KNOWLEDGE
MEDIA
DESIGN
INSTITUTE

KMDI REPORTS

A Search Engine for Structured Health Data

Authors:
Professor Mark Chignell and
Mahsa Rouzbahman, PhD Candidate

KMD-2014-3

A Search Engine for Structured Health Data

Mark Chignell* and Mahsa Rouzbahman**
* Professor, ** PhD Candidate
Department of Mechanical and Industrial Engineering, University Toronto
5 King's College Road
Toronto, ON M5S 3G8, Canada
** Chignel@mie.utoronto.ca, 416-978-8951

Abstract

This paper presents the architecture of a health data search engine, along with preliminary findings that demonstrate the feasibility of the approach taken. The work is motivated by the need to incorporate information about similar patients into clinical decision making, and by the need to develop a tool that can search for similar patients in health data repositories. Central to the design of the search engine is the use of clustering analysis within health data repositories to ensure that responses to queries consist of data summaries that do not violate the confidentiality of patient records. Recent results concerning the feasibility of this search engine approach are reviewed. These results speak to the relative ease of creating clinically meaningful summaries of patient types, and to the accuracy of predictions made using the summarized data. The paper concludes with a brief discussion of further work required to implement a health data search engine and to demonstrate its effectiveness.

Keywords: Privacy; Data Mining; Health Data Search Engine; Clinical Decision Support

1. Introduction

We envision a search engine for structured data that would send queries to multiple health data repositories and aggregate the results into summaries of related patient types. The resulting patient types could then used by applications such as clinical decision support. Within each repository a query sent by the search engine, from outside the repository firewall, would match a set of cases, which would then undergo a clustering process. Non-confidential clustered summaries of relevant patient types existing in the repository would then be sent back to the health data search engine for merging with results from other health repositories. The search engine results could then be browsed directly, or could be passed to other applications for further processing (e.g., predictive analytics that predict values or ranges of unknown data fields in the current patients health record). Further research is needed to determine the detailed design of components of this search engine. The resulting search engine should preserve privacy, be sufficiently fast to be useful in real-time decision making, and should provide useful functionality and acceptable results. In this paper we introduce the general architecture of the health data search engine and present some initial results concerning its feasibility.

2. Background

Increasingly, patient data is being stored in the form of electronic health records and personal health records. Massive amounts of health data are potentially available for informing clinical decision making. Data mining and text mining are being applied within large healthcare data warehouses (Koh and Tan, 2011). Past research has typically focused on looking for novel patterns in healthcare data such as risk factors for poor control of blood sugar levels in diabetics (Breault et al., 2002) or heart attack prediction (Patil and Kumaraswamy, 2009). We would like to extend this approach so that data mining can be carried

out on the fly in response to a wide range of queries representing different types of patient, and medical context.

The traditional medical model has relied on high quality clinical evidence based on experimental studies, and randomized control trials in particular. In recent years, models of treatment and management using evidence-based medicine (e.g., Sackett, 1997) have become dominant. Yet evidence alone is often not sufficient (Sackett et al., 1996, p. 72): "Good doctors use both individual clinical expertise and the best available external evidence, and neither alone is enough. Without clinical expertise, practice risks becoming tyrannized by evidence, for even excellent external evidence may be inapplicable to or inappropriate for an individual patient." What is to be done in situations where relevant clinical evidence is unavailable, where available clinical expertise is insufficient, or where the current case involves a complex set of interactions or co-morbidities for which current guidelines, and the physician's prior clinical experience, are inadequate?

One strategy for better supporting clinical decision making is to make clinical evidence more readily available by bringing it to the point of care. When this is done, physicians frequently change their decisions (Straus, 1999). This suggests that unsupported clinical decision making may often be sub-optimal. An extensive literature on medical decision making has documented inconsistencies in diagnoses and decisions. For instance, Kaplan and Frosch (2005) cited one example where cardiologists disagreed with each other 60% of the time when interpreting a feature in an angiogram, and a second example where cardiologists who were asked to interpret the same angiogram at two different times disagreed with their first assessment between 8% and 37% of the time.

Unsupported medical decision making is prone to inconsistency and occasional error, and the right clinical evidence for a particular case is not always available at the right place and at the right time. In view of these considerations, an alternative strategy is to carry out data mining on similar patients and then present the results, to the physician, as summaries or predictions. This strategy has only become feasible in the past few years, due to increasing computational power at reasonable cost, and the widespread storage of large numbers of health records in electronic data repositories. Ideally, clinical decision support should be available on demand, but without distracting the physician when it is not wanted or needed. The physicians that we have talked to about increased use of data mining-derived suggestions for clinical decision support say that they would be welcome as long as they were sufficiently accurate, relevant, and not "in your face".

Our initial interest in developing a health data search engine arose from our study of emergency physicians (Yu et al, 2010). Yu et al. found that emergency physicians have a strong need for decision support, but have little time to carry out searches or to look at sources of clinical evidence when making decisions. Physicians frequently resort to a general purpose search engine like Google when looking for clinical information (Tang and Ng, 2006). According to Hughes et al., 2009 (p. 651) "…tools such as Wikipedia or Google are used three times more in our sample than PubMed, the 'official' best evidence tool introduced in medical school." Traditional sources of clinical evidence are not well suited to the demands of real-time clinical decision making (Takeshita et al, 2002). The data mining approach described later in this paper is an interesting supplement to the current emphasis on utilizing guidelines and recommendations derived from knowledge translation and distillation of best available evidence into practice guidelines (Straus et al., 2009, 2013).

The decision making process of physicians is often influenced by previous cases they have seen. This is particularly true for older doctors (Choudhry et al., 2005). Physicians often use case based reasoning, so it seems natural to develop clinical decision support tools based on patients who are similar to the current patient (Chan, 2010). Browsing information about similar patients allows the physician to explore hypotheses about the current patient (Natarajan et al. 2010). Fortunately, the presentation of information about similar patients can be done through a background search process that doesn't require additional cognitive effort from the physician (Ebadollahi et al., 2010).

Search in EHR repositories is often difficult because the data is heterogeneous, making it difficult to formulate queries. Nevertheless systems have been developed for implementing search on EHR systems. For instance, EMERSE (The Electronic Medical Record Search Engine) and StarTracker (An Integrated, Web-based Clinical Search Engine) provide search capabilities for the free-text portions of electronic medical records (Hanauer et al. 2006; Gregg et al. 2003; Groves, 1980). QPID (Queriable Patient Interface Dossier) is another search engine attached to an EHR system (Zheng et al. 2011). However, these search

engines are designed to work within repositories and are not designed for ready integration into applications like decision support.

Electronic health records contain massive amounts of patient data that are relevant for patient care (by physicians), health care planning and administrative purposes (by hospital managers, staff and insurance companies) and last but not least, for clinical and epidemiological research (Haux, 2006). Once we shift from search within healthcare repositories, to more generalized search and research across a number of healthcare databases, privacy problems associated with data mining have to be addressed (Kobsa, 2007). Medical facilities are extremely cautious about providing access to data. There is a tradeoff that exists between the need for research and data mining on healthcare databases and the need for privacy Hyman (2000). De-identification is a strategy for maintaining privacy in data sets, but it has its limits (Rothstein, 2010) although some concerns about re-identification can be mitigated with appropriate algorithms El Emam (2011). K-anonymity methods (Sweeney, 2002) have been used to try and guarantee that only groups of a minimum size k can be identified, rather than individuals. However, even after using k-anonymity methods, a solution can be open to attacks (Aggarwal et al., 2006).

How can we extract the important knowledge in large confidential data sets, so as to support physicians, researchers and hospital managers, without violating the privacy of the people whose data is being utilized? In principle, it may be possible to apply a "safety in numbers" approach where large groups of similar patients are clustered and summarized so that the summaries can be communicated to outside sources without providing any information about individuals, or about small groups of patients. This is the strategy adopted in the health data search engine proposed in this paper.

Data mining and text mining have been applied to healthcare date (Koh and Tan, 2011). For instance, a text mining algorithm was applied to predict disease status from discharge summaries (Yang et al. 2009). Among various data mining and machine learning methods mentioned in the research literature (e.g., Everitt et al., 2011; Lucas, P., 2004), clustering methods are most suitable when there is no pre-existing classificatory system available (Duda et al., 2012). Cluster analysis has been used in identifying patient types (Chignell and Stacey, 1981), although it can be challenging to use (Chignell and Stacey, 1980). According to Xu and Wunsch (2005, p. 647), "Cluster analysis is not a one-shot process. In many circumstances, it needs a series of trials and repetitions. Moreover, there are no universal and effective criteria to guide the selection of features and clustering schemes."

While there are many clustering methods, the partitioning approach known as k-means analysis is a particularly efficient technique that can scale up to extremely large data sets. Shailesh et al. (2011) applied both text mining methods (for pre-processing) and traditional data mining methods (k-means analysis) to classify discharge summaries based on length of stay and diagnosis of patients in a hospital. In another study, k-means analysis was used to cluster a heart disease database so as to assist in future prediction of heart attacks (Patil & Kumaraswamy, 2009).

In summary, new methods of clinical decision support are needed, but they should respect privacy norms and concerns. Summarized views of groups of similar patients seem to be to be a promising approach, and will be used as the basis for the health data search engine outlined below. Our particular strategy for construction a health data search engine is based on the following observations:

- Physicians frequently compare the current patient to other patients seen previously
- Large health data sets can be clustered into groups of similar patient types very quickly using efficient methods such as k-means analysis
- Appropriately structured summaries of relatively large groups of patients should preserve a satisfactory level of health data confidentiality

3. Prerequisites for a Health Data Search Engine

The first prerequisite for a health data search engine is the availability of large amounts of open health data that can be used for data mining. In general, the more data that is available, the more implicit knowledge will be embedded within that data in the form of statistical relationships, and the higher the quantity and quality of predictions that will be possible with it.

As of this writing, there seems to be a trend towards open health data which seems likely to continue and grow. For instance, the physionet.org MIMIC II database provides data for over 30,000 ICU and NICU patients from a Boston Hospital. Physionet claimed that, each month, about 45,000 visitors worldwide use PhysioNet, retrieving about 4 terabytes of data, as of May 2014. As a second example, the Heritage Health Prize (HHP) was a recent (as of this writing) global data mining competition to predict, by using claims data, the number of days patients will be hospitalized in a subsequent year. El Amami et al (2012) performed an evaluation of re-identification risk on the HHP data using simulated attacks and matching. They found that the probability of re-identification was acceptably low when it was released to a global user community in support of the analytics competition.

While developments in open health data are encouraging, at current rates of database creation and population it would take many years before there was a sufficient critical mass of data to support a general health data search engine that could service queries referring to a wide range of patient types. To be fully effective, a health data search engine must have access to many thousands of health data repositories and health records from millions of patients, just as Google has access to the contents of millions of Web servers. But organizations will only connect their health data repositories to a search engine if the privacy of patient data can be guaranteed. Thus, in our model of the health data search engine, data stays in the repository, behind the firewall, and only non-confidential summaries of patient data are communicated through the firewall to the search engine, in response to each query.

In formulating the search engine, we assume that we can get non-confidential patient summaries from health data repositories in response to queries. We also assume that it is possible to generate meaningful patients summaries of data within a health repository in close to real time (in the order of no more than a few seconds). In contrast to the situation with a search engine like Google, the search engine cannot crawl the data repository and index the data ahead of time (due to privacy concerns). So the search engine has to rely on an efficient clustering process within the data repository (and behind the firewall) to generate the non-confidential summarized data that the search engine needs on the fly (and in a timely fashion) in response to each query that it submits to the repository.

A third assumption is that the summarized data collected by the search engine from the health data repositories will actually be useful. While there are different ways in which the summarized data might be used, one application that we are focusing on is the use of the summarized data to make predictions and to support clinical decision making.

4. Architecture of a Health Data Search Engine

Above all, a health data search engine has to preserve the confidentiality of patient data. In the health data search engine envisioned here, this requirement is met by only dealing with summarized data, where any summarized data exported outside its “home” data repository represents a minimum of at least 100 patients, and where the clustering process that produces each summarized grouping is “opaque” in the sense that any mapping between cluster summaries and individual patient records cannot be inferred.

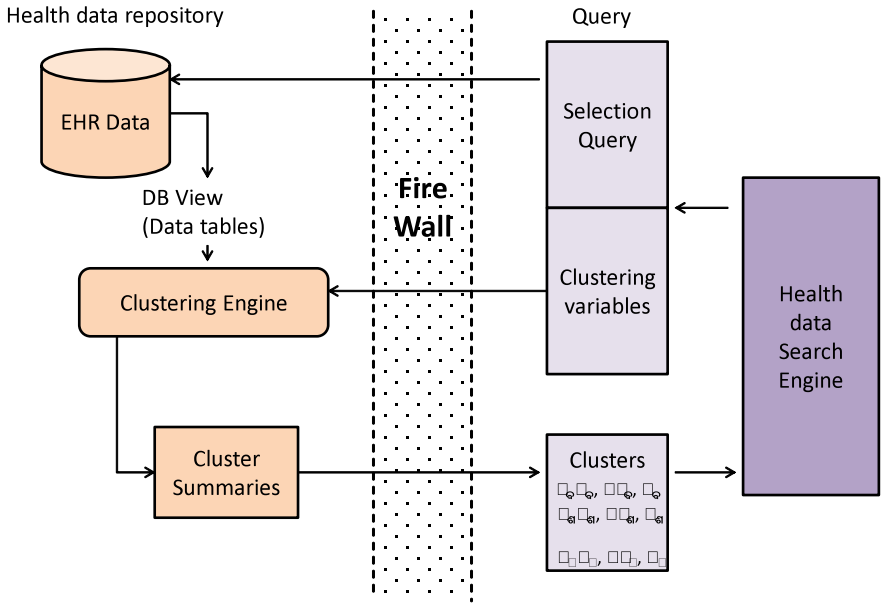

Fig 1. Clustering of Query Results within a Health Data Repository

Figure 1 shows the basis clustering process on which the health data search engine is built. Instead of indexing the contents of pages on Web servers, as is done in a text search engine like Google or Bing, the health data search engine relies on summarized clusters that are produced at query run-time within each one of a number of health data repositories (conceptually equivalent to web servers) in response to the queries. Within each repository a master database table (view) is generated (Figure 1) based on the set of patient records that match the selection query sent from the search engine. This table of data is then sent to the clustering engine, which also receives the list of variables that are to be clustered (passed on from the search query). The output of the clustering is then a set of cluster summaries (consisting of means, standard deviations, and inter-correlation matrices of variables for all the clusters) that are sent back to the search engine.

While there are many clustering methods available, in our design of the search engine we use k-means analysis, since it has yielded good results in our feasibility studies. A particular search could be carried out across many health data repositories and in that case matching clusters from different repositories could be merged (e.g., based on similarity of cluster centroids) in order to create an overall representation of patient types that are relevant to the current query.

K-means analysis is extremely efficient and can be carried out quickly even on very large data sets. It has the desirable property of having only a linear increase in computation time as the number of cases, or variables being clustered, increases. The input to the cluster analysis is the set of health records, within the repository, that match the query that is input through the search engine. Figure 2 shows a representation of the incoming query as having two appended components, the selection query (e.g., women between the ages of 60 and 70 who are in intensive care) that is used to select relevant data records, and the vector of variables to be used in the cluster analysis.

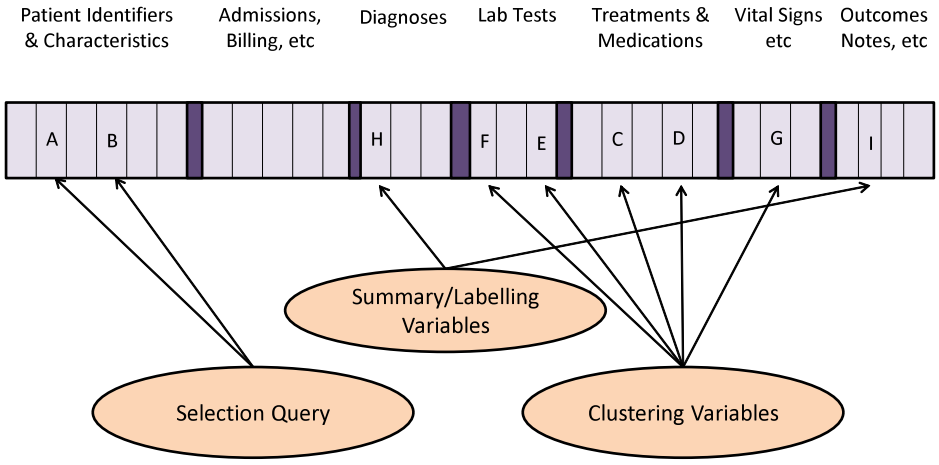

Fig 2. The health Record as a query

The choice of variables and values to include in the selection query will depend on the patient and the context. In the example shown in Figure 2, Value A (e.g., female) for variable X3 (e.g., gender) and value B (e.g, 60-70) for variable X5 (age) are used in the selection. In another query one or more diagnoses might also be used in the selection query, but in this case diagnosis H (as well as outcome I) will be used to summarize and label the clusters that are identified. Once the set of patients is returned based on the selection query shown in Figure 2, lab tests F and E, treatments C and D, and vital signs G are used as variables in the cluster analysis. While only a few variables are shown in the example represented in Figure 2, in real situations there may be hundreds of variables used in the clustering.

Once the k-means cluster analysis is carried out on the group of data records defined by the selection query, the resulting clusters are then summarized in terms of the means, standard deviations and inter-correlations of the corresponding features/variables.

Figure 3 shows a schematic representation of the overall search process (due to space considerations in the figure the generated patient types for repository j are not shown but they would have similar structure to those shown for repositories i and k). Note also that the labels for patient types 1 to N as shown in Figure 3 are local to each repository (e.g., for one repository the total number of clusters might be 110 while for another repository the number N of clusters returned from the cluster analysis might be 120). Each of the data repositories connected to the health data search engine has an embedded clustering engine that works in the way shown in Figure 1. Each of the repositories sends the results of the cluster analyses back to the search engine (one set of cluster summaries per data repository). There could potentially be thousands of sets of cluster summaries if there are thousands of repositories being used to process queries. We describe a proposed method for merging clusters below. However, as the number of repositories being used increases, the merging process might become computationally expensive. Thus it may be simpler and more efficient to simply find the k nearest clusters to the current patient of interest and then make predictions and provide decision support based on the summaries for those clusters, somewhat similar to the way in which millions of hits in a Web search engine are presented in rank order and most people only view the first (and maybe sometimes second) page of hits. Once large numbers of repositories become available for use with a health data search engine, research should be carried out to determine an appropriate value of k for the k-th nearest cluster method for making predictions based on a particular patient of interest. This process of continuously revising the algorithms and user interface to make search engines more effective has been

going on ever since early Web search engines such as OpenText, AltaVista, and HotBot, and we should also expect it to happen with health data search engines once they become available.

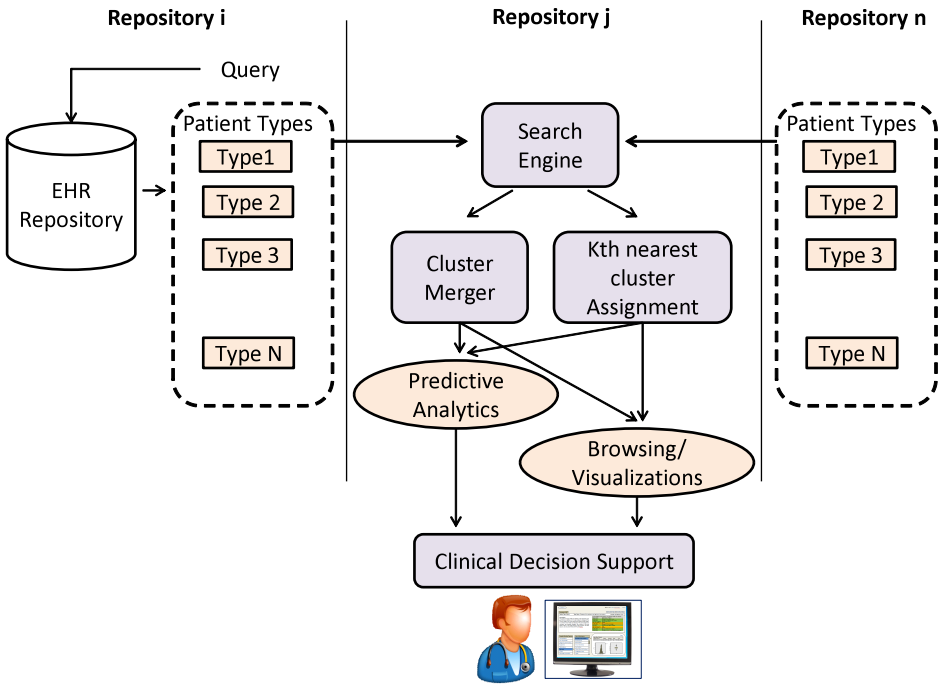

Fig 3. Overall Architecture of a Health Data Search Engine Connected to a Clinical Decision Support System.

Once a set of matching clusters is identified for the current patient of interest (whether by merging clusters or by using k-th cluster assignment), the resulting cluster summaries can be used by predictive analytics (e.g., regression modeling) or may be passed to visualization or browsing tools that would be used by the physician who is trying to make decisions about the current patient. Ideally, the results of predictive analytics, as well as tools for browsing and visualization of summarized clusters of matching patients, would be shown within the user interface of a usable clinical decision support tool, perhaps embedded within the electronic health record of the current patient, as indicated schematically at the bottom of Figure 3.

In summary, the basic steps in the health data search engine model envisioned here are:

1. Select the Query Seed (e.g., the current patient health record)
2. Formulate the query (which will have a selection component that selects relevant patient records plus a clustering component which indicates which variables are to be used in clustering the patient)
3. Pass the Query to a set of health data repositories
4. Wait for the clustering process to be carried out by the health data repositories
5. Receive the set of summarized clusters from each of the health data repositories that report back within some pre-set time (where that time could be a property of the search engine or could be specified by a service/application that calls the search engine)
6. Create a merged representation of all the summarized clusters from the different health data repositories
7. Pass the summarized clusters back to the service/application that submitted the query to the search engine

During Step 4 above, a clustering engine within a participating health data repository would do the following:

1. Select all the patients that match the query submitted by the search engine
2. Cluster the selected patients
3. Select the set of clusters that exceed some minimum size (e.g, 100, to minimize the possibility of re-identification of patient data)
4. Create a summary representation of each selected cluster
5. Pass the summarized versions of the selected clusters back to the search engine

We currently conceive of the health data search engine not as a general search engine that is queried by the general public, but rather as a secure service that is accessed by privileged services and users. In the first instance access to the search engine might be restricted to users within particular institutions (e.g., hospitals, universities, and government ministries). There are, no doubt, many different types of service and application that could utilize the health data search engine envisaged here. One use case that we utilized in formulating the search engine is a form of clinical decision support where unknown fields are populated by predictions made on the basis of the summarized clusters returned by the search engine. In this case regression analysis would be carried out on the summarized clusters and then the current patient's health record would be annotated by suggestions based on the resulting predictions. Annotations might include suggested lab tests, possible diagnoses, and treatments or therapies to consider. Note also that a similar health data search engine could be constructed for personal health records. However, in this paper we focus on electronic health records where the data is created and "owned" by healthcare organizations such as hospitals.

5. Feasibility Evaluation

Prior to proposing the health data search engine in this paper we carried out research on its feasibility using the MIMIC II ICU data set. We used this data set because it is fairly substantial (covering over 20,000 adult patients) and it was readily available (at the time of writing) from physionet.org. In our first feasibility analysis we showed that it is possible to cluster patients (Chignell et al., 2013) from this database automatically (quickly) and blindly (without using any specialized medical knowledge) and still get fairly good results (based on interpretation of the clusters by a physician).

In a further feasibility study we asked whether summarized clusters returned in response to queries could be useful. Rouzbahman and Chignell (2014) showed that predictions based on summarized data (prediction of death status in the ICU in this case) could be more accurate than predictions made on raw/individual data. Thus, for the MIMIC II data set (and presumably other data sets or repositories as well) prediction of unknown values based on summarized health data is in fact feasible.

Table 1. Accuracy of Prediction of Death Status for Summarized Regression vs. Three Standard Methods Applied to the Case Level (Raw) Data (from Rouzbahman and Chignell (2014, Table V1).

Prediction Method	Acc.
Linear Regression	75.4
Logistic Regression	77.2
Discriminant Analysis	77.2
Summarized Regression	81.9

Table 1 summarizes the overall results obtained by Rouzbahman and Chignell (2014). For the MIMIC II data set that they used, prediction of death in the ICU was around 5% more accurate for regression analysis based on the summarized clusters than it was for logistic regression or discriminant analysis carried out on the raw data. Rouzbahman and Chignell explained this possibly surprising result in the following way:

> *"How can predictions based on summarized data outperform predictions based on the original, un-summarized data? Assigning people to clusters likely boosts accuracy because differentiation of patients into clusters effectively partitions the high dimensional multivariate space of predictor variables into subspaces that are likely to be governed by different regression models within each subspace. If regression relationships differ within different regions of the multivariate space, then separately fitting different regression models within different subspaces should lead to more accurate fitting than trying to fit a single model over the entire high dimensional space."*

s

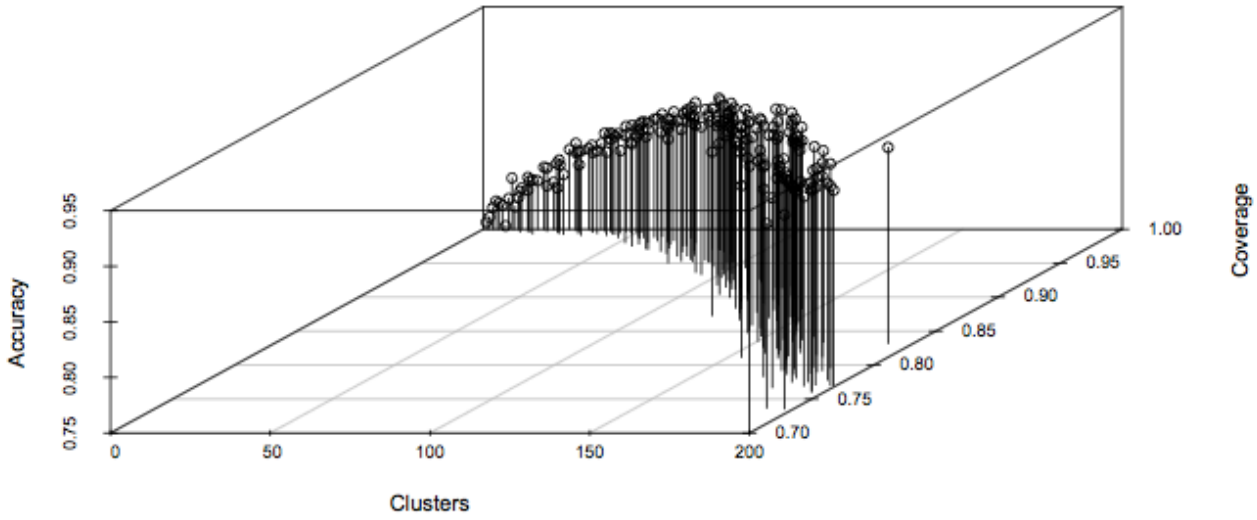

Figure 4. Observed relationship between number of cluster partitions, patient coverage, and accuracy when predicting death in the MIMIC II ICU data set.

There are a number of technical issues in implementing a clustering engine inside a health data repository. These include dealing with sparse variables (missing data), dealing with hierarchical categories (e.g., diagnostic codes), handling outliers, and dealing with multi-collinearity (highly correlated variables). Some of these issues are discussed in more detail by Rouzbahman and Chignell (2014) or are covered in statistical texts such as the one by Field, Miles, and Field (2012). There are also issues concerning the relative merits of different clustering approaches (e.g., k-means analysis vs. hierarchical clustering methods) that are outside the scope of this paper. In our approach we use k-means analysis on normalized (z-score) data with a Euclidean distance proximity measure.

Recent research in our lab suggests that the accuracy estimate of 82% for predicting death in the MIMIC II ICU data set (Table 1) is in fact a significant underestimate of the accuracy that can be achieved when using regression analysis on summarized patient types. Chignell and Rouzbahman (2014) chose k=40 after visually inspecting the results from a number of different values of k, where k varied between 1 and 50. However, when larger numbers of k (partitions) were subsequently considered there was a tendency for accuracy to keep increasing as the number of clusters increased until a point was reached where coverage (the proportion of patients who were in clusters with at least 100 members) dropped off precipitously (Figure 4). Figure 4 was created by running 199 k-means cluster analyses on the MIMIC II data (with k varying from 2 to 200) and by then using summarized regression on the results of each cluster analysis to

obtain accuracy scores. The percentage of patients (coverage) who were members of clusters of more than 100 patients in each analysis was also calculated, and then plotted, along with the number of partitions, and resulting predictive accuracy, associated with each of the clustering analyses.

For practical implementation of the search engine, automated methods are needed to choose the right value of k. One approach is to cluster using a number of different values of k and to then choose a resulting cluster analysis that represents around 90% of people (after applying the constraint that people/patients have to be within clusters with at least 100 people in them). We are currently conducting research on developing more formal methods for choosing an appropriate value for k in clustering patient data. Our initial results suggest that (for the specific case of predicting death within the MIMIC II database) it is possible to choose values of k that cover around 90% of selected patients while still making good predictions (around 90% accuracy in the case of predicting death in the ICU) and ensuring that a minimum sample size of at least 100 patients per cluster is adhered to. Since the purpose of the search engine is to uncover trends and relationships in similar patient data, clusters that cover more than 85-90% of matching patients while providing a reasonable level of accuracy should be more than adequate and may not create too much bias in the sample of patients used in making predictions. Note that "reasonable accuracy" in this case would reflect a level of accuracy that is significantly better than that achieved by unaided clinical decision makers, and/or that is relatively close to, if not better than the accuracy of alternative decision support methods that are less efficient, less readily available when needed, or inconvenient to use.

The pattern of results shown in Figure 4, where accuracy increases as the number of clusters increases up to a point where coverage drops to unacceptable levels, is likely to hold for other sets of data and other predicted outcomes, although the levels of accuracy and coverage achieved are likely to vary widely for different clinical contexts. While we observed a global pattern of increasing accuracy with increasing number of partitions until coverage starts to suffer, the local pattern of accuracy and coverage in adjacent partition sizes is much more variable, as shown in Figure 5. Figure 4 shows a generally increasing trend in accuracy as the number of clusters increases. Nevertheless, within the region of the space that represents the "Sweet Spot" of relatively high accuracy but with good coverage, levels of accuracy and coverage may bounce around quite a bit from one partition size to the next. Further research is needed to explore the relationship between number of partitions, prediction accuracy, and patient coverage in different contexts.

Our current view is that selecting a small set of cluster analyses, with three or four well chosen values of k, should be enough to create a solution that provides good coverage of the selected patients while also subdividing the multivariate space of patient data in a way that is likely to generate relatively good predictions in subsequent regression analysis. The issue of how to specify the clustering process within repositories to optimize prediction is beyond the scope of this paper. However, we believe that a solution should be possible and that it may involve running a small number of cluster analyses within a predicted sweet spot where a given level of coverage (e.g., around 90%) is achieved. These clusters (from within a single repository) could then be merged (using a process similar to that described below for merging clusters obtained from different repositories) and then passed back to the search engine as a single merged set of clusters.

Given the efficiency of the k-means technique and the ready availability of computational power we do not expect that running a few cluster analyses per query within a repository will be particularly onerous. Further research is needed to determine how to select the number of cluster partitions k in different contexts and how best to merge solutions from multiple clusterings generated by different values of k. One further idea that could be borrowed from existing search engines is to feed back results from use of the search engine into revisions of the clustering and search engine algorithms. For instance, relevance feedback implied by user selections of hits to view may be used for improving the performance of Web search engines (Joachims et al., 2007). In the case of health data search engines, click-throughs on recommendations and suggestions made in the context of clinical decision support, or on patient types made available for browsing or visualization, might be referred to machine learning methods, to improve the overall search performance and prediction accuracy. The machine learning could be based on parameters used in the clustering and merging processes and parameter assignments could be specialized for particular contexts/diagnoses. Relevant parameters that could be tuned using feedback from user behaviour might include the number of partitions, k, the level of required accuracy, the minimum cluster size, etc.

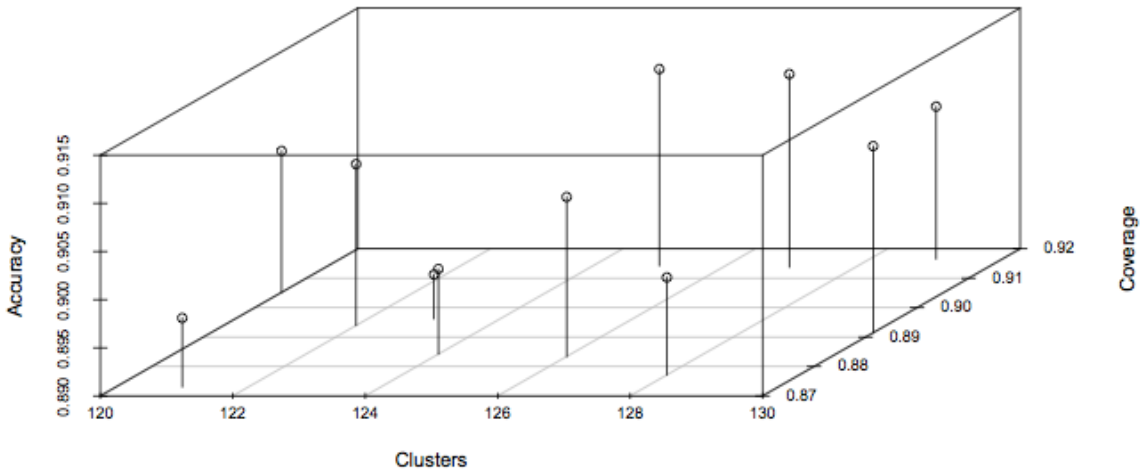

Figure 5. Variability in Accuracy in Death Prediction and Patient Coverage For Nmber of Cluster Partitions Ranging between 120 and 130.

We do not claim that the clustering approach described above is optimal, but we have found that it yields results that are good enough to demonstrate the feasibility of our search engine approach. While more work needs to be done to fine tune the approach, in the first case study that we carried out we were able to identify clustering solutions that predicted death for around 90% of ICU patients with around 90% accuracy (in the revised analyses summarized in Figures 4 and 5).

6. Merging Clusters

When general Web search engines like Google or Bing perform a search they query many web servers, but present the results of the search as a single merged list. When clusters are returned from multiple health data repositories, how should the results be merged into a single set of matching patient types?

We propose a simple cluster merging procedure that is within the spirit of the k-means analysis clustering procedure. Imagine that we have sets of summarized clusters from a total of 100 different health data repositories. As a simplifying assumption, in the heuristic merging process described here we use the cluster centroids (the vector of variable means) as the basis for merging in the following steps:

1. Pick five of the repositories as the seeds for the merging process.
2. For each repository, and for each cluster within the repository, go through the clusters and repositories in a single pass (considering each cluster only once). For each cluster, find the closest (using Euclidean distance) cluster within the four other repositories and merge with it, where mergers may also occur with previously merged clusters (similar to the average linkage method of hierarchical clustering. but where there is only a single pass through all the clusters in all the repositories). This will result in a set of merged clusters containing either two, three, four or five of the original clusters.
3. Select for further processing (across the cluster sets representing the remaining repositories) those clusters that contain at least three of the original clusters (i.e., discard combined clusters where only two of the original clusters were merged since they are indicative of a degree of "idiosyncracy" or disagreement in how patients should be clustered between repositories).
4. Continue with a single pass through all the clusters returned from the remaining repositories, matching to the merged clusters retained from the analysis of the set of seed repositories, and merging clusters with revised centroids as each closest matching (merged or un-merged) cluster is identified. For each selected merged cluster calculate the combined centroid as the weighted mean of centroids of the component clusters within the merged cluster (where weighting is done based on the number of people represented in each of the clusters). Note that different versions of the

proposed merging algorithm can be developed depending on whether or not (and how) new clusters are created in addition to the merged clusters developed from the first seed set of repositories.

5. Once all the clusters have been merged, recalculate the cluster centroids by again calculating the weighted mean centroid based on the centroids of the clusters that have been merged, and adjusting for the number of people represented by each cluster.
6. Adjust the summary of standard deviations and inter-correlations within each cluster using a weighted averaging process (the details of this last step are outside the scope of this paper and should be based on research that explores this issue specifically).

The merging process described above is meant to be a rough description of the process rather than a definitive algorithm. Since we didn't have access to data from multiple repositories in carrying out the research reported here there was no way to verify that methods such as those listed above will work well in practice (although we expect that some iterative process similar to that described above will likely work well).

One remaining challenge in terms of feasibility of the current approach is to show that the kind of information made available through health data search in clinical decision support, or other applications, is useful. We are currently working on the development of motivating examples to show how the results of summarized prediction using a health data search engine can be viewed and interacted with using tablet-based access to a clinical decision support tool embedded within a patient information system or electronic health record.

7. Steps Towards Implementation of the Health Data Search Engine

In this section we list a possible set of steps towards the development of the health data search engine, recognizing that the order and composition of steps may change over time as experience is gained about the tasks and processes associated with implementing clustering engines and health data search.

Step 1. The health data search engine envisaged here assumes the existence of clustering engines within health data repositories so that query results can be clustered and summarized on the fly. A natural first step would be the implementation of a clustering engine within a single large health data repository. The first version of the health data search engine could then query that health repository.

Step 2. Develop Use Cases that motivate use of this first version of the Health Data Search Engine.

Step 3. Evaluate the first version of the health data search engine and revise the search engine and the clustering engine to address shortcomings observed and any associated privacy concerns.

Step 4. Implement clustering engines within a set of data repositories and link those repositories to a revised version of the search engine that handles multiple repositories.

Step 5. Develop methods and training materials to help the technical administrators of an unlimited number of health data repositories to install their own clustering engines.

Step 6. Develop a validation process to ensure that clustering within a new health data repository is producing sufficiently accuracy results before it is linked to the health search engine data network for the first time.

Step 7. Add on applications and services that can utilize the health data search engine to improve a variety of activities including clinical decision making.

8. Conclusions

It seems possible, and perhaps likely, that health data search engines of the type envisaged in this paper could revolutionize many aspects of healthcare. Data mining of the voluminous collections of electronic health data now available should be a valuable supplement to high quality clinical evidence, and clinical expertise.

Up until now large scale data mining of electronic health data across different health data repositories has seemed infeasible because of concerns about the privacy and security of health data. Researchers and developers have been faced with an admirable, but restrictive, policy where the privacy of the individual is paramount (Jonas, 1969) and this policy is also reflected in policy. Relevant policies include Canada's Tri-Council Policy Statement on ethical requirements for human subjects research, and the Canadian PIPEDA (Personal Information Protection and Electronic Documents Act) and the US HIPAA (Health Insurance Portability and Accountability Act) regulations. However, the potential benefits of large scale health data search and data mining are great, and methods are needed to allow new methods of clinical decision support to be explored while still preserving ethical and privacy requirements. Ideally, methods should be developed that provide useful clinical decision support based on data mining of similar cases, but without compromising the confidentiality of patient records or violating current privacy regulations.

We propose the development of a health data search engine that uses clustered views of patient health data repositories where the summarized clusters are large enough, and the clustering process opaque enough, to ensure that the confidentiality of patient health records is sufficiently protected. Such a search engine could be useful for a variety of tasks, including clinical decision support via access to similar patient types.

The relevant summaries of patient types returned by the health data search engine could be used for predictive analytics and clinical decision support, all without exposing private data outside the hospital firewalls. We are currently using minimum cluster sizes of at least 100 patients per cluster, and the summaries of each cluster consist of the means, standard deviations, and inter-correlations between variables. Since the cluster summaries do not contain identifying information relating to individual patients, and the summaries refer to unknown groups of at least 100 patients per cluster it would be difficult if not impossible, for a malicious agent to infer anything about individual patient records based on cluster summaries provided in response to queries. Further safeguards for preserving privacy when using this type of health data search might include only generating responses from a data repository in cases where at least 1,000 patients are clustered (i.e., are selected by the query), and using different random variations of the clustering procedure at different times to further confuse attackers and to remove any possibility of reverse engineering of the clustering process. Access to the health data engine could also be restricted to qualified users working within specific institutions such as hospitals.

The results of feasibility studies cited in this paper show that it is possible to automatically generate meaningful clusters (patient types) in health data repositories, and that relatively accurate predictions can be made based on summaries of those clusters. While we have demonstrated the feasibility of our approach in the context of the MIMIC II data set it remains to be seen if similar patterns of results will be found with other data sets as well. We are not claiming that our approach to health data search will work for all queries and clinical decisions, but we believe that it should work for a significant fraction of the clinical decisions that physicians face. It remains to be seen whether our vision of a general health data search engine can be implemented successfully, or whether more specialized versions of the health data search engine are needed that work in specific domains.

In this paper we described the architecture of the search engine and some initial results that demonstrate its feasibility and usefulness. While we believe that the search engine can also be implemented so as to provide strong safeguards for the privacy of individual health records, a detailed analysis of the privacy implications of the search engine is beyond the scope of this paper.

Acknowledgment

This research was supported by a Google Faculty Research Award, and by a Discovery Grant from the National Science and Engineering Research Council of Canada (NSERC), both to the first author.

References

Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., & Zhu, A. (2006, June). Achieving anonymity via clustering. In Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 153-162). ACM.

Breault, J. L., Goodall, C. R., & Fos, P. J. (2002). Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine*, 26(1), 37-54.

Chan, L.W.C (2010). Machine Learning of Patient Similarity, 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops.

Choudhry, N.K. et al. (2005). Systematic Review: The Relationship Between Clinical Experience and Quality of Health Care. *Annals of Internal Medicine*, No. 142, Issue 4.

Chignell, M.H. and Stacey, B.G., (1981), "The Classification of Patients into Diagnostic Groups", *Journal of Clinical Psychology*, 37, pp. 151-153.

Chignell, M.H. and Stacey, B.G., (1980a), "Exploratory Cluster Analysis of Attitude Structure in the New Zealand Electorate", *Psychological Reports*, 47, p. 258.

Chignell, M.H. and Stacey, B.G., (1980b), "Practical Problems Associated with the Use of Cluster Analysis", *Psychological Reports*, 46, pp. 131-134.

R.O. Duda, P. E, Hart and D. G. Stork, "Pattern classification," John Wiley & Sons, 2012.

Ebadollahi, S. Sun, J. Gotz, D. Hu, J. and Sow, D. Neti, Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognosis. Proceedings of AMIA 2010, Nov. 2010.

El Emam, K., Arbuckle, L., Koru, G., Eze, B., Gaudette, L., Neri, E., & Gluck, J. (2012). De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. *Journal of medical Internet research*, *14*(1).

Everitt, B.S., Landau, S. Leese, M., & Stahl, D. 2011. Cluster analysis. London: Wiley.

Field, Andy, Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. London: SAGE Publications.

Gregg, W., Jirjis, J., Lorenzi, N.M. and Giuse, D. StarTracker: An Integrated, Web-based Clinical Search Engine. Department of Biomedical Informatics Vanderbilt University Medical Center, Nashville, Tennessee, AMIA 2003 Symposium Proceedings − Page 855.

Groves, W.E. Storage and retrieval of coded patient diagnoses and data on a clinical laboratory computer system. *Comput Programs Biomed*. 1980 Dec; 12(2-3):225-29.

Hanauer, D.A. EMERSE: The Electronic Medical Record Search Engine. Department of Pediatrics, Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, Michigan, AMIA 2006 Symposium Proceedings Page – 941.

Haux, R. Health information systems_past, present, future. International Journal of Medical Informatics (2006) 75, 268-281.

Hughes, B., Joshi, I., Lemonde, H., and Wareham, J. (2009). Junior physician's use of Web 2.0 for information seeking and medical education: A qualitative study, *International Journal of Medical Informatics*, 78(10), 645-655.

Hyman, S. E. (2000). The needs for database research and for privacy collide. *American Journal of Psychiatry*, 157(11), 1723-1724.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, *25*(2), 7.

Kobsa, A. (2007). Privacy-Enhanced Personalization. Communications of the ACM (50:8), pp. 24-33.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Managemen*t—Vol, 19(2), 64-72.

Lucas, P. (2004), Bayesian analysis, pattern analysis, and data mining in health care. *Curr Opin Crit Care* 10:399–403.

Natarajan, K., Stein, D., Jain, S. and Elhadad, N (2010). An analysis of clinical queries in an electronic health record search utility. *Int J Med Inform*. 79(7):515-22.

Patil, S.B. and Kumaraswamy, Y.S. Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network. *European Journal of Scientific Research*, Vol.31 No.4 (2009), 642-656.

Rothstein, M. A. (2010). Is deidentification sufficient to protect health privacy in research?. *The American Journal of Bioethics*, 10(9), 3-11.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ: British Medical Journal*, *312*(7023), 71.

Sackett, D. L. (1997, February). Evidence-based medicine. In *Seminars in perinatology* (Vol. 21, No. 1, pp. 3-5). WB Saunders.

Shailesh, T., Acharya, D. And Shailesh, K.R. Computerized storing of electronic health care records using rext mining. *JPBMS*, 2011, 5 (04).

Straus, S. E. (1999). Bringing evidence to the point of care. *Evidence Based Medicine*, *4*(3), 70-71.

Straus, S., Tetroe, J., & Graham, I. D. (Eds.). (2013). *Knowledge translation in health care: moving from evidence to practice*. John Wiley & Sons.

Straus, S. E., Tetroe, J., & Graham, I. (2009). Defining knowledge translation. *Canadian Medical Association Journal*, *181*(3-4), 165-168.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty*, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Takeshita, H., Davis, D., & Straus, S. E. (2002, September). Clinical evidence at the point of care in acute medicine: a handheld usability case study. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 46, No. 16, pp. 1409-1413). SAGE Publications.

Tang H and Ng J.H.K. (2006). Googling for diagnosis—the use of Google as a diagnostic aid: internet based study. *BMJ* 2006; 333:1143–1145.

Van Couvering, E. (2008) "The History of the Internet Search Engine: Navigational Media and the Traffic Commodity," in Spink, A., and Zimmer, M., (eds.), Web Search: Multidisciplinary Perspectives, Berlin: Springer, pp. 177-206.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. Neural Networks, IEEE Transactions on, 16(3), 645-678.

Yang, H., Spasic, I., Keane, J.A. and Nenadic, G. A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries. *J Am Med Inform Assoc*. 2009;16: 596–600.

E.Yu, R. Kealey, M. Chignell, J. Ng and J Lo. (2010). Smarter Healthcare: An Emergency Physician View of the Problem. In M. Chignell, J. Cordy, J. Ng., and Y. Yesha, (Eds). The Smart Internet: Current Research and Future Applications. Lecture Notes in Computer Science, 6400. Berlin: Springer.

Zheng. K. et al. (2011). Collaborative search in electronic health record, *J Am Med Assoc*, 18, 282-291.