luo, G., Tang, C., and Thomas, S.B., Intelligent personal health record: experience and open issues. *JMS* 36(4): 2111-2128, 2012.
92. Luo, G., Thomas, S.B., and Tang, C., Automatic home medical product recommendation. *JMS* 36(2): 383-398, 2012.
93. LeClerc, C.M., Wells, D.L., and Craig, D., Falling short of the mark: tales of life after hospital discharge. *Clin Nurs Res*. 11(3): 242-263, 2002.
94. Medical Home Portal homepage. http://www.medicalhomeportal.org/, 2012.
95. Medi-Span Clinical (including a drug disease contraindications API). http://www.medi-span.com/Common/PDF/Medi-Span_SellSheet_Long_FIN.pdf, 2012.
96. MedlinePlus homepage. http://www.nlm.nih.gov/medlineplus/, 2012.
97. Microsoft HealthVault homepage. http://www.healthvault.com, 2012.
98. My HealtheVet of the U.S. Department of Veteran Affairs. http://www.va.gov/eauth/My_HealtheVet.asp, 2012.
99. Moens, M., *Information Extraction: Algorithms and Prospects in a Retrieval Context*. Springer, 2006.
100. Miura, Y., Aramaki, E., and Ohkuma, T. *et al*., Adverse-effect relations extraction from massive clinical records. *Proceedings of NLPIX'10*, pp. 75-83, 2010.
101. Mitra, M., Bogen, K., and Long-Bellil, L.M. *et al*., Unmet needs for home and community-based services among persons with disabilities in Massachusetts. *Disabil Health J*. 4(4): 219-228, 2011.
102. Markowetz, A., Chen, Y., and Suel, T. *et al*., Design and implementation of a geographic search engine. *Proceedings of WebDB'05*, pp. 19-24, 2005.
103. Meyer, M.M., and Derr, P., *The Comfort of Home: a Complete Guide for Caregivers*, 3rd ed. CareTrust Publications LLC, 2007.
104. Meystre, S., and Haug, P.J., Automation of a problem list using natural language processing. *BMC Med Inform Decis Mak.* 5:30, 2005.
105. Mandl, K.D., and Kohane, I.S., Tectonic shifts in the health information economy. *N Engl J Med*. 358(16): 1732-1737, 2008.
106. Matsuyama, R.K., Kuhn, L.A., and Molisani, A. *et al*., Cancer patients' information needs the first nine months after diagnosis. *Patient Educ Couns*. 90(1): 96-102, 2013.
107. Mandl, K.D., Mandel, J.C., and Murphy, S.N. *et al*., The SMART Platform: early experience enabling substitutable applications for electronic health records. *JAMIA* 19(4): 597-603, 2012.
108. Meystre, S.M., Savova, G.K., and Kipper-Schuler, K.C. *et al*., Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008: 128-144.
109. NC Health Info homepage. http://www.nchealthinfo.org, 2012.
110. Nicolaije, K.A., Husson, O., and Ezendam, N.P. *et al*., Endometrial cancer survivors are unsatisfied with received information about diagnosis, treatment and follow-up: a study from the population-based PROFILES registry. *Patient Educ Couns*. 88(3): 427-435, 2012.
111. Neuvirth, H., Ozery-Flato, M., and Hu, J. *et al*., Toward personalized care management of patients at risk: the diabetes case study. *Proceedings of KDD'11*, pp. 395-403, 2011.
112. Office Ally personal health record homepage. https://www.patientally.com/Main, 2012.
113. Overhage, J.M., Mamlin, B., and Warvel, J. *et al*., A tool for provider interaction during patient care: G-CARE. *Proc Annu Symp Comput Appl Med Care*. 1995: 178-182.
114. Page, L., Brin, S., and Motwani, R. *et al*., The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.
115. Ponte, J.M., and Croft, B.W., A language modeling approach to information retrieval. *Proceedings of SIGIR'98*, pp. 275-281, 1998.
116. RAND Health. Consumers and health care quality information: need, availability, utility. http://www.chcf.org/~/media/MEDIA%20LIBRARY%20Files/PDF/C/PDF%20ConsumersAndHealthCareQualityInformation.pdf, 2001.

117. Reiter, E., and Dale, R., *Building Natural Language Generation Systems*. Cambridge University Press, 2000.
118. Rindflesch, T.C., and Fiszman, M., The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *JBI* 36(6): 462-477, 2003.
119. Radley, D.C., How, S.K., and Fryer, A.K. *et al*., Rising to the challenge: Results from a scorecard on local health system performance. http://www.commonwealthfund.org/~/media/Files/Publications/Fund%20Report/2012/Mar/Local%20Scorecard/1578_Commission_rising_to_challenge_local_scorecard_2012_FINALv2.pdf, 2012.
120. Rubrichi, S., Quaglini, S., and Spengler, A. *et al*., A system for the extraction and representation of summary of product characteristics content. *Artif Intell Med*. http://dx.doi.org/10.1016/j.artmed.2012.08.004, 2012.
121. Ricci, F., Rokach, L., and Shapira, B. *et al*., *Recommender Systems Handbook*. Springer, 2010.
122. Seniors blue book homepage. http://www.seniorsbluebook.com/, 2012.
123. Sadeghi, B., How health care consumers use quality of care information to choose health coverage. Ph.D. thesis, University of California - Davis, 2007.
124. Sarawagi, S., Information extraction. *Foundations and Trends in Databases* 1(3): 261-377, 2008.
125. Schneider, J.M., Electronic and personal health records: VA's key to patient safety. *J Consum Health Internet* 14(1): 12-22, 2010.
126. Smith, C.A., Introduction to the Go Local special issue. *J Consum Health Internet* 15(3): 235-245, 2011.
127. Stephens, D., Healthy local search. http://searchenginewatch.com/article/2064187/Healthy-Local-Search, 2011.
128. Stewart, M.A., Effective physician-patient communication and health outcomes: a review. *CMAJ* 152(9): 1423-1433, 1995.
129. Ash, A.S., Shwartz, M., and Peköz, E.A. *et al*., Comparing outcomes across providers. In: Iezzoni, L.I., editor. *Risk adjustment for measuring healthcare outcomes*, 4th ed. Health Administration Press, 2012.
130. Sontag, D., Collins-Thompson, K., Bennett, P.N. *et al*., Probabilistic models for personalizing web search. *Proceedings of WSDM'12*, pp. 433-442, 2012.
131. Segura-Bedmar, I., Crespo, M., and Pablo-Sánchez, C. *et al*., Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. *BMC Bioinformatics* 11(S-2): 1, 2010.
132. Schmittdiel, J., Grumbach, K., and Selby, J.V. *et al*., Effect of physician and patient gender concordance on patient satisfaction and preventive care practices. *J Gen Intern Med*. 15(11): 761-769, 2000.
133. Stone, E.M., Heinold, J.W., and Ewing, L.M. *et al*., Accessing physician information on the Internet. http://www.commonwealthfund.org/~/media/Files/Publications/Fund%20Report/2002/Jan/Accessing%20Physician%20Information%20on%20the%20Internet/stone_mdinternet_503%20pdf.pdf, 2002.
134. Silbajoris, C., and McDuffee, D., Location matters: The Go Local developers' perspective. *J Consum Health Internet* 15(3): 246-255, 2011.
135. Segura-Bedmar, I., Martínez, P., and de Pablo-Sánchez, C., Using a shallow linguistic kernel for drug-drug interactions extraction. *JBI* 44(5): 789-804, 2011.
136. Schoen, C., Osborn, R., and How, S. *et al*., In chronic condition: Experiences of patients with complex health care needs, in eight countries, 2008. *Health Aff* 28(1), w1–w16, 2009.
137. Schoen, C., Osborn, R., and Huynh, P.T. *et al*., Primary care and health system performance: Adults' experiences in five countries. *Health Aff*. Suppl Web Exclusives: W4-487-503, 2004.
138. Schmittdiel, J.A., Selby, J.V., and Grumbach, K. *et al*., Choice of personal physician and patient satisfaction in a health maintenance organization. *JAMA* 278: 1596-1599, 1997.
139. Safran, D.G., Taira, D.A., and Rogers, W.H. *et al*., Linking primary care performance to outcomes of care. *J Fam Pract*. 47(3): 213-220, 1998.
140. Tari, L., Anwar, S., and Liang, S. *et al*., Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics* 26(18): i547-i553, 2010.
141. Taylor, K., and Currow, D., A prospective study of patient identified unmet activity of daily living needs among cancer patients at a comprehensive cancer care centre. *Australian Occupational Therapy Journal* 50(2): 79-85, 2003.
142. Thom, D.H., Ribisl, K.M., and Stewart, A.L. *et al*., Further validation and reliability testing of the trust in physician scale. *Med Care.* 37(5): 510-517, 1999.
143. Taylora, M.V., and Stephensonb, P.L., Self-management of chronic disease: A Webliography. *J Consum Health Internet* 12(4): 349-360, 2008.
144. UCompareHealthCare. http://www.ucomparehealthcare.com/physicians_start.html, 2012.
145. Vogeli, C., Shields, A.E., and Lee, T.A. *et al*., Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. *JGIM* 22 (Suppl 3): 391–395, 2007.
146. WebMD personal health record homepage. http://www.webmd.com/phr, 2012.
147. White, S., Bissell, P., and Anderson, C., Patients' perspectives on cardiac rehabilitation, lifestyle change and taking medicines: implications for service development. *J Health Serv Res Policy*. 15 Suppl 2: 47-53, 2010.

148. Wang, X., Hripcsak, G., and Markatou, M. *et al*., Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *JAMIA* 16(3): 328-337, 2009.
149. Washington, K.T., Meadows, S.E., and Elliott, S.G., Information needs of informal caregivers of older adults with chronic health conditions. *Patient Educ Couns*. 83(1): 37-44, 2011.
150. Workmana, T.E., and Stoddartb, J.M., Building online health resources using freely available tools: The goLocalUtah experience. *J Consum Health Internet* 11(1): 15-31, 2007.
151. Wang, F., Sun, J., and Ebadollahi, S., Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. *Statistical Analysis and Data Mining* 5(1): 54-69, 2012.
152. Wang, A.Y., Sable, J.H., and Spackman, K.A., The SNOMED clinical terms development process: refinement and analysis of content. *AMIA Annu Symp Proc.* 2002: 845-849.
153. Xu, R., Supekar, K., and Morgan, A. *et al*., Unsupervised method for automatic construction of a disease dictionary from a large free text collection. *AMIA Annu Symp Proc.* 2008: 820-824.
154. Yelp homepage. http://www.yelp.com, 2012.
155. Garcia-Molina, H., Ullman, J.D., and Widom, J., *Database Systems: The Complete Book*, 2nd ed. Prentice Hall, 2008.
156. Witten, I.H., Frank, E., and Hall, M.A., *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Morgan Kaufmann, 2011.

Appendix A: List of acronyms

AIRS	Alliance of Information and Referral Systems
HRS	health-related service
ICD	International Classification of Diseases
IHP	individual healthcare provider
iPHR	intelligent personal health record
PHR	personal health record
RCT	randomized clinical trial
SCA	self-care activity
SMART	Substitutable Medical Applications, Reusable Technologies
SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms
SPC	Summary of Product Characteristics

Appendix B: List of symbols

α_H	a positive constant controlling the degree of discounting over time for the health issue H
A	SCA
B	business or organization
B_1	the business or organization in the set $S_{remaining}$ with the largest relevance score
c	criterion
C	the complete set of search guide information for all health issues of concern by the user
d_H	health issue weight discount factor
d_T	type weight discount factor
D_T	the text displayed as the name of the type T of HRS in the navigation hierarchy of the navigation output interface
F	the set of criteria used in computing the degree of matching between an IHP's profile and the user's needs
F'	the set of criteria in F that are relevant to the user and have non-empty utilities
G	search guide phrase
H, H_1, H_2	health issue
I_1, I_2	IHP
J	the number of top businesses and organizations re-ranked for search result diversification
K_b	health knowledge base
L	the user's location
L_h	the list of health issues of concern by the user
L_u	the user's location at a specific level of geographical granularity
n_w_H	the normalized weight of the health issue H
$n\_w_{T,H}$	the normalized weight of the type T of HRS for the health issue H
N	a constant to control the amount of time spent on search result diversification
N_h	the user's general need of HRSs regardless of his location
P_B	the profile of the business or organization B
Q_c	conceptual query representing the user's overall need
$Q_{G,L}$	the query formed from the combination of the search guide phrase G and the user's location L
R_{all}	the complete set of retrieved businesses and organizations
$rank(B, Q_{G,L})$	the rank of the business or organization B among all businesses and organizations that the query $Q_{G,L}$ retrieves from the local Web search engine
$score(B, Q_{G,L})$	the business or organization B's ranking score that the local Web search engine computes for the query $Q_{G,L}$

$score(B, Q_c)$	the business or organization B's relevance score computed for the conceptual query Q_c
s_G	the map scale that the local Web search engine returns for displaying the set $S_{G,L}$ of businesses and organizations on a map in the output interface
$S_{G,L}$	the set of businesses and organizations retrieved by the query $Q_{G,L}$
S_H	the list of types of HRSs relevant to the health issue H
$S_{remaining}$	the set of businesses and organizations remaining to be returned to the user
$S_{returned}$	the set of businesses and organizations already returned to the user
S_T	the set of phrases pre-compiled for the type T of HRS
$S_{T,H}$	the set of search guide phrases formed for the type T of HRS when T is linked to the health issue H
t_c	the current time
t_l	the most recent time when an IHP saw a patient with a particular health issue
T	type of HRS
$T_{T,H}$	the text explaining why the type T of HRS is relevant to the health issue H
u_c	the utility computed for the criterion c
w_c	the weight of the criterion c
w_H	the weight of the health issue H
$w_{T,H}$	the weight of the type T of HRS for the health issue H