Big Healthcare Data Analytics: Challenges and Applications

Chonho Lee leech@cmc.osaka-u.ac.jp*[3],
Zhaojing Luo zhaojing@comp.nus.edu.sg[1],
Kee Yuan Ngiam kee_yuan_ngiam@nuhs.edu.sg[1,2],
Meihui Zhang meihui_zhang@sutd.edu.sg[4],
Kaiping Zheng kaiping@comp.nus.edu.sg[1],
Gang Chen cg@zju.edu.cn[5],
Beng Chin Ooi ooibc@comp.nus.edu.sg[1], and
Wei Luen James Yip james_yip@nuhs.edu.sg[1,2]

[1]National University of Singapore
[2]National University Hospital
[3]Osaka University
[4]Singapore University of Technology and Design
[5]Zhejiang University

*Chonho Lee's work was done while he was at National University of Singapore

Abstract

Increasing demand and costs for healthcare, exacerbated by ageing populations and a great shortage of doctors, are serious concerns worldwide. Consequently, this has generated a great amount of motivation in providing better healthcare through smarter healthcare systems. Management and processing of healthcare data are challenging due to various factors that are inherent in the data itself such as high-dimensionality, irregularity and sparsity. A long stream of research has been proposed to address these problems and provide more efficient and scalable healthcare systems and solutions. In this chapter, we shall examine the challenges in designing algorithms and systems for healthcare analytics and applications, followed by a survey on various relevant solutions. We shall also discuss next-generation healthcare applications, services and systems, that are related to big healthcare data analytics.

Key Words:
Healthcare, Data Analytics, Big Data, Machine Learning.

1 Introduction

Large amounts of heterogeneous medical data have become available in various healthcare organizations and sensors (e.g. wearable devices). Such data, which is called Electronic Healthcare Records (EHR), is the fundamental resource to support medical practice or help derive healthcare insights. Previously, most of the medical practices were completed by medical professionals backed by their experiences, and clinical researches were conducted by researchers via painstakingly designed and costly experiments. However, nowadays the rapidly increasing availability of EHR is becoming the driving force for the adoption of data-driven approaches, bringing the opportunities to automate healthcare related tasks. The benefits may include earlier disease detection, more accurate prognosis, faster clinical research advance and better fit for patient management.

While the promise of Big Healthcare Analytics is materializing, there is still a non-negligible gap between its potential and usability in practice. Heterogeneity, timeliness, complexity, noise and incompleteness with big data impede the progress of creating value from data. Big Healthcare Analytics is no different in general. To make the best from EHR, all the information in EHR must be collected, integrated, cleaned, stored, analyzed and interpreted in a suitable manner. The whole process is a data analysis pipeline where different algorithms or systems focus on different specific targets and are coupled together to deliver an end-to-end solution. It can also be viewed as a software stack where in each phase there are multiple solutions and the actual choice depends on the data type (e.g. sensor data or text data) or application requirements (e.g. predictive models or cohort analysis).

There are mainly two types of EHR data, namely electronic medical records (EMR) and sensor data. There are two major directions of the advancement of Big Healthcare Analytics related to EMR data and sensor data respectively. One is to provide better understanding and interpretation about the basic EMR from hospitals. The key challenges are to detect the specific characteristics of EMR data and build customized solutions for every phase of the data analysis pipeline. The other is to benefit from the development of new technologies of sensors (e.g. capturing devices, wearable sensors, and mobile devices) by getting more medical related data sources. The key challenges are to support real time data processing and real time predictive models.

EMR Data: With the development of electronic healthcare information systems, more and more EMR data is collected from hospitals and ready to be analyzed. EMR data is time series data that records patients' visits to hospitals. As shown in Figure 1, EMR data typically includes socio-

demographic information, patients' medical history and heterogeneous medical features such as diagnoses, lab tests, medications, procedures, unstructured text data (e.g., doctors' notes), image data (e.g., magnetic resonance imaging (MRI) data) and so on. The effective use of EMR can be extremely helpful in data analytics tasks such as disease progression modeling, phenotyping, similar patient and code clustering [54] and so on. However, mining from EMR data is challenging due to the following reasons. First, EMR data is high-dimensional as a large number of medical features have to be captured. Second, EMR data is often dirty or incomplete due to the collection being done over a long period of time; consequently, this data has to be treated before it can be used. Third, EMR data is typically collected irregularly by hospitals as patients tend to visit the hospital only when necessary. Consequently, we have to address challenges such as high-dimensionality, sparsity, noise, missing data, irregularity and bias when we design analytics solutions.

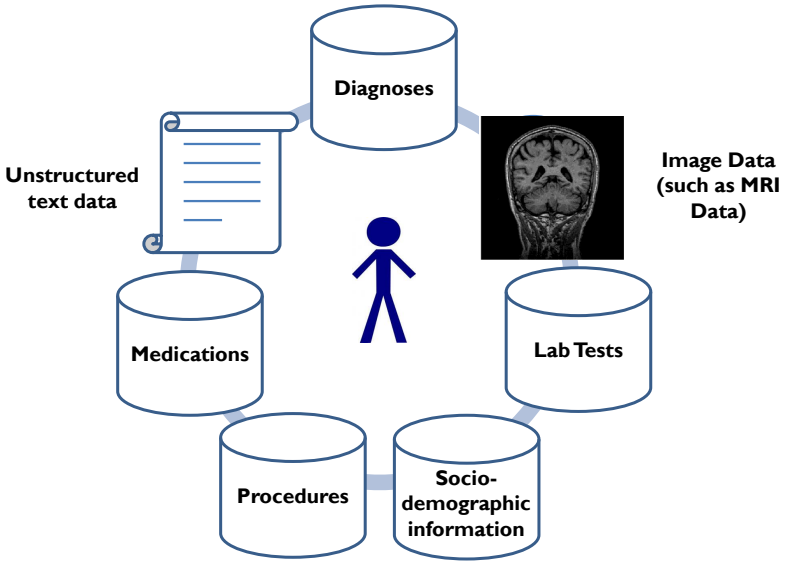

Figure 1: EMR data consisting of structured data, unstructured text data and image data etc.

Sensor Data: With the wide use of sensors in collecting data for monitoring and better response to the situational needs, sensor signals or data streams are also common in healthcare data. From a big data perspective, such sensor signals exhibit some unique characteristics. The signals originate from millions of users and sensor/mobile devices, form an extremely large volume of heterogeneous data streams in real time. Figure 2 shows example networks with various sensors/mobile devices, where the data streams are generated.

With the advancement in sensor technology and miniaturization of sensor devices, various types of tiny, energy-efficient and low-cost sensors are expected to be widely used for

improving healthcare [2, 15, 29]. These sensors form wireless networks such as Internet of Things [21], wearable networks [5] and in-body nano-scale networks [67, 69], and generate massive and various types of data streams. Monitoring and analyzing such multi-modal data streams are useful for understanding the physical, psychological and physiological health conditions of patients. For examples, surveillance cameras, microphones, pressure sensors installed in a house can track the daily activities of elderly people remotely and can help detect falls *; EEG and ECG sensors can capture changes in patient's emotions and help control the severity of stress and depression [91, 89, 116]; Carbon nano-tube sensors measuring oxygen saturation and pH of the body, which are bio-markers to react against cancer tissues, help doctors to manage patients [103, 55]. For many healthcare applications, such data must be acquired, stored and processed in a real-time manner. However, there are limitations in implementing the real time processing of enormous data streams with a conventional centralized solution that does not scale well to process trillions of tuples on-the-fly [21]. Instead, distributed architectures [1, 60, 47, 35, 75, 115] are more amenable to scalability and elasticity to cater to different workloads.

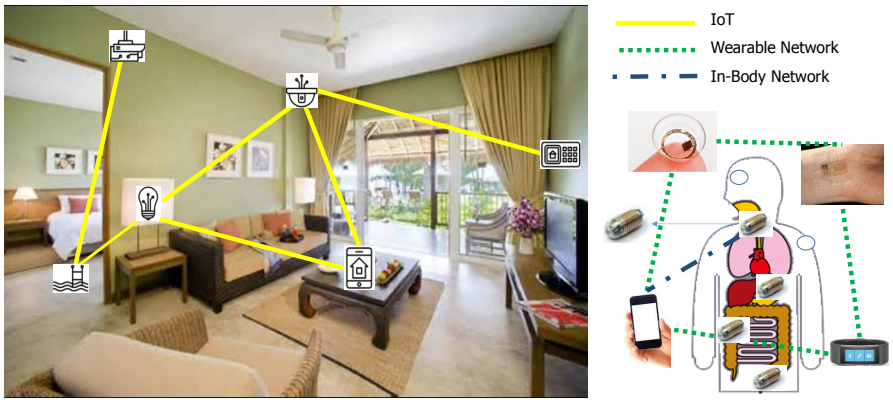

Figure 2: Network of interconnected sensors (e.g., mobile phones, cameras, microphones, ambient sensors, smart watches, smart lenses, skin-embedded sensors [114], intestinal gas capsules [48]) that produce healthcare data streams.

Implementing the next-generation smart healthcare systems, especially those for supporting Big Healthcare Analytics, requires us to carefully examine every phase in the data analysis pipeline, and adjust the methods by modeling the specific medical context. An overview of existing solutions would be of value to those who want to implement a new solution or application. With this in mind, we hereby provide an overview of healthcare data analytics and systems in this chapter. Based on the different types of EHR data and their characteristics introduced

*www.toptenreviews.com/health/senior-care

earlier, we next outline several challenges in big healthcare data analytics and review various proposals with respect to these challenges in Section 2. Section 3 describes several key steps for processing healthcare data before doing data analytics. Section 4 presents various healthcare applications and services that can be supported by data analytics, and various healthcare systems. We summarize and discuss potential directions in Section 5.

2 Challenges

Mining EMR data is challenging because of the following reasons: high-dimensionality, irregularity, missing data as well as sparsity, noise and bias. Figure 3 shows a real-life patient matrix to help readers better understand different challenges in EMR data. Each challenge will be described in detail.

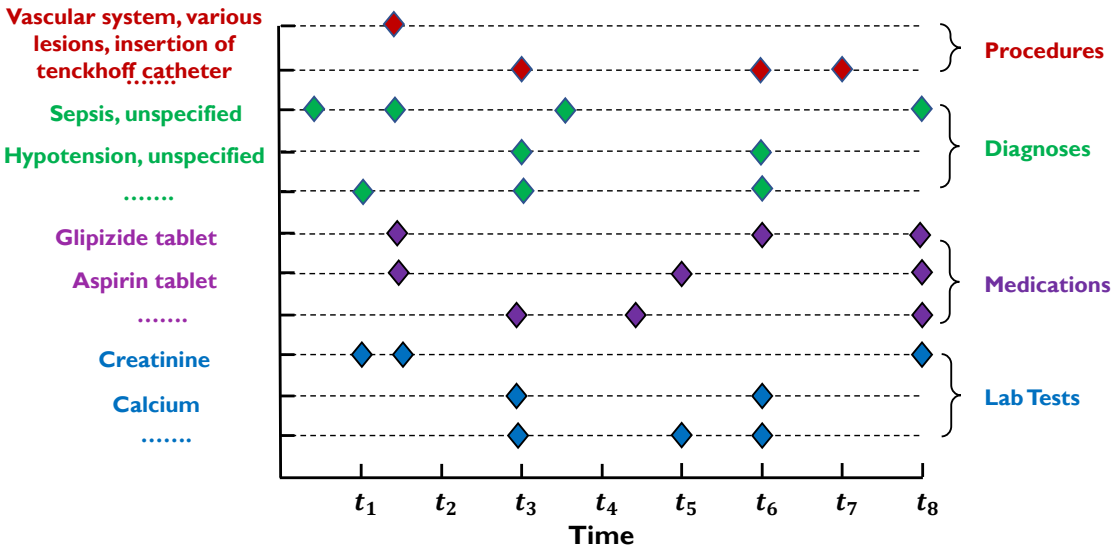

Figure 3: EMR data of patients

2.1 High-dimensionality

EMR data typically consists of hundreds to thousands of medical features from multiple sources. This gives rise to the high-dimensionality problem. To illustrate, in a sample data set from the real-world longitudinal medical database of National University Hospital, for 10000 patients over a one year period, there are 4143 distinct diagnosis codes. However, nearly 80% of the patients have fewer than 10 diagnosis codes and about 70% of them have fewer than four visits to the hospital, which makes each patient's feature vector high-dimensional and sparse. Similar characteristics are observed from public data sets. In a diabetes readmission data set from UCI

Machine learning Repository[†], there are about 900 distinct diagnosis codes, but most patients are associated with fewer than three diagnosis codes. In a subsample of 10000 patients, extracted from a data set provided by Centers for Medicare and Medicaid Services (CMS) 2008-2010, we find nearly 88% of the patients have fewer than four diagnosis codes, although there are 153 distinct diagnosis codes in total.

Dealing with very high-dimensional data is challenging, as it introduces more parameters into the model, making the model much more complex. Also, high-dimensional data is highly likely to be associated with noise and sparsity problems. To address the high-dimensionality problem, there are two main categories of dimensionality reduction methods, namely feature selection and feature extraction [‡].

2.1.1 Feature Selection

Feature selection is the process of selecting a subset of relevant predictive features for model construction [34]. Common feature selection methods include filter methods, wrapper methods and embedded methods [34]. Filter methods select significant features independent of models. These methods will rank the features according to their relations to the predicted features and are usually univariate. Filter methods are computationally efficient and robust to over-fitting but the relations between features are neglected. Different from filter methods, wrapper methods take the relationships between features into consideration. A predictive model will be built to evaluate the combinations of features and a score will be assigned to each set of feature combinations based on the model accuracy. Wrapper methods take a much longer time since they need to search a large number of combinations of features. Also, if the data is not enough, this method will have over-fitting problem. Embedded feature selection methods shift the process of feature selection into the building process of the model. Embedded methods have the advantage of the previous two methods, fast and robust to over-fitting as well as considering relationships between features. Unfortunately, these methods are not generic as they are designed for specific tasks with certain underlying assumptions. For healthcare analytics, univariate analysis and stepwise regression are widely adopted. These two methods belong to filter methods and wrapper methods respectively.

In [68], a univariate analysis as well as a multivariate logistic regression with stepwise forward variable selection are implemented to perform feature selection. Among the initial 20 or so manually selected features, five of them are finally found to be significantly associated with readmission within 30 days for a population of general medicine patients in Singapore and are

[†]https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

[‡]http://www.kdd.org/kdd2016/topics/view/dimensionality-reduction

included in the final model. These features include age, Charlson comorbidity index, white cell count, serum albumin and number of emergency department (ED) visits in previous 6 months. In [53], a modified stepwise logistic regression is performed to do feature selection in order to predict heart failure readmissions. In this work, with the help of domain experts, 95 condition categories (CC), two demographic variables (age and gender) and two procedure codes are included as candidate features. After feature selection, 37 features are considered in the final model. In [106], a backward stepping feature selection method is used to select significant features for the final model. 48 patient-level and admission-level features are collected from 4812 patients that are discharged in Ontario. Among these variables, only 4 of them, namely, length of stay in days, acute (emergent) admission, comorbidity (Charlson comorbidity index score) as well as number of ED visits during previous 6 months, are finally found out to be significant to the readmission prediction task.

2.1.2 Feature Extraction

Apart from feature selection methods, we may perform feature extraction to learn low-dimensional latent representations of original features to reduce dimensionality. The main idea of feature extraction is to embed original features in a lower-dimensional space where each dimension corresponds to a combination of original features. Compared to the features derived by feature selection methods, the features learned by feature extraction are much more difficult to interpret. There are mainly two categories of feature extraction methods, depending on whether the transforming methods are linear or non-linear. Linear transforming methods may struggle in discovering complex non-linear relationships between the original features while non-linear transforming methods are much more difficult to optimize and are more likely to be trapped in local optima.

In [57], Gaussian process regression is used to infer longitudinal probability densities for uric acid sequences. Following this transforming step, an auto-encoder is then used to infer meaningful features from the transformed probability densities. When configuring the hidden layer of the deep learning model, the dimension of the hidden layer could be set smaller than the visible layer so as to avoid learning the identity transformation.

In [105], a modified Restricted Boltzmann Machine (RBM) is trained to embed medical objects in a low-dimensional vector space which works as a new representation for the raw high-dimensional medical feature vector. This new low-dimensional representation is then used for assessing suicide risk.

In addition to learning non-linear low-dimensional hidden representations using deep learning models, dimensionality reduction can also be achieved through principal component analysis

(PCA). A stochastic convex sparse PCA method is developed in [7] to effectively perform sparse PCA on EMR data so that the derived representation is both low-dimensional and interpretable.

2.2 Irregularity

Irregularity is one of the bothersome characteristics of EMR data and provides challenges for EMR data analytics. Irregularity is caused by the fact that patients will only have EMR data recorded when they visit the hospital. As a consequence, patients' EMR data is organized into a "longitudinal patient matrix" where one dimension represents various medical features and the other is time [118, 108], and the consecutive patients' EMR records will be scattered within uneven-spaced time spans. Moreover, for different patients, the granularity of medical records varies significantly and the time periods between visits also vary a lot.

Generally, there are three categories of methods to alleviate this irregularity issue. The details are demonstrated as follows.

2.2.1 Use of Baseline Features

The first kind of methods is to utilize patients' "baseline" features (i.e., the data recorded when patients visit the hospital to perform examinations for the first time) for EMR data analytics tasks.

For instance, baseline MRI scans [100] are used to predict patients' clinical scores including Mini-Mental State Examination (MMSE), Dementia Rating Scale (DRS), Auditory Verbal Learning Test (AVLT) and Alzheimer's disease Assessment Scale - Cognitive Subtest (ADAS-Cog). In this work, a relevance vector regression, a novel sparse kernel method in a Bayesian network, is employed. Similarly, patients' baseline MRI features (together with baseline MMSE features, and some demographic features) [27] are used to predict the one-year changes in the MMSE feature. The whole process entails data collection and extraction from MRI data, feature dimensionality reduction via PCA, prediction of future MMSE changes via robust linear regression modelling. In [107], the association between patients' baseline features and changes in severity-related indicators is examined via linear mixed-effects models and the baseline features are used to predict the conversion time from amnestic mild cognitive impairment (aMCI) to Alzheimer's disease via Cox proportional hazards models. In [95], a risk score based on patients' baseline features and demographic features is proposed to predict the probability of developing Type 2 diabetes. Specifically, a multivariate Cox regression model is used to assign weights to different variables. In [26], Alzheimer's disease patients' baseline features are used to predict their probability for different class memberships representing different severity levels

based on a multivariate ordinal regression model using Gaussian process that is implemented in a Bayesian network.

Another line of research focuses on multi-task learning [14]. Several works [119, 117, 80] choose Alzheimer's disease patients as the cohort and predict their future severity in terms of MMSE values and ADAS-Cog values in multiple timepoints. The prediction in each timepoint is modeled as a regression task. These works utilize patients' baseline features and employ multi-task learning to capture the relationships between tasks (i.e. the prediction tasks in multiple future timepoints), where all these tasks are trained together with constraints on the changes within consecutive timepoints. However, there are several minor differences between these two methods. Besides predicting patients' future severity, [119] manages to select a common set of features that are significant to all prediction tasks via a $l_{2,1}$-norm penalty term. [117] extends [119] in that it not only selects a common set of features for all tasks, but also selects task-specific bio-markers via a l_1-norm penalty term. [80] proposes a further improvement of regression performance to consider the consistency for prediction utilizing multi-modal data (i.e., multiple sources/forms of medical features), and handles the missing data in both modality data and label data via an adaptive matrix factorization approach.

The prediction performance of this category may be limited by only making use of baseline features. This is due to under-utilization of time-related features. Since patients' health conditions tend to change along with time, it is of vital importance to utilize as many time-related features available as possible other than just baseline features. Another limitation specific to multi-task learning methods is that they can only deal with linear relationships among features. However, in the medical area, relationships between medical features, relationships between medical features and labels can be quite complicated and may not be described using simple linear relationships.

2.2.2 Data Transformation

In regularly sampled series, lots of successful algorithms have been developed. However, there remain many challenging problems in handling irregular data. In the medical area, we are faced with longitudinal, irregularly collected EMR data. To alleviate this problem, some existing works organize patients' EMR data along with time and have divided such longitudinal data into "windows". For instance, in [110], a probabilistic disease progression model based on Markov jump process is proposed to model the transition of disease states for Chronic Obstructive Pulmonary Disease (COPD) patients. The EMR data is processed by segmenting the time dimension into non-overlapping windows (i.e., encounters) with a length of 90 days, and the regularly reorganized data is then used for further modelling and analysis.

Similarly, in [18], two kinds of features are used: daily recorded features and static features. These two kinds of features are exploited to distill knowledge from deep learning models (including Stacked Denoising Auto-encoder and Long Short-Term Memory (LSTM)) by making use of Gradient Boosting Trees.

In [62], training data is processed by resampling to an hourly rate, where the mean measurement is applied in each hourly window. The application task is to classify 128 medical diagnoses employing a LSTM model [40] to capture the dynamic patterns in input features. In [16], the dynamic changing trends are captured using an alternative approach. After preprocessing data into overlapping "windows", the occurrence of a certain disease is predicted based on Multi-Layer Perceptron (MLP) with prior domain knowledge.

While transforming irregular data into regular time series allows us to employ some efficient methods (such as linear algebra) directly, we need to be aware of the side effects associated with such method. For instance, the resampling method may possibly lead to the sparsity and missing data problems because for some features, there could be no observations during certain time windows. Moreover, by dividing longitudinal data into windows, the model may be less sensitive to capturing short-time feature patterns.

2.2.3 Direct Use of Irregular Data

Contrary to the methods mentioned above, there are approaches that make use of medical features with irregular, accurate time information directly. In [86], the computation of LSTM model is adapted by incorporating the time spans between consecutive medical features to handle the irregularity. The proposed model is applied to model disease progression, recommend interventions and predict patients' future risks. Similarly, models based on Gated Recurrent Units (GRU) [19] have been proposed which simultaneously consider the masking and time durations between consecutive medical features in a decay term [17]. Through this decay term, the proposed method is designed to handle irregular data directly.

This category of methods demonstrates the possibility of fully utilizing available data. However, when parameterizing time between consecutive medical features, these methods model the decay term using a heuristic method, such as a monotonically non-increasing function based on logarithm or a parametric method to learn time weight matrix [86]. Such heuristic methods may cause either under-parameterization or over-parameterization.

2.3 Missing Data and Data Sparsity

Typically, missing EMR data can be caused by either insufficient data collection or lack of documentation. In data collection problem, patients are not checked specifically for a certain medical feature. In documentation problem, patients are checked for a certain feature, but either their outcomes are negative, which means that they are not needed to be documented, or the outcomes are positive but are not recorded due to human errors [112]. The missing data problem is further exacerbated by data sparsity due to the fact that most patients only pay a few visits to the hospital, and in most visits, only a couple of medical features are recorded. Fortunately, missing data and sparsity problems share many common techniques for solving them.

In [93], various imputation methods for handling missing data are described and broadly categorized into two categories. The first category is under the *missing at random* assumption, including methods from simple ones such as case deletion, mean imputation, to the advanced ones such as maximum likelihood and multiple imputation. The second category is under the assumption of *missing not at random*, which mainly includes selection models and pattern-mixture models.

In [18], a simple imputation strategy is adopted in solving missing temporal features: for features with binary values, the majority value is used for filling; for features with numerical values, the mean values are used for imputation. In [62], the forward-filling and back-filling method is proposed to fill the missing data during a resampling process. For a feature that is totally missing, the clinically normal value suggested by medical experts is used instead.

Apart from solving the missing data problem through imputation in preprocessing phase, a recent work [17] addresses missing data by incorporating two missing patterns: masking and time duration inside the objective function of the deep learning model structure. The proposed method is designed to capture the informative missing EMR data.

For sparsity, the above-mentioned missing data imputation methods, such as mean imputation, forward-filling and back-filling, are also widely used to get a dense patient matrix in order to solve the sparsity problem. Matrix densification/completion is another method to solve the sparsity problem [118]. The basic idea of matrix completion is to recover unknown data from a few observed entries. The algorithm assumes that the completed data for each patient has the factorization form of two matrices. Thus the data can be completed by multiplying these two derived matrices to densify the raw patient matrix.

2.4 Noise

EMR data is usually noisy due to various reasons, such as coding inaccuracies, inconsistent

naming conventions, etc. Many machine learning researchers tend to learn latent representations to derive more robust representations in order to solve this problem. These methods include Gaussian regression models, topic models or factorization-based methods. A Gaussian process regression is proposed in [57] to transform the raw noisy data (uric acid sequences) into a continuous longitudinal probability density function. This transforming step assumes that each uric acid sequence is a set of possibly noisy samples taken from the source function. Afterwards, instead of operating on the noisy raw data, an auto-encoder takes the mean vector of the learned Gaussian distribution to derive hidden representations. In [39], the noise problem is resolved by learning meaningful medical concepts (or phenotypes) in the form of tensors. Its insight is to map raw EMR data into medical concepts, learn latent bases of the raw data and perform predictive tasks in the latent space. Similar to [39], [111] factorizes the raw patient tensor into several interaction tensors, each representing a phenotype. Experimental results suggest that this method is robust to noise because it not only depends on the observed tensor but also on various other constraints to derive the phenotype tensors. Another latent variable model to solve the noise problem is proposed in [88], which leverages a topic modeling technique to handle noise in the raw EMR data by modelling the relationships between observations that are implicit in the raw data.

2.5 Bias

Bias is also an outstanding characteristic of EMR data, which is regarded as a non-negligible issue in healthcare data analytics [87, 37, 41, 42]. Bias is often considered as biased sampling, which means that the sampling rate is dependent on patients' states, and also dependent on doctors' judgment on patients. Consequently, patients are sampled more frequently when ill, but are sampled less frequently when comparatively healthier [42]. Other sources of bias include (i) the same patient may visit different healthcare organizations for medical help and different organizations do not share information between each other; (ii) patients fail to follow up in the whole medical examination process; (iii) the recorded data in one specific healthcare organization is incomplete [37].

In [87], bias in the lab tests of EMR data is modeled by examining the relationships between concrete lab test values and time intervals between consecutive tests, and exploiting the lab test time patterns to provide additional information. Furthermore, different lab test time patterns are identified so that they can be separately modeled when EMR analytics and experiments are performed. The limitation of this method is that it can only model the bias based on coarse-grained patterns, and the intra-pattern biases remain to be unsolved.

The influence of time parametrization in EMR data analytics is studied in[42], in which

three methods of parameterizing time are compared: sequence time (i.e., the sequence of measurements' occurrences after a specified start time), clock time (i.e., the absolute time of measurements) and an intermediate warped time which is a trade-off between the previous two. The study finds that the sequence time could perform the best among three methods, perhaps due to clinicians' tendency to change sampling rate according to patients' severity. However, the proposed time parameterization methods are heuristic in nature and may cause under-parameterization or over-parameterization.

2.6 Knowledge Base

Over the years, a large number of knowledge sources in the healthcare domain have been built and maintained to provide people with easy access to comprehensive medical information. Common knowledge sources include International Classification of Diseases (ICD-9 or ICD-10) for diagnoses, Healthcare Common Procedure Coding System (HCPCS) for procedures, Logical Observation Identifiers Names and Codes (LOINC) for laboratory tests. Besides, Unified Medical Language System (UMLS) maintains useful relationship knowledge, and Lab Tests online explains the relationships between lab tests. Incorporating structured medical knowledge provides a good basis to construct intelligent predictive models, which can then be used to improve healthcare data analytics in terms of interpretability and predictive ability. In [111], existing medical knowledge is incorporated into the tensor factorization algorithm in order to derive more fine-grained phenotypes. In particular, the algorithm can derive different sub-phenotypes which fall into a broader phenotype. This can help to stratify patients into more specific subgroups.

Since knowledge has been successfully incorporated into deep learning models in natural language processing field [9], many deep learning researchers are heading towards incorporating existing medical knowledge into deep learning models in order to improve interpretability as well as performance of the model. In [16], medical ontology knowledge is incorporated into the MLP as the regularization term so as to improve the performance of the model. A similar approach is developed in [105], in which structural smoothness is incorporated into the RBM model via a regularization term. Its basic underlying assumption is that two diseases which share the same parent in the taxonomy are likely to possess similar characteristics.

3 Key Steps for Processing

Before EHR data (including EMR data and sensor data) is input into various models for analysis, data needs to go through several steps of processing. Figure 4 illustrates the pipeline for

big data analysis [45, 22]. Firstly, EHR data needs to be recorded, accessed and acquired. Secondly, obtained raw EHR data is probably heterogeneous, composed of structured data, free-text data (such as doctors' notes), image data (such as MRI images) and sensor data. Hence, data extraction is of great concern for further analysis. Furthermore, data cleansing is needed to remove inconsistencies and errors, and data annotation with medical experts' assistance contributes to effectiveness and efficiency of this whole process from acquisition to extraction and finally cleansing. Thirdly, data integration is employed to combine various sources of data, such as different hospitals' data for the same patient. Finally, processed EHR data is modeled and analyzed, and then analytics results are interpreted and visualized. In this section, several key steps for processing EHR data, namely, data annotation, data cleansing, data integration and data analytics/modelling are described in detail respectively.

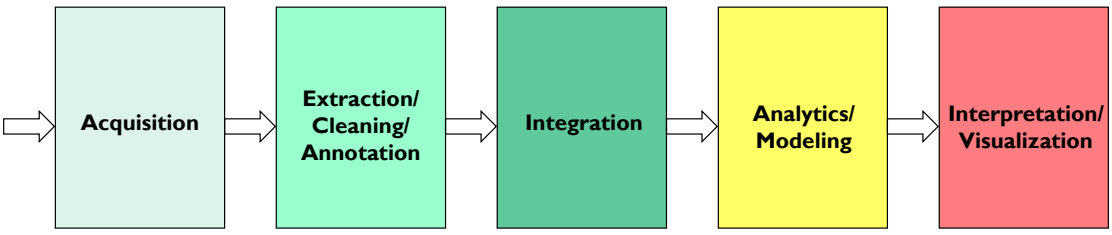

Figure 4: The big data analysis pipeline [45]

3.1 Data Annotation

Incompleteness is the leading data quality issue when using EHR data to build a learning model [10], since many study variables have missing values to various degrees. The uncertainty of EHR data can be resolved by model inference using various learning techniques [110]. However, the rationale of most healthcare problems can be too complex to be inferred by machines simply using limited EHR data. In such cases, enriching and annotating EHR data by medical experts are the only choice to help the machine to interpret EHR data correctly.

The acquisition of supervised information requires annotations by experts, resulting in a costlier exploitation of data. To reduce the cost involved in data annotation, voluminous research works have been conducted. In general, most of the research issues belong to active learning, which aims to only annotate those important data instances while inferring others and thereby the total number of annotated data is significantly reduced. The key idea of active learning is that learning algorithms can achieve higher accuracy with fewer training labels if they can choose the data from which they learn. The general solutions of active learning include reducing the uncertainty in training models [59], differentiating hypotheses which are consistent with the current learning set (i.e. Query-By-Committee) [97, 78, 102], maximizing the expected model

change after receiving a new sample [96], minimizing the expectation [90] or variance [20] of the empirical loss, maximizing the information density among the whole query space [96] and etc.

However, in current status, all these methods have limitations in real healthcare applications. The fundamental reason is that the supervised information in some complex analytics tasks may be hard to be quantified by a human. Since most easy annotating tasks can usually be well recognized by simply using machine efforts, the required tasks for expert annotation are usually complex jobs such as inference flow in a medical concept graph. These categories of supervised information can hardly be annotated via quantified labels which are well studied in the active learning community and integrated to the healthcare analytics system.

3.2 Data Cleansing

In this section, we discuss the importance of data cleansing for EHR data (including EMR data and sensor data). As mentioned in Section 2.4, EMR data is typically noisy due to several reasons, for example, coding inaccuracies, erroneous inputs, etc. Before raw EMR data is ready for use, we should develop data cleansing techniques. This requires us to understand the healthcare background of the dirty EMR data and work with domain experts to achieve better cleansing performance. Data cleansing is quite challenging when we consider sensor data. Data from sensor/mobile devices is inherently uncertain due to lack of precisions, failures of transmissions and instability of battery life, etc. Thus, it is essentially required to (i) identify and remove inaccurate, redundant, incomplete and irrelevant records from collected data and (ii) replace or interpolate incorrect and missing records with reasonably assigned values. These processes are expected to improve data quality assessed by its accuracy, validity and integrity, which lead to reliable analytics results.

3.3 Data Integration

Data integration is the process of combining heterogeneous data from multiple sources to provide users with a unified view. [36, 32, 25] explore the progress that has been made by the data integration community and some principles, as well as theoretical issues, are introduced in [58, 24].

Data integration techniques for EMR data and sensor data have different characteristics. For EMR data, we need to integrate heterogeneous EMR data from different sources including structured data such as diagnoses, lab tests, medications, unstructured free-text data like discharge summary, image data like MRI, etc. Different from EMR data, sensor data is generated by var-

ious types of sensor/mobile devices at different sampling rates. The heterogeneity of abundant data types brings another challenge when we integrate data streams due to a tradeoff between the data processing speed and the quality of data analytics. The high degree of multi-modality increases the reliability of analytics results, but it requires longer data processing time. The lower degree of multi-modality will improve data processing speed but degrade the interpretability of data analytics results. The efficient data integration helps reduce the size of data to be analyzed without dropping the analytics performance (e.g., accuracy).

3.4 Data Analytics and Modelling

After the three processing steps described above, we focus on EHR data analytics and modelling part. We have proposed a healthcare analytics framework as shown in Figure 5. This framework is composed of four phases which can give a better representation of medical features, and exploit the intrinsic information in EHR data and therefore, benefit further data analytics performance. The key idea for each phase is demonstrated as follows.

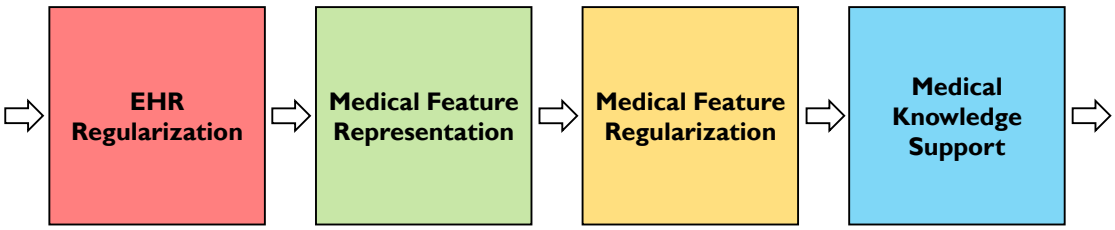

Figure 5: Our healthcare analytics framework

EHR Regularization: In this step, we focus on transforming the EHR data into a multivariate time series, solving the problems of irregularity, missing data and data sparsity, and bias as discussed in Section 2. The output of this phase is an unbiased, regularly sampled EHR time series.

Medical Feature Representation: In this phase, we aim to represent the medical features to reflect their feature-time relationships. To be specific, we learn for each medical feature whether this feature has influence after a certain time period and which features it poses influence on.

Medical Feature Regularization: After regularizing EHR data into a more suitable format for analytics and representing features to reveal underlying relationships, we now turn to re-weight medical features for better analytics results. This re-weighting can be achieved by trading-off features' confidence and significance and differentiating common/rare, significant/noisy features.

Medical Knowledge Support: In this phase, we propose to instil medical knowledge into typical machine learning and deep learning models for better analytics performance. This will involve finding the best structures to represent existing medical knowledge (i.e., domain knowledge) and developing the model training scheme using such knowledge.

4 Healthcare Applications

This section presents several healthcare applications, services and systems that are supported by data analytics in EMR data and sensor data. Figure 6 illustrates some of them using EMR data.

4.1 Applications for EMR Data

4.1.1 Clustering

Clustering can help detect similar patients or diseases. Since the raw healthcare data is not clean, there are usually two kinds of approaches for researchers to derive meaningful clusters. The first approach tends to learn robust latent representations first, followed by clustering methods while the other approach adopts probabilistic clustering models which can deal with raw healthcare data effectively. In [105], diseases are first embedded into 200-dimension using a modified RBM model, eNRBM model. These latent 200-dimension hidden vectors are then projected into 2D space using t-SNE. In this 2D space, we can see several meaningful groups consisting of related diagnoses and procedures. Similar to [105], [79] embeds raw patient vectors into latent vectors using a modified RBM, then patient clustering is performed on these latent vectors. Experiments show some groups of patients are closely related to a specific disease condition (say Type I diabetes). [98] identifies multivariate patterns of perceptions using cluster analysis. Five different patient clusters are finally identified and statistically significant inter-cluster differences are found. In [72], a probabilistic clustering model is applied for multi-dimensional, sparse physiological time series data. It shows that different clusters of patients have large differences in mortality rates. Moreover, this clustering model can be used to construct high-quality predictive models. Similarly, in [94], a probabilistic sub-typing model is proposed to cluster time series of clinical markers in order to identify homogeneous patient subgroups.

4.1.2 Phenotyping

Computational phenotyping has become a hot topic recently and has attracted the attention of a large number of researchers because it can help learn robust representations from sparse, high-dimensional, noisy raw EMR data. It has several kinds of forms including (i) rules/algorithms

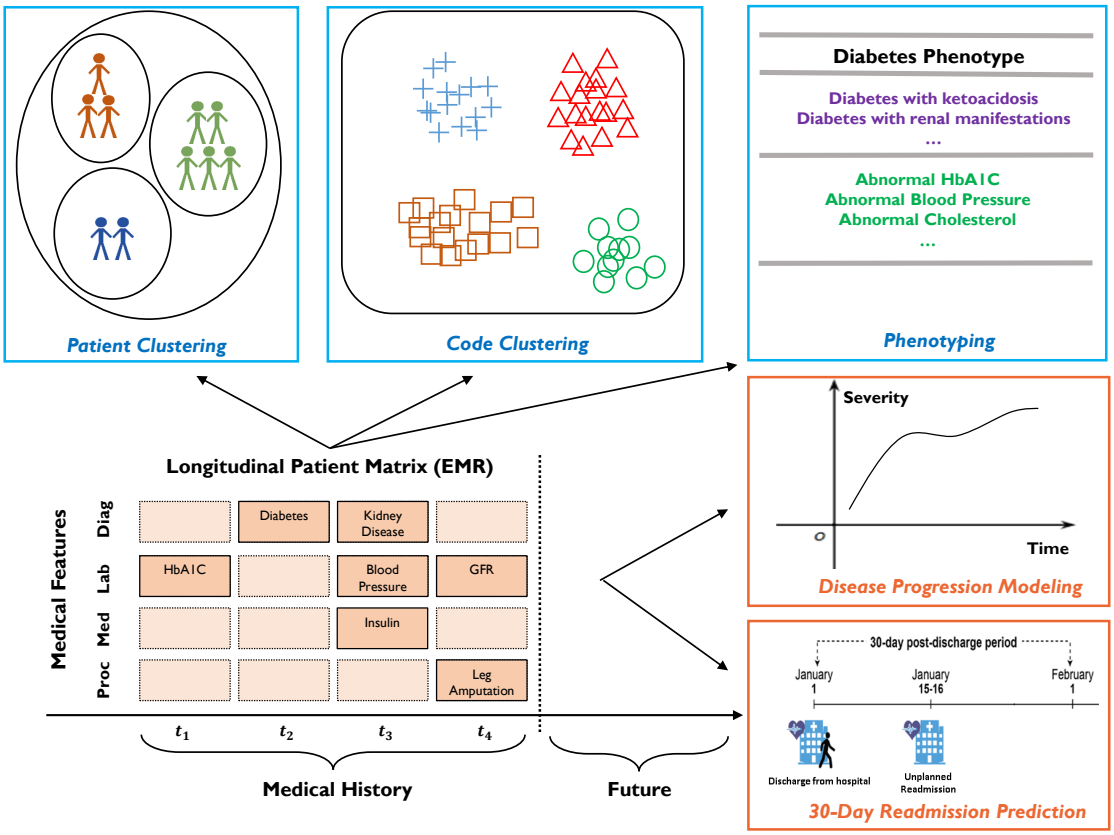

Figure 6: An illustration of some applications using EMR data analytics. From medical history, we can perform patient clustering, code clustering and phenotyping tasks, while regarding prediction of patients' future, we can do disease progression modeling and 30-day readmission prediction [53].

that define diagnostic inclusion criteria (ii) latent factors or latent bases for medical features [49].

Traditionally, doctors regard phenotyping as rules that define diagnostic or inclusion criteria. The task of finding phenotypes is achieved by a supervised task [73]. A number of features are first chosen by domain experts, then statistical methods such as logistic regression or chi-square test are performed to identify the significant features for developing acute kidney injury during hospital admissions. PheKB[§] is a phenotype knowledge base that shows many rules for different diseases and medical conditions. Traditional methods using statistical models are easier to be implemented and interpreted, but they may require a large amount of human intervention.

Recently, machine learning researchers are working on high-throughput methods to derive more meaningful phenotypes. These works mainly discover latent factors or bases as phenotypes. [39] first constructs a three-dimensional tensor which includes patients, diagnoses as

[§]https://phekb.org/

well as procedures to represent the raw input data. Then this tensor is split into several interaction tensors and a bias tensor. Each interaction tensor is a phenotype and the non-zero features in each tensor can be regarded as the features of the corresponding phenotype. [111] is similar to [39], and it represents phenotypes in the form of interaction tensors. However, different from [39], it emphasizes on imposing knowledge into the learned phenotypes and proposes to derive distinct phenotypes. In [118], raw patient data is represented using a two-dimensional longitudinal patient matrix with one axis being time and the other being medical features. Then the algorithm decomposes this longitudinal patient matrix into a latent medical concept mapping matrix and a concept evolution matrix. Phenotypes can then be obtained from the latent medical concept mapping matrix by discovering feature groups inside the matrix. Different from traditional statistical methods, phenotyping algorithms based on high-throughput machine learning methods can generate a number of phenotypes at the same time. Moreover, some of the unsupervised algorithms can derive phenotypes which are independent of prediction tasks and are more general.

Deep learning achieves record-breaking performance in a number of image and speech recognition tasks for its distinguished ability to detect complex non-linear relations from raw data and the ability to learn robust high-level abstractions [8, 50]. Since body system itself is complex and highly non-linear, it may be potential for us to utilize deep learning methods to perform phenotyping tasks. [57] is an early work that applies deep learning models in computational phenotyping. It first applies Gaussian process regression to transform the uric acid sequence to a probability density function. Then an auto-encoder is used to learn the hidden representations of Gaussian distribution's mean vectors. The learned weights of the auto-encoder are regarded as phenotypes and the learned features are also visualized. Similar to [57], [105] utilizes a simple two-layer unsupervised deep learning model, RBM, to learn hidden representations of patients' raw input vectors (aggregated counts of medical features, such as diagnoses, procedures). Each unit of this RBM's hidden layer is regarded as a phenotype and this hidden vector is then used for clustering and classification tasks. Different from [57] and [105] which employ an unsupervised model, [16] utilizes a supervised MLP model to extract phenotypes from ICU time-series data. In order to visualize MLP's ability to disentangle factors of variation, the authors apply tools from causal inference to analyze the learned phenotypes quantitatively and qualitatively.

4.1.3 Disease Progression Modelling

Disease progression modelling (DPM) is to employ computational methods to model the progression of a specific disease [76]. With the help of DPM, we can detect a certain disease early

and therefore, manage the disease better. For chronic diseases, using DPM can effectively delay patients' deterioration and improve patients' healthcare outcomes. Therefore, we can provide helpful reference information to doctors for their judgment and benefit patients in the long run.

4.1.3.1 Statistical Regression Methods

Traditionally, many related works employ statistical regression methods for DPM, which can model the correlation between patients' medical features and patients' condition indicators [27, 95]. Then, via such correlation, we can have access to the progression of patients with patients' features. For example, in [95] , an accurate risk score model through a multivariate Cox regression model is proposed for predicting patients' probability of developing diabetes within 5 years. Similarly, in [27], a robust linear regression model is employed to predict clinically probable Alzheimer's disease patients' MMSE changes in one year.

Another line of research focuses on "survival analysis", which is to link patients' disease progression to the time before a certain outcome. The linking is accomplished via a survival function. For instance, in [85], a disease progression model is proposed to predict liver transplant patients' long-term survival. The objective is to stratify patients into clusters according to their survival characteristics and then assign different intervention strategies to different patient clusters. Similarly, in [107], the time of patients' progression from amnestic mild cognitive impairment to Alzheimer's disease is studied.

While statistical regression methods have shown to be efficient due to their simple models and computation, we should note that this is accomplished with an underlying assumption that the progression (i.e. medical time-series data) of a disease follows a certain distribution. However, for real-life applications, this assumption may not be true, and the performance of statistical regression methods would suffer. Therefore, it could be difficult to generalize such methods to most clinical applications where the disease progression cannot be abstracted by a certain simple distribution.

4.1.3.2 Machine Learning Methods

Existing works which employ machine learning methods to solve DPM problem are quite various, from graphical models including Markov models [110, 44], to multi-task learning methods [119, 117, 80] and to artificial neural networks [101].

In [110], a Markov jump process is employed to model COPD patients' transition behaviour between disease stages. In [44], a multi-state Markov model is proposed for predicting the progression between different stages for abdominal aortic aneurysm patients considering the

probability of misclassification at the same time. Due to the structure as directed graphs, these methods have the advantages of good causality and interpretability. However, medical experts need to be involved to determine the causal relationships during model construction.

Another category of methods is to employ multi-task learning. In [119, 117], the DPM problem is formalized in the multi-task learning setting as predicting patients' future severity in multiple timepoints and select informative features of progression. Also with a multi-task learning method, in [80], the consistency between multiple modalities is considered in the objective function and missing data problem is handled. The limitations of multi-task learning methods are two-fold. First, they only make use of medical features corresponding to patients' first visits to the hospital instead of time-related features. Second, they can only deal with linear relationships in the model.

4.1.3.3 Deep Learning Methods

In [101], an artificial neural network is employed to predict the recurrence of breast cancer after surgery. With a deeper neural network than this, deep learning models become more widely applicable with its great power in representation and abstraction due to its non-linear activation functions inside. For instance, in [86], a variant of LSTM is employed to model the progression of both diabetes cohort and mental health cohort. They use "Precision at K" as the metric to evaluate the performance of models. However, the lack of interpretability is a possible limitation of these deep learning methods. Furthermore, more training data is of vital significance in order to improve deep learning models' performance.

4.1.4 Image Data Analysis

MRI is widely used to form images of the body using strong magnetic fields, radio waves, and field gradients. Analyzing these images is beneficial for many medical diagnoses and a wide range of studies focus on MRI image data classification or segmentation tasks. In [52], a novel classification method that combines both fractal and GLCM features is proven to be more effective for MRI and CT Scan Medical image classification than previous models which only utilize GLCM features. A model that combines deep learning algorithms and deformable models is developed in [3] for fully automatic segmentation of the left ventricle from cardiac MRI datasets. Experiments show that by incorporating deformable models, this method can achieve better accuracy and robustness of the segmentation. In [4], a review of recent methods for brain MRI image segmentation is presented.

4.2 Applications for Sensor data

4.2.1 Mobile Healthcare

Healthcare for ageing population has become a major focus, especially in developed countries. Due to the shortage of clinical manpower, there has been a drive toward using ICT (information and communication technology), called mobile healthcare or mHealth[¶]. With the advanced technologies including machine learning and high-performance computing, personalized healthcare services will be provided remotely, and diagnoses, medications and treatments will be fine-tuned for individuals on the basis of spatio-temporal and/or psycho-physiological conditions.

Human activity recognition (HAR) is one of the key technologies for mHealth. HAR research is mainly classified into two groups in terms of approaches, namely the video-based and the wearable device-based. The video-based approach continuously tracks human activities through cameras deployed in rooms; however, it raises privacy issues and requires the targeted person to remain within the vicinity of the camera [56]. Moreover, the feature extraction from the captured video/images requires complex computations for further analytics [28]. Because of these limitations, there has been a shift towards the use of wearable sensors requiring less data processing.

Nowadays, the activity recognition is implemented on smart devices for online processing [13, 99, 104] while it is done offline using machine learning tools in backend machines or servers. It has enabled smart healthcare applications such as fitness assessment [64], life logging[‖], and rehabilitation [74] where the user activities can be tracked anytime and anywhere.

From the data analytics perspective, [33] discusses the feature extraction algorithm for HAR using only a single tri-axial accelerometer. Relevant and robust features are successfully selected and the data size is reduced; thereby, the processing speed increases without degrading accuracy.

Retrieved features correspond to activities specify patterns, and the patterns are used for classification or modelling. Sliding window methods are typically used for static or periodic activities while sporadic activities can be recognized using template matching approaches [71] or Hidden Markov Modelling (HMM) [84, 12]. In [83], a deep learning model is designed using convolutional neural networks and LSTM recurrent neural networks, which captures spatio-temporal patterns of signals from wearable accelerometers and gyroscopes.

[¶]www.mobilehealthsummit.ca

[‖]http://www.sonymobile.com/global-en/apps-services/lifelog

4.2.2 Environment Monitoring

Another interesting healthcare application integrates chemical sensors [6, 70, 66] for detecting the presence of specific molecules in the environment. For example, we can collect Pollutant Standards Index (PSI) data that reflects six pollutants (e.g., sulfur dioxide (SO2), particulate matter (PM10) and fine particulate matter (PM2.5), nitrogen dioxide (NO2), carbon monoxide (CO) and ozone (O3)), from individual users and construct a fine-grained pollution map together with images and location information. The environmental monitoring for haze, sewage water and smog emission etc. has become a significant worldwide problem. Combined with the cloud computing technology, a large number of smart mobile devices make a distributed data collection infrastructure possible, and the recent scalable, parallel, resource efficient, real-time data mining technologies have enabled smart device-based data analysis [120, 65].

[77] proposes the Personal Environmental Impact Report (PEIR) system that uses location information sampled from smartphones and calculates personalized estimates of environmental impact and exposure. The running PEIR system, which runs GPS data collection at mobile devices and the HMM-based activity classification at servers before computing the PEI values, is evaluated. A big contribution of their work is that this platform can be used for various targets such as traffic condition measuring, environmental pollution monitoring, and vehicle emission estimating.

4.2.3 Disease Detection

Biochemical-sensors deployed in/on the body can detect particular volatile organic compounds (VOCs). Many studies [92, 23, 11] have unveiled the relationships between VOCs and particular diseases responding to VOCs, as summarized in Table 1. The big potential of such sensor devices and the big data analytics of VOCs will revolutionize healthcare both at home and in hospital.

Developments of nano-sensor arrays and micro electro mechanical systems have enabled artificial olfactory sensors, called electronic noses [30, 66], as tiny, energy efficient, portable devices. [30] discusses the essential cause of obesity from over-eating and an intake of high-calorie food, and presents the way to compute energy expenditure from exhaled breath.

In [51], nano-enabling electrochemical sensing technology is introduced, which rapidly detects beta-amyloid peptides, potential bio-markers to diagnose Alzheimer's disease, and a tool is developed to facilitate fast personalized healthcare for AD monitoring.

Table 1: List of volatile organic compounds related to particular diseases

Volatile organic compound	Relevant disease
acetoin, 1-butanol	Lung cancer
aceton	Diabetes
etan, pentan	Asthma
ammonia	Hepatic encephalopathy
hydrogen, metan	Maldigestion syndrome
toluen	Thinner addiction
trimethylamine	Renal failure

4.3 Healthcare Systems

Instead of solving individual problems, a number of healthcare systems have been designed and built to serve as platforms for solving the problems described above. Now we shall discuss several representative healthcare systems.

HARVEST [38] is a summarizer for doctors to view patients' longitudinal EMR data at the point of care. It is composed of two key parts: a front-end for better visualization; a distributed back-end which can process patients' various types of EMR data and extract informative problem concepts from patients' free text data measuring each concept via "salience weights".

miniTUBA [113] is designed to assist clinical researchers to employ dynamic Bayesian networks (DBN) for data analytics in temporal datasets. The pipeline of miniTUBA includes logging in the website, inputting data as well as managing project, constructing DBN models, analyzing results and doing prediction in the end. Users can use miniTUBA to discover informative causal relationships for better inference or prediction.

In [43], a system which focuses on data-driven analytics for personalized healthcare is proposed. The applications supported in this system include analyzing patient similarity, constructing predictive models, stratifying patients, analyzing cohorts and modelling diseases. The target is to achieve personalized healthcare resource utilization and deliver care services at low costs. Cohort analysis has a wide range of healthcare applications, such as testing the hypothesis of a new treatment, seeing how similar patients in a hospital database are doing compared with the specific indexed patient, etc.

To provide better support for Big Healthcare Analytics, we have been implementing various software systems that form an end-to-end pipeline from data acquisition and cleansing to visualization. We call the system GEMINI [61], whose software stack is depicted in Figure 7. We are addressing various healthcare analytics problems, such as phentotyping, disease progressing

modeling, treatment recommendation etc. We shall introduce each component of our software stack via the example process of doing EMR data analytics in the following.

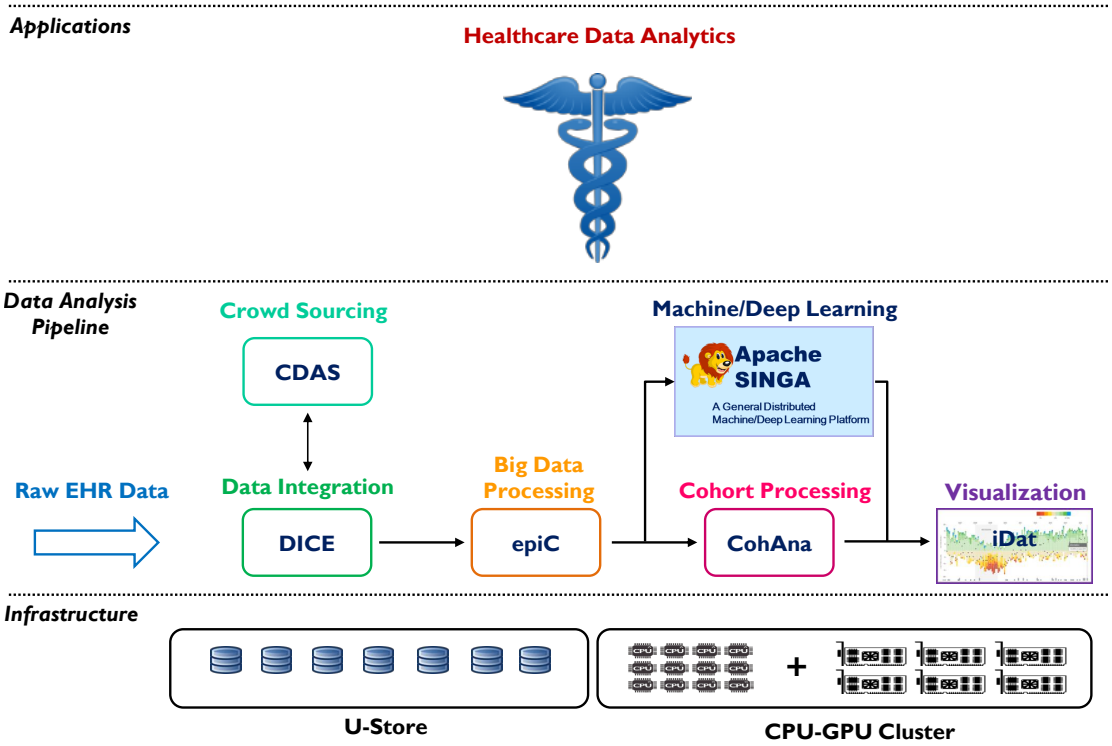

Figure 7: GEMINI healthcare software stack.

We work on the longitudinal EMR dataset from the National University Hospital. We encounter the various challenges as discussed in Section 2; hence, we need to process the data, through data cleansing and data integration, before we can conduct any data analytics. Even though we want to automate the processing and relieve the burden on doctors, EMR data cannot be cleaned and integrated effectively without doctors' assistance. Hence, we leverage automatic methods and doctors' participation with their expertise domain knowledge. DICE is our general data cleansing and integration platform that exploits both doctors' expertise and knowledge base. Additionally, to assist in the data cleansing and integration, CDAS [63] which is a crowdsourcing system, selects meaningful representative tasks for the clinicians to resolve so as to reduce the overall efforts and costs. Ultimately, we tap onto the clinicians who are the subject matter experts for their knowledge, without over imposing on their time, to improve the quality of the data and the analytics process [61, 81].

Due to the value of the data and the need to maintain it for a long period of time, we have designed and implemented UStore, which is a universal immutable storage system, to store the data. We process the data in epiC [47], which is a distributed and scalable data processing

system based on Actor-like concurrent programming model. By separating computation and communication, this system can process different kinds of data (structured data, unstructured data and graph data) effectively, and also supports different computation models. However, epiC provides only database-centric processing and analytics such as aggregation and summarization. In order to provide deep analytics, we have implemented a generic distributed machine learning/deep learning platform, called Apache SINGA [82, 109]. We are implementing our deep learning models on top of Apache SINGA for analytics on various diseases.

For behavioural analysis of patients, we employ "cohort analysis" which was originally introduced in social science [31]. For our applications, we have built CohAna [46], a column-based cohort analysis engine with an extended relation for modelling user activity data and a few new operators to support efficient cohort query processing.

To enable the clinicians to visualize the data and analytics results, we have developed iDAT, an exploratory front-end tool that allows interactive drill down and exploration.

5 Summary and Discussions

In this chapter, we summarize the challenges of Big Healthcare Analytics and their solutions to relevant applications from both EMR data and sensor data. The challenges mainly consist of high-dimensionality, irregularity, missing data, sparsity, noise and bias. Besides the basic model construction for analytics, we discuss four necessary steps for data processing, namely data annotation, data cleansing, data integration and data analytics/modelling. Based on an examination of various types of healthcare analytics on both EMR data and sensor data, the data analytics pipeline is still the foundation for most healthcare applications. However, specific algorithms which are adopted must be adjusted by modelling the unique characteristics of medical features. With recent advancement in hardware and other technologies, smart healthcare analytics is gaining traction, and like other application domains, we are likely to experience a sharp leap in healthcare technologies and systems in the near future.

Next-generation healthcare systems are expected to integrate various types of EHR data and provide a holistic data-driven approach to predict and pre-empt illnesses, improve patient care and treatment, and ease the burden of clinicians by providing timely and assistive recommendations. Below, we discuss several applications that are likely to attract attention and interest in the near future.

Treatment recommendation system for doctors: Through various levels of automation in diagnosis model and prognosis prediction, the system may improve the medical treatment process

in different degrees from helping doctors to make decisions (e.g. visualize cohort information) to outperforming doctors in treatment planning and recommendation.

Treatment explanation system for patients: Doctors may not have sufficient time to explain treatment plans to the patients in detail or may not be able to express clearly to the patients. An automatic treatment explanation system may be able to improve patient treatment compliance as well as the transparency of healthcare. Further, patients can review the plans anytime anywhere.

Real-time surgical operation suggestion: Lots of emergency situations may happen during surgical operations. Armed with real-time sensors and reinforcement learning models, the machine may be able to deliver a better contingency plan in a much shorter time and with more accurate decision making.

Data-driven drug combination study: Drug combination discoveries used to be considered as a hard problem due to the insufficiency of clinical data. An integrated system with better EHR data analytics may be able to quantify the effect of drug combinations and also discover more valuable drug combination patterns. This study can be very useful for personalized medicine recommendations, which can further help to provide a more effective healthcare.

We are looking forward to more clinical advances and healthcare products being brought to the table by both the data science and medical communities. After all, with more data and higher computational capacity, deep analytics can lead to deeper insights and hence better decisions.

6 Acknowledgments

This work is supported by National Research Foundation, Prime Ministers Office, Singapore under its Competitive Research Programme (CRP Award No. NRF-CRP8-2011-08). Gang Chen's work is supported by National Natural Science Foundation of China (NSFC) Grant No. 61472348. Meihui Zhang is supported by SUTD Start-up Research Grant under Project No. SRG ISTD 2014 084. We would like to thank Jinyang Gao and Gerald Koh for the discussion and useful suggestions that help to improve the chapter.

Bibliography

[1] Apache storm. `http://storm.apache.org`.

[2] H. Alemdar and C. Ersoy. Wireless sensor networks for healthcare: A survey. *Computer Networks*, 54(15):2688–2710, 2010.

[3] M. R. Avendi, A. Kheradvar, and H. Jafarkhani. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri. *Medical image analysis*, 30:108–119, 2016.

[4] M. A. Balafar, A. R. Ramli, M. I. Saripan, and S. Mashohor. Review of brain MRI image segmentation methods. *Artificial Intelligence Review*, 33(3):261–274, 2010.

[5] H. Banaee, M. U. Ahmed, and A. Lout. Data mining for wearable sensors in health monitoring systems: A review of recent trends and challenges. *Sensors*, 13(12), 2013.

[6] A. J. Bandodkar, I. Jeerapan, and J. Wang. Wearable chemical sensors: Present challenges and future prospects. *ACS Sensors*, 1:464–482, 2016.

[7] I. M. Baytas, K. Lin, F. Wang, et al. Stochastic convex sparse principal component analysis. *EURASIP Journal on Bioinformatics and Systems Biology*, 2016(1):1–11, 2016.

[8] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[9] J. Bian, B. Gao, and T.-Y. Liu. Knowledge-powered deep learning for word embedding. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 132–148, 2014.

[10] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng. Secondary use of ehr: data quality issues and informatics opportunities. *AMIA Summits Transl Sci Proc*, 2010:1–5, 2010.

[11] Y. Y. Broza and H. Haick. Nanomaterial-based sensors for detection of disease by volatile organic compounds. *Nanomedicine (Lond)*, 8(5):785–806, 2013.

[12] A. Bulling, U. Blanke, and B. Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Survey*, 46(3):1–33, 2014.

[13] N. A. Capela, E. D. Lemaire, N. Baddour, et al. Evaluation of a smartphone human activity recognition application with able-bodied and stroke participants. *NeuroEngineering and Rehabilitation*, 13(5), 2016.

[14] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[15] R. D. Caytiles and S. Park. A study of the design of wireless medical sensor netork based u-healthcare system. *International Journal of Bio-Science and Bio-Technology*, 6(3):91–96, 2014.

[16] Z. Che, D. C. Kale, W. Li, et al. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516, 2015.

[17] Z. Che, S. Purushotham, K. Cho, et al. Recurrent neural networks for multivariate time series with missing values. *arXiv preprint arXiv:1606.01865*, 2016.

[18] Z. Che, S. Purushotham, R. Khemani, et al. Distilling knowledge from deep networks with applications to healthcare domain. *arXiv preprint arXiv:1512.03542*, 2015.

[19] K. Cho, B. Van Merriënboer, C. Gulcehre, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[20] D. A. Cohn. Neural network exploration using optimal experiment design. In *NIPS*, 1994.

[21] R. Cort, X. Bonnaire, O. Marin, et al. Stream processing of healthcare sensor data: studying user traces to identify challenges from a big data perspective. In *Proceedings of the 4th International Workshop on Body Area Sensor Networks*, 2015.

[22] B. Cui, H. Mei, and B. C. Ooi. Big data: the driver for innovation in databases. *National Science Review*, 1(1):27–30, 2014.

[23] A. G. Dent, T. G. Sutedja, and P. V. Zimmerman. Exhaled breath analysis for lung cancer. *Journal of thoracic disease*, 5:S540, 2013.

[24] A. Doan, A. Halevy, and Z. Ives. *Principles of data integration*. Elsevier, 2012.

[25] X. L. Dong and D. Srivastava. Big data integration. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 1245–1248, 2013.

[26] O. M. Doyle, E. Westman, A. F. Marquand, et al. Predicting progression of alzheimers disease using ordinal regression. *PloS one*, 9(8):e105542, 2014.

[27] S. Duchesne, A. Caroli, C. Geroldi, et al. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. *Neuroimage*, 47(4):1363–1370, 2009.

[28] A. S. Evani, B. Sreenivasan, J. S. Sudesh, et al. Activity recognition using wearable sensors for healthcare. In *Proceedings of the 7th International Conference on Sensor Technologies and Appplications*, 2013.

[29] L. Filipe, F. Fdez-Riverola, N. Costa, et al. Wireless body area networks for healthcare applications: Protocol stack review. *International Journal of Distributed Sensor Networks*, 2015:1:1–1:1, 2015.

[30] J. W. Gardner and T. A. Vincent. Electronic noses for well-being: Breath analysis and energy expenditure. *Sensors*, 16(7):947, 2016.

[31] N. D. Glenn. *Cohort analysis*. Sage, 2005.

[32] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér. Data integration in the era of omics: current and future challenges. *BMC Systems Biology*, 8(2):1–10, 2014.

[33] P. Gupta and T. Dallas. Feature selection and activity recognition system using a single triaxial accelerometer. *IEEE Transactions on Biomedical Engineering*, 61(6):1780–1786, 2014.

[34] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3:1157–1182, 2003.

[35] M. Haghighi, P. Woznowski, N. Zhu, et al. Agent-based decentralised data-acquisition and time-synchronisation in critical healthcare applications. In *Proceedings of the IEEE 2nd World Forum on Inernet of Things*, 2015.

[36] A. Halevy, A. Rajaraman, and J. Ordille. Data integration: The teenage years. In *Proceedings of the 32Nd International Conference on Very Large Data Bases*, pages 9–16, 2006.

[37] W. R. Hersh, M. G. Weiner, P. J. Embi, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical care*, 51:S30–S37, 2013.

[38] J. S. Hirsch, J. S. Tanenbaum, S. Lipsky Gorman, et al. Harvest, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274, 2014.

[39] J. C. Ho, J. Ghosh, and J. Sun. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 115–124, 2014.

[40] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[41] G. Hripcsak and D. J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

[42] G. Hripcsak, D. J. Albers, and A. Perotte. Parameterizing time in electronic health record studies. *Journal of the American Medical Informatics Association*, 22(4):794–804, 2015.

[43] J. Hu, A. Perer, and F. Wang. Data driven analytics for personalized healthcare. In *Healthcare Information Management Systems*, pages 529–554. Springer, 2016.

[44] C. H. Jackson, L. D. Sharples, S. G. Thompson, et al. Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(2):193–209, 2003.

[45] H. Jagadish. Challenges and opportunities with big data, 2012.

[46] D. Jiang, Q. Cai, G. Chen, et al. Cohort query processing. *Proceedings of the VLDB Endowment*, 10(1), 2017.

[47] D. Jiang, G. Chen, B. C. Ooi, et al. epic: an extensible and scalable system for processing big data. *Proceedings of the VLDB Endowment*, 7(7):541–552, 2014.

[48] K. Kalantar-Zadeh, C. K. Yao, K. J. Berean, et al. Intestinal gas capsules: A proof-of-concept demonstration. *Gastroenterology*, 150(1):37–39, 2016.

[49] D. C. Kale, Z. Che, M. T. Bahadori, et al. Causal phenotype discovery via deep networks. In *AMIA Annual Symposium Proceedings*, pages 677–686, 2015.

[50] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2014.

[51] A. Kaushik, R. D. Jayant, S. Tiwari, et al. Nano-biosensors to detect beta-amyloid for alzheimer's disease management. *Biosensors and Bioelectronics*, 80(15):273–287, 2016.

[52] R. Korchiyne, S. M. Farssi, A. Sbihi, R. Touahni, and M. T. Alaoui. A combined method of fractal and GLCM features for MRI and CT scan images classification. *arXiv preprint arXiv:1409.4559*, 2014.

[53] H. Krumholz, S.-L. Normand, P. Keenan, et al. 30-day heart failure readmission measure methodology. Technical report, Yale University/Yale-New Haven Hospital Center for Outcomes Research And Evaluation (YNHH-CORE), 2008.

[54] Z. Kuang, J. Thomson, M. Caldwell, et al. Computational drug repositioning using continuous self-controlled case series. *arXiv preprint arXiv:1604.05976*, 2016.

[55] S. Kumar, M. Willander, J. G. Sharma, et al. A solution processed carbon nanotube modified conducting paper sensor for cancer detection. *Journal of Materials Chemistry B*, 3:9305–9314, 2015.

[56] O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 15(3):1192–1209, 2013.

[57] T. A. Lasko, J. C. Denny, and M. A. Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):1–13, 2013.

[58] M. Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 233–246. ACM, 2002.

[59] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and*

Development in Information Retrieval, SIGIR '94, pages 3–12, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[60] Q. Lin, B. C. Ooi, Z. Wang, et al. Scalable distributed stream join processing. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 811–825, 2015.

[61] Z. J. Ling, Q. T. Tran, J. Fan, et al. GEMINI: An integrative healthcare analytics system. *Proceedings of the VLDB Endowment*, 7(13):1766–1771, 2014.

[62] Z. C. Lipton, D. C. Kale, C. Elkan, et al. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[63] X. Liu, M. Lu, B. C. Ooi, et al. CDAS: a crowdsourcing data analytics system. *Proceedings of the VLDB Endowment*, 5(10):1040–1051, 2012.

[64] J. W. Lockhart, T. Pulickal, and G. M. Weiss. Applications of mobile activity recognition. In *ACM Conference on Ubiquitous Computing*, pages 1054–1058, 2012.

[65] J. W. Lockhart, G. M. Weiss, J. C. Xue, et al. Design considerations for the wisdm smart phone-based sensor mining architecture. In *Proceedings of the 5th International Workshop on Knowledge Discovery from Sensor Data*, pages 25–33, 2011.

[66] P. Lorwongtragool, E. Sowade, N. Watthanawisuth, et al. A novel wearable electronic nose for healthcare based on flexible printed chemical sensor array. *Sensors*, 14(10):19700, 2014.

[67] V. Loscrí, L. Matekovits, I. Peter, et al. In-body network biomedical applications: From modeling to experimentation. *IEEE Transactions on Nanobioscience*, 15(1):53–61, 2016.

[68] L. L. Low, K. H. Lee, M. E. Hock Ong, et al. Predicting 30-day readmissions: performance of the lace index compared with a regression model among general medicine patients in singapore. *BioMed research international*, 2015, 2015.

[69] D. Malak and O. B. Akan. Molecular communication nanonetworks inside human body. *Nano Communication Networks*, 3(1):19–35, 2012.

[70] C. Manjarrs, D. Garizado, M. Obregon, et al. Chemical sensor network for ph monitoring. *Journal of Applied Research and Technology*, 14(1):1–8, 2016.

[71] J. Margarito, R. Helaoui, A. M. Bianchi, et al. User-independent recognition of sports activities from a single wrist-worn accelerometer: A template-matching-based approach. *IEEE Transactions on Biomedical Engineering*, 63(4):788–796, 2016.

[72] B. M. Marlin, D. C. Kale, R. G. Khemani, et al. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398, 2012.

[73] M. E. Matheny, R. A. Miller, T. A. Ikizler, et al. Development of inpatient risk stratification models of acute kidney injury for use in electronic health records. *Medical Decision Making*, 30(6):639–650, 2010.

[74] A. McLeod, E. M. Bochniewicz, P. S. Lum, et al. Using wearable sensors and machine learning models to separate functional upper extremity use from walking-associated arm movements. *Physical Medicine and Rehabilitation.*, 97(2):224–231, 2016.

[75] N. Q. Mehmood, R. Culmone, and L. Mostarda. A flexible and scalable architecture for real-time ANT+ sensor data acquisition and nosql storage. *International Journal of Distributed Sensor Networks*, 12(5), 2016.

[76] D. Mould. Models for disease progression: new approaches and uses. *Clinical Pharmacology & Therapeutics*, 92(1):125–131, 2012.

[77] M. Mun, S. Reddy, K. Shilton, et al. Peir, the personal environmental impact report, as a platform for participatory sensing systems research. In *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*, pages 55–68, 2009.

[78] I. Muslea, S. Minton, and C. A. Knoblock. Selective sampling with redundant views. In *AAAI/IAAI*, pages 621–626, 2000.

[79] T. D. Nguyen, T. Tran, D. Phung, et al. Latent patient profile modelling and applications with mixed-variate restricted boltzmann machine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 123–135, 2013.

[80] L. Nie, L. Zhang, Y. Yang, et al. Beyond doctors: Future health prediction from multimedia and multimodal observations. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 591–600, 2015.

[81] B. C. Ooi, K. L. Tan, Q. T. Tran, et al. Contextual crowd intelligence. *ACM SIGKDD Explorations Newsletter*, 16(1):39–46, 2014.

[82] B. C. Ooi, K.-L. Tan, S. Wang, et al. SINGA: A distributed deep learning platform. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 685–688, 2015.

[83] F. J. Ordez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors (Basel, Switzerland)*, 16(1):115, 2016.

[84] F. J. Ordonez, G. Englebienne, P. de Toledo, et al. In-home activity recognition: Bayesian inference for hidden markov models. *IEEE Pervasive Computing*, 13(3):67–75, 2014.

[85] R. K. Pearson, R. J. Kingan, and A. Hochberg. Disease progression modeling from historical clinical databases. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 788–793, 2005.

[86] T. Pham, T. Tran, D. Phung, et al. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 30–41, 2016.

[87] R. Pivovarov, D. J. Albers, J. L. Sepulveda, et al. Identifying and mitigating biases in ehr laboratory tests. *Journal of biomedical informatics*, 51:24–34, 2014.

[88] R. Pivovarov, A. J. Perotte, E. Grave, et al. Learning probabilistic phenotypes from heterogeneous ehr data. *Journal of biomedical informatics*, 58:156–165, 2015.

[89] S. R. and C. L. Stress detection using physiological sensors. *IEEE Computer*, 48(10):26–33, 2015.

[90] N. Roy and A. McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

[91] M. Salai, I. Vassnyi, and I. Ksa. Stress detection using low cost heart rate sensors. *Journal of Healthcare Engineering*, 2, 2016.

[92] Y. Sasaya and T. Nakamoto. Study of halitosis-substance sensing at low concentration using an electrochemical sensor array combined with a preconcentrator. *IEEE Journal of Transactions on Sensors and Micromachines*, 126, 2006.

[93] J. L. Schafer and J. W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[94] P. Schulam, F. Wigley, and S. Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2956–2964, 2015.

[95] M. B. Schulze, K. Hoffmann, H. Boeing, et al. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. *Diabetes care*, 30(3):510–515, 2007.

[96] B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1070–1079, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[97] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 287–294, New York, NY, USA, 1992. ACM.

[98] M. J. Sewitch, K. Leffondré, and P. L. Dobkin. Clustering patients according to health perceptions: relationships to psychosocial characteristics and medication nonadherence. *Journal of psychosomatic research*, 56(3):323–332, 2004.

[99] M. Shoaib, S. Bosch, O. D. Incel, et al. A survey of online activity recognition using mobile phones. *Sensors*, 15(1):2059–2085, 2015.

[100] C. M. Stonnington, C. Chu, S. Klöppel, et al. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *Neuroimage*, 51(4):1405–1413, 2010.

[101] N. Street. A neural network model for prognostic prediction. In *Proceedings of the 15th International Conference on Machine Learning*, pages 540–546, 1998.

[102] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, Mar. 2002.

[103] S. N. Topkaya and D. Ozkan-Ariksoysal. Prostate cancer biomarker detection with carbon nanotubes modified screen printed electrodes. *Electroanalysis*, 28(5), 2016.

[104] C. Torres-Huitzil and A. Alvarez-Landero. *Accelerometer-Based Human Activity Recognition in Smartphones for Healthcare Services*, pages 147–169. Springer, 2015.

[105] T. Tran, T. D. Nguyen, D. Phung, et al. Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm). *Journal of biomedical informatics*, pages 96–105, 2015.

[106] C. van Walraven, I. A. Dhalla, C. Bell, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal*, 182(6):551–557, 2010.

[107] P. Vemuri, H. Wiste, S. Weigand, et al. MRI and CSF biomarkers in normal, MCI, and AD subjects predicting future clinical change. *Neurology*, 73(4):294–301, 2009.

[108] F. Wang, N. Lee, J. Hu, et al. Towards heterogeneous temporal clinical event pattern discovery: a convolutional approach. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 453–461, 2012.

[109] W. Wang, G. Chen, A. T. T. Dinh, et al. SINGA: Putting deep learning in the hands of multimedia users. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 25–34, 2015.

[110] X. Wang, D. Sontag, and F. Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94, 2014.

[111] Y. Wang, R. Chen, J. Ghosh, et al. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1265–1274, 2015.

[112] B. J. Wells, A. S. Nowacki, K. Chagin, et al. Strategies for handling missing data in electronic health record derived data. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 1(3):7, 2013.

[113] Z. Xiang, R. M. Minter, X. Bi, et al. minituba: medical inference by network integration of temporal data using bayesian analysis. *Bioinformatics*, 23(18):2423–2432, 2007.

[114] T. Yokota, P. Zalar, M. Kaltenbrunner, et al. Ultraflexible organic photonic skin. *Science Advances Online Edition*, 2(4), 2016.

[115] H. Zhang, G. Chen, B. C. Ooi, et al. In-memory big data management and processing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 27(7):1920–1948, 2015.

[116] X. Zhang, B. Hu, L. Zhou, et al. An eeg based pervasive depression detection for females. In *Proceedings of the 2012 International Conference on Pervasive Computing and the Networked World*, pages 848–861, 2013.

[117] J. Zhou, J. Liu, V. A. Narayan, et al. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1095–1103, 2012.

[118] J. Zhou, F. Wang, J. Hu, et al. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 135–144, 2014.

[119] J. Zhou, L. Yuan, J. Liu, et al. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 814–822, 2011.

[120] T. Zhu, S. Xiao, Q. Zhang, et al. Emergent technologies in big data sensing: A survey. *International Journal of Distributed Sensor Networks*, 2015(8):1–13, 2015.