A Systemic, Participative Design of Decision Support Services for Clinical Research:

Article - June 2014
DOI: 10.4018/ijitsa.2014070102

CITATIONS
2

READS
320

6 authors, including:

Alexandra Pomares Quimbaya
Pontificia Universidad Javeriana
69 PUBLICATIONS 140 CITATIONS

Rafael A Gonzalez
Pontificia Universidad Javeriana
79 PUBLICATIONS 366 CITATIONS

Ricardo Bohórquez
Pontificia Universidad Javeriana
25 PUBLICATIONS 45 CITATIONS

Oscar Mauricio Muñoz
Pontificia Universidad Javeriana
94 PUBLICATIONS 363 CITATIONS

Some of the authors of this publication are also working on these related projects:

Clinical Cardiology View project
CAOBA - Center of Excellence in Big Data and Data Analytics View project

All content following this page was uploaded by Oscar Mauricio Muñoz on 06 November 2014.
The user has requested enhancement of the downloaded file.
A Systemic, Participative Design of Decision Support Services for Clinical Research

Alexandra Pomares Quimbaya, Pontificia Universidad Javeriana, Bogotá, Colombia
Rafael A. González, Pontificia Universidad Javeriana, Bogotá, Colombia
Wilson Ricardo Bohórquez, Pontificia Universidad Javeriana, Bogotá, Colombia and Hospital Universitario San Ignacio, Bogotá, Colombia
Oscar Muñoz, Pontificia Universidad Javeriana, Bogotá, Colombia and Hospital Universitario San Ignacio, Bogotá, Colombia
Olga Milena García, Hospital Universitario San Ignacio, Bogotá, Colombia
Dario Londoño, Pontificia Universidad Javeriana, Bogotá, Colombia and Hospital Universitario San Ignacio, Bogotá, Colombia

ABSTRACT

Development of IT-based services to support decision-making in healthcare should be guided by the following considerations: rigor, relevance, user-centered participation and inclusion of the best practices for IT-based service systems. In this paper, the balance between rigor and relevance is achieved by following the design science research methodology; user-centered participation is tackled from the socio-technical tradition in information systems; best practices considered in the planning, design and implementation of the services are informed by the MOF framework. Moreover, and considering the premise that these pillars should holistically converge, this research has been approached from a systemic stance where iterative, participative, socio-technical activities have allowed the effective collaboration between information systems researchers, clinical researchers, medical staff and administrative hospital personnel. This paper argues for a move towards enhancing systemic, participative, design-centered service systems engineering by reporting a case which applies these concepts for providing decision-support services, enabled by data and text mining techniques, to contribute to clinical research and administration by being able to search electronic health records where narrative text hides meaningful information that would otherwise require a time-consuming human revision of these records.

Keywords: Decision Making, IT, MOF, Methodology, Science

DOI: 10.4018/ijitsa.2014070102
1. INTRODUCTION

IT in healthcare has had a prominent place within applied information systems research and has a strong tradition, ranging from the very deeply technical and algorithmic – such as classic and modern uses of artificial intelligence and expert systems (Gresh, Rabenhorst, Shabo, & Slavin, 2002; Lisboa, 2002; McCauley & Ala, 1992) – to more socio-technically minded interventions – such as the use of Checkland’s Soft Systems Methodology mostly within the context of the UK’s NHS (e.g. Kalim, Carson, & Cramp, 2004, 2006). Despite the progress, real impact has not been felt significantly where it matters most. As Hesse and Shneiderman (2007) argue, it has probably been a matter of not asking the right questions: rather than focusing on what technology can do, we should be focusing on what people can do. This follows a general trend in information systems which pays attention to user-centered, participative design (Carroll & Rosson, 2007; Mao, Vredenburg, Smith, & Carey, 2005) as well on situated, context-dependent uses of technology (Orlikowski, 2000). The popularization of many user-oriented information technologies has shown that user experience and involvement in the design, appropriation and evolution of IT exceed consumer electronics and applications and are indeed morally and pragmatically desirable for information systems development in general (McCarthy & Wright, 2004).

This suggests that the development of IT-based services to support decision-making in healthcare should be guided by the following considerations: rigor, relevance, user-centered participation and inclusion of the best practices for IT-based service systems. These requirements are not mutually exclusive but dependent on each other. As such, this paper will present the ongoing development of a set of decision-support services for clinical research, based on a socio-technical, systemic, design-centered approach. The balance between rigor (transparency, validation, formalism, systematic use of existing knowledge) and relevance (utility, novelty) is achieved by following the design science research methodology (Hevner, March, Park, & Ram, 2004). User-centered participation is tackled from the socio-technical tradition in information systems (Stahl, 2007). Some best practices considered in the planning, design and implementation of the services are informed by the MOF framework (Pultorak, 2008). Moreover, and considering the premise that these pillars should holistically converge, this research has been approached from a systemic stance where iterative, participative, socio-technical activities have allowed the effective collaboration between information systems researchers, clinical researchers, medical staff and administrative hospital personnel.

Through this effort, we have been able to obtain meaningful findings in terms of the technical contribution that data and text mining can have for decision-support in clinical research and hospital administration as well as to explore the integration of participative mechanisms and IT service systems best practices into the growing body of work around development of artifacts through a design science research perspective. In sum, this paper argues for a move towards enhancing systemic, participative, design-centered service systems engineering. It does so by reporting a case which applies these concepts for providing decision-support services, enabled by data and text mining techniques, to contribute to clinical research and administration by being able to search electronic health records where narrative text hides meaningful information that would otherwise require a time-consuming human revision of these records.

One of the main sources for supporting decision-making in hospitals has been the creation and use of electronic health records (EHR), a rich source of data when properly exploited. Nonetheless, in practice, the use of EHR is more complex due in part to the lack of having considered the three pillars of rigor, relevance and user-centeredness. Often,
healthcare professionals are invited (or forced) to adapt to the systems that keep track of patient records, rather than having the system support the professionals in their activities. Of course, striking that balance is not easy, but neglecting it creates problems, such as the tendency to use open text fields to input information that should otherwise be input into structured fields. This narrative approach to recording patient information fits many healthcare practices but unfortunately does not exactly match the logic and technologies offered by current decision-support systems.

One specific approach that has garnered increased attention within the healthcare domain has been data mining (Bellazzi & Zupan, 2008; Windle, 2004). Through the various technologies that can be used for data mining, clinical research and practice can be considerably improved and new patterns can be extracted to help with diagnosis, treatment, and cost-benefit analysis, among others. However, as stated above, when data mining relies on electronic health records, the use of narrative text, makes traditional data mining approaches limited. This paper discusses the design of a support tool for clinical research (at a first stage) and administrative decision-making (at a second stage), using data mining technologies and considering the restrictions imposed by a dataset that is not structured and relies on narrative text.

This paper is structured as follows. Section 2 describes the research methodology to go onto Section 3, which presents related works surrounding the use of data mining techniques for patient identification and visualization of electronic health records. Section 4 then presents some of the potential uses of these techniques in supporting decision making both for clinical research and health administration. Section 5 goes on to present the proposal of system, dubbed DISEarch, to be used for patient identification using health records from a university hospital information system. Some early results of the prototype are discussed in Section 6, and then some conclusions and future work are mentioned in Section 7.

2. RESEARCH METHODOLOGY

This research follows a design science research philosophy (Hevner et al., 2004), which has gained increasing support from information systems researchers, given its open goal of providing a framework for research that is both relevant and rigorous. The tension between these often conflicting aspects is dealt with through a design-centered paradigm, where the relevance of solving real-world problems is achieved through a specific design, which in Herbert Simon’s tradition is no less than problem-solving itself, i.e. design is problem-solving as it fills the gap between a present situation and a desired one (Simon, 1996). In addition, design also involves the use of applicable knowledge which becomes embedded in the design process and product, thus being rigorous insofar as this knowledge is applied transparently and systematically. As a result, design science research offers a tripartite framework, relevance-design-rigor, onto which this paper further elaborates the recommendation that the core design be participative in nature, making it more relevant, as the beneficiaries (be they users, beneficiaries or customers) become co-designers. It is in the multi-disciplinary interaction between this research group and an information systems research group that a user-centered participative design has been followed. Together with the participation of other potential users, a prototype has been developed as a proof of concept of the underlying data (text) mining models and algorithms and as a source of medical validation with respect to the quality of the results.

The research approach has followed a problem-initiated process, guided by a three-cycle design science research approach (see Hevner, 2007), as shown in Figure 1. An initial relevance cycle has been used to identify requirements, potential users, associated processes (both for clinical research and for administrative purposes), existing technology (the hospital’s information system, the underlying databases, the users’ capabilities), as well as identifying
other hospitals that, given similar technologies, could also benefit from the resulting artifact. A rigor cycle has helped uncover the applicable knowledge by doing a literature. Although it could appear as if the rigor cycle follows after the relevance cycle, in reality both cycles have moved mostly in parallel. This simultaneous cycling through rigor and relevance is desirable and natural, since by uncovering requirements the researchers have been led to revise and refine the search space within the knowledge base. Conversely, and given the fact that the initial problem was ill-defined and thus still open, as applicable knowledge is found and shared with the medical researchers, this feeds back on the relevance cycle, by making explicit the possibilities offered by existing methods, tools and technologies.

Once the relevance and rigor cycles have offered sufficiently clear requirements and applicable knowledge, they meet inside a design cycle, which iterates between actual design and continued evaluation. This process is akin to classic information systems design, where initial mockups, forms and flow charts are built, refined and evaluated both from a technical point of view and from a potential user’s point of view, until these preliminary models can be codified, following a traditional data mining process. Inspired by CRISP-DM (Shearer, 2000), understanding the business is followed by understanding the data, after which such data is prepared (to make it amenable for treatment by data mining techniques), models are created to process the data and finally the solution is evaluated and deployed. As can be seen, this data mining process fits naturally within the design science research approach and simply gives it a specific flavor.

Evaluation, as stated above, is an iterative process in parallel with the design and refinement of the data mining models and the resulting software prototype. The aim has been to use a test copy of the hospitals EHR database in order to produce the required results and then sharing such results with the complete research team so that evaluation is carried out not only from an information systems (data mining) point of view, but crucially from a medical perspective as well. Final validation of the resulting artifact is achieved by using the prototype as a proof of concept and analyzing the results so that the desired EHRs are found with precision and recall criteria, or in clinical terms, with enough sensitivity and specificity. The EHRs to be found respond to the requirements, which stem from the following research question.

The main problem and research question has been how to identify patients associated with a specific diagnosis within the set of electronic health records contained in the hospital information system. Though at first this may seem a straightforward question amenable to treatment through simple database queries, the complexity of an actual diagnostic process (for instance, the fact that it is not a point in time, but the result of
several events) and the aforementioned nature of the data (unstructured narrative text) make using queries limited and indeed impractical. For example, one may query the system for patients whose EHR contains “diabetes” in a specific field for diagnosis, but very often the diagnosis is not contained in this field in such a clear-cut fashion. Moreover, extending the query to other fields does not solve the problem because the narrative is nuanced and dynamic (one could find, for instance, “discard diabetes”).

Having access to the list of patients with an identified diagnosis is useful given that it is the basis for many of the clinical research studies being carried out by the chronic disease research group involved in this project, but of course of other groups as well. Furthermore, this data is also invaluable for administrative purposes in assessing fitness to treatment protocols, costs and patient distributions.

According to Hevner (2007), design science research goes through relevance, rigor and design cycles. This cyclical nature also emphasizes the fact that the cycles are connected. One should thus expect that being inside the “design cycle” still will require a connection and parallel advancement in rigor (by continuously revising knowledge applicable to the design) and in relevance (by clarifying and getting feedback from the real problem environment). This is the way this research has proceeded in general, given its iterative and participative nature. However, the paper structure benefits from a following a more linear model of design science research, as proposed by (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). Accordingly, the rest of this paper presents the problem (Section 2) and related solutions (Section 3) first and subsequently the design of the artifact (Section 4) and its evaluation and validation (Section 6) are explained.

3. RELATED WORKS

Today, healthcare organizations are carrying out significant efforts to correctly include the use of information technologies and collect data in an electronic format, for example in Electronic Health Records. Nonetheless, such accumulation of data has overcome the capacity of healthcare organizations to use this data in profitable ways and support business decisions. In order to truly leverage digital clinical data to its full potential, investment in Healthcare Analytics solutions must be carried out. Ferranti et al. (2010) state: “hospitals have lagged in adopting thoughtful analytical approaches that would allow operational leaders and providers to capitalize upon existing data stores… We believe that such active investment in health analytics will prove essential to realizing the full promise of investments in electronic clinical systems”.

Healthcare analytics is increasingly supported by data mining strategies and their associated techniques. In order to uncover the evolution of these strategies, this section discusses the results of a literature review focused on identifying some of the main contributions, institutions and authors centered on this topic. Subsequently, the second part of this section describes some outstanding recent contributions for the improvement of medical data visualization to support healthcare analytics, offering a view of their strengths and opportunities for improvement that have been taken into consideration in designing the proposal brought forward in this paper.

The development of tools, techniques and models for identifying (that is, searching and retrieving EHR according to a specific information need) and analyzing (that is, looking for patterns, classifying or processing EHR for decision making) electronic health records in the last thirty years has been studied through systematic queries in the Scopus database using keywords related to patient, medical or health records, coupled to data preparation, data classification, data and text mining, prioritization, and decision or regression trees. Iteratively, different queries have been used in order to refine the coverage and relevance of the results. In addition, since
medical data is particularly sensitive, anonymization, data protection and privacy protection have also been coupled to EHR.

Although the main goal of this research is centered on data mining (and related) strategies for clinical research, most of the work found is aimed at aiding in diagnosis and contributing to effectiveness analysis of particular treatments (or drugs), whether from a medical or an economic point of view. There are few contributions directly related to identifying or analyzing a set of health records for which a specific diagnosis has already been registered for use in clinical research on that specific disease, and given a context in which such diagnosis is hidden in free or narrative text, as described elsewhere in this paper.

By looking at some of the most cited papers, it can be seen that, for instance in (Gaspar et al., 1997), cited 774 times at the time of query, patient records are partitioned according to their being treated with radiation therapy, in order to estimate the effectiveness of the treatment through decision trees. However, as pointed out, these patient records have already been identified as belonging to the interest group, prior to the study. Furthermore, the actual analysis is purely statistical in nature and although most data mining techniques are indeed statistics-based, in building the model proposed in this research, the interest is on techniques supported directly by software. In (Highet, Forrest, Ballow, & Schentag, 1999), cited 45 times, a similar contribution is related this time to determining the effectiveness of an antibiotic through regression analysis.

Another important project, more closely related to informatics or computer science techniques, can be seen in (Burbidge, Trotter, Buxton, & Holden, 2001), with 265 citations. This time it deviates from the clinical or medical domain into chemistry (drug research), by reporting a data mining algorithm which, through machine learning, can be used to predict the effect of a particular drug, according to its chemical composition. A similar effort is found in (Kroes et al., 2004), except this time the emphasis is on predicting the toxicity of the compound.

Moreover, the work presented in (Suppes et al., 2005), cited 176 times, report on algorithms that are used in decision making support for treating bipolar disorder; in this case, given the disorder, certain treatment protocols are designed. For the purposes of the present work, the opposite would be more appropriate; that is, tracing the compliance with a given treatment protocol on a set of patients that have to be identified as having been diagnosed with a given disease—the idea would then be to extract treatment patterns from the identified set and then comparing those patterns to the already established protocol.

One of many cases found reporting a cost-benefit analysis is presented in (Hill et al., 2000), cited 156 times. This time related to using prophylactic antiemetic therapy for reducing postoperative nausea and vomiting. Similarly, (Adams, Gregor, Kertesz, & Valberg, 1995), with 124 citations, evaluates the effectiveness of carrying out a hemocromatosis test in blood donors in order to demonstrate the economic benefit of doing so for a national health system.

One paper potentially more closely related to the purpose of this paper’s contribution is (Hess, Abbruzzese, Lenzi, Raber, & Abbruzzese, 1999), cited 130 times. Their work is aimed at classifying patients into specific groups, according to a set of medical variables. The difference lies in that the starting set of patient records is already known to be diagnosed with unknown primary carcinoma, while our work starts one step behind: identifying patients diagnosed with a specific condition from a complete set of records of diverse patients. Furthermore, Hess et al. only consider quantitative, structured variables, while our work considers also unstructured narrative text.

In terms of the sources of these related works, something worth noting is that they are
mostly originating in hospitals, medical centers and pharmaceutical labs. Few of the results are explicitly affiliated to collaboration between hospitals and universities or between university hospitals and other departments (especially those related to computer science).

Other recent contributions explore the analysis of narrative texts to identify special associations as in the case of (Iyer, Harpaz, Lependu, Bauer-Mehren, & Shah, 2013) which analyzes drug-event associations based on the identification of drug mentions in clinical notes. Another work presents CLIX (Clinithink, s. f.), an engine that analyzes medical narrative texts and transforms them into structured content based on a specific standard code like SNOMED CT, ICD-9, ICD-10, among others. Although in both cases the input is narrative text contained in a medical record and the output is a structured content, they cannot act, by themselves, as a decision support system that can improve the work of clinical research and administration. However, these works can be useful to enrich a decision support system.

In sum, a detailed exploration of relevant works suggests that there have been many contributions to patient identification using data mining techniques and that these have been growing rapidly. However, recalling the introduction of this paper, there is still work to be done in order to generalize these works into decision-support tools to be used widely for clinical research and health administration. Significantly, it has been found that the majority of works are aimed at sets of data for which a diagnosis, disease or condition has already been determined. In other words, the population for study has already been defined. There is a clear opportunity to contribute works that support the preceding stage; that is, we have found that in practice most clinical research group and health administration must go through a time-consuming, labor-intensive, statistically supported, iteratively queried process of determining patients for study.

4. IMPROVING DECISION-MAKING FOR CLINICAL RESEARCH AND HEALTH ADMINISTRATION

As the previous section discusses, electronic health record systems (EHR) have improved the access to patient information by health care providers. They are a rich source of knowledge widely used to improve health care activities such as diagnostic and treatment definition. In addition, they have been also used to enhance health research processes and administrative tasks in health institutions; however, their use for these purposes is limited due to different factors such as confidentiality, heterogeneity of information and incompleteness of medical data. As a consequence, most of decision-making in health sector does not take full advantage of the vast source of information from EHR because of the difficulty of obtaining the adequate information at the right moment. This section aims to analyze how can be improved the decision making process specifically for clinical research and health administration using as a source of information EHR.

The problems around decision-making were studied in a general hospital in Bogotá, Colombia. This hospital has a main information system called SAHI that includes modules to manage the Electronic Health Record, Contracts, Human Resources, Client Service, Budgeting, Purchases and Supply, etc. This system allows physicians to obtain the EHR of each patient during the medical attention; however, their use for research purposes has been limited due to the fact that important information is stored in narrative texts, intended for human beings that are difficult to search and analyze automatically. One of the common requirements of medical research is to find the medical records of patients that have been diagnosed with a specific disease. This task that should be easily done using classical queries (e.g. using SQL) is very time-consuming. This is because diagnosis is
frequently hidden in narrative texts (e.g. medical notes, progress notes), hindering the possibility of automatically detecting relevant records and requiring the participation of an expert in the analysis. Similarly, the administration of the Hospital frequently requires analyzing costs and efficiency of medical treatments. Even if some of this information is well structured, as medications and laboratory orders, the complete sequence of events related to a patient is hidden in narrative texts. In summary, the main requirement of this hospital decision support system is to recognize which medical records are useful for clinical and administrative research, taking into account all the information in the EHR, including the one that is in narrative texts. Particularly, the decision support system must support biomedical research and quality analysis and service delivery.

Currently, the processes for assessing adherence to international recommendations and in general to evaluate medical decisions include long term activities of analysis of EHR. Figure 2 illustrates a classic EHR analysis process from the point of view of data requirements. As it can be seen an important and currently time-consuming task is the exploration and identification of relevant EHR. IT staff members are the owners of the EHR and they are frequently a bottleneck during research projects. In addition, the effort to validate the relevance of the provided EHR implies time between five and forty minutes for reading each one of the attentions of patients. Providing a facility to obtain these EHR automatically will improve process time metrics. In addition, providing visual analysis of the relevant EHR enriches the quality of the decision process due to the improvement of user empowerment.

5. DISEARCH: A SYSTEM FOR ELECTRONIC MEDICAL RECORD ANALYSIS

This section describes the decision support system created in order to support the requirements of clinical research and health administration, mentioned in the previous section, using as a source of information EHR. The main principle of this system called DISearch is to combine the analysis of structured and unstructured information contained in EHR to enhance decision-making.

Considering the main users of DISEArch are medical doctors, during the development of this system the definition of functionalities and the user interface were made using a participatory design approach. In this approach users cooperate with designers and developers during the different phases of the project. The design followed an evolutionary cycling where the system was discussed mainly from a clinical and practical point of view rather than a technical perspective.

In what follows, this section presents the general architecture of the system and explains in detail its main components.

*Figure 2. EHR analysis process*
5.1. DISEArch Architecture

The architecture of DISEArch is illustrated in Figure 3. The principle of the system is to provide different capabilities of visualization and analysis to enhance biomedical research as well as quality and services delivery analysis.

The components are divided in three layers. The components of the Data Layer are in charge of store all the information used or generated by the system. The Knowledge base component manages the relevant taxonomies and definitions required to analyze EHR. For instance, it stores the taxonomies of diseases and medications that are used to categorize a record. The DISEarch database stores all the temporal results and the analysis results obtained by the components of the business level. The Electronic Health System component represents the database of the Hospital Information System that contains the EHR. Even if this last component is external, DISEArch contains all the information required to extract the EHR from it.

The Business Layer components contain the main capabilities of the decision system. The Extraction Manager is responsible for extracting the set of medical records according to the pre-selection defined by the users (e.g. gender, age, date). This is the only component that interacts directly with the Electronic Health System component, and contains the logic for wrapping the schema of the original source of health records. In addition, when it is required, this component launches an anonymization service that assures the de-identification of identifiers and pseudo-identifiers from EHR. The Analysis Engine is in charge of analyzing the structured and non-structured elements of EHR to identify and prioritize which of them are useful for a specific study (e.g. descriptive study). The Knowledge Engine goal is to extract knowledge from different points of view. First, the LifeLine Generator presents a 360-degree view of a patient visualizing all the events where he/she has been involved. The event extraction includes the identification of events that were...
written in narrative text as well as those that are well structured. The second component is the Pattern Extractor; this component applies data mining techniques over a set of EHR to obtain natural clusters, association rules and sequence patterns. The Treatment Guideline Analyzer allows structuring a disease treatment guideline and comparing it to a specific medical record or a group of medical records. Using this component a medical institution may know at what level its patients follow the recommended treatment guideline. This component is especially important for quality and service delivery analysis.

Finally, the presentation layer contains the graphical user interface to query the different components of the business layer.

5.2. Prototype

DISEArch architecture was developed in Java language using different libraries to improve the analysis of EHR, specially the narrative texts. The components implemented in DISEArch 1.0 are presented in Figure 4.

In the Presentation Layer, the researcher can describe the characteristics of the EHR that

Figure 4. DISEArch prototype
he/she is interested on using the EHR analysis form. As presented in Figure 5, this form allows specifying the characteristics of the EHR required for the analysis like the vocabulary of the disease, the related lab exams, etc. The Knowledge Engine handles the knowledge base that allows the enrichment of the description of the disease defined in the EHR analysis form. This enrichment increases the definitions made by the researcher adding synonyms or related words to the initial description.

The Extraction and Pre-selection component is in charge of the extraction and initial preprocessing of medical records from the EHR system. This component is parameterized according to the characteristics of the system and extracts the records according to the definition of initial parameters, such as date of admission, gender or age of patients. This component was built using the analysis services of SQL Server 2008.

The EHR Analysis Engine is the core of the analysis. It uses natural language processing to analyze EHR. It implements Stemming using Porter Stemmer algorithm, simple string tokenization, sentence splitting, POS tagging using Probabilistic Part-of-Speech Tagging Using Decision Trees (Schmid, 1994) for annotating text with part-of-speech and lemma information and finally gazetteer lookup using regular expressions. This component has a coordinator that calls each of the search engines. The Narrative Text Analysis is in charge of the analysis of natural language and was developed using the GATE API (Cunningham et al., 2011). This API enables the inclusion of all the language processing functionality within DISEArch. In addition, we use Tree tagger (Schmid, 1994), a Pearl implementation which provides tokenization and Part of the Speech tagger.

The Structured Analysis Engine is in charge of searching the disease over the structured attributes. Finally, the Prioritization component integrates the results using the semantic rules and prioritizes the set of records.

6. RESULTS

One of the main results of this project is the reduction of time and the improvement of the accuracy in the results of retrospective medical

Figure 5. DISEArch presentation
research. The design as proposed and implemented in the DISEArch software prototype proves to be useful to promote a systematic approach to query EHR including an important portion of narrative texts. These characteristics aim to fill the gap found during the analysis of the evolution of tools, techniques and models related to EHR analysis in the last thirty years. This section presents first the evaluation method followed to demonstrate this improvement; the second part illustrates the limitations that were found in DISEArch during this evaluation.

6.1. Evaluation Results

The aim of this project is to allow the hospital administrators and medical researchers to recognize the utility of the information contained in EHR to improve quality of service and research processes. To evaluate this goal we developed a process to measure the utility of the system, the quality of the results (in terms of the accuracy of obtained records) and the efficacy of the system to change the related processes.

The team that carried out the evaluation combined medical doctors, students of medical specializations and the university hospital IT staff. For evaluation purposes, three diseases were selected as case studies: diabetes mellitus, heart failure and pulmonary hypertension. These diseases were selected taking into consideration their common characteristic of being chronic diseases. This alone makes the use and evaluation of the system easier to analyze, because it is more likely that such patients have continued visits to the hospital, thus making their EHRs a richer source of data. In addition, those diseases also exhibit differences related to their ease of detection, their prevalence and incidence proportions, which helps in triangulating the evaluation results.

The utility and quality of results were measured using the EHR from the SAHI system. During the first phase a group of medical doctors analyzed manually 400 EHR selected randomly from patients in the department of Internal Medicine. From this analysis we obtained three sets of EHR, each one with the patients with the selected diseases. In essence, this is used as the control dataset, obtained with the current manual method for obtaining EHRs related to a disease.

The second phase consisted in the use of DISEARCh by a group of medical doctors specialized in each one of the diseases. They described the diseases using DISEArch and ask for the relevant EHR. The utility was measured evaluating the consistency of the manual selection results with respect to the selected EHR by DISEArch. Considering that DISEArch generates a list of prioritized EHR, in order to compare it with the control dataset, we set three partitions within the results: EHRs with a score between 70% and 100% are in the set A, EHR with a score between 40% and 69% are in the set B and those EHR with a score below 40% are in the set C. The hypothesis of the process developed by DISEArch is that set A should contain mostly EHR of patients with the disease; set B should include patients whose EHR match some literals of the disease query, but not necessarily have the disease, in this case the user must analyze the findings of DISEArch to make a decision; and set C contain primarily EHR with some coincidences, but with a low probability of having the disease.

The comparison of DISEArch results with respect to the manual analysis is presented in Figures 6, 7 and 8. The x axe represents partitions according to DISEArch results; the y axe represents the percentage of EHR. The set % Yes means the EHR was classified manually as Yes, The set %No means the EHR was classified manually as No.

As expected by the hypothesis of DISEArch, the EHR in set A correspond to patients with the analyzed disease. This fact was true for prevalent diseases like Heart Failure and Diabetes Mellitus.

On the contrary, Pulmonary Hypertension had another behavior; in this case the ratio of Yes and No founded was similar. The analysis of the result details for this disease revealed that most of the percentage of No EHR in partition A are from patients who have had a lot of
Figure 6. Heart failure comparison

![Heart failure comparison chart](image)

Figure 7. Pulmonary hypertension

![Pulmonary hypertension comparison chart](image)

Figure 8. Diabetes mellitus comparison

![Diabetes mellitus comparison chart](image)
related and confirmatory tests, but were never diagnosed with the disease. This fact allowed us to conclude that uncommon diseases require the analysis of the result details to select or discard an EHR for the analysis.

With respect to Partitions B and C, the results were as expected; EHR in Partition C can be discarded most of the times, except for those analyses that required being exhaustive. Finally, Partition B requires the analysis of DISearch result details in order to identify the set of relevant EHR. Even if this implies an effort, the number of details that should be analyzed is not representative with respect to the original amount of data.

These results prove the utility of DISearch to identify the EHR required for the analysis. However, even if lower, the not-matching results in partition B were analyzed to identify what kind of coincidences they have.

Medical doctors analyzed once again all the events associated with the “problematic” EHRs. The new analysis identified that there are patients with an early stage of a disease and because of that the manual analysis determined it was not relevant; however, the evolution of the patient would probably end in a Heart Failure diagnosis. According to the doctors, neither the medical doctor that did the early analysis, nor DISearch were wrong: the difference was the level of evolution that each one of them aimed to use to identify EHR. This is worth noting because neither the manual process nor DISearch are free from error. In fact, as discussed in an evaluation meeting, some medical doctors view the situation as if each individual actually carries all diseases potentially in him or her, and if we lived long enough we would all eventually develop the diseases. As such, selecting patients will always leave out some that could have been included and it is a human decision whether to include those with borderline conditions.

The other set of EHR analyzed correspond to patients with a large number of service records (visits, tests, hospitalizations, etc). Some of these included early medical statements discarding the disease, but more recent ones confirmed its presence. These cases represent a great challenge to DISearch because the analysis involves a new variable of analysis that initially was not taken into account; the next section develops this issue further.

The efficacy of the system to renew the process of analysis was measured comparing the previous (manual) process with the new (DISearch-supported) one, and the effects it has in time invested and user empowerment. Figure 9 illustrates the new process. In this new process the final user is not dependent on IT staff, avoiding the bottleneck that this implied; In addition, Table 1 illustrates improvements on the time required for the analysis of EHR.

6.2. Implications for IT-Based Service Systems (ITSS)

In order to assure that the new system would deliver the expected results, the project followed the guidelines of the MOF framework (Pultorak, 2008). The Plan, Deliver, Operate and Manage phases of the IT service lifecycle allowed us to

![Figure 9. EHR Analysis Process with DISearch](image)
bring together information and people; however, in some of these phases we had to take important decisions that worth mentioning.

During the Plan phase, the formation of the project team was extremely time consuming because of the misalignment between the IT area and the legal and human resources department. Even if these problems were solved, for new projects it will be essential to do a pre-analysis of the financial structure of the project to prevent time consumption in administrative tasks.

Considering this project followed a participatory design approach, during the Manage phase we needed to take into account an open strategy for change management. In this strategy we defined formal mechanisms to register required changes on the system, but an open participatory mechanism to evaluate the consequences on time and effort to deliver these changes. IT analysts worked together with final users to prioritize changes according to the effects it will have on the improvement of the business value. In such a way, we avoided frequent iterations that will impact project time.

Since the initial problem on this project was ill-structured, the Envision function of the Deliver phase did not include a clearly documented vision and scope of the project. Nevertheless, the fact that the project followed a participatory design approach allowed to reduce the risks of this shallow original definition.

6.3. Discussion

In fact, a long term goal of this research is to gradually improve the quality of patient treatments based on formal evidence gathered through the specific analysis of each disease, especially chronic diseases.

The process used to design and create DISEArch followed an integration of rigor and relevance reinforced through user experience and involvement during the design that enabled the appropriation and evolution of the information system created. This approach demonstrates the value of participatory design to enrich the requirements phase during all the process to assure the quality of the final design and the fulfillment of user expectations. However, the drawback of the approach is the high time consumed to undertake the process, particularly during the design of the user interface. This may not be a disadvantage but the evidence that time typically used for this activity was considerably lower than required.

7. CONCLUSION AND FUTURE WORK

This paper presents a health decision-support system called DISEArch that allows the identification and analysis of relevant EHR for decision-making. It uses structured and non-structured data, and provides analytical as well as visualization facilities over individual or sets of EHR. DISEArch proves to be useful to empower researchers during analysis processes and to reduce considerably the time required to obtain relevant EHR for a study. Technically speaking, the main novel contribution of DISEArch is its ability to analyze multiple EHRs simultaneously prioritizing the most relevant to improve decision making, using a hybrid approach that combines text mining techniques and structured analysis. Contrary to other related works, DISEARch provides a complete set of components that can be used

<table>
<thead>
<tr>
<th>Table 1. DISEArch time performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Manual Process</td>
</tr>
<tr>
<td>Number of EHR</td>
</tr>
<tr>
<td>Time required</td>
</tr>
</tbody>
</table>


to manage, visualize and analyze EHRs. From a methodological point of view, this work represents a successful case of using participative, design-centered service systems engineering to produce a rigorous and relevant solution in the health care domain.

This work opens important research issues. Among them, the extension of DISEARCh design to include temporal analysis over events of EHR with the same diagnosis. This kind of analysis involves the recognition of events from narrative texts and the application of time series analysis to identify temporal patterns. Such event-based patterns should contribute to matching actual treatment against proposed protocols and guidelines, open up traceability for a clinical and administrative use, and provide alternative views on health records that can enhance daily medical practice. In addition, the use of text mining techniques for creating abstract summaries of EHR in DISEArch is promising. The combination of such techniques with the analysis of semantic distance between EHR should also be further developed.

Future research also involves the use of our proposal in other medical institutions leading probably to other analysis requirements and technical issues related to other hospital information systems.

 Probably the most important research perspective we consider is the further use of participatory design for the enrichment of health decision-making system. It is complex but necessary to strike a balance and generate a fruitful discussion from very different sets of expertise. As with any information systems project, a conversation needs to be put in place to realize the full potential that IT-based systems offer for people, in this case within the medical domain. It is a mutual learning experience that requires constant translations, frequent prototype discussions, grounding of new IT-based support in current practices and clear identification of existing problems and future opportunities that are opened up in order to enrich the momentum of the project, enlarge the community of early adopters and guaranteeing the continued financial, scientific and administrative support for the project from management stakeholders. Our experience is very positive and we intend to further pursue this approach and extract lessons learned for similar projects.

ACKNOWLEDGMENT

This work is part of the project entitled “Identificación semiautomática de pacientes con enfermedades crónicas a partir de la exploración retrospectiva de las historias clínicas electrónicas registradas en el sistema SAHI del Hospital San Ignacio” funded by Hospital Universitario San Ignacio and Pontificia Universidad Javeriana.

REFERENCES


Alexandra Pomares is associate professor at the Systems Engineering Department of Pontificia Universidad Javeriana, Bogotá, Colombia, where she performs research related to information integration and knowledge extraction through data mining and text mining techniques. She received her PhD in Informatics from the Université de Grenoble, Grenoble, France and in Engineering from Universidad de los Andes, Bogotá, Colombia in August 2010. Dr. Pomares has been also a consultant in Business Intelligence for different private and public companies in Colombia.

Rafael A. Gonzalez is a Systems Engineer from Javeriana University (Bogotá) with an MSc in Computer Science and a PhD in Systems Engineering (cum laude) from Delft University of Technology (The Netherlands). His research interests are focused on development of information systems with a design science approach, centered on the issues of coordination, complexity and the interplay between ICT and society. His is currently Associate Professor and Chair of the Systems Engineering Department at Javeriana University.

Dario Londoño is Physician and surgeon at the Bosque University, specialist in internal medicine and Pulmonary Medicine at Pontificia Universidad Javeriana, Bogota, Colombia. Master degree in Health Economics and Clinical Epidemiology at McMaster University, Hamilton, Ontario, Canada. His clinical and research work focuses are venous thromboembolic disease, chronic obstructive pulmonary disease and data mining. Currently, he is working at the Department of Internal Medicine, and he is the Pulmonary Medicine Unit director at San Ignacio University hospital and Titular Professor, school of medicine, Pontificia Universidad Javeriana.

Oscar Muñoz is an Assistant Professor at the Department of Internal Medicine, School of Medicine, Pontificia Universidad Javeriana at Bogotá Colombia. Currently is Chief of General Internal Medicine Unit at San Ignacio’s Hospital at Bogotá, Colombia. Dr. Muñoz attended National University of Colombia, where he received an M.D. degree in 1998. His subsequent career included residency in internal medicine at San Ignacio’s Hospital and a Master degree in Clinical Epidemiology at Javeriana University. His research has focused on cardiovascular risk factor, and on development and implementation of Clinical Practice Guidelines in the Colombian Health System.
Olga Milena García Morales is physician and surgeon at the University of Rosario, specialist in internal medicine at the Pontificia Universidad Javeriana. Her clinical work focuses on the management of patients with chronic diseases in hospital areas. Her research has focused on venous thromboembolic disease, chronic obstructive pulmonary disease and data mining over electronic health records. She has multiple scientific publications on national and international journals and books. Currently, she works for the Department of Internal Medicine at San Ignacio hospital and is professor ad Honorem at Pontificia Universidad Javeriana.

Ricardo R. Bohórquez physician and surgeon from the Pontificia Universidad Javeriana School of Medicine class of 1984; specialist in Internal Medicine from Pontificia Universidad Javeriana from the class of 1988; became a specialist in Clinical Cardiology at Nueva Granada Military University in 1990. His principal fields of research include Clinical Cardiology, Acute Coronary Syndrome, venous thromboembolic diseases, echocardiography and data mining over electronic health records. Currently, Dr. Bohórquez is a Senior Professor of Medicine at Pontificia Universidad Javeriana, Clinical Cardiologist at San Ignacio University Hospital and is member of the Internal Medicine Colombian Association and Colombian Cardiology Society.
APPENDIX A

Compliance with Design Science Research Guidelines

In Hevner et al (2004), seven guidelines are proposed to be addressed by all design science research in order for it to be complete and effective. As a reflective assessment of the present work, we provide a table indicating where and how this research addresses each of the guidelines (See Table 2).

Table 2. Design science observations

<table>
<thead>
<tr>
<th>Design Science Guideline</th>
<th>Observation On How it is Followed in this Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design as an artifact</td>
<td>The product of design is always an artifact, which in information systems may be classified as either a concept, model, method or instantiation. This paper describes the way in which a model is developed to enable identifying and prioritizing electronic health records (EHRs). This model is instantiated in a particular software system which contributes to decision support in clinical research.</td>
</tr>
<tr>
<td>Problem relevance</td>
<td>In striking a balance between rigor and relevance, all design science research in information systems, creates a technical artifact to solve a business problem; in this case, the problem is how to identify and prioritize EHRs to aid decision-making in clinical research, where the EHRs contain unstructured text which holds significant but hidden information.</td>
</tr>
<tr>
<td>Design evaluation</td>
<td>Section 6 in this article describes the evaluation carried out to test the quality and efficacy of the model, as instantiated in a first software system prototype, showing a significant improvement in efficiency, when compared to the current process as well as better accuracy in the results.</td>
</tr>
<tr>
<td>Research contributions</td>
<td>Transparent description of the process behind this research is intended to provide design process knowledge related to the method, procedures and overall participative aspects of this project in order for similar problem settings to be treated in a similar fashion. It is precisely through the integration of data and text mining techniques, together with participative, iterative design cycles that this project was able to articulate medical and informatics expertise resulting in a decision support artifact for clinical research. In the same vein, describing the resulting model, its architecture and its application domain, this paper is able to contribute design product knowledge, extending the solution space for similar scenarios within medical informatics and decision-support information systems, especially when the problem data involves unstructured text.</td>
</tr>
<tr>
<td>Research Rigor</td>
<td>On the other side of relevance, a sound design science research project should also be rigorous. In this Project, rigor is anchored in the systematic study and incorporation of applicable knowledge, as described in Section 3. This rigor is followed through by employing an ordered design process and executing several tests on the resulting system.</td>
</tr>
<tr>
<td>Design as a search process</td>
<td>Having a systemic, iterative approach guiding the design process enabled the construction of an artifact that does not merely employ useful knowledge and existing data and text mining techniques and tools, but also configures them and creates a user experience in tune with the exiting technological, organizational and user capabilities, increasing the chance of its acceptance beyond the already mentioned technical quality and effectiveness attributes embedded in the system.</td>
</tr>
<tr>
<td>Communication of Research</td>
<td>The early stages of this project have already been shared with the academic community in conferences and book chapters. However, in line with design science research, it has been also presented in professional (in this case clinical research) contexts and innovation forums, enabling the research team to find not only valuable feedback and insights, but also potential partners for further deployment and extensions.</td>
</tr>
</tbody>
</table>
APPENDIX B

System Development Process

The development process conducted for creating DISEArch follows an iterative approach. The main feature of this approach is the system construction through iterations allowing the system to be gradually refined. This process promotes a permanent communication with final users, who play an important role on the evaluation of the iteration progress.

As illustrated in Figure 10, in order to assure the robustness of the system, the first phase consists of the abstract specification, which objective is to specify the general characteristics of the system and design its global architecture. During this specification the team decides the number and the objective of each one of the iterations.

An iteration is developed based on an initial technical design that ensures it will fit in the general architecture. Then, it is delivered to the medical part of the team for being evaluated. If the evaluation result is successful, the next iteration is developed until the whole system is suitable for the users. In case of having a problem with an iteration, its design is refined and the changes are developed to begin a new evaluation process.

Figure 10. EHR analysis process with DISEArch