Journal of Big Data
a SpringerOpen Journal

SURVEY PAPER Open Access

A review of data mining using big data in health informatics

Matthew Herland, Taghi M Khoshgoftaar and Randall Wald*

*Correspondence: rwald1@fau.edu
Florida Atlantic University, 777 Glades Road, Boca Raton, FL, USA

Abstract

The amount of data produced within Health Informatics has grown to be quite vast, and analysis of this Big Data grants potentially limitless possibilities for knowledge to be gained. In addition, this information can improve the quality of healthcare offered to patients. However, there are a number of issues that arise when dealing with these vast quantities of data, especially how to analyze this data in a reliable manner. The basic goal of Health Informatics is to take in real world medical data from all levels of human existence to help advance our understanding of medicine and medical practice. This paper will present recent research using Big Data tools and approaches for the analysis of Health Informatics data gathered at multiple levels, including the molecular, tissue, patient, and population levels. In addition to gathering data at multiple levels, multiple levels of questions are addressed: human-scale biology, clinical-scale, and epidemic-scale. We will also analyze and examine possible future work for each of these areas, as well as how combining data from each level may provide the most promising approach to gain the most knowledge in Health Informatics.

Keywords: Big data; Health informatics; Bioinformatics; Neuroinformatics; Clinical informatics; Public health informatics; Social media

Introduction

The field of Health Informatics is on the cusp of its most exciting period to date, entering a new era where technology is starting to handle Big Data, bringing about unlimited potential for information growth. Data mining and Big Data analytics are helping to realize the goals of diagnosing, treating, helping, and healing all patients in need of healthcare, with the end goal of this domain being improved Health Care Output (HCO), or the quality of care that healthcare can provide to end users (i.e. patients).

Health Informatics is a combination of information science and computer science within the realm of healthcare. There are numerous current areas of research within the field of Health Informatics, including Bioinformatics, Image Informatics (e.g. Neuroinformatics), Clinical Informatics, Public Health Informatics, and also Translational BioInformatics (TBI). Research done in Health Informatics (as in all its subfields) can range from data acquisition, retrieval, storage, analytics employing data mining techniques, and so on. However, the scope of this study will be research that uses data mining in order to answer questions throughout the various levels of health.

Each of the studies done in a particular subfield of Health Informatics utilizes data from a particular level of human existence [1]: Bioinformatics uses molecular level data,

Neuroinformatics employs tissue level data, Clinical Informatics applies patient level data, and Public Health Informatics utilizes population data (either from the population or on the population). These subfields do sometimes overlap (for example, a single study might consider data from two adjacent levels), but in the interest of minimizing confusion we in this work will classify a study based on the highest data level used (as this paper will be structured according to data usage). In addition, within a given data level we will break down studies based on the type (i.e., level) of question a study attempts to answer, where each question level is of a relatively comparable scope to one of the data levels. The tissue level is of analogous scope to the human-scale biology questions, the scope of patient data is related to clinical questions and the scope of the population data is comparable to the epidemic-level questions.

The scope of data used by the subfield TBI, on the other hand, exploits data from each of these levels, from the molecular level to entire populations [1]. In particular, TBI is specifically focused on integrating data from the Bioinformatics level with the higher levels, because traditionally this level has been isolated in the laboratory and separated from the more patient-facing levels (Neuroinformatics, Clinical Informatics, and Population Informatics). TBI and the idea of combining data from all levels of human existence is a popular new direction in Health Informatics. The main level of questions that TBI ultimately tries to answer are on the clinical level, as such answers can help improve HCO for patients. Research throughout all levels of accessible data, using various data mining and analytical techniques, can be used to help the healthcare system make decisions faster, more accurately, and more efficiently, all in a more cost-effective manner than without using such methods.

This paper is organized as follows: Section “Big data in health informatics” provides a general background on Big Data in Health Informatics. Section “Levels of health informatics data” delineates and discusses the subfields: Bioinformatics, Neuroinformatics, Clinical Informatics, and Public Health Informatics, as well as discussing their corresponding data and question levels (used to structure Sections “1” through 1). Section “Using micro level data – Molecules” looks at studies using data from the micro (molecular) level while the studies in Section “Using tissue level data” use data from the tissue level, Section “Using patient level data” covers research being done on the patient level, and the research efforts in Section “Using population level data – Social media” employ data gathered at the population level. The sub-field of TBI is discussed in Section “Translational bioinformatics”. Section “Analysis and future works” analyzes the works shown and discusses both possible future work and the possible directions each line of research could take. Section “Conclusion” contains our conclusion.

Big data in health informatics

The term Big Data is a vague term with a definition that is not universally agreed upon. According to [2], a rough definition would be any data that is around a petabyte (10^{15} bytes) or more in size. In Health Informatics research though, Big Data of this size is quite rare; therefore, a more encompassing definition will be used here to incorporate more studies, specifically a definition by Demchenko et al. [3] who define Big Data by five V’s: Volume, Velocity, Variety, Veracity, and Value. Volume pertains to vast amounts of data, Velocity applies to the high pace at which new data is generated, Variety pertains to the level of complexity of the data, Veracity measures the genuineness of the

data, and Value evaluates how good the quality of the data is in reference to the intended results.

Data gathered for Health Informatics research does exhibit many of these qualities. Big Volume comes from large amounts of records stored for patients: for example, in some datasets each instance is quite large (e.g. datasets using MRI images or gene microarrays for each patient), while others have a large pool with which to gather data (such as social media data gathered from a population). Big Velocity occurs when new data is coming in at high speeds, which can be seen when trying to monitor real-time events whether that be monitoring a patient's current condition through medical sensors or attempting to track an epidemic through multitudes of incoming web posts (such as from Twitter). Big Variety pertains to datasets with a large amount of varying types of independent attributes, datasets that are gathered from many sources (e.g. search query data comes from many different age groups that use a search engine), or any dataset that is complex and thus needs to be seen at many levels of data throughout Health Informatics. High Veracity of data in Health Informatics, as in any field using analytics, is a concern when working with possibly noisy, incomplete, or erroneous data (as could be seen from faulty clinical sensors, gene microarrays, or from patient information stored in databases) where such data needs to be properly evaluated and dealt with. High Value of data is seen all throughout Health Informatics as the goal is to improve HCO. Although data gathered by traditional methods (such as in a clinical setting) is widely regarded as High Value, the value of data gathered by social media (data submitted by anyone) may be in question; however, as shown in Section "Using population level data – Social media", this can also have High Value.

It should be mentioned that not all the studies covered here or in the field of Health Informatics fit all 5 of the qualities in Demchenko et al.'s. definition of Big Data. Even so, many still impose significant computational constraints that need to be addressed in one way or another. Offline storage of datasets such as Electronic Health Records (EHR) can be difficult even if the data does not exhibit Big Velocity or Variety, and high-throughput processing seen in real-time continuous data requires capable and efficient techniques even when each individual data instance does not have Big Volume. Data that has Big Value without Big Veracity may need complex methods to find a consensus among various models, or which require time-consuming adjustments to the data which could expand the size of the dataset. Therefore, even datasets lacking Big Volume can still have Big Data problems, meaning that the Big Data definitions mainly focusing on Volume and Velocity may not be considering enough qualities of the dataset to fully characterize it.

It is noted in [4] and [5] that just in the United States, using data mining in Health Informatics can save the healthcare industry up to $450 billion each year. This is because the field of Health Informatics generates a large and growing amount of data. As of 2011, health care organizations had generated over 150 exabytes of data [4] (one exabyte is 1000 petabytes). This data needs to be sifted through and efficiently analyzed in order to be of any use to the health care system. As mentioned, health information can be stored in EHRs, which can store 44+ petabytes of patient data, which on top of this health data can come in many other forms. This explosion of data that is seen in Health Informatics has also been noticed in Bioinformatics as well, where genomic sequencing can generate many terabytes of data. With data coming from many different places and in many different forms, it is up to the Health Informatics community to find ways of dealing with

all this data. It would seem that it is becoming more and more popular to integrate and combine different sources of data, even across different subfields (i.e. Translational Bioinformatics), and even across Health Informatics and Bioinformatics. Successful integration of this huge amount of data could lead to a huge improvement for the end users of the health care system, i.e. patients.

Levels of health informatics data

This section will be describing various subfields of Health Informatics: Bioinformatics, Neuroinformatics, Clinical Informatics, and Public Health Informatics. The lines between each subfield of Health Informatics can be blurred in terms of definition, confusing which subfield a study should fall under; therefore, this paper will be deciding subfield membership by the highest level of data used for research and will be the organizing factor for Sections "1" through 1 . The works from the subfield of Bioinformatics discussed in this study consist of research done with molecular data (Section "Using micro level data – Molecules"), Neuroinformatics is a form of Medical Image Informatics which uses image data of the brain, and thus it falls under tissue data (Section "Using tissue level data"), Clinical Informatics here uses patient data (Section "Using patient level data"), and Public Health Informatics makes use of data either about the population or from the population (Section "Using population level data – Social media").

In Health Informatics research, there are two sets of levels which must be considered: the level from which the data is collected, and the level at which the research question is being posed. The four subfields discussed in this study correspond to the data levels, but the question level in a given work may be different from its data level. These question levels are of similar scope to the data levels: the tissue level data is of similar scope to human-scale biology questions, the patient level data is of comparable scope to clinical questions, and the population level data is of proportionate scope to epidemic-scale questions. Each section will be further sub-sectioned by question level starting with the lowest to the highest. Table 1 summarizes the breakdown of Sections "1" through "Using population level data – Social media" including the sections (by data levels used), subsections (by question level used), and the questions the studies discussed are attempting to answer.

Bioinformatics

Research in Bioinformatics may not be considered as part of traditional Health Informatics, but the research done in Bioinformatics is an important source of health information at various levels. Bioinformatics focuses on analytical research in order to learn how the human body works using molecular level data in addition to developing methods of effectively handling said data. The increasing amount of data here has greatly increased the importance of developing data mining and analysis techniques which are efficient, sensitive, and better able to handle Big Data.

Data in Bioinformatics, such as gene expression data, is continually growing (due to technology being able to generate more molecular data per individual), and is certainly classifiable as Big Volume. The issue of Big Volume within molecular data leads to research such as McDonald et al. [24] who created a Bioinformatics suite of software tools they call khmer. This suite seeks to solve hardware computational problems through software. The tools in this suite pre-process Big Volume genomic sequence data by breaking

Table 1 Summary of Studies Covered

Sections	Data level(s) used	Subsections	Question level(s) answered	Questions to be answered
Using Micro Level Data – Molecules	Molecular	Using Gene Expression Data to Make Clinical Predictions	Clinical	What sub-type of cancer does a patient have? [6] Will a patient have a relapse of cancer? [7]
Using Tissue Level Data	Tissue	Creating a Connectivity Map of the Brain Using Brain Images	Human-Scale Biology	Can a full connectivity map of the brain be made [8,9]?
	Patient	Using MRI Data for Clinical Prediction	Clinical	Do particular areas of the brain correlate to clinical events? [10] What level of Alzheimer's disease does a patient have? [11]
Using Patient Level Data	Patient	Prediction of ICU Readmission and Mortality Rate	Clinical	Should a patient be released from the ICU, or would they benefit from a longer stay? [12-14] What is the 5 year expectancy of a patient over the age of 50? [15]
		Real-Time Predictions Using Data Streams	Clinical	What ailment does a patient have (real-time prediction) [16,17] Is an infant experiencing a cardiorespiratory spell (real-time)? [18]
Using Population Level Data – Social Media	Population	Using Message Board Data to Help Patients Obtain Medical Information	Clinical	Can message post data be used for dispersing clinically reliable information? [19,20]
		Tracking Epidemics Using Search Query Data	Epidemic-Scale	Can search query data be used to accurately track epidemics throughout a population? [5,21]
		Tracking Epidemics Using Twitter Post Data	Epidemic-Scale	Can Twitter post data be used to accurately track epidemics throughout a population? [22,23]

up long sequences into relatively short strings which can be stored in a Bloom filter-based hash table, helping both the ability and efficiency of analysis of Bioinformatics data.

Neuroinformatics

Neuroinformatics research is a young subfield, as each data instance (such as MRIs) is quite large leading to datasets with Big Volume. Only recently can computational power keep up with the demands of such research. Neuroinformatics concentrates its research on analysis of brain image data (tissue level) in order to: learn how the brain works, find correlations between information gathered from brain images to medical events, etc., all with the goal of furthering medical knowledge at various levels. We chose the field of Neuroinformatics to represent the broader domain of Medical Image Informatics because by limiting the scope to brain images, more in-depth research may be performed while still gathering enough information to constitute Big Data. From this point on Neuroinformatics research using tissue level data will be referenced by data level rather than the subfield.

Clinical informatics

Clinical Informatics research involves making predictions that can help physicians make better, faster, more accurate decisions about their patients through analysis of patient data. Clinical questions are the most important question level in Health Informatics as it works directly with the patient. This is where a confusion can arise with the term "clinical" when found in research, as all Health Informatics research is performed with the eventual goal of predicting "clinical" events (directly or indirectly). This confusion is the reason for defining Clinical Informatics as only research which directly uses patient data. With this, data used by Clinical Informatics research has Big Value.

Even with all research eventually helping answer clinical realm events, according to Bennett et al. [25] there is about a 15 ± 2 year gap between clinical research and the actual clinical care used in practice. Decisions these days are made mostly on general information that has worked before, or based on what experts have found to work in the past. Through all the research presented here as well as with all the research being done in Health Informatics, the healthcare system can embrace new ways that can be more accurate, reliable, and efficient.

Public health informatics – Social media

Public Health Informatics applies data mining and analytics to population data, in order to gain medical insight. Data in Public Health Informatics is from the population, gathered either from "traditional" means (experts or hospitals) or gathered from the population (social media). In either event, population data has Big Volume, along with Big Velocity and Big Variety. Data gathered from the population through social media could possibly have low Veracity leading to low Value, but techniques for extracting the useful information from social media (such as Twitter posts), this line of data can also have Big Value. cdot

Using micro level data – Molecules

The studies covered in this section use data gathered from the molecular level, and answer the clinical question of using gene expression data for prediction of clinical

outcome. Molecule-level data frequently experiences the problem of "high dimensionality," where the data has a large number of independent attributes; this is because molecule-level data tends to have thousands (or tens of thousands) of possible molecules, configurations of molecules, or molecule-molecule interactions, and these are represented in datasets as features. This high dimensionality can stymie approaches which do not consider feature selection to address this form of Big Data. Although we only focus on one form of clinical question in this paper, other applications of micro-level data exist, such as cheminformatics [26], high-throughput screening [27], and DNA sequence analysis [28].

Using gene expression data for prediction of clinical outcome

As stated, the studies in this subsection use gene expression data to answer clinical questions. Two research efforts are reviewed in this subsection, both of which focus on cancer: the first uses gene expression profiling to categorize leukemia into two different subclasses, while the second study uses gene expression data to predict relapse among patients in the early stages of colorectal cancer (CRC). Both of these studies (as well as similar studies) can help physicians guide, advise and treat their cancer patients.

Haferlach et al. [6] formulated a gene expression profiling classifier to place patients into 18 different subclasses of either myeloid or lymphoid leukemia. This study used 3,334 patients where about two-thirds were used for training (2,143 patients) and the rest one-third for testing (1,191 patients), and from each patient 54,630 gene probe set samples were taken ($3{,}334 \times 54{,}630 \approx 182$ million). The authors chose to use an all-pairwise classification design using the trimmed mean of the difference between perfect match and mismatch intensities with quantile normalization, all to handle the multiclass nature of this research. This technique, called Difference of Quantile Normalized values (DQN), is explained in greater detail in [29] by Liu et al. There are 153 class pairs created due to there being 18 distinct classes ($(18 \times 17)/2 = 153$) where for each pair there will be a linear binary classifier created using Support Vector Machines (SVM) [30]. They tested their method on the training pool using 30-fold cross-validation, where each of the 30 runs used the top 100 probe sets with the highest t-statistic for each class pair. The results of this testing yielded a mean specificity of 99.7% and an accuracy of 92.2%.

The second part of their study was to get the results for the testing pool of patients where Haferlach et al. achieve a median specificity of 99.8% and a median sensitivity of 95.6% for the classification of 14 subclasses of acute leukemia where 6 are lymphoid and 8 are myeloid. The authors also claim that using their methods with microarray data, on 57% of the discrepant instances they were able to achieve better results than with routine diagnostic methods. This research has shown promise that using microarray gene expression data patients can be reliably classified into different forms of leukemia, yet this study could have done more by exploiting any number of feature selection techniques (or in this case gene probe selection). Using feature selection techniques in this study could also lead to determining which gene probes have a strong correlation between different forms of leukemia.

Salazar et al. [7] construct a gene expression classifier seeking to predict, within a five-year period, whether or not a patient will relapse back into having CRC. The authors gathered a total of 394 patients where 188 were used for training and 206 were used for validation. The training pool was accumulated over a 19 year period (1983-2002) from

three different institutions from different counties, and the validation pool was assembled over 8 years (1996-2004) from another institution from a different country. They decided on a set of 33,834 gene probes that were found to have variation within the patients from the test group ($33{,}834 \times 188 \approx 6.4$ million not counting the probes that showed no correlation). The authors decided for feature selection (gene probe selection) to use a leave-one-out cross-validation method to determine which gene probes were strongly correlated with the 5-year distant metastasis-free survival (DMFS) with a t-test as the deciding factor. After this selection method they ended up with an optimal set of 18 gene probes.

For classification, Salazar et al. used a Nearest Centroid-Base Classifier (NCBC) named ColoPrint. A basic explanation of a centroid based classification system would be: when a new instance is ready to be classified it is assigned to the class consisting of the training samples whose average centroid is closest to the new instance. ColoPrint also incorporates feature reduction to alleviate the problem of high dimensionality, by using a predefined set of 18 genes which were previously found to be useful in identifying CRC. With ColoPrint, a patient can be classified as either low or high risk. The authors note that in a similar study (for breast cancer prognostic tests [31]) that centroid-based classifiers have shown to give reliable results. The authors compared ColoPrint to an assortment of clinical factors through multivariate analysis to see which could best predict Relapse Free Survival, and ColoPrint was found to be one of the most significant factors with a hazard ratio of 2.69 at a confidence interval of 95% and a P value of 0.003. The hazard ratio shows the proportion of relapse rates of predicted-relapse and predicted-RFS patients). Through this study, they were able to find which patients were having a generally higher probability of being marked for high risk than other patients (e.g. males are more often marked as high risk than females).

The approach the authors took by gathering data from different sources for the training set (giving variation to the learning) and even another source for validation (giving variation to the testing) is a notably good strategy and is a strategy future research should employ. One place variation could have been added, though, is using more probe selection techniques: this would help test if the ColoPrint feature subset was indeed "optimal", especially since there were 33,834 probes to start with.

Both of the studies discussed in this subsection are showing the usefulness of microarray gene expression data as they can be used for determining both: if a patient will relapse back into cancer as well as which subtype of cancer a patient has. The results of these research efforts lead to the idea that microarray data could give similar results if these procedures were applied to other types of cancers in order to help physicians both diagnose and begin to treat their patients.

Using tissue level data

All the studies covered in this section will cover data at the tissue level and venture to answer human-scale biology questions including: creating a full connectivity map of the brain, and predicting clinical outcomes by using MRI data. This level, which incorporates imaging data, brings in a number of additional Big Data challenges, such as feature extraction and managing complex images. Studies which combine imaging data with other data sources also exemplify the Variety aspect of Big Data, by building models which incorporate diverse data sources.

Creating a connectivity map of the brain using brain images

This subsection will be covering two studies with the goal of answering human-scale biology questions attempting to develop a comprehensive connectivity diagram of the human brain. A diagram such as this can provide many opportunities for health information gain for physicians for prognosis, diagnosis, treatments, etc. through the execution of analysis and novel data mining.

There is a huge and ongoing project called the Human Connectome Project (HCP) led by the WU-Minn HCP consortium where the goal is to eventually map the human brain by making a comprehensive connectivity diagram. According to [9] HCP is looking to find a map of the neural pathways that make up the brain in order to advance current knowledge of how the brain functions and behaves region-to-region. The project is broken into two phases: 1.) in the first 2 years (Fall 2010- Spring 2012), methods for data acquisition and analysis were improved, 2.) in the last 3 years (summer 2012-Summer 2015), these methods will be applied to 1200 healthy adults (between the ages of 22 and 35 from varying ethnic groups) using top of the line methods of noninvasive neuroimaging. The subjects used are twins and non-twin siblings in order to also determine variability and heritability factors in brain structure and connectivity throughout such cases. This project is a five year long project that was started in fall 2010 and should be finished in 2015. The HCP data generated is being made freely accessible to the public, and the first and second quarterly data are now available at: http://www.humanconnectome.org/ containing the data generated from a total of 148 of the 1200 participants (about 12% of the total). The data available includes image data (T1w and T2w MRI, rfMRI, tfMRI, dMRI) as well as behavioral measures for a current total of about 4.5 TB. Data will be further released at a quarterly rate with each quarter including about 100 new participants.

According to Van Essen et al. [9], the project looks very promising because new and extremely important information can be gleaned from mining the HCP data coming out from the HCP consortium's research. Creating a full connectivity map of the brain could lead to information that could help in determining the reasons why people have certain brain disorders at a level previously unattainable, giving physician a possibility for easier diagnosis, early detection of future illnesses or maybe even prevention of mental or physical ailments. Considering this data is only recently released, the research applying this data is all in the future, but with the technology being so advanced there is endless possibility for studies employing this data for heath information gain. Once all the data for the 1200 patients have been generated, there could be similar data created on patients with various ailments and various ages to find the differences between such brains through data mining and analysis.

Annese [8] develops a system to link MRI (primarily diffusion tensor imaging (DTI)) measurements to that of an actual brain sample gathered from histological methods (physical studying of tissue) arguing that it is not enough to just depend on neuroimaging for making a comprehensive connectivity map of brain connectivity, but should also include histological methods of studying of actual brain tissue. The author mentions that the difficulties of matching up MRI measurements with that of anatomical measurements is a considerable issue interfering with making a comprehensive connectivity model of the human brain. MRIs are large and have high resolution, a higher resolution then histological methods which give results that are neither the same size nor quality. Annese also comments that it would be highly beneficial to validate MRIs with studies of actual

brain tissue. The study was carried out on one patient that had died at the age of 88, and as the patient had a pacemaker (which prevented the acquisition of MRI images while the patient was alive), all images of the brain were taken after death and compared to the histological study of the brain tissue using the author's method.

To ensure that the histological images match the Big Volume found in typical MRI images, Annese takes stained histological slices of the brain at 20x magnification, with a resolution of around 0.4 μm/pixel. This creates 334,500 × 266,200 images. In addition, multiple MRI images were taken of the brain *in situ* prior to the histological imaging. The author was able to show that MRIs (and DTIs in particular) do show tissue properties that can be correlated to stained tissue (histological images), making comparison possible. The author argues that histological comparison to MRIs can help validate MRIs, localize neuropathlogical phenomena that show as MRI abnormalities, and create the full connectivity map of the human brain. Data mining and analysis of both types of images in parallel could offer more gains than simply using MRI data alone, offering a much more powerful set of data.

As technology only recently could handle the endeavor of creating a full connectivity map of the brain this line of research is very new. The HCP is creating and releasing Big Volumes of data with various amounts of different MRIs and behavioral measurements data, allowing for more analysis and novel data mining to be executed. Along with the HCP, there is discussion and testing being conducted for comparing MRIs to histological data to help validate MRI data, to help create the connectivity map of the brain and offer more power to the datasets being created for novel data mining. The HCP could benefit from employing a comparison to histological image data.

Using MRI data for clinical prediction

This subsection will be covering two studies with the goal of answering clinical level questions. The first study uses both MRI data and a list of clinical features with the goal to find correlations between physical ailments to that of different locations of the brain. The second looks to take MRI data and determine the amount a patient has Alzheimer's disease. Research using MRIs can be beneficial to clinical diagnosis and predictions giving physicians another option with which to make decisions.

Yoshida et al. [10] propose a novel method combining patients' clinical features with that of MRI image intensities consisting of millions of voxels (an element of volume representing a point on a grid in three dimensional space). The method the authors create is based on the algorithm of radial basis function-sparse partial least squares (RBF-sPLS) giving their method an advantage over similar methods granting the ability to select not only clinical characteristics but also determine effective brain regions. This is to say that by creating sparse, linear combinations of explanatory variables, the developed approach concurrently performs feature selection and dimensionality reduction (that is, creating new, more condensed features). Simultaneously doing both of these tasks is problematic for other techniques especially for vast amounts of data, but employing RBF-sPLS allows the authors to manage large amount of data efficiently. Yoshida et al. do a comparison of their RBF-sPLS to that of the original sPLS on a simulated dataset, and demonstrate that their technique outperforms the original in terms of sensitivity, specificity and c-index scores. The simulated dataset does appear to be the only testing of the RBF-sPLS and even with the results garnered being quite promising this method

should be tested on real-world data so that their results can be considered clinically significant.

They do present an example of their prospective method using a dataset of 102 chronic kidney patients with the goal of attempting to identify if any locations of the brain are correlated to patients with chronic kidney disease. For each of these 102 patients they gathered 73 clinical features and around 2.1 million voxels from the MRI data. Through experiments Yoshida et al. found that there is a strong correlation between clinical variables related to chronic kidney disease and the bilateral temporal lobe area of the brain. They also determined that the bilateral temporal lobes are closely related to aging and arterial stiffness, while the occipital lobes correspond to clinical markers for anemia. As this is only a preliminary study, further testing will be needed to confirm the accuracy of these results.

Estella et al. [11] introduce a method with the goal of predicting to what degree a patient has Alzheimer's disease with three levels of classification: completely healthy, Mild Cognitive Impairment (MCI), and already has Alzheimer's. They gathered around 240GB of brain image data for 1200 patients stored by the Alzheimer's Disease Neuroimaging Initiative (ADNI). There are a number of steps to their devised method, which include spatial normalization, extraction of features, feature selection and patient classification. After the feature extraction step, the authors found two different subgroups of features: morphological and mathematical where 332 and 108 were respectively gathered. Examples of features from the morphological subgroup are Area Centroid, Major Axis Length, Whole Matter Volumes, etc. Examples of features from the mathematical subgroup are Mean, Cosine Transform Coefficients, Euclidean distance, etc. The authors also decided to have a third group that consists of features from both subgroups and deemed this group as "mixed". For the feature selection step they used a method based on Mutual Information (MI) along with some influence from the minimal Redundancy-Maximal-Relevance criterion (mRMR). MI will assist in determining the dependence between two given variables while the mRMR looks to acquire applicable features correlated to the final prediction while simultaneously removing redundancies from the model.

Fuzzy Decision Tree (FDT) classifiers work best with Estella et al.'s feature extraction and feature selection combination, getting better results than other classifiers tested in terms of classifying efficiency (able to make reliable diagnosis on a minimal set of features). FDT is an extension to the traditional decisions with the additional ability to handle fuzzy data. The authors found that by using their method they were able to get a classifying efficiency of 0.94 when using 75% of the mixed feature subset and when only using 10% of the mixed features could still get a reliable efficiency of 0.75. Estella et al. conclude that using their method while using only a minimal number of both morphological and mathematical variables they can efficiently and reliably classify patients into three levels of Alzheimer using only MRI data. Even though this study compared various classification techniques there was only one feature selection technique tested; there could be other feature selection methods that could have worked better with the FDT or one of the other classification techniques tested.

The studies shown in this section using MRI data have shown that they can be useful in answering clinical questions as well as making clinical predictions. More research (on real world data) will be needed before the brain regions, determined by Yoshida et al. [10] to have correlation with kidney disease, anemia and aging, can be determined as clinically

significant. In Estella et al. [11], various features extracted from MRIs were shown here to have the ability to classify patients into varying degrees of dementia. This process could eventually be improved upon with the goal of adding more classification levels for Alzheimer patients leading to such patients being detected both earlier and more efficiently. This is an interesting line of research that could be very beneficial if a physician could one day have methods that can look at MRIs of the brain and be able to determine whether a patient has kidney disease, more likely to have kidney disease, determining how far along a patient is down the road to dementia, etc. With the promising results shown in these studies, further research should extend these research efforts to see what other correlations between MRIs and other diseases can be discovered, giving physicians one more tool to diagnose and treat a patient earlier and more accurately.

Using patient level data

All the studies presented in this section will cover data at the patient level and venture to answer clinical level questions including: Prediction of ICU readmission, prediction of patient mortality rate (after ICU discharge and 5 years) and making clinical predictions using data streams. Due to this use of streaming data, ICU data exemplifies the Velocity aspect of Big Data, as well as Variety, as clinical information can contain many different types of features. As with the molecular-level data, feature selection can help choose the most important features, which also helps in making quicker clinical decisions.

Prediction of ICU readmission and mortality rate

The focus of the research covered in this subsection will be on predicting Intensive Care Unit (ICU) readmission, mortality rate after ICU discharge as well as predicting a 5 year life expectancy rate. The five year life expectancy rate will be looking to see how likely a patient will survive within a 5 year period. This is a useful line of research in that it can potentially help physicians know what to look for in their patients, determine which patients should have their ICU stay extended, and better tell which patients should receive particular treatments.

The study done by Campbell et al. [12] focuses on ICU patients that were discharged and expected to both live and not return too early afterwards. Their research used 4376 ICU (out of 6208 total) admissions from a database containing admissions from one ICU from January 1995 to January 2005 (a 10 year period). There was a total of 16 attributes chosen for each ICU admission, where among these attributes there were a few well-known scores used for the prediction of ICU readmission and death after discharge including: Acute Physiology and Chronic Health Evaluation II (APAHACE II) score [32], Simplified Acute Physiology Score II (SAPS II) [33], and the updated Therapeutic Intervention Scoring System (TISS) [34] score. The APACHE II score is a popular and well tested score based severity of disease classification system (SoDCS), described in [32] by Knaus et al., which uses a simple set of 12 physiological variables for prediction. The SAPS II score is another popular and well tested SoDCS using 17 total features, including age, 12 physiology features, admission type, and 3 disease type features, which is discussed further by Gall et al. [33]. The TISS score is a third popular and well tested SoDCS where originally 57 therapeutic intervention measurements were used but was updated where some features were added and some removed, while test results stayed the same. The updated TISS score is covered by Keene et al. [34].

Campbell et al. decided upon three research directives for prediction: death after ICU discharge (but before hospital discharge), readmission to the ICU within 48 hours of ICU discharge (again, before hospital discharge), and readmission to the ICU at any point after ICU discharge (and before hospital discharge). It should be noted that each patient could potentially fall into more than one of these categories, but this is not an issue due to there being three separate binary models built. The importance these prediction models could potentially have would be to help physicians determine which of these patients fall into these three groups and most significantly why they fall into these groups. If physicians can know why then they can determine which patients need their ICU stay extended. The feature selection method decided upon was simple logistic regression for all three models to determine which of the 16 attributes had a strong correlation to each prediction ($P \leq 0.2$). Multiple Logistic regression was chosen for building the prediction models. All three of their models were tested using the Hosmer and Lemeshow goodness-of-fit (HLgof) test [35] and determined that the calibration was good and no more calibration was necessary. The HLgof is used to determine if logistic regression models have sufficient calibration (discussed by Hosmer et al. [35]).

The Area Under the (ROC) Curve (AUC) was used in this study to determine the classification and discrimination performance of the three models. According to Bradley [36], AUC is the best criteria for measuring the classification performance of a binary classifier such as logistic regression. AUC is a fitting metric to use on ICU admission data because the positive-class instances are a small percentage (i.e. only 3.3% of the patients were readmitted to the ICU within 48 hours). For determining the quality of the three models (using the chosen set of features) the AUC results for each model are compared to results obtained from the APACHE II score for each prediction. The first model, predicting death after ICU discharge (before hospital discharge) had an AUC value of 0.74 compared to AUC garnered from APACHE II of 0.69. The second model, for predicting readmission (before hospital discharge) obtained an AUC of 0.67 while APACHE II received an AUC of 0.63. The third model for predicting readmission within 48 hours of ICU discharge acquired an AUC value of 0.62 while APACHE II earned an AUC of 0.59.

The three models with the chosen set of features only achieved minimal improvement over APACHE II alone for the prediction of ICU readmission and mortality rate for ICU patients. Campbell et al. note that this minimal improvement could be because the APACHE II score already uses physiological variables, and it is these variables which are generally most useful for predicting ICU readmission and the mortality rate after ICU discharge. However, one non-physiological variable was shown to be highly correlated to both ICU readmission and death predictions: increasing age.

Only about 23.3% of patients used fall into one of these three model's positive class and should not have been released from the ICU where if their ICU stay was extended, maybe some of these patients could have been saved. One item to note is that Campbell et al. only used one feature selection method as well as only one prediction model for their research. There is a possibility if other methods were tested there could have been models built that were able to outperform APACHE II. Even though the data was gathered from a database containing a 10 year's worth of ICU patients, the data was still only from one ICU, and for the purpose of validation a larger variation (collecting from various ICUs) would have been beneficial.

Ouanes et al. [14] conducted their research with the goal of predicting whether a patient would die or return to the ICU within the first week after ICU discharge. Research was performed on 3462 patients (out of 5014) admitted to an ICU for a minimum of 24 hours, gathered from 4 different ICUs from the Outcomerea database. As Campbell et al. found in their data, the positive class is very small compared to the overall population with only about 3% (where 0.8% died and 2.1% were readmitted within 7 days). The feature selection method chosen was univariate analysis selecting variables with ($P < 0.2$) to add to the final model, and used the Akaike Information Criterion (AIC) [37] to identify the best model. AIC is a metric used to verify the overall quality of statistical models. The model created was then subjugated to a few validation and verification steps for testing (clinical relevancy, variable inter-correlation and co-linearity between variables) in order to end up with their final set of 6 variables from the original 41. The six variables chosen were age, SAPS II, the need for a central venous catheter, SIRS score during ICU stay, SOFA score, and discharge at night.

These variables were used to make the final prediction model using multivariate logistic regression, which will be used to develop their Minimizing ICU Readmission (MIR) score. The MIR score will be a quantitative measurement for determining whether a patient should be discharged from an ICU or not. The predictive results of MIR are compared to the results garnered from both SAPS II and Stability and Workload Index for Transfer (SWIFT) [38], another SoDCS using a small set of commonly available variables. Through MIR, Ouanes et al. were able to achieve good results with good calibration decided by the HLgof test and an AUC of 0.74 at a 95% confidence interval. The result SAPS II received for AUC was 0.64 and SWIFT getting an AUC score if 0.61, which shows that MIR performed considerably better. In this research only one feature selection method as well as only one classification method is used. The MIR score possibly could have yielded better results if more than one feature selection technique was used to determine which of the 41 variables would make the best model. In conjunction, MIR could possibly be further benefitted by testing out a number of other classifiers to develop the final model with the highest predictive power for this line of research. To fully test these results, one more comparison that should be made is their MIR score to that of either APACHE II or APACHE III scores, as shown in Campbell et al. [12] and Failho et al. [13].

Failho et al. [13] also seek to predict patients that will be readmitted to the ICU, but with the goal of only using a small amount of physiological variables, following the findings of Campbell et al. [12]. The prediction that this research is focusing on making is determining if patients will be readmitted within 3 days after discharge. The dataset they gathered was from the MIMIC II database [39] choosing only the patients over the age of 15 having an ICU stay of more than 24 hours giving a sample of 19,075 adults that were admitted to one of four ICUs. These 19,075 patients were further reduced by considering only patients for whom the researchers have all the variables available, leaving 3,034 patients. Further preprocessing reduced this to 1,267 patients. Finally, out of these 1,267 patients only 1,028 survived, giving a final dataset with 1,028 instances (and 13 members of the positive (readmittance) class).

Failho et al. decided upon two possible methods for feature selection: Sequential Forward Selection (SFS), or bottom-up approach, and Sequential Backwards Elimination (SBE), or top-down feature selection where both are discussed in [40]. SFS works on the original pool of features and starts with one feature then during each iteration one more

feature will be added and the iterations will be stopped when the current model is deemed the best possible. SBE takes a different approach starting with the whole set of the original features and removing one feature at each iteration until the model is deemed the best possible. Failho et al. compared both of these methods in combination with their classifier, and found SFS to have better results than SBE in terms of AUC. Thus, it was chosen as their feature selection method. Using SFS, 6 out of the original 24 physiological variables were chosen: mean heart rate, mean temperature, mean platelets, mean blood pressure, mean SpO2, and mean lactic acid.

The classifier method used in this study was fuzzy modeling with sequential forward selection [40,41], specifically Takagi-Sugeno (TS) fuzzy modeling. Fuzzy models use linguistic interpretations to formulate rules and logical connectives in order to make connections between features and the final prediction. With the use of linguistic interpretation, fuzzy modeling is a good choice for this line of research as clinical data needs to be interpreted, as a physician would do in a clinical setting. The only worry would be that with the rule-based side of the models there could possibly be too much rigidity causing the absence of physician discernment in the final model.

Failho et al. compared the results they got from their set of 6 physiological variables (determined by SFS) in conjunction with TS to that of the sets determined by APACHE II and APACHE III [42] scores (also in conjunction with TS). APACHE III is another SoDCS, and was created with the goal of improving some of the problems in APACHE III as discussed in [42] by Knaus et al. The results show a significant advantage in favor of Failho et al.'s. set securing an AUC score of 0.72 ± 0.04 while APACHE II and APACHE scored an AUC of 0.62 ± 0.03 and 0.64 ± 0.04 respectively. The SFS set also scored better in terms of specificity, sensitivity, and accuracy. This result shows that good prediction performance can be reached by using a small set of physiological variables. Even with the promising results there is more that could have been done in this research, one being that more variables, either physiological or not, could have been added to the original pool to see if results could have been improved. It might also have been beneficial to see if other feature selection techniques (tree-based or otherwise) may have improved upon the results achieved by SFS. Failho et al. do mention that fuzzy modeling has been shown to work comparably well to other classification methods for medical data yet there still could be benefit to this research if other classifier methods were tested to see if TS does yield the best results.

The research objective of Mathias et al. [15] is along a slightly different line where instead of trying to predict ICU readmission they look to predict a 5 year mortality rate through the construction of an Ensemble Index (EI). They used a group of 7463 patients taken from an Electronic Health Record (EHR) along with 980 attributes for each patient. There were two requirements for the patients to be used in this study: must be over the age of 50 (due to increasing age being a huge prediction factor for this line of study) and had at least 1 hospital visit within the year 2003. Due to the large amount of attributes in the original set of variables the feature selection method chosen was Correlation Feature Selection (CFS) [43] along with greedy stepwise search which will be used to create their EI. The CFS method finds variables that are both strongly correlated to the final prediction and that are weakly correlated between them. The CFS method with greedy stepwise search found a subset of 52 features which was broken down further by manual reduction followed by another round of CFS bringing the subset down to 23. This subset was

then populated by one variable (gender) giving the final EI subset of 24 variables. The top 6 attributes in the EI (ranked by information gain) are age, comorbidity count, amount of hospitalization a year prior to admission, high blood urea nitrogen levels, low calcium, and mean albumin.

Rotation Forest Ensembling (RFE) [44] with Alternating Decision Tree (ADT) [45] was used to create their predictive model and was evaluated with tenfold cross-validation. Mathias et al. tested this technique with many other methods and found this technique to perform better (the reason it was used). The RFE algorithm is an ensemble of decision trees creating variation by assigning each tree a subset of features randomly chosen where Principle Component Analysis (PCA) is applied to each subset before each tree model is built. The ADT is a decision tree which instead of having a single class prediction located in its base leaf nodes, has a "probability of class membership" prior to each terminal node, and the sum of all these values are along an instance's whole path in order to predict its class value. RFE and ADT is a good combination as RFE brings accuracy and diversity to the model, and ADT allows for more information to be gained about an instance as it goes along the tree.

The Ensemble Index was able to achieve quite good results scoring better in recall, precision c-statistic, etc. than both the modified Walter Life Expectancy Index (WLEI) [46] and the modified Charlson Comorbidity Index (CCI) [47]. The modified WLE and CCI are two well-tested and better-known life expectancy indices that are used for prediction in similar research. Even though these good results were achieved for the EI, there was only one feature selection process used where if more techniques were used a better subset could have been created, which is possible with there being 980 attributes in the original set of features. Again more variation of data could be added to this research as all the data was gathered from one source.

The studies shown in this section have the potential to improve clinical discharge procedure, determining which patients should be released from the ICU and which patients should receive a particular line of treatment. The goal of these studies is to find which attributes are the most correlated to why patients return to ICUs early or do not survive after discharge. Looking at the research efforts of Campbell et al. [12], Failho et al. [13], and Ouanes et al. [14] that covered prediction of ICU readmission and death rate after discharge, the top variables of why are age, APACHE scores, various physiological variables (e.g. heart rate), amount of organ dysfunction, as well as a few others. If physicians can better predict which of their patients will return to the ICU or not survive then they would know which patients to keep in the ICU longer and to give more focused care potentially saving, if not, at least prolonging a life. According to Ouanes et al., between day one and day seven after discharge the readmission rate and death rate go down drastically, meaning that keeping a patient just a little longer could be beneficial, but could also take away an ICU bed from another patient that needs intensive care. These are the reasons that studies attempting to figure out why patients return early or die soon after ICU discharge are quite important as lives can potentially be saved. The target of this research of course is on those patients that have preventable death as not all death will be preventable.

The variables that were shown to be most telling for the 5 year survival rate (Mathias et al. [15]) happened to be similar with age and physiological variables being at the top of the list. One benefit of this research is by looking well into a patient's future can help physicians advise their patients better as far a treatment options. An example of this would

be a physician could advise a patient whether or not to go through a particular rough line of treatment when they may not live long enough to reap the benefits of such treatment. This research is especially important for patients with increasing age as the older a patient is the less likely a harsh treatment would be beneficial.

Real-time predictions using data streams

The studies covered in this section are also on data gathered from the patient level and again have the intention of answering clinical level questions. Instead of predicting the patient's condition in the future (i.e. ICU readmission or 5 year survival), the research here will be using data streams in order to predict patient's conditions in real-time. Data streams are never ending torrents of data that requires continuous analysis giving the possibility for real-time results (a feature not available when using static data sets). This section will sample two different categories of data steam studies: making prognosis and diagnosis predictions for patients, and detecting if a new born is experiencing a cardiorespiratory spell both in a real time. The researchers here are attempting to develop methods that use these constant streams of data and make predictions in a continuous manner while keeping satisfactory accuracy and precision.

Zhang et al. [17] develop a clinical support system using data stream mining with the goal of analyzing patient data in order to make real-time prognosis and diagnosis. In order to handle the continuous stream of data an algorithm that can handle high-throughput data will be necessary leading the authors to choose Very Fast Decision Tree (VFDT). The VFDT algorithm is quite efficient as it was built to handle thousands of instances per second using basic hardware (discussed by Domingos et al. [48]). They discuss that VFDT has many advantages over other methods (e.g., rule based, neural networks, other decision trees, Bayesian networks) such as VFDT can make prediction both diagnostically and prognostically, can handle a changing non-static dataset, not using rigid rules (can be difficult for experts to put their knowledge into rules).

VFDT alone, though, is not able to give future predictions of a patient's status only the current status; therefore, Zhang et al. decided to modify VFDT. For the modified VFDT, one or more pointer(s) were added to each of the terminating leaf nodes, where each base node corresponds to a distinct medical condition and each pointer corresponds to one medical records of a previous patient. To connect each stored medical record to its corresponding pointer, the authors created a mapping table so when the VFDT runs through and ends up on a base node the map will connect the leaf to its pointer(s) (corresponding medical records). These medical records, through Natural Language Processing (sentence and semantic similarity [49,50]), will then be used to make a prediction about the patient's future and give physicians the ability to better treat and advise their patients based on previous similar situations. The VFDT and the mapping table are updated as necessary (i.e. when a physician makes a new diagnosis or when a new medical record is added to the map).

Zhang et al. compare their method to that of IBM's similar data stream mining technique covered by Sun et al. [51]. To test their method, IBM used 1500 ICU patients from the MIMIC II database along with various physiological waveforms (taken from an assortment of medical devices) and clinical data on each patient. IBM's method consists of three main parts: 1.) physiological stream processing, 2.) offline analysis, and 3.) online analysis. For the steam processing part, a correlation base technique was chosen as such

techniques are able to correlate well among sensors and are able to efficiently handle missing data (estimating missing values by way of linear regression models using other sensors during that period of time). Better results could possibly have been attained if techniques other than linear regression models were used to estimate missing values. A correlation based technique was not the only technique tested, they also tried a window based technique which estimates missing values for a sensor by using an averaged value during a small window of time from that sensor and imputing that value for the missing time. The correlation based technique found better results and therefore was used in the final method.

During offline analysis, Sun et al. use a method they created called Locally Supervised Metric Learning (LSML), which learns an adjustable distance metric by using knowledge from the current domain (in this study, clinical knowledge). The last step of online analysis takes place when a new patient is ready for prediction of prognosis where the system will find the set of similar cases by way of temporal alignment, and this is followed by applying a regression model to account for the uniqueness of patients. Sun et al. ran a comparison of their LSML to that of PCA and Linear Discriminant Analysis (LDA) in terms of both precision and accuracy with LSML scoring considerably better in both.

Mentioned by Zhang et al., their system will use fewer computer resources to run compared to IBMs as offline analysis will not needed. This comes down to a comparison of the complexity between LSML and VFDT, where LSML is quite complex and does not allow the real-time nature that can be offered by VDFT (the most complex calculation of Zhang et al.'s. method). Zhang et al.'s. method is not tested on real world data and would need to be before its usefulness can be determined and can be legitimately compared to IBM's method. The real-time nature does offer the ability for making quick prognosis for patients, and if the results can be similar to the results found in IBM's method, then predictions could not just be made quickly but with good accuracy and precision.

Thommandram et al. [18] use data streams with a different goal attempting to detect and eventually classify neonatal cardiorespiratory spells (a condition that can be greatly helped by being detected and classified in real time). A cardiorespiratory spell is classified as some combination of a pause in breathing, drop in blood oxygen saturation, and a decrease in heart rate. The name of their system is called Artemis and is designed to use a steady stream of physiological data from the new born patient and both detect and ascertain which type of cardiorespiratory spell the patient is experiencing all in real-time. The real-time manner could potentially save the lives of these infants giving physicians more time to fix what is wrong as the need for human diagnosis will be less. The actual stream processing part of Artemis is handled by the middleware system developed by IBM called InfoSphere Streams [16], built to handle multiple high-throughput data streams. Middleware is software that works to connect two programs that are otherwise not connected. Three different data streams (from three different sensors) are used, which correspond to the three different conditions for a cardiorespiratory spell: a respiratory impedance wave, a decrease in blood oxygen saturation, and a decrease in heart rate.

The authors wanted to develop a system that will improve upon current machines used today that use the absolute change method (i.e. if a machine detects heart rate under or over a cutoff, then notify physicians). Artemis will use a relative change method, where instead of a cutoff; there will be a sliding baseline that will be continuously updated in real-time as the patients "normal" readings change over time. The sliding baseline method

can give a more reliable reading as it adapts to the unique reading of each patient as well as allow for more accurate spell detection. For classification, the reading from the three streams will be analyzed through a hierarchal rule based temporal model to determine which of the many cardiorespiratory spells the infant is experiencing.

The detection part of Thommandram et al.'s. system was tested on one patient in a Neonatal ICU during a 24 hour period where the sliding baseline method was found to alert physicians as often as found in the cutoff method for both heart rate and SpO2 readings. From the results for this one tested patient, they showed that their sliding baseline method could achieve clinically significant results for heart rate detection, with a specificity of 98.9% and a sensitivity of 100%. They did not test their classification method as mentioned by the authors as their future work. The detection part of this method will need further testing on many more patients before this method can be considered as clinically acceptable. When the classification part is tested it could be beneficial if they were to compare other methods to their hierarchal rule based temporal model.

Data stream mining presented here has shown the potential to be beneficial for clinical practice as it can be extended to be used in real time by use of efficient algorithms and methods (that are not previously used in the clinic). By using these data stream diagnosis for prognosis and spell detection, physicians could make faster and more accurate decisions and start solving the problem without spending as much time developing a plan. This line of research is fairly young and more studies will be needed through the development and testing of various new methods for using data stream mining for medical data with the goal of outputting results in real-time.

Using population level data – Social media

All the studies covered in this section will use data at the population level, specifically social media data (data that can be from anyone) and venture to answer both clinical questions and epidemic-scale questions. Data in Health Informatics is "traditionally" gathered from the doctors, clinics, hospitals and such, but recently people all around the world are starting to document health information all over the internet. This data could be from Twitter, internet query data (e.g. Google search data), message boards, or anywhere else people put information on the internet. This form of Big Data, which brings additional challenges such as text mining and handling potentially malicious noise, could possibly lead to finding many new breakthroughs in the field of medicine.

The first line of research presented is determining whether message board data can be useful to help patients find information on a given ailment. The second line of study is testing if using search query data or Twitter post data can effectively track an epidemic across a given population (in real-time). As previously mentioned above, the challenge with social media data is that although it is clearly High Volume, Velocity, and Variety, it could have both low Veracity and Value (data coming in could be unreliable, as discussed by Hay et al. [52]). However, through analyzing and mining the data, the useful parts can be extracted and used to gain medical insight.

There is much that can be learned by employing research on social media data including but not limited to: spatiotemporal information of disease outbreaks, real-time tracking of a harmful and infectious diseases, increasing the knowledge of global distribution for various diseases, and creating an extremely accessible way of letting people get information about any medical questions they might have.

Using message board data to help patients

Research in this subsection concentrated on determining if message board entries could possibly be a useful source of data for helping people find health information that is beneficial in appropriating reliable medical knowledge (answering clinical questions). Even though this line of research does not appear to be all that popular, it could offer keys to unlocking boundless reliable health information (as 59% of US adults use look for health information on the internet [19]), where one in five use social media).

Ashish et al. [19] created a platform called Smart Health Informatics Program (SHIP) with the goal of helping patients connect to the medical experiences of other patients posted throughout the internet via message boards (i.e. forums, blogs etc.). This study used a pool of 50,000 discussions including over 400,000 posts from four different message board websites (inspire.com, medhelp.com, and 2 others). SHIP uses a pipeline structure taking in all the message board entries (discussions and posts) from these four websites and go through various steps to find the entries with useful medical information and store them in a database for patient retrieval.

The first step of SHIP's pipeline is Elementary Extraction, which will execute some basic text processing for each entry: parsing through the HTML and extracting information on each entry (number of replies, when discussion was last updated, etc.) and giving a unique ID to each discussion and post. This is followed by the Entity Extraction step looking to determine which of the entries have medical significance related to health (such as treatments, side-effects, hospitals, drugs, etc.). Ashish et al. decided to use the XAR system [53] (an information extraction system of free text), which they expanded by incorporating ontologies from UMLS (Unified Medical Language System) [54] as well as other vocabularies and ontologies. Then SHIP enters Expression Distillation which looks at each post and determines whether or not the post includes any number of many expression of interests broken into 5 categories: 1.) personal experience, 2.) advice, 3.) information, 4.) support, 5.) outcome. This step requires that 5 classifiers be built for each expression of interest and they decided upon the J48 decision tree algorithm using the WEKA tool [55]. The next step is Aggregation using the data from the previous steps (done at the post level) and aggregates them at the discussion level. After the entries have gone through this process the numerous facts and expressions (over 20 million) found in each entry are stored in a database.

For retrieval from the database the authors extended Lucene [56], an open-source Java-based text search engine library, by adding optimizations for indexing, dynamic score boosting, and local caching. They tested their system on a test case of a patient experiencing severe cough since starting Tarceva (a type of chemotherapy). This was chosen as a test case due to coughing not being listed as a possible side-effect on the Tarceva website [57] or on another website depicting a clinical trial for this drug. Through the authors system (website) a search of Tarceva shows that cough is actually the third most common side-effect. This system could also find advice (through search filters) on how to help with the cough due to Tarceva. As this system has been shown to be beneficial for patients there could also be benefits to physicians as they may be looking to clinical trials for answers they could use this system (employing message board data) to find other cases similar to their patients and find information helping to make diagnosis, treatment options, etc. The authors tested extraction accuracies for each of the 5 expression of interest categories in terms of precision and accuracy: personal experience (0.87 and 0.82),

advise (0.91 and 0.62), Information (0.93 and 0.91), support (0.89 and 0.90), and outcome (0.80 and 0.58).

There will need to be more testing with more test cases to see if the data they are presenting through search is correct rather than just speculation as this data can be posted by anyone and even if it is predominantly posted by medical professionals, this does not automatically make the information correct. Martin [2] mentions that traditional health data should be used in the development of SHIP. This would help determine whether the data they gathered has medical importance or not, rather than relying on testing SHIP after it is constructed. From a data mining stand point the step of expression distillation could have been improved by using more classifiers other than J48 with which they could have improved the results they got for extraction accuracies.

Rolia et al. [20] devise a new system to use social health forums to help patients learn about their condition from posts by other patients with similar conditions. Their method consists of three steps: 1.) determining the patient's current medical condition from the personal health record (PHR), 2.) the system will ascertain which other users have a similar condition, 3.) a metric will be implemented evaluating and ranking the forum topics to determine the most relevant to present to the user. They describe their system implementation for a test case of type II diabetes mellitus (DM II). Due to privacy issues it was decided to use synthesized electronic medical health records (EMR) and PHRs with help from a medical professional. Also created is an aggregated clinical pathway that a patient can follow from the moment the patient is diagnosed onward where the path is determined by clinical factors (e.g. glucose levels) depicting a sequence of clinical events and was based on information found in medical research [58]. Using the aggregated clinical pathway they created 1000 synthetic patients (including their path and a few clinical variables) akin to ones that could be found in typically PHRs and EMRs as well as manually selecting 100 relevant posts from the websites: www.diabetesforum.com and www.diabetesdaily.com.

The system starts by creating and assigning the states to the synthesized patients and for DM II they decided upon 6 states (also described in medical research [58]). Assigning the patients into these six classes is done through rules (created by using NICE Pathways [59] and a medical expert who worked with Rolia et al.) guiding the patient along the pathway. The next step consists of appointing weights to forum topics (as for this example they only used the topic and not the content within). A medical expert will give the original set of weights to topics (for each of the six states) as percentages of likelihood that a patient with a given state would be benefitted by that topic, which will give the initial population model. This system will also offer the new user the ability to determine how relevant they want posts to be compared to their current state, which is implemented by calculating the cosine similarity between each state using the weights determined by each of the 100 posts. Now the system moves onto ranking the forum topics by correlation to the current user's state. The authors, in order to rank the topics used a metric deemed adjusted weights using the similarity between classes, patient preference (0 through 1), and the weights between post and clinical state; the higher the adjusted weight the higher the rank. The best feature about this system is its ability to learn as user feedback can adjust the weights of a post has to the given level of the current user (assuming using a similar method used by the expert). The user will also be able to change their current state as well as modify their patient preference.

The authors did build a prototype of this system and found that their system did behave as it was designed. One issue is that this system was not tested nor appears to be built with real data. A system such as this one is difficult to test outside letting real patients test this system in numerous field tests. As the authors mentioned they could possibly improve upon the cosine similarity by testing other such methods for defining weights. Another note, is that instead of a medical expert deciding the initial weights for posts through rules they could use a technique that would need to be tested and could then be approved by an expert, which could help with efficiency as well as the quality of the initial weights. This efficiency could help production of their classes as there are numerous diseases and ailments out there, and this research only discusses the construction for one.

Social media and the internet are becoming more and more popular for looking up and sharing medical data as mentioned in both [19] and [20]. The results found in these studies do show that there is a possibility that message board data could be used, but there is no real field testing to show that it works in the real world. If the research in these studies are found to be useful they could even be extended to help advise physician's diagnosis and treatments of their patients as other physicians and other patients with similar experiences to their current patients are located in these message board discussions.

Using search query data to track epidemics

The focus of the studies covered in this subsection will be discussing research done using search query data gathered from two popular search engines: Google (google.com) and Baidu (baidu.com) with the goal of predicting whether such data can be used to predict the occurrence and movement of Epidemics in a given population (all in real-time). The Center of Disease Control and Prevention (CDC) produces their results for influenza-like illness (ILI) epidemics, but there is generally a one to two week delay for this information. The research here is attempting to use search query data to get ILI epidemic information out to the public quicker than by the traditional method of the CDC reports. This research can help physicians and hospital to know both when and where an Epidemic is happening with real-time updates allowing them to act quicker in stopping the spread of the disease as well as help the patients already infected.

Ginsburg et al. [21] developed an automated method that can analyze a Big Volume of search queries from Google with the goal of tracking ILI within a given population. They conducted their research on Google search queries taken from historical logs during a 5 year period (between 2003 and 2008) using 50 million of the most popular searches as well as using data from the CDC historical data. These queries were taken directly, without combination, spelling correction, translation, or any other modification. They built and trained their model based on data from the years 2003-2007 and the validation was done on data from 2008. As a note, the CDC splits the US into 9 regions and this study looked to make predictions also using these regions of separation.

The model the authors built looks to find the probability that a patient visiting a physician is related to an ILI for a particular region using a single explanatory variable: the probability that a given search query is related to an ILI within the same region. To accomplish this the authors fit a linear model using both the log-odds for ILI physician visits and ILI related search queries giving: $logit(I(t)) = \alpha \times logit(Q(t)) + \epsilon$, where $I(t)$ is the percentage of ILI physician visits, α is a coefficient, $Q(t)$ stands for the fraction of queries related to ILI at time t, ϵ stands for the error in the formula (and $logit(p) = ln(p/1-p)$).

The single explanatory variable $Q(t)$ is determined through an automated technique that does not need any prior knowledge of influenza. The authors tested each of the 50 million stored queries alone as $Q(t)$ to see which queries fit best with the CDC ILI visit percentage for each region (presumably univariate analysis). The top 45 search queries, sorted by Z-transformed correlation throughout the nine regions, were chosen to belong to $Q(t)$ as the top 45 scored the best after they tested (through cross-validation) the top 1 search query through the top 100 search queries. Examination of these top queries showed many connections to influenza symptoms, complications, and remedies, consistent with searches made by an individual affected by influenza. The model was trained using weekly ILI percentages from 2003 through part of 2007 for all nine regions jointly to end up with a coefficient α that is region independent. Ginsburg et al. were able to obtain good fit compared to that of the reported CDC ILI percentages scoring a mean correlation of 0.90 throughout all nine regions.

Validation was carried out on the data gathered for part of 2007 through 2008 on 42 points per region and was able to achieve a mean correlation of 0.97 compared to the reported CDC ILI percentages. This correlation is quite high and through the authors' method was able to be reported 1 to 2 weeks prior to the CDC reports; thus, showing that search query data can be used to determine an ILI epidemic in a more real-time manner. This study only used one feature selection method narrowing down the 50 million most popular search queries down to 45; the results achieved maybe could have been improved upon if other techniques were used to determine an optimal set other than the chosen 45.

Yuan et al. [5] developed a similar system using search query data, but this study uses search queries gathered from Baidu (baidu.com) with the goal of tracking ILI epidemics across China. The author gathered their data from Baidu's database (http://index.baidu.com/) which stores the online search query since June 2006. For this study they only gathered data from March 2009 to August 2012, which was during the H1N1 epidemic and compare their results to that of China's Ministry of Health (MOH). Similar to the United State's CDC the MOH also releases their data with a 1 to 2 week delay. Baidu releases their search query data on a daily basis allowing for methods exploiting this data the ability to give answers in near real-time.

Yuan et al.'s system is split into four main parts: 1.) choosing keywords, 2.) filtering these keywords, 3.) defining weights and composite search index, and 4.) fitting the regression model with the keyword index to that of the influenza case data. For choosing keywords they reference Ginsburg et al. mentioning the benefit of choosing an optimal set of keywords, but want a more efficient way of determining such a set. Yuan et al. determined the key words using a tool from a Chinese website (http://tool.chinaz.com/baidu/words.aspx) which determined keywords through recommendations from Baidu as well as from others mined using semantic correlation analysis from sources such as portal websites, online reports, and blogs. The source keyword put into the tool was the Chinese symbol for Flu giving a total of 94 related keywords (assuming a total of 95 including Flu is the current set). The current set of keywords is then taken into the filtering step where three conditions must be met: (i) the keyword must include aspects that could impact the influenza epidemic, (ii) the data for a keyword must be presented with a time series with a resolution (daily, weekly, or monthly), and finally, (iii) the time series data for the chosen keywords must have a maximum cross-correlation coefficient no less than 0.4 to that of the influenza case data.

The next step starts with defining the weights for the set of keywords that passed filtering. The authors tested two techniques for determining weights for the keywords: systematic assessment (SA) [60,61] and strength of the correlation coefficient (SCC) [21,62]. SA entails using the principal of prior evaluations to rate the keywords and assign these ratings as weights, while SCC compares the influenza epidemic curve to that of the keyword frequency curve determining the correlation coefficient between these and assigning the weights accordingly. Yuan et al. decide to determine the optimal set of keywords with the formula: $y = \alpha_0 + \alpha_1 \times index_j + \epsilon$, where α_0 is the intercept, α_1 is the coefficient, and ϵ is the error. The $index_j$ is equal to $\sum_{i=1}^{j} \omega_i x_i^l$, where ω_i is the weight for the ith keyword and x_i^l denotes the sequence after alignment (not used in this study, presumably set to 1). Each keyword will be brought into the model in a stepwise manner in an order determined by their correlation coefficient and then use a partial F test to evaluate goodness of fit after each keyword is brought into the model (this will be repeated until the goodness of fit stops improving). After all the keyword selection processes are concluded they are left with 8 keywords left for the optimal set.

The authors decided to use the following regression model: $ICD[t] = \beta_0 \times ICD[t-1] + \beta_1 \times index[t] + \beta_2 \times index[t-1] + \epsilon$, where ICD is the influenza case data, β_0, β_1, and β_2 are coefficients, *index* represents the optimal set, and again ϵ represents the error. The variable t in this formula stands for time and can be broken into month time blocks. The formula therefore would be estimating ICD using the ICD from the previous month and on the optimal index for the current and previous months. The model was trained on the data from March 2009 to December 2011 and validated on the time period of January 2012 to August 2012. They achieved an R-squared of 0.95, an AIC of 18.50 and found that autocorrelation was not an issue due to the Durbin-Watson test results of 1.89. The validation of the model tested over the validation time period resulted in a mean absolute error of 10.6%, but in a more real-time manner of up to 1-2 weeks earlier. One area that could have improved in this study is the starting pool was very small compared to the study done by Ginsburg et al. [21], depending on other sources and not their techniques to determine the staring pool of keywords.

These studies have shown that search query data can be a useful tool for quickly and accurately detecting the occurrence of an ILI epidemic which could even be extended to tracking an epidemic. Additionally, it would be interesting to see if the formula created on this study could work as well for ILI epidemics that happen many years in the future from the time period used to train the model or in a different population that uses the same language. This line of research can help health officials, physicians, hospitals in reacting to epidemics faster and work to stop them better (faster) than with traditional methods used today such as the CDC or MOH reports.

Using twitter post data to track epidemics

The studies discussed in this subsection have a similar research goal to the previous subsection of attempting to detect and track ILI epidemics, but instead of using search query data the researchers use Twitter post data. One advantage to Twitter data over search query data is that Twitter posts come with context [23] (as opposed to not knowing why a person is searching for a topic). Twitter is a social networking site that allows its user to post any messages they like in 140 words or less (known as tweets), and currently has about 554 hundred million members worldwide with around 58 million tweets per

day [63]. With this Big Volume of people (sensors) there is a high probability that there can be useful ILI epidemic information being posted, but, of course, there will be noisy sensors and only through data mining techniques and analysis can the useful information be found. Another issue that could make Twitter post data potentially not reliable is that it is difficult to directly ascertain the age of a given Twitter user, and in fact a user may be discussing a disease symptom from a family member who differs in age from the user anyway, but research shows [22] that this is not a deterrent to age-based prediction.

Signorini et al. [23], in their research, employ Twitter post data across the United States by searching through particular spatiotemporal areas and analyzing the data in order to predicate weekly ILI levels both across and within these regions (CDC's ILI regions). The focus of their efforts is on the time period when the H1N1 epidemic was happening in the US as they gathered a large amount of Tweets from October 1, 2009 - May 20, 2010 using Twitter's streaming application programmer's interface (API) [64]. The tweets were sifted through looking for posts containing a preset of key words correlation to H1N1 (*h1n1*, *flu*, *swine*, *influenza*). Twitter's posts stream does come with its own filters according to the API documentation; therefore, not all tweets in the United States were used, only the subset to come through Twitter's filtered stream.

Tweets containing the following attributes were not used for analysis: if located outside the United States, from a user with a time zone outside the US, containing less than five words, not in English, not containing ASCII characters, and those submitted through the "API". The tweets leftover are used to create a dictionary of English words, from which items such as (#hashtags, @user, and links) are not used and also words were brought to their root form through Porter's Stemming Algorithm [65] to make the dictionary as efficient to use as possible. Using this dictionary Signori et al. gathered daily and weekly statistics (such as amount of tweets a word is present within) for each word both in the dictionary throughout the US and within each of the CDC's 10 regions.

The authors use the weekly statistics in order to estimate the weekly ILI epidemic status through a more general class of SVM called Support Vector Regression [66]. This is a generalization of Support Vector Machines, a type of classifier which attempts to find a minimal-margin separator, which is a hyperplane in the space of instances such that one class is on one side of the hyperplane and the other class is on the other side, with the distance between each class's instances and the hyperplane being maximized. As data is rarely linearly separable in feature space, a kernel function is used to transform the data into a higher-dimensional space, which generally assures that such a hyperplane exists. Signorini et al. used a polynomial kernel function for this purpose. To extend SVMs to the problem of regression (as opposed to classification), the model disregards any training data points which are already within a threshold ϵ of the model prediction (just as a SVM classification model disregards points which lie outside such a margin, as those points cannot help in determining the optimal hyperplane), and then builds a nonlinear model to minimize a preselected linear-error-cost function. In this model, each point (or instance) is a tweet, and the features each represent dictionary terms which occur more than 10 times per week. The value of each feature is the fraction of total tweets within the given week which contain the corresponding dictionary word (after stemming).

Determining if Twitter data can indeed detect ILI epidemics by accurately estimating CDC ILI values was done on a weekly basis on a national level and a regional level

(where region 2 is presented by the authors). For the national estimation they trained their method using 1 million of the tweets from October 1, 2009 - May 20, 2010 throughout the United States, with the objective being CDC's ILI values throughout the nation. Leave-one-out cross-validation was used to determine the accuracy of the model. The results found for the national level were quite accurate, scoring an average error of 0.28% with a standard deviation of 0.23%. In order to estimate the ILI levels in a particular region, with the goal being real-time prediction (as again, the CDC has a 1 to 2 week delay), they took the tweets that had geolocation information and fit these tweets to CDC region ILI readings from 9 of the 10 regions in order to construct the model to determine the results for last region (region 2). The authors argue that perhaps the smaller amount of tweets containing the geolocation information could have generated the slightly higher error rating of 0.37% with a standard deviation of 0.26%.

This study could have incorporated more words to include in their tweet searches rather than just the 4 they used as well as use methods to determine the most affective set. Also there was one prediction model applied, where as with more tested they could have found that a different model generated better results with less error across the US and within the tested region. As a note for this paper, Signorini et al. [23] also (and primarily) used the tweets to follow public concern for ILI epidemics throughout daily and monthly trends of tweets, but the scope of this survey does not look to cover such findings.

Achrekar et al. [22] devises a system deemed Social Network Enabled Flu Trends (SNEFT) that continuously monitors tweets with the goal of detecting and tracking the spread of ILI epidemics. This study uses a dataset of tweets and profile details of the Twitter users who have commented on flu keywords starting on October 18, 2009. The SNEFT network uses an OSN Crawler (bot that systematically searches online social networks) they developed to retrieve tweets from the internet using keywords *flu*, *H1N1*, and *swine flu* storing important information about the tweets (e.g. location, further determined by Google to determine geolocation), and relative keyword frequency in a spatio-temporal database. Also gathered are CDC ILI reports along with other influenza related data are downloaded from the internet.

The authors found that in one of their previous works using tweets from 2009–2010 (when influenza was a larger issue), the correlation coefficient was 0.98, but a huge drop was seen in this measure since for 2010-2011 at 0.47. Due to the H1N1 being less of an issue the noisy tweets are much more prominent resulting in a lower correlation coefficient, which leads to give something to think about: maybe this measure could be used to detect ILI in an area where the higher it is (without using methods to correct such noise) the more likely it is that an ILI epidemic is spreading through that region. The authors do try to alleviate the noise using text classification determining whether or not a tweet is correlated to a flu event or not.

For text classification they tested three different methods: decision tree, SVMs, and Naive Bayes, along with a few configurations for efficient learning. SVM was determined as the best classification method, beating the other two methods in both precision and recall. These classifiers were trained on a dataset of 25,000 tweets manually classified by the use of Amazon Mechanical Turk (an internet marketplace to perform such tasks by the coordinated use of human intelligence). The results of adding this text classifier caused the correlation coefficient to rise back up for the US as a whole and within each region. The authors also performed data cleansing from both retweets and multiple

tweets posted by the same user during a single bout with the same illness (different time windows were tested for the length of a single bout). Results found that only retweets should be considered.

The prediction model they used is Logistic Autoregression with exogenous inputs (ARX), which has the goal of predicting the CDC ILI statistic during a given week using both the tweets details and the CDC data (percent of physician visits from preceding weeks). The autoregression part of the model is the prediction of current ILI activity employing ILI activity from past weeks, and the exogenous inputs come from the tweets from the previous weeks. As simple Linear ARX could not handle the fact that the number of Twitter users is only bounded below by 0 (rather than being completely bounded between 0% and 100%), and so the authors introduce a logit link function for the CDC data and use a logarithmic transformation of the Twitter data. The purpose of the model is to produce timely updates estimating the percentage of physician visits for the week. The model can optionally use two distinct variables to determine how many previous weeks of data are used for each data type, with m referring to CDC data and n referring to Twitter data. The authors used 10-fold cross-validation on 33 weeks of data (from October 3, 2010 - May 15, 2011) to determine the (m, n) model that gave the least root mean squared error, and the (2,1) model (using only the current twitter data and the two most recent instances of CDC data) gave the best results.

Using their methods they were able to achieve a higher correlation to the actual CDC percent of visits ILI-related compared to the Twitter data alone for the 33 week period in 2010-2011, yet the authors do mention that Twitter data alone would still predict higher toward the beginning and end of the annual flu season (and, of course, during an epidemic). Achrekar et al. also found that they could get good results for regions 1, 6, and 9 determining that Twitter data and percentage of visits ILI-related are correlated across regions. One issue with this research is that only one prediction method is used; it would have been beneficial to see if the jump in correlation could also be seen with other prediction models or that results could be better with other models.

Results shown here are impressive, showing that Twitter data can be used to detect and possibly track Influenza like epidemics in real time. Both research efforts have shown that they can get very little error for predicting ongoing ILI epidemics, and produce results up to 1-2 weeks faster than the CDC posts their results. One issue not looked at by either paper discussed here, but discussed by Doan et al. [67], is that results could be improved if the keywords used were more broad, using a more knowledge-based method with 37 symptom keywords under respiratory syndromes from the BioCaster Ontology (BCO) [79] plus the word *flu*. This leads to wondering what results could be found for ILI prediction if other keyword selection methods were chosen. It would be interesting if more areas across the globe were researched in order to see if the results shown here can be achieved worldwide. An example of such research could be to see if similar results can be garnered using a set of keywords used to achieve good prediction results in a given population could get the same results in another. This line of research, with the promising results shown, could possibly one day create a system that could continually be tracking twitter posts in order to create a worldwide ILI epidemic map in real time helping people and the health care systems stay one step ahead.

Translational bioinformatics

Many authors argue that although it is a young field, Translational BioInformatics (TBI) is the way of the future for Health Informatics. TBI is an interdisciplinary subfield that deals with High Volumes of biomedical data and genomic data, where current research areas include developing new techniques for integrating biological data and clinical data as well as improving clinical methodology by including findings from biological research [68]. According to Chen et al. [1], the scope of TBI encompasses all the same levels of Health Informatics in general: Micro Level (i.e. Molecules), Tissue Level, Patient Level, and Macro Level (i.e. Population). The main goal of TBI is answering various questions at the clinical level. However, TBI does not seem to have a universally accepted definition. Also confused is the dividing line between what is included as clinical information and overall Health Informatics (where both seem to be used interchangeably as biomedical). When clinical information is referenced, it does seem to include all Health Informatics' levels from the tissue, patient, and population levels (biomedical data), as they are all connected to health. TBI is the field to bridge the gap between these fields and the molecular level, by way of developing tools that can better link this disparate data as well as develop and test techniques that can efficiently and accurately analyze such data together with the end goal of improving HCO [69].

TBI looks to use the health information from the discussed levels, combine the aggregated information in order to provide the most health gains, and help in offering the best of modern health practices. It makes sense that any field of research containing numerous subfields (including Health Informatics) would eventually converge into "translation" between its subfields, and this is what appears to have happened in the construction of TBI. To fully or even partially encompass TBI, many discussions and research examples would need to be presented, but for the scope of this paper the essence of "translation" is what will be shown. Thus, this section will present a small sample of discussions (such as editorials, perspectives, and highlights) from JAMIA (the Journal of the American Medical Informatics Association) to give the overall feel of TBI.

Butte et al. [70], in their 2011 editorial, discuss several Translational Bioinformatics studies featured in JAMIA which combine biological data (e.g. microarray data) with medical records to achieve medical gains as more data angles are tested in tandem. The authors acknowledge that TBI started from research done by a small group who found how to bridge the gap between computational biology and medicine. They also comment that they expect TBI to exploit Big Data in the not-too-distant future, with the goal of answering various clinical level questions and leading to a number of clinically applicable decisions.

Sarkar et al. [69] in a perspective article (released in 2011) discuss three areas as being the primary research objectives of TBI: 1.) determining the molecular level (genotype) impacts on the evolution of diseases, 2.) learning the impact of therapeutic procedures as can be measured by molecular biomarkers, and 3.) understanding the overall consistency between the molecular, the phenotype, and environmental correlations across different populations. The authors believe that with the explosion of both the molecular level data and biomedical data, as well as with the technological advances, TBI is in a prime position to possibly determine many of the mysteries of complex diseases or any of the other primary research objectives mentioned.

Shah et al. [71], in their 2012 editorial, discuss JAMIA's focus of Big Data in TBI. The goal of their 2012 summit (as well as TBI in general) was bringing molecular level data into Health Informatics, which is now possible due to the explosion of computational power now available. With this new computational power, molecular level data can now be used to improve medical understanding, as the farther one deviates from the patient level, the more data will be generated. This editorial acknowledges many works that are implementing TBI methodologies with success, including Liu et al. [72], who combined chemical, biological, and phenotype properties of drugs to improve predictions of adverse drug reactions by 5%, compared to only using chemical properties of the drugs. The authors argue that data for TBI should include a Big Volume of molecular data and a small amount of patient measurements from a Big Volume of patients. This leads to us wondering why are the authors limiting the scope of TBI research to only these two levels, and why not (as shown in [1]) include all levels of accessible human existence. The authors also mention, in closing, that the TBI Summit and the Clinical Research Informatics Summit are increasing their synergy, which is a good start, but eventually study groups from all levels of accessible human existence should come together in the same way.

Analysis and future works

This section will cover analysis and future work that could be beneficial from the lines of research presented in this survey.

Covering molecule-level

For the subsection "Using Gene Expression Data to Make Clinical Predictions" (covering studies [6] and [7]), the main challenge is handling the Big Volume of the features (probes). A particular concern, with gene expression microarray data only getting larger, is developing and testing probe selection techniques. Future work will need to be put into making these probe selection techniques able to handle these vast amounts of gene probes and select the subset that has the best correlation to the final prediction in a way that is as fast, accurate, and efficient as possible.

Covering tissue-level

For the "Creating a Connectivity Map of the Brain Using Brain Images" subsection (presented by the studies [8,66], and [9]), actual data mining analysis of the connectivity map remains entirely in the scope of future work. With the Big Volume of data being created by the HCP and the study done by Annese [8] there is a lot of opportunity for novel data mining techniques to be employed, allowing for the possible discovery of previously-unattainable knowledge about the brain and how it connects to the health of the human body, leading to possibly giving physicians more accurate diagnosis methods, earlier detection of diseases, etc.

For the subsection "Using MRI Data for Clinical Prediction" (which covered research efforts [11] and [10]), the difficulty is handling the Big Volume of the high-resolution MRI data. As each instance in an MRI dataset is quite large, and numerous MRI samples are needed to perform acceptable research, attempting to analyze the whole brain MRI datasets can be of quite a Big Volume challenge; luckily, recent computational power has entered an era that can handle such data (meaning this line of research is fairly young). More testing will be needed here, and the way forward for using MRI data for clinical

predictions is to create and test new machine learning methods that can accurately locate brain regions that best correlate to specific ailments to help physicians make more reliable predictions and diagnosis.

Covering patient-level

For the subsection "Prediction of ICU Readmission and Mortality Rate" (presented by the studies [12,13,15], and [14]), although all the studies had similar goals, none started with the same pool of variables; it could be beneficial in future work if all these variables and even variables from other data levels could be used in this area of research (as data from all 4 levels discussed here can be used to answer clinical questions). Also, data from the other studies presented previously could be used to expand the size of the starting pool (such as MRI data seen in [10], using the relative baseline method from [18], or even message board data from [19] and [20]). Another item to note is that throughout these studies only one feature selection technique was tested with minimal overlap; therefore, future work should look to test multiple feature selection techniques in order to find which one works the best with medical data. Classifiers that are employed for this line of research should be able to make decisions as a physician would do, that is, be able to look at a patient's medical attributes and make subjective decisions. Thus, there are three main concerns for making these predictions (starting variables, feature selection, and learning algorithms) and future research should be to test as many combinations of these three to find the best one that can make the most accurate and efficient predictions.

For the "Real-Time Predictions Using Data Streams" subsection (covered by the research efforts [16,18], and [17]), we see that this area of research is relatively new due to technology only very recently being able to handle the high-throughput processing necessary from the Big Velocity of data streams. As such, research is lacking testing and validation of developed methods as seen for Zhang et al. [17] and Thommandram et al. [18]. As sensors are not perfect (creating missing or erroneous data during a given time period), especially when being used for real-time analysis, future work will need to focus on developing and testing methods that can handle such data in the most reliable and efficient way. Another issue for future work in this line research is the need to devise and test numerous classification methods to find the best for both prognosis and acute problem detection (knowing that the best choices could be different for different ailments). It could also be beneficial if numerous sensors were checked to find the best set of sensors for each ailment prediction.

Covering population-level

For the "Using Message Board Data to Help Patients Obtain Medical Information" subsection (presented here by [19] and [20]), we find that the existing work does show that message board data does have the potential to supply patients with reliable medical information. More real world testing will be needed before this question can be fully answered, however. Message board data could be helpful if used in conjunction with studies such as Zhang et al.'s. [17] research on real-time diagnosis and prognosis, if it was found to give reliable medical data.

For the subsection "Tracking Epidemics Using Search Query Data" (shown by [21] and [5]), the results shown are promising as they could detect the occurrence of ILI epidemics in a given region in a relatively real-time manner with minimal error. Future work for this

line of research will need to find ways of finding the optimal set of keywords/queries to use for predicting the occurrence of an ILI epidemic. Also work should be done to determine whether research done in one area of the world can be translated to another (e.g. could the results found from Ginsburg et al. [21] translate to China's region or visa-versa), or across different languages or different areas with the same language.

For the "Tracking Epidemics Using Twitter Post Data" subsection (covered by [22] and [23]), we see that Twitter data can be quite useful for detecting or even tracking ILI epidemics. In the future there should be more work on developing methods to best determine what keywords to use for study (filtering methods), as well as testing more text classification methods in order to reduce the noisy tweets in the collected datasets. More research should also be concentrated on determining the best model for using Twitter post data to predict the CDC's percentage of ILI related visits, as according to [22] this is one of the ways the CDC monitors ILI epidemics and Twitter is quite highly correlated to this statistic. This line of research, with the promising results shown, could be possibly one day help create a system that could continually track Twitter posts in order to create a worldwide ILI epidemic map in real time, helping people and the health care systems stay one step ahead. Also, this leads one to wonder what else Twitter data can be used to determine for the health care system.

Covering translational bioinformatics

As discussed by reviews and editorials by [1,69,70], and [71], the way forward for TBI and Health Informatics as a whole is for the translational approach to increasingly envelop the entire scope of Health Informatics and continue to combine data from all levels of human existence (as diseases and ailments handled by the healthcare system are very complex). Through this combination, questions throughout all levels can be more precisely answered and results can be validated both more quickly and more accurately. All future work in Health Informatics should look to take the translational approach shown by TBI, not just focusing on combining the molecular level with the other levels, but attempting to make connections across as many levels of data as possible. This combination of data would offer Big Volume, Velocity, Variety, Veracity, and, of course, Value, which could provide an unprecedented degree of medical knowledge gain.

Conclusion

This survey discussed a number of recent studies being done within the most popular sub branches of Health Informatics, using Big Data from all accessible levels of human existence to answer questions throughout all levels. Analyzing Big Data of this scope has only been possible extremely recently, due to the increasing capability of both computational resources and the algorithms which take advantage of these resources. Research on using these tools and techniques for Health Informatics is critical, because this domain requires a great deal of testing and confirmation before new techniques can be applied for making real world decisions across all levels. The fact that computational power has reached the ability to handle Big Data through efficient algorithms (as well as hardware advances, of course) lets data mining handle the Big Volume, Velocity, Variety, Veracity, and Value of the data generated by Health Informatics (traditional or otherwise). The use of Big Data provides advantages to Health Informatics by allowing for more tests cases or more features for research, leading to both quicker validation of studies and the ability to accrue

enough instances for training when only a small fraction of instances exist within the positive class. The way forward for Health Informatics is definitely exploiting the Big Data created throughout all the various levels of medical data and finding ways to best analyze, mine and answer as many medical questions as possible.

As mentioned, the overall goal of answering any medical question, whether it be on the level of human-scale biology, clinical itself, or population, is to eventually improve healthcare for patients. The human body is a compound and complex system, containing many levels (not all of which may even be accessible as of now). Therefore, research needs to be done on data at all of these levels in order to answer the ever-growing list of medical questions on all of these levels. The studies covered here seek to answer questions on all four levels from data through analysis of data from the molecular, tissue, patient, and population levels.

TBI has the scope to use all these data levels with the goal of answering clinical questions, but this objective could benefit from not just attempting to answer clinical questions but questions on all levels. Perhaps future research in TBI and Health Informatics overall could focus on using data from all levels in order to find correlations and connections between them, possibly giving physicians more ways of diagnosing, treating, and helping their patients. All future work in Health Informatics should have a translational approach of using data from all levels of human existence.

All of the techniques discussed here show promise, provide inspiration for future work, and show the importance of using all accessible levels of data in Health Informatics. Each of the techniques presented in these studies can be both expanded and further tested using datasets with Big Volume and Variety (potentially on data from other levels) to see if their results are the same through different populations.

These studies are only a taste of the future possibilities that could be achieved through data mining and analysis of Big Data for Health Informatics. As computational power increases, more efficient and accurate methods will be developed. This could lead to possibly new levels of human existence becoming available for analysis, such as below the level of molecular data (e.g. the atomic scale) where the data gathered from each patient would be Really Big Data (RBD). Who knows, maybe a missing electron in a given chromosome could indicate a patient will be susceptible to cancer.

Even with the research discussed here and even with the promising results shown, it is up to future research to improve medical knowledge and create more advanced methods, as well as the readiness of the healthcare field to accept and apply these findings and techniques to improve HCO.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MH performed the primary literature review and analysis for this work, and also drafted the manuscript. RW worked with MH to develop the article's framework and focus. TMK introduced this topic to MH and RW, and coordinated the other authors to complete and finalize this work. All authors read and approved the final manuscript.

Received: 12 October 2013 Accepted: 20 February 2014
Published: 24 June 2014

References

1. Chen J, Qian F, Yan W, Shen B (2013) Translational biomedical informatics in the cloud: present and future. BioMed Res Int 2013: 8. [http://dx.doi.org/10.1155/2013/658925]
2. Martin M (2013) Big Cdata/social media combo poised to advance healthcare. HPC Source: 33–35. http://www.scientificcomputing.com/digital-editions/2013/04/hpc-source-big-data-beyond

3. Demchenko Y, Zhao Z, Grosso P, Wibisono A, de Laat C (2012) Addressing Big Data challenges for Scientific Data Infrastructure In: IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom 2012). IEEE Computing Society, based in California, USA, Taipei, Taiwan, pp 614–617
4. Huan JL, Pai V, Teredesai AM, Yu S(Eds) (2013) IEEE Workshop on BigData In Bioinformatics and Health Care Informatics. http://www.ittc.ku.edu/~jhuan/BBH/
5. Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS (2013) Monitoring influenza epidemics in China with search query from Baidu. PLoS ONE 8(5): e64323. [doi: 10.1371/journal.pone.0064323]
6. Haferlach T, Kohlmann A, Wieczorek L, Basso G, Kronnie GT, Béné MC, De Vos J, Hernández JM, Hofmann WK, Mills KI, Gilkes A, Chiaretti S, Shurtleff SA, Kipps TJ, Rassenti LZ, Yeoh AE, Papenhausen PR, Wm Liu, Williams PM, Fo R (2010) Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the international microarray innovations in leukemia study group. J Clin Oncol 28(15): 2529–2537. [http://jco.ascopubs.org/content/28/15/2529.abstract]
7. Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J, Bruin S, Kerr D, Kuppen P, van de Velde C, Morreau H, Van Velthuysen L, Glas AM, Van't Veer LJ, Tollenaar R (2011) Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol 29: 17–24. [http://jco.ascopubs.org/content/29/1/17.abstract]
8. Annese J (2012) The importance of combining MRI and large-scale digital histology in neuroimaging studies of brain connectivity and disease. Front Neuroinform 6: 13. [http://europepmc.org/abstract/MED/22536182]
9. Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K (2013) The WU-Minn human connectome project: an overview. NeuroImage 80(0): 62–79. [http://www.sciencedirect.com/science/article/pii/S1053811913005351]. [Mapping the Connectome]
10. Yoshida H, Kawaguchi A, Tsuruya K (2013) Radial basis function-sparse partial least squares for application to brain imaging data. Comput Math Methods Med 2013: 7. [http://dx.doi.org/10.1155/2013/591032]
11. Estella F, Delgado-Marquez BL, Rojas P, Valenzuela O, San Roman B, Rojas I (2012) Advanced system for automously classify brain MRI in neurodegenerative disease In: International Conference on Multimedia Computing and Systems (ICMCS 2012). IEEE, based in New York, USA, Tangiers, Morocco, pp 250–255
12. Campbell AJ, Cook JA, Adey G, Cuthbertson BH (2008) Predicting death and readmission after intensive care discharge. British J Anaesth 100(5): 656–662. [http://europepmc.org/abstract/MED/18385264]
13. Fialho AS, Cismondi F, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN (2012) Data mining using clinical physiology at discharge to predict ICU readmissions. Expert Syst Appl 39(18): 13158–13165. [http://www.sciencedirect.com/science/article/pii/S0957417412008020]
14. Ouanes I, Schwebel C, Franais A, Bruel C, Philippart F, Vesin A, Soufir L, Adrie C, Garrouste-Orgeas M, Timsit JF, Misset B (2012) A model to predict short-term death or readmission after intensive care unit discharge. J Crit Care 27(4): 422.e1–422.e9. [http://www.sciencedirect.com/science/article/pii/S0883944111003790]
15. Mathias JS, Agrawal A, Feinglass J, Cooper AJ, Baker DW, Choudhary A (2013) Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. J Am Med Inform Assoc 20(e1): e118–e124. [http://jamia.bmj.com/content/20/e1/e118.abstract]
16. Ballard C, Foster K, Frenkiel A, Gedik B, Koranda MP, Nathan S, Rajan D, Rea R, Spicer M, Williams B, Zoubov VN (2011) IBM Infosphere Streams: Assembling Continuous Insight in the Information Revolution. [http://www.redbooks.ibm.com/abstracts/sg.pages=247970html]
17. Zhang Y, Fong S, Fiaidhi J, Mohammed S (2012) Real-time clinical decision support system with data stream mining. J Biomed Biotechnol 2012: 8. [http://dx.doi.org/10.1155/2012/580186]
18. Thommandram A, Pugh JE, Eklund JM, McGregor C, James AG (2013) Classifying neonatal spells using real-time temporal analysis of physiological data streams: Algorithm development In: IEEE Point-of-Care Healthcare Technologies (PHT 2013). IEEE, based in New York, USA, Bangalore, India, pp 240–243
19. Ashish N, Biswas A, Das S, Nag S, Pratap R (2012) The Abzooba smart health informatics platform (SHIP)™– from patient experiences to big data to insights. CoRR abs/1203.3764: 1–3
20. Rolia J, Yao W, Basu S, Lee WN, Singhal S, Kumar A, Sabella S (2013) Tell me what i don't know - making the most of social health forums. Tech. Rep: HPL-2013–43. Hewlett Packard Labs [https://www.hpl.hp.com/techreports/2013/HPL-2013-43.pdf]
21. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009) Detecting influenza epidemics using search engine query data. Nature 457(7232): 1012–1014. [http://dx.doi.org/10.1038/nature07634]
22. Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B (2012) Twitter improves seasonal influenza prediction In: International Conference on Health Informatics (HEALTHINF'12). Nature Publishing Group, based in London, UK, Vilamoura, Portugal, pp 61–70
23. Signorini A, Segre AM, Polgreen PM (2011) The use of twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS ONE 6(5): e19467. doi:10.1371/journal.pone.0019467
24. McDonald E, Brown CT (2013) khmer: Working with big data in Bioinformatics. CoRR abs/1303.2223: 1–18
25. Bennett C, Doub T (2011) Data mining and electronic health records: selecting optimal clinical treatments in practice. CoRR abs/1112: 1668
26. Ertl P (2003) Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. J Chem Inform Comput Sci 43(2): 374–38. [http://pubs.acs.org/doi/abs/10.1021/ci0255782]. [PMID: 12653499].
27. Zhang J, Chung TDY, Oldenburg KR (1999) Validation of high throughput screening assays. J Biomolecular Screening 4(2): 67–73
28. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) Software version 4.0. Mol Biol Evol 24(8): 1596–1599. [http://mbe.oxfordjournals.org/content/24/8/1596.abstract]
29. Liu W, Li R, Sun JZ, Wang J, Tsai J, Wen W, Kohlmann A, Williams PM (2006) PQN and DQN: Algorithms for expression microarrays. J Theor Biol 243(2): 273–278. [http://www.sciencedirect.com/science/article/pii/S0022519306002530]
30. Bennett KP, Campbell C (2000) Support vector machines: hype or hallelujah SIGKDD Explor Newslett 2(2): 1–13. [http://doi.acm.org/10.1145/380995.380999]

31. Glas A, Floore A, Delahaye L, Witteveen A, Pover R, Bakx N, Lahti-Domenici J, Bruinsma T, Warmoes M, Bernards R, Wessels L, Van 't Veer L (2006) Converting a breast cancer microarray signature into a high-throughput diagnostic test. BMC Genomics 7: 278. [http://www.biomedcentral.com/1471-2164/7/278]
32. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: A severity of disease classification system. Crit Care Med 00003246-198510000-00009 13(10): 818–829. [http://journals.lww.com/ccmjournal/Fulltext/1985/10000/APACHE_II__A_severity_of_disease_classification.9.aspx]
33. Le Gall J, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 270(24): 2957–2963. [http://dx.doi.org/10.1001/jama.1993.03510240069035]
34. Keene AR, Cullen DJ (1983) Therapeutic intervention scoring system: Update 1983. Crit Care Med 00003246-198301000-00001 11: 1–3. [http://journals.lww.com/ccmjournal/Fulltext/1983/01000/Therapeutic_Intervention_Scoring_System__Update.1.aspx]
35. Hosmer DW, Lemesbow S (1980) Goodness of fit tests for the multiple logistic regression model. Commun Stat - Theory Methods 9(10): 1043–1069. [http://www.tandfonline.com/doi/abs/10.1080/03610928008827941]
36. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30(7): 1145–1159. [http://www.sciencedirect.com/science/article/pii/S0031320396001422]
37. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automatic Control 19(6): 716–723
38. Gajic O, Malinchoc M, Comfere TB, Harris MR, Achouiti A, Yilmaz M, Schultz MJ, Hubmayr RD, Afessa B, Farmer JC (2008) The stability and workload index for transfer score predicts unplanned intensive care unit patient readmission: initial development and validation *. Crit Care Med 36(3): 676–682. [http://journals.lww.com/ccmjournal/Fulltext/2008/03000/The_Stability_and_Workload_Index_for_Transfer.2.aspx]
39. Saeed M, Lieu C, Raber G, Mark R (2002) MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring In: Computers in Cardiology. IEEE Computer Society, based in California, USA, Memphis, Tennesee, USA, pp 641–644
40. Mendonça LF, Vieira SM, Sousa JMC (2007) Decision tree search methods in fuzzy modeling and classification. Int J Approximate Reason 44(2): 106–123. [http://www.sciencedirect.com/science/article/pii/S0888613X06000843]. [Fuzzy Decision-Making Applications]
41. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybernet SMC-15: 116–132
42. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A (1991) The APACHE III prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults. CHEST Journal 100(6): 1619–1636. [http://dx.doi.org/10.1378/chest.100.6.1619]
43. Hall M (1997) Correlation-based feature selection for machine learning. PhD thesis. The University of Waikato, Hamilton, New Zealand
44. Rodriguez JJ, Kuncheva LI, Alonso CJ (2006) Rotation forest: a new classifier ensemble method. IEEE Trans Pattern Anal Mach Intell 28(10): 1619–1630
45. Freund Y, Mason L (1999) The alternating decision tree learning algorithm In: Proceedings of the Sixteenth International Conference on Machine Learning, ICML '99. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp 124–133. [http://dl.acm.org/citation.cfm?id=.pages=645528657623]
46. Perkins AJ, Kroenke K, Unützer J, Katon W, Williams JW, Hope C, Callahan CM (2004) Common comorbidity scales were similar in their ability to predict health care costs and mortality. J Clin Epidemiol 57(10): 1040–1048. [http://www.sciencedirect.com/science/article/pii/S0895435604000812]
47. Walter LC, Covinsky KE (2001) Cancer screening in elderly patients: A framework for individualized decision making. JAMA 285(21): 2750–2756. [http://dx.doi.org/10.1001/jama.285.21.2750]
48. Domingos P, Hulten G (2000) Mining high-speed data streams In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '00. ACM, New York, NY, USA, pp 71–80. [http://doi.acm.org/10.1145/347090.347107]
49. Achananuparp P, Hu X, Shen X (2008) The evaluation of sentence similarity measures. In: Song IY, Eder J, Nguyen T (eds) Data Warehousing and, Knowledge Discovery, Volume 5182 of Lecture Notes in Computer Science. Springer Berlin, Heidelberg, pp 305–316. [http://dx.doi.org/10.1007/978-3-540-85836-2_29]
50. Thiagarajan R, Manjunath G, Stumptne M (2008) Computing semantic similarity using Ontologies. Tech. Rep.: HPL-2008–87. Hewlett Packard Labs [http://www.hpl.hp.com/techreports/2008/HPL-2008-87.pdf]
51. Sun J, Sow D, Hu J, Ebadollahi S (2010) A system for mining temporal physiological data streams for advanced prognostic decision support In: IEEE 10th International Conference on Data Mining (ICDM 2010), pp 1061–1066. doi:10.1109/ICDM.2010.102
52. Hay SI, George DB, Moyes CL, Brownstein JS (2013) Big data opportunities for global infectious disease surveillance. PLoS Med 10(4): e1001413. doi:10.1371/journal.pmed.1001413
53. Ashish N, Mehrotra S (2009) XAR An integrated framework for semantic extraction and annotation. In: Kalfoglou Y, IGI Global (eds) Cases on Semantic Interoperability for Information Systems Integration: Practices and Applications, Hershey, PA, USA, pp 235–254. [http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60566-894-9.ch011]
54. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32(suppl 1): D267–270. [http://nar.oxfordjournals.org/content/32/suppl_1/D267.abstract]
55. Hall MA, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: An update. SIGKDD Explor Newslett 11: 10–18
56. The Apache Software Foundation (2013) Apache Lucene. [http://lucene.apache.org/]. [Accessed: 2013-9-18]
57. OSI Pharmaceuticals (2013) Tarceva®(erlotinib) tablets advanced-stage non-small cell lung cancer treatment possible risks and side effects. [http://www.tarceva.com/patient/considering/effects.jsp]. [Accessed: 2013-9-18]
58. Centers for Disease Control and Prevention (2012) Diabetes report card 2012. Tech. rep. Centers for Disease Control and Prevention, US Department of Health and Human Services, Atlanta, GA. [http://www.cdc.gov/diabetes/pubs/pdf/diabetesreportcard.pdf]

59. National Institute for HealthandCareExcellence (2013) NICE pathways. [http://pathways.nice.org.uk/]. [Accessed: 2013-9-18]
60. Boehm EA (2001) The contribution of economic indicator analysis to understanding and forecasting business cycles. Ind Econ Rev 36: 1–36. [http://www.jstor.org/stable/29794223]
61. Moore GH, Shiskin J (1967) Indicators of Business Expansions and Contractions. National Bureau of, Economic Research. [http://papers.nber.org/books/moor67-2]
62. Liu Y, Lv B, Peng G, Yuan Q (2012) A preprocessing method of internet search data for prediction improvement: application to Chinese stock market In: Proceedings of the Data Mining and Intelligent Knowledge Management Workshop, DM-IKM '12. ACM, New York, NY, USA, pp 3:1–3:7. [http://doi.acm.org/10.1145/2462130.2462133]
63. Statistic Brain ResearchInstitute publishing as Statistic Brain (2013) Twitter statistics – statistic brain. [http://www.statisticbrain.com/twitter-statistics/]. [Accessed: 2013-9-18]
64. Twitter Inc (2013) The streaming APIs. [https://dev.twitter.com/docs/streaming-apis]. [Accessed: 2013-9-18]
65. van Rijsbergen CJ, Robertson SE, Porter MF (1980) New Models in Probabilistic Information Retrieval. British Library research & development reports, Computer Laboratory, University of Cambridge. [http://books.google.com/books?id=WDZ3bwAACAAJ]
66. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. In: Mozer MC, Jordan MI, Petsche T (eds) Advances in neural information processing systems. MIT Press, Cambridge, MA, pp 155–161
67. Doan S, Ohno-Machado L, Collier N (2012) Enhancing twitter data analysis with simple semantic filtering: example in tracking influenza-like illnesses. IEEE Computing Society, based in California, USA, La Jolla, California, USA
68. American Medical Informatics Association (2013) Translational Bioinformatics. [http://www.amia.org/applications-informatics/translational-bioinformatics]. [Accessed: 2013-9-18]
69. Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, Ohno-Machado L (2011) Translational bioinformatics: linking knowledge across biological and clinical realms. J Am Med Inform Assoc 18(4): 354–357. [http://jamia.bmj.com/content/18/4/354.abstract]
70. Butte AJ, Shah NH (2011) Computationally translating molecular discoveries into tools for medicine: translational bioinformatics articles now featured in JAMIA. J Am Med Inform Assoc 18(4): 352–353. [http://jamia.bmj.com/content/18/4/352.short]
71. Shah NH, Tenenbaum JD (2012) The coming age of data-driven medicine: translational bioinformatics' next frontier. J Am Med Inform Assoc 19(e1): e2–e4. [http://jamia.bmj.com/content/19/e1/e2.short]
72. Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Xw Chen, Matheny ME, Xu H (2012) Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J Am Med Inform Assoc 19(e1): e28–e35. [http://jamia.bmj.com/content/19/e1/e28.abstract]

doi:10.1186/2196-1115-1-2
Cite this article as: Herland *et al.*: **A review of data mining using big data in health informatics.** *Journal of Big Data* 2014 **1**:2.