International Journal of Studies in Nursing; Vol. 1, No. 1; 2016
ISSN 2424-9653
Published by July Press

Big Data Analytics for Medication Management in Diabetes Mellitus

Lidong Wang[1] & Cheryl Ann Alexander[2]

[1] Department of Engineering Technology, Mississippi Valley State University, Itta Bena, USA

[2] Technology and Healthcare Solutions, Itta Bena, USA

Correspondence: Lidong Wang, Ph.D., Associate Professor, Department of Engineering Technology, Mississippi Valley State University, Itta Bena, USA.

Received: November 20, 2016 Accepted: December 5, 2016 Online Published: December 18, 2016

doi:10.20849/ijsn.v1i1.99 URL: http://dx.doi.org/10.20849/ijsn.v1i1.99

Abstract

Medication development plays a prominent role in the fight against chronic illness such as hypertension, diabetes mellitus, asthma, etc. Without proper testing and methods for management of drug data, the disease management would fail. Providers rely on pharmaceutical companies to provide research data in widespread formats and pharmaceutical companies rely on hospitals for electronic medical record data (EHR) and for pharmacy refill records from insurance companies. Big Data Analytics (BDA) provides an excellent basis to examine and manage terabytes of data that comprises drug data and can manage all aspects of drug development. This survey paper examines the current literature to determine what is current practice in the area of Big Data analytics and medication management.

Keywords: medication management, Big Data Analytics, genomics, diabetes, health system, health promotion

1. Introduction

Diabetes is a diverse group of disease grouped around Type 1 Diabetes Mellitus (DM) common among juveniles and Type 2 Diabetes common among older adults, which we will focus on in this paper. Each type of Diabetes has a different pathophysiology and phenotype; the exact mechanism underlying the disease process being unknown, the cause being thought to be an inability of the body to utilize insulin efficiently. The progression of the disease is still unclear, and so are the complications although DM is related to many comorbid conditions such as vascular disease, eye and retina disease, hypertension, renal failure, neuropathy, cognitive disorders, heart disease, retinopathy, etc. Besides focusing on the specific and individual DM comorbidities and how to relate to patient age and sex, it is important to compare the strength of these relations among each other and order them according to significance. A standardized testing comorbidity procedure to obtain a complete comorbidity profile for Type 2DM is now available using medical claims data. Numerous disorders associated to DM can be credited to complications. DM comorbidities are the rule rather than the exception and their treatment must always address age and sex. Although a risk factor for certain diabetic complications, sex may also influence and to a certain degree even determine the mechanisms causing the disease progressions (Klimek *et al*., 2015).

The popularization of genomic high-throughput technologies is producing a revolution in biomedical research and, predominantly, is transforming the arena of drug discovery. Accordingly, modeling and computational data analysis will have progressively more important roles in drug detection. Massive volumes of genomic information have three main associated complications: (a) the technical problem of big data management and storage, (b) the problem of data primary processing (process by which raw data is converted into comprehensive descriptions of variation of the genomes or transcriptomes), and (c) the most intricate challenge of data interpretation, when discrepancies at genomic or/and transcriptomic level must be correlated to traits. The cloud is gradually more reachable and practical and delivers computation and storage at practical prices as well as software for genomic data analysis as virtual machines. It is essential that data be transferred to the cloud and there are also important issues regarding data security, which can create critical obstacles for its use. Pharmacogenomics is becoming a crucial factor of the concept of personalized medicine. Blockbusters, drugs established for the principal possible populations (e.g. diabetes, among others), were fundamental to the classical approach of pharmaceutical companies for drug development. However, the heterogeneity in drug responses indicates that manufacturing a new blockbuster is beyond a difficult task. Therefore, drug repositioning, or discovering new uses for current drugs, rises in its importance as an essential part of drug discovery stratagems

(Dopazo, 2013). Pharmacogenomics has expanded beyond reviewing individuals' drug response based on genome features (e.g., copy number variations and somatic mutations) and now combines added transcriptomic and metabolic qualities such as gene expression, considering factors that influence the concentration of a drug reaching its targets and aspects connected with the drug targets. Since the gene expression profiles of cell lines are recognized as varying substantially in the development of extended culture under different culture environments and techniques, the use of gene expression from cell lines for prediction of drug response in the patient is presently debatable.

Connecting diverse metadata with structures extricated from image modalities is key to illustrate the construction, purpose, and evolution of diseases. Explaining this challenge offers an exceptional chance for linking the semantic gap between images and more operative prediction, diagnosis, and treatment of diseases. Dimensionality reduction and feature selection can aid us to handle with the curse of dimensionality. Some machine learning methods, such as deep learning, encompass learning quite a lot of layered transformations of the data to discover the best high-level abstraction for the question at hand, imitating the way neuroscience describes knowledge (Feldman *et al.*, 2012). Innovations in data administration, predominantly virtualization and cloud computing, are simplifying the expansion of platforms for more efficient capture, storage and manipulation of great capacities of data. Pharmaceutical developers can assimilate population clinical data sets with genomics data, they may move closer to getting more and better drugs accepted in the first place, and more notably, to getting the right drug to the right patient at the right time. The endless stream of new data amassing at unparalleled rates poses new challenges. Just as the volume and variety of data that is gathered and warehoused has transformed, so too has the velocity at which it is produced and the speed required to recover, investigate, evaluate, and reach decisions using the output. Most health care data have traditionally been quite inert—paper files, X-ray films, scrips. But in some medical situations, real-time data (trauma monitoring for blood pressure, operating room monitors for anesthesia, bedside heart monitors, etc.) become a matter of life or death. In between are the medium-velocity data of multiple daily diabetic glucose measurements (or more continuous control by insulin pumps), blood pressure readings, and EKGs. Data quality problems are a specific anxiety in health care for two reasons: (a) it matters—life or death conclusions are contingent on having the right information (b) the excellence of health care data, specifically unstructured data, is decidedly variable and all too often inappropriate. Illegible handwritten prescriptions are possibly the most infamous example. Refining management of care, evading errors and cutting costs are contingent on high-quality data, as are developments in drug safety and efficacy, diagnostic accuracy and more meticulous targeting of disease methods by actions. But, high Variety and Velocity obstruct the capability to cleanse data before analyzing it and making decisions, raising questions of data "trust " (Fang *et al.,* 2016). The US health care data alone arrived at 150 exabytes (1018) in 2011 and it will exceed the zettabyte (1021) and the yottabyte (1024) sometime soon. Some of the contributing factors to the letdown of conventional arrangements in managing these datasets consist of the vast assortment of structured and unstructured data such as handwritten doctor notes, medical records, medical diagnostic images (magnetic resonance imaging (MRI), computed tomography (CT)), and radiographic films; presence of noisy, heterogeneous, complex, longitudinal, diverse, and great datasets in health care informatics (Gurrin *et al.*, 2014). Table 1 (Fang *et al.,* 2016) summarizes the types and features of healthcare data.

Table 1. Healthcare data types and features

Data Type	Data Format	Structured/ Unstructured	Examples
Human-generated data	ASCII/text	Structured & unstructured	Physicians' notes, email, and paper documents
Machine-generated data	Relational tables	Structured & unstructured	Readings from various monitoring devices
Biometric data	ASCII/text, images	Structured & unstructured	Genomics, genetics, heart rate, blood pressure, x-ray, fingerprints, etc.
Transaction data	Relational tables	Semi-structured & structured	Billing records and healthcare claims
Publications	Text	Unstructured	Clinical research and medical reference material
Social media data	Text, images, videos	Unstructured	Interaction data from social websites

"Big data" is an often-misused term and is wrongly associated with vast volumes of information, therefore the use of the term "big". In fact, "big data" isn't just about volume, it is equally about veracity (the precision and accuracy of data which may have been eroded due to things like calibration drift in sensors), velocity (the shifting patterns and changes in data over time) and variety (the heterogeneous sources from which data is collected). Big data is a contemporary problem and is about mining and cross-referencing information from diverse sources to discover new knowledge. Primary data comprises sources such as physiological data from wearable sensors (heart rate, respiration rate, galvanic skin response, etc.), movement data from wearable accelerometers, location data, nearby Bluetooth devices, Wi-Fi networks and signal strengths, temperature sensors, communication activities, data activities, environmental context, images or video from wearable cameras, and that doesn't take into account the secondary data that can be originated from this primary lifelog data through semantic scrutiny (Viceconti *et al*., 2015). The expansion of mathematical models accomplished as exactly envisaging what will happen to a biological system is obligatory to undertake this vast task; multifaceted research is required. Modern big data technologies make it probable in a short time to analyze a big assortment of data from thousands of patients, categorize clusters and correlations, and finally form predictive models utilizing statistical or machine-learning methods. In this original perspective, it would be practical to take all the data gathered in all previous epidemiology studies and keep on enriching them with original studies where not only new patients are included, but dissimilar categories of data are gathered. Another mechanism is the normalization of digital medical images to conventional space-time reference systems, using elastic registration methods followed by the treatment of the quantities (Ohm, 2015). Big data is now the grist for knowledge creation. Instead of handcrafted knowledge bases for diagnosis and prediction, we have lots of data at the individual level from health-care system use, clinical trials, real-time monitoring, and various other sources.

Machine learning algorithms can determine beneficial blueprints for prediction and clarification as well as cost reduction. Undeniably, this area has been a hotbed of activity over the last few years. It is entirely conceivable that the new information bases revealed throughout data will lead to a renaissance in knowledge-based systems beyond the identification of causal associations that can be utilized to ''reason'' about complications in the spirit of expert systems such as Internist. For people projected to be at higher risk, we could perhaps contemplate a more assertive outreach and testing, including tracking compliance of people on medication, specifically those displaying comorbidity. Information technologies are making it easier to apply these various levels of touch and can funnel more fine-grained information at the individual level into the predictive system (Harrison, 2012). This paper examines more than one aspect of computational analytics and big data analytics in medication adherence among Type 2 Diabetes Mellitus. Lifelogging, network pharmacology, and privacy related concerns related to Big Data analytics are among some of the topics discussed.

2. Big Data Analytics for Diabetes and Clinical Trial Modelling

There are many benefits to clinical trial modelling, such as diminished cost and time savings in circumstances where the use of prevailing data can precisely simulate a trial and its capacity outcome without the need for an actual trial. Modelling and simulation is not a new tactic to drug expansion — for instance, it has been studied in circumstances where data are sparse, such as in pediatrics and other populations with a limited number of patients. It is stimulating to conduct actual Type 2 Diabetes (T2D) clinical trials for countless reasons. For instance, T2D is a heterogeneous disease, so there is a broad spectrum of patient reactions seen in clinical trials. In addition, T2D is a progressive disease, so persons face dissimilar degrees of disease progression, and this will influence their reactions to a drug in a clinical trial. Therefore, modelling could demonstrate an advantageous side for T2D clinical trials: these between-patient transformations could be netted in a model to forecast the trajectory of a patient's progress or reaction to a drug. But some characteristics of T2D clinical trials might be hard to model, such as patient reactions that are not seized by biomarker end points. Biomarkers — such as HbA1c (glycosylated hemoglobin) to determine glycemic control — only quantify the totality of the therapeutic effect and do not enlighten about fluctuations in blood glucose throughout the day and night. Such results would be problematic to incorporate in a model. What happens to a patient before they are drafted to a clinical trial (such as the health status of the patient before recruitment: for example, blood pressure and levels of blood glucose and lipids) can impact what ensues during a trial (including a patient's response to therapy), and it is very hard to factor this into a model (Krumholz, 2014).

The analytic methods for big data classically deviate from traditional statistics and hypothesis testing; they integrate techniques such as machine learning, normally for prediction and discovery. The current medical research enterprise cannot keep pace with the information requirements of patients, clinicians, administrators, and policy makers. Clinical trials, for instance, often exclude complicated patients—those who may have quite a few medical ailments and multifaceted treatment regimens—who are representative of patients that are seen in

medical offices. These studies are most generally focused on a single question, are often expensive, and take years to complete. Moreover, most studies are poorly equipped to explore how various factors may interrelate to shape the result for a specific patient. Meanwhile, data produced every day, for an assortment of practical reasons, could serve as a practically inexhaustible source of information to fuel a learning health care system. Big Data methods can hold the complexity of the data and illuminate the ways that biological, demographic, clinical, and environmental factors interact with each other to influence risk and outcomes. Machine learning can also circumvent the confirmation bias that can pollute investigations directed by scientists with sturdy, preexisting ideas. Machine learning can provide input like that of a truly independent expert. The trial is to develop algorithms such as those that indorse an intelligence about which of the hundreds or thousands of dimensions of data are correlated unpredictably, which are correlated trivially, and which are not interrelated at all. Better characterization of patient profiles could elevate researchers' ability to control for factors that can confuse studies and lead to false conclusions. With better profiling of individuals, scholars could match patients who are and are not treated in certain ways to determine the effect of specific tactics. Given that it is not conceivable to conduct enough trials to cover all types of patients, approaches that leverage actual knowledge could be critical to producing evidence applicable to the full range of patients (Han *et al*., 2015).

Traditional database approaches are not well-matched for data discovery because they are augmented for faster retrieval and summarization of data provided this is what the user wants to ask (i.e. a query), not discovery of patterns in massive swaths of data as soon as the operator does not have a sound formulated query. Machine learning "works" in the sense that these methods perceive elusive structure in data relatively easily without having to make strong suppositions about linearity, monotonicity, or parameters of dispersals. The disadvantage is that these approaches also gather up the noise in data and habitually have no means of distinguishing amid what is signal and what is noise. The universal argument is that when the data are great and multidimensional, it is virtually unattainable for us to know a priori that a query such as the one above is a good one, that is, one that offers a possibly fascinating intuition. Appropriately considered machine learning help find such patterns for us. Similarly, notably, these patterns must be predictive. For handling, very large datasets, nevertheless, typical database systems constructed on the relational data model has severe limitations. The recent move towards Hadoop for dealing with mammoth datasets motions for a new set of required skills for data scientists. The first type is from misspecification of a model. For example, a linear model that struggles to match a nonlinear phenomenon will most likely produce an error purely since the linear model imposes an inappropriate bias on the challenge. The second cause of error is from the use of samples for estimating parameters. The third is due to randomness, even when the model is perfectly specified. The theoretical limitation of observational data, regardless of how big it is, is that it is generally "passive," representing what happened in contrast to the multitude of things that could have happened had the circumstances been different. In the health care example, it is like having observed the use of the health care system passively, and now gaining the chance of understand it in retrospect and extract predictive patterns from it. A restriction of this is that we are controlled in our capability to affect the future through intervention that is conceivable when the causal mechanisms are well understood. The second constraint of predictive modeling with causation is that various models that seem unalike on the exterior might characterize the same fundamental causal structure, but there is no way to know this. For example, in the diabetes example, there could be numerous uncorrelated robust patterns that forecast obstacles (Bottles and Begoli, 2014).

But some Big Data projects will also be a front-runner to bad outcomes, like invasions of privacy and hard-to-detect invidious discrimination. One likely advantage of this distinction is that it might help us guide scientists in the direction of the kind of Big Data lessons that do not jeopardize privacy. There are sufficient enormously significant complications that we can answer in a method constant with individual privacy that it seems a shame that many investigators dedicate so much energy to privacy invasion. Second, we should distinguish that many of the benefits we care most deeply about, including most medical research, originate in research institutions with a well-known track record of respecting personal privacy. For example, hospitals and other entities that hold electronic health records (EHR) *should* cooperate with computer scientists to investigate their substantial datasets to come to new and improved results. But they should do this within the ambit of accountable research; directed by accomplished scientists; and subject to controls and protocols of trust and monitoring, most importantly those rules designed to protect data subjects, such as human subjects review and the Common Rule (Harrison, 2012).

Physicians are competent to create hypotheses that can be verified by the double-blind clinical trial that uses randomization to guarantee that the only transformation between the control group and the treated group is the therapy or process under investigation. Evidence-based medicine concentrates on treatments that have endured

this rigorous and expensive way of doing things. Big data's focus on correlations, not causality, is problematic for physicians biased toward the biomedical model, where the emphasis is discovery of the cause of the disease to successfully treat it. By analyzing the 1,200 data points per second from the wireless sensors fastened to the babies, scholars could diagnose infections 24 hours before fever and higher white blood cell count changed the disease to clinically evident. The actionable correlation exposed by the data was that very stable, continuous vital signs signified impending infection, not well-being. By intermittently having his blood tested for about 40,000 proteins, an integrative personal genomics profile consisting of 30 terabytes of data about how body functions was established. Diabetes prediction was astonishing because the patient was slender, had no family history of diabetes, and had never had raised blood glucose readings. A three-hour fasting blood sugar test exposed a raised initial blood glucose level of 127, and later hemoglobin A1C tests were elevated at 6.4 percent and 6.7 percent launching a diagnosis of diabetes.

Cultured procedures now exist that health care providers can use to reduce per capita costs and increase the quality of the care they deliver. The current progress of open source big data analytic platforms and the increased affordability of cloud computing answer the affluent obstructions hospital executives have faced in the past of owning their own data warehouse. Concealed biases in both the gathering and scrutiny phases present extensive risks and are as important to the big data equation as the numbers themselves. The keystone of privacy laws has been "notice and consent" where people are told at the time of compilation what data is being gathered and the reason for which it will be used. The actionable correlations at the focus of big data count on the secondary use of material such as search, wireless sensor data and social media.

The information contained in Figure 1 illustrates how Big Data sensing is all around us every day (Andreu-Perez *et al*. 2015). It is tough for the individual to give informed consent for secondary uses that are not even imagined when the data are first composed. Big data poses problems if it's anonymous. While stripping out personal identifiers thrives in the setting of small data, the truth of big data means re-identification is conceivable. The impression that big data correlations can entirely exchange deeper comprehension of the basis of problems also necessitates amendment. An alternative pitfall of thoughtlessly tolerating the big data hype is to disregard the fact that additional data means more struggle extricating the noise from the signal. It also means additional false information (Tene and Polonetsky, 2013).

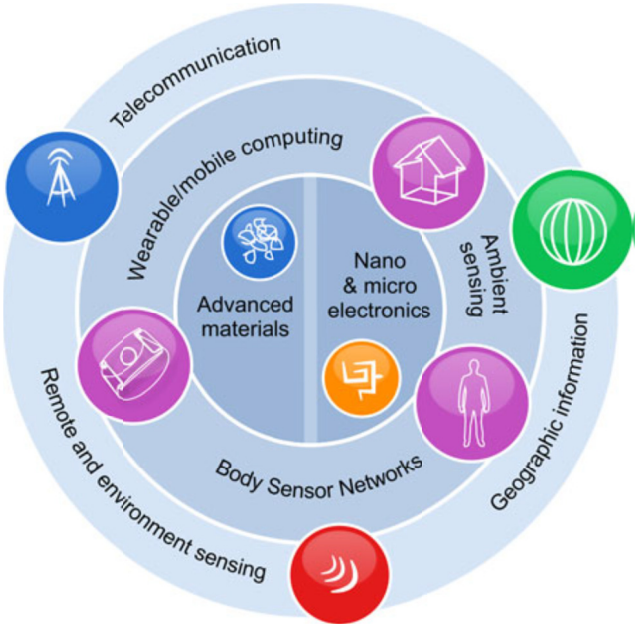

Figure 1. Illustrates the daily exchange of sensor information and telecommunication

To craft equilibrium between positive uses of data and individual privacy, policymakers must tackle some of the most fundamental concepts of privacy law, including the definition of "personally some of the most difficult evidence," the role of individual power, and the ethics of data minimization and purpose limitation. In addition, since in a Big Data world it is often not the data but rather the extrapolations lured from them that give cause for apprehension, establishments: it should be compulsory to disclose their decisional criteria. The inclination is

motivated by condensed costs of storing information and moving it around in conjunction with amplified volume to instantaneously analyze piles of unstructured data utilizing modern experimental methods, observational and longitudinal studies, and large scale simulations. The big data paradigm challenges some of these fundamental principles, including the scope of the framework (often addressed by framing the term personally identifiable information (PII)), the concepts of data minimization (focused collection) and consent (individual control and respect for context), and the right of individual access (access and accuracy). The potential advantages of big data analytics within the medical field have resulted in public policy initiatives to mine and leverage such data. Big data poses big privacy risks. The harvesting of large sets of personal data and the use of state of the art analytics implicate growing privacy concerns. Big data may simplify predictive scrutiny with stark inferences for individuals vulnerable to disease Predictive analysis is predominantly challenging when grounded on delicate types of data, such as health, race, or sexuality. A supplementary apprehension elevated by big data is that it tilts an already uneven scale in favor of organizations and against individuals. Conventionally, de-identification was regarded as a silver bullet allowing organizations to reap the benefits of analytics while preserving individuals' privacy. Organizations used numerous approaches of de-identification (anonymization, pseudonymization, encryption, key-coding, data sharing) to distance data from personal identities. Yet, over the past few years, computer scientists have repeatedly shown that even anonymized data can typically be re-identified and associated with specific individuals. De-identified data, in other words, is a temporary state rather than a stable category More importantly, many beneficial uses of data would be severely curtailed if information, ostensibly not about individuals, comes under full remit of privacy laws based on a remote possibility of being linked to an individual at some point in time through some conceivable method, no matter how unlikely to be used. Where prospective data uses are highly beneficial and privacy risks minimal, the cogency of dispensation should be presumed even if individuals decline (or are not asked) to consent. Information asymmetries and well-documented cognitive biases cast a shadow on the authenticity of individuals' privacy choices. Privacy suffers not only when individuals are unaware of data practices, but also when they are uninterested or disengaged. Such an environment, regardless of the regulatory mechanisms in place, provides insufficient checks on data collection and use (Dendrou *et al*., 2013).

Sources can be either quantitative (e.g., sensor data, images, gene arrays, laboratory tests) or qualitative (e.g., free text, demographics). The objectives underlying this data challenge are to support the basis for observational evidence to answer clinical questions, which would not otherwise have been solved via studies based on randomized trials alone. In addition, the issue of generalizing results based on a narrow spectrum of participants may be solved by taking advantage of the potential of Big Data for deploying longitudinal studies. Veracity is important for Big Data as personal health records may contain typographical errors, abbreviations, and cryptic notes. Ambulatory measurements are sometimes taken within less reliable, uncontrolled environments compared to clinical data, which are collected by trained practitioners. The use of spontaneous unmanaged data, such as those from social media, can lead to wrong predictions as the data context is not always known. For individual patient reports, the use of natural language processing plays an essential role for systematic analysis and indexing of the underlying semantic contents. Mining EHRs is a valuable tool for improving clinical knowledge and supporting clinical research. Mining local information included in EHR data has already been proven to be effective for a wide range of healthcare challenges, such as disease management support pharmacovigilance building models for predicting health risk assessment enhancing knowledge about survival rates therapeutic recommendation discovering comorbidities, and building support systems for the recruitment of patients for new clinical trials (Feldman *et al.*, 2012).

3. Big Data Analytics as Precision Analytics

For over a century, medicine has depended on veteran physicians witnessing a clinical phenotype—usually attained at the bedside—to classify patients before settling on a course of therapy. However, limitations associated with this method have become increasingly apparent. The appreciation that classical, observational medicine is constrained by its inability to define disease with the precision required for optimal diagnosis and subsequent therapeutic benefit and prognosis has spurred conceptual and technological advances to elucidate underlying disease mechanisms. The improved classification of disease through the integration of observational medicine with information derived from experimental findings is paving the way toward more personalized medicine. Research into disease etiology and new therapeutic avenues encompasses genetics and genomics, proteomics, microbiomics and stem cell research, among others, resulting in an explosion of information. Therefore, a key challenge now is to avoid being lost in translation—big data outputs need to be interpreted and converted into improved knowledge of a disease that can then be effectively incorporated into clinical decision making. It is evident that the emerging disease pathways could not have been deduced from the purely

observational assessment of patients. Based on this paradigm, translational research integrating genetic data and molecular and cellular analytic methodologies in the context of common diseases can facilitate appropriate drug administration to prevent adverse responses, aid the repositioning of safe drugs approved for one condition to treat another and spur *de novo* drug discovery. As medicine and translational research have been advancing on parallel tracks, the integration of the two is leading to substantial improvements in disease diagnosis and definition and in subsequent prevention, therapy and prognostication. The resulting change in medicine may find the classical emphasis on observational phenotyping and taxonomy superseded by a more integrative clinical practice that also incorporates an appreciation of disease mechanism to achieve more informed and personalized patient management (Lupton, 2014).

To exploit the full potential of 'Big Data' for medical sciences the development of novel, quantitative methods to extract clinically relevant features from large datasets of electronic health records (EHR) is necessary. First efforts in this direction have proven to be extremely fruitful by developing or improving data-driven comorbidity indices to predict mortality rates, or by studying healthcare utilization and outcome measures of specific patient cohorts. Large-scale analyses of comorbidities using EHR data have demonstrated that human disease phenotypes can be related to each other in highly connected networks with strong pairwise correlations between diseases This work shows the enormous potential that large-scale analyses of EHR data offer for the medical sciences (Klimek *et al*., 2015).

An emphasis has progressed towards the embracing of the new digital media technologies that have been enabled by Web 2.0 as a means of producing and sharing such data, both by healthcare providers and patients in what has been variously described as 'e-health', 'Health 2.0', 'Medicine 2.0' or 'digital health' initiatives. Digital media technologies are now promoted for use in patient self-care and self-monitoring, conducting medical encounters remotely and collecting data about healthcare use. The term 'e-scaped medicine' was employed in relation to Web 1.0 technologies to describe the ways in which medical information and knowledge had apparently moved beyond the boundaries of the clinic. In the Web 2.0 era, further technological developments have brought with them even greater opportunities for lay people to not only seek information across an ever-growing array of websites and blogs directed at health and medical issues but also to engage in patient support and activism communities, the evaluation of medical care and contribute to the aggregation of data about medical procedures and drug therapies for specific illnesses and diseases. Through social media platforms dedicated to specific illnesses or conditions such as Facebook pages, Twitter hashtags and YouTube videos of patient experiences and medical techniques and therapies, as well as the more traditional format of online discussion sites, information can be exchanged, discussion facilitated and activism mobilized across the globe in real-time. The new digital health technologies participate in the growing accumulation of what has been termed 'big data' or 'transactional data': that is, the vast quantities of data, both quantitative and qualitative, that are the digital traces or by-products of users' interactions and transactions with digital media technologies. These digital traces include the data that are gathered on users' activities when they visit websites, including the products they buy, the telephone numbers they call and the government agencies and commercial entities with which they interact. It also includes 'user-generated content', or data that have been intentionally uploaded to social media platforms by users as part of their participation in these sites. These data are particularly valued because they are collected as a by-product of behavior rather than directly via purposive surveys or interviews, and because they can be collected in real time. The lure and potential of big data have had a major impact upon healthcare policy. There is now much focus on and discussion concerning the power of large masses of data gathered by digital technologies both to inform patients about their own bodies and health states and to provide information to healthcare providers about the health states of populations and the use of healthcare. In the past, pharmaceutical companies have established or financially supported some patient support websites. The arena of clinical trials for new drugs is one form of medical knowledge generation where crowdsourcing via patient-focused social media platforms has been employed for some years as an alternative to the expensive traditional format of the standard clinical trial. The accumulation of big data that is afforded by the new digital media technologies is positioned as an innovative way forward for health care, supposedly providing better, more informed and more economically efficient medical treatment (Yensen and Naylor, 2016).

A systematic data-processing pipeline for generic big data in health informatics, covering data capturing, storing, sharing, analyzing, searching, and decision support. Natural Language Processing (NLP) is utilized to extract key elements from unstructured data, such as notes and reports. Healthcare data could be classified into unstructured, structured, and semi-structured. Historically, most unstructured data usually come from office medical records, handwritten notes, paper prescriptions, MRI, CT, and so on. The structured and semi structured data refers to electronic accounting and billings, actuarial data, laboratory instrument readings, and EMR data

converted from paper records. Leveraging heterogeneous datasets and securely linking them have the potential to improve healthcare by identifying the right treatment for the right individual. The increase in the volume and variety of healthcare data is highly related to the velocity at which it is produced and the speed needed to retrieve and analyze the data for timely decision making. The high velocity of healthcare data poses another big challenge for big data analytics. An initial attempt includes the utilization of the Hadoop platform for running analytics across massive volumes of data using a batch process. It is not uncommon that the healthcare data contains biases, noise, and abnormalities, which poses a potential threat to proper decision-making processes and treatments to patients. High-quality data can not only ensure the correctness of information but also reduce the cost of data processing. It is highly desirable to clean data in advance of analyzing it and using it to make life-or-death decisions. However, the variety and velocity of healthcare data raise difficulties in generating trusted information. Low latency is a highly desired property for stream processing as a big data technology, while scaling data integration is critical for adapting to the high-volume and high-velocity nature of big data. Apache Hadoop coupled with existing integration software and the Hadoop Map/Reduce framework could provide the computing environment for parallel processing. More recently, Hadoop Spark a successor system that is more powerful and flexible than Hadoop MapReduce, is getting more and more attention due to its lower-latency queries, iterative computation, and real-time processing. Specialists and physicians use analyzed data to search for systematic patterns in patients' information, which helps them in having a more precise diagnosis and treatment. Data mining is an analytic process that is designed to search and explore large-scale data (Big Data) to discover consistent and systematic patterns. One of the main challenges in big data mining in the medical domain is searching through unstructured and structured medical data to find a useful pattern from patients' information. The privacy of data is another big concern of future Big Data analytics in healthcare informatics. In addition, there are a range of other issues, such as data protection, data security, data safety, and protection of doctors against responsibility derived from manipulated data, that require special big data analytics to handle these complex restrictions (Gurrin *et al*., 2014).

The limitations of modern healthcare have been ascribed as causative agents in the diabetes crisis. The current healthcare system tends to provide a reactive response to patient symptoms, with a subsequent diagnosis and corresponding treatment of the specific disease. More recently a rapid improvement in OMIC analyses, bioinformatics and knowledge management tools, as well as the emergence of big data analytics, and systems biology have led to a better understanding of the profound, dynamic complexity and variability of individuals and human populations as they undertake their daily activities. These developments in conjunction with escalating healthcare costs and relatively poor disease treatment efficacies have fermented a rethink in how we execute current medical practice. This has led to the emergence of "P-medicine" which includes personalized and precision medicine. P-medicine is still in a fledgling and evolutionary phase and there has been considerable debate over its status and future trajectory, as well as its ability to affect the runaway crises of pandemic diabetes. Some have argued that as personalized medicine has morphed into precision medicine (PM) we are just realizing the tip of the PM iceberg. There appears to be a chronic lack of confidence in global healthcare systems. Stakeholder expectations have been fueled by repetitive media reports of spectacular advances in the diagnosis and treatment of the major diseases that afflict patients. However, those same patients perceive a lack of delivered value from their healthcare provider. In part this is predicated on the transition of patients into consumers, and their accompanying expectations, particularly in the developed world. Many of these patient complaints involve diagnostic and prognostic inaccuracies, poor treatment efficacies for a specific disease indication, and lack of timely access to patient care. The general dissatisfaction is apparent regardless of the specific healthcare delivery system. The current *modus operandi* of modern medicine is based on the determination of an individual's symptoms, along with an associated diagnosis and subsequent response to a specific treatment. These data for the individual are compared to a statistically similar and disease-relevant patient population dataset. There is also a focus on a specific disease indication as it pertains to compartmentalized tissue and/or organs involving, in many cases, a highly specialized clinician. The current healthcare system tends to be reactive, providing treatment post-onset of the disease, with limited efforts focused on prediction and prevention. This reliance on the comparative analysis of an individual compared to a defined population tends to neglect and disregard human individuality, complexity and variability. It also fails to recognize the systems level interconnectedness of human molecular biology, biochemistry, metabolism and physiology in the form of systems, network and pathway biology (Özdemir *et al*., 2015).

Proteomics is a Big Data technology and a next generation biomarker, supporting novel system diagnostics and therapeutics in psychiatry. Proteomics technology is, in fact, much older than genomics and dates to the 1970s, well before the launch of the International Human Genome Project. Big Data has multiple meanings; it refers not only to the enormous size of contemporary datasets, but also to the rapid rate by which data can move around

different locations, time zones and application contexts worldwide. Big Data includes a diverse array of unprecedentedly large datasets created by omics biotechnologies (genomics, proteomics, metabolomics, metagenomics), biosensors, electronic health records, simulation experiments, social media, crowdfunding platforms and the Internet, to mention but a few Big Data-driven fields such as proteomics bring about vast uncertainties about their societal trajectory and how social systems might in turn influence the development of proteomics science (Zhang, 2016).

Therefore, new distributed computing solutions such as MapReduce and heterogeneous computational environments, in which conventional CPUs are merged with specialized accelerators such as graphics processing units (GPUs) or field-programmable gate array (FPGA) that can speedup calculations several orders of magnitude, will be soon common for genomic big data analysis (Dopazo, 2013). Clearly, newer approaches are needed for the redefinition of disease using not only the traditional signs and symptoms but also the underlying genomic/proteomic causes and other factors (Bottles and Begoli, 2014). The voxel values of the scan then become another medical dataset, potentially to be correlated with average blood pressure, body weight, age, or any other clinical information. Using statistical modeling or machine learning techniques we may obtain good predictors valid for the range of the datasets analyzed; if a database contains outcome observables for a subset of patients, we will be able to compute automatically the accuracy of such a predictor. Typically, the result of this process would be a potential clinical tool with known accuracy; in some cases, the result would provide a predictive accuracy sufficient for clinical purposes, in others a higher accuracy might be desirable. Most Big Data applications deal with data that do not refer to an individual person. This does not exclude the possibility that their aggregated information content might not be socially sensitive, but very rarely is it possible to reconnect such content to the identity of an individual. In the cases where sensitive data are involved, it is usually possible to collect and analyze the data at a single location; this becomes a problem of computer security; within the secure box, the treatment of the data is identical to that of non-sensitive data (Kolker *et.al*. (2014).

4. New Challenges of Big Data Analytics in Medication Development and Management

Network pharmacology is devoted to understanding the pharmacological mechanism of drug action in the network perspective. Network pharmacology aims to understand diseases at the systematic level, and to know the interaction between the drug and the body based on equilibrium theory of biological networks. It is substantially bringing the significant changes of theory and methodology in drug design. Network pharmacology is an interdisciplinary science based on pharmacology, network biology, systems biology, bioinformatics, computational science, and other related scientific disciplines. Network pharmacology aims to understand the network interactions between a living organism and drugs that affect normal or abnormal biochemical function. It tries to exploit the pharmacological mechanism of drug action in the biological network, and helps to find drug targets and enhance the drug's efficacy. The scope of network pharmacology covers but not limits to: (1) theories, algorithms, models and software of network pharmacology; (2) network construction and interactions prediction; (3) theories and methods on dynamics, optimization and control of pharmacological networks (here generally refer to disease network, disease - disease, disease - drug, drug - drug, drug - target network, network targets - disease, and drug targets - disease network, etc.); (4) network analysis of pharmacological networks, including flow (flux) balance analysis, topological analysis, network stability, etc.; (5) various pharmacological networks and interactions; (6) factors that affect drug metabolism; (7) network approach for searching targets and discovering medicines (including medicinal plants, etc.); (8) Big Data analytics of network pharmacology, etc. Pharmacology is the branch of medicine and biology on drug action where a drug can be broadly defined as any man-made, natural, or endogenous molecule which exerts a biochemical and/or physiological effect on the cell, tissue, organ, or organism. In the perspective of network pharmacology There are two sources of fundamental data for research in network pharmacology, public databases and experimental verification. With big data analytics, e.g., high-performance data mining, predictive analytics, text mining, forecasting and optimization, we can analyze huge volumes of data that conventional analytics cannot handle. In addition, machine learning techniques are ideally suited to addressing Big Data needs. Many problems in network pharmacology, network construction, interactions prediction, etc. are also expected to be addressed by using big data analytics. Most pharmacological networks are unknown or imperfect. Therefore, how to construct a pharmacological network is a prerequisite for such diseases. Among them, the networks of disease related protein interactions are the most important. The most used methods to find such interactions and construct pharmacological networks include phylogenetic profile gene neighborhood gene fusion event mirror tree correlated mutation correlated evolutionary rate prediction from primary structure and homologus structural complexities etc. Among them, phylogenetic profile method is particularly useful for construction of networks and prediction of large scale interactions. Network models are the foundation to understand interactions within complex networks. Various random graph models produce network structures that may be used in comparison to real complex networks Network pharmacology helps to better understand the influence of behaviors of cells and

organs on functional phenotypes to understand the mechanism of drug functioning to provide theoretical basis and technical support for drug design and for rational use of drugs. It helps explain and predict drug interactions and optimize the use of drugs and find factors that affect drug efficacy and safety and to quickly discover biomarkers and targets. It guides the drug's clinical use. Currently drug resistance has become a common phenomenon. Because network pharmacology targets more than one molecule or gene and examines multiple areas of drug use, it is more useful than traditional pharmacology methods where only one target is examined. Besides being limited by deficient methodology of network biology and systems biology, network pharmacology will face such challenges as the limited knowledge and technology for identification of drug targets, fewer multi-target drugs, and poor database quality, etc. Nevertheless, network pharmacology is an emerging branch of pharmacology built on massive -omics data and multiple sciences. It is expected to greatly develop in the future (Pasquale, 2013).

Biological processes are fundamentally driven by complex interactions between biomolecules. Integrated high throughput omics studies enable multifaceted views of cells, organisms, or their communities. With the advent of new post-genomics technologies, omics studies are becoming increasingly prevalent; yet the full impact of these studies can only be realized through data harmonization, sharing, meta-analysis, and integrated research. These essential steps require consistent generation, capture, and distribution of metadata. For ensuring transparency, facilitating data harmonization, and maximizing reproducibility and usability of life sciences studies, a simple common omics metadata checklist was proposed. The omics metadata checklist and data publications will create efficient linkages between omics data and knowledge-based life sciences innovation and, importantly, allow for appropriate attribution to data generators and infrastructure science builders in the post-genomics era. Modern life science technologies enable rapid and efficient acquisition of omics data (Kuo *et al*., 2014). These data comprehensively measure multilayered molecular networks and provide a snapshot of biological processes in a cell, organism, or their communities. Collected on the same sample at the same time, omics data provide information on the functioning of biomolecules and their interactions. Omics studies are essential for the systemic investigation of biological systems—an endeavor that is crucial to improve our ability to manage and cure diseases, identify drug targets, understand regulatory cascades, and predict ecosystem responses to environmental changes. The use, integration, and reuse of data require accurate and comprehensive capture of the associated metadata, including details describing experimental design, sample acquisition and preparation, instrument protocols, and processing steps. The data and metadata must be captured together in a rigorous and consistent manner to allow the integration of data across omics experiments. complexities not only make reproducible, integrative, accurate, and comprehensive capture of data and metadata an intricate challenge that must be overcome but also place an excessive burden on researchers trying to convey metadata Because of the large amount of data and the complexity of data acquisition, it is exceedingly difficult to capture, disseminate, and interpret the metadata. In its short, structured form, the checklist captures important experimental parameters and strikes a balance between comprehensiveness and ease of use (El-Gayar and Timsina, 2014).

By siloing data, health insurers and providers have impeded the types of large-scale analysis common in other industries. Providers have kept vital information about price, quality, and access secret to maintain a competitive advantage or hide shortcomings. Each major drug company's data exclusivity may mean that rivals waste vast amounts of money pursuing leads that have already proven to be dead ends. Health information technology systems may not be interoperable, leaving them unable to "talk to one another" and share data. Federal and state agencies need to require providers and insurers to reveal key data in exchange for government support, while minimizing the possibility of improper uses of that data. The challenge is to rationalize complex, often conflicting legal frameworks as the stakes rise (Dhar, 2014). Unfortunately, trade secrecy and some other IP protections now prevent the realization of the full scale of efficiencies possible in an era of big data. Health care, however, is one of many areas where intermediaries consider information gathering either a commodifiable service or an aspect of their own competitive strategy. There is an important divide between researchers who have access to critical medical research and those who do not. massive misallocation of resources may be attributed, in part, to failures to act on current data, but it also occurs because useful data is not available, does not exist, or is actively hidden. The emerging field of agnotology studies such lacunae by examining the structural production of ignorance, its diverse causes and conformations, whether brought about by neglect, forgetfulness, myopia, extinction, secrecy, or suppression There is a remarkable amount of undisclosed health data about the effects of pharmaceuticals. Companies push to keep exclusive access to their own data, even when serious concerns arise about their products. There are serious deficiencies in America's system of pharmacovigilance—namely the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem. Fraud and abuse have long been a scourge of the U.S. health care system. Stories abound of diluted medications, unlicensed providers, cosmetic surgery misrepresented as 'medically necessary,' and kickbacks designed to bilk CMS (Kofke, 2014).

Health is a product of genome and environment. The cost of DNA sequencing is plummeting. The impact of genomics in medicine includes (1) to understand and treat disease, (2) pharmacogenomics driving drug therapy, and (3) managing health care in healthy people. This will include development for possibly billions of individuals' characterization of the genome, epigenome, transcriptome, proteome, cytokine-ome, metabolome, auto antibody-ome, and microbiome (gut, urine, nose, tongue, and skin). With this we will have lots of impressively big data (e.g., a million genes and other –omic information in millions and billions of people). One hoped-for consequence of this will be an ability to predict disease and monitor diseases. This should lead to new therapeutic strategies for diseases. He envisions a world where genomic and other -omic information is derived at birth from each person, who at that early time will have a unique –omics profile which will then be used to predict and prevent disease, monitor disease progress, and guide treatment. He also mentioned the problem of overlapping meta-analysis on the same topic (e.g., use of statins after cardiac surgery). He then discussed the need for international large scale collaboration and agreements to combine data from big data sets (Ohm, 2015; Rahimi *et al*., 2014].

Epidemiology, pharmacoepidemiology & pharmacoeconomic researchers seek multiple diverse global data sources, but need information on their use & limitations. Electronic data sources expand the breadth of data inquiry (Kuo *et al*., 2014). Greater awareness & standardization will support development of new, more useful DBs for public health and practice uses. Support for the practice of multi-country studies and standardization will facilitate meta-analyses focused product development, planning for risk management and surveying post marketing for use, risk & benefit (Krumholz, 2014).

In lifelogging, all these varied sources merge and combine together to form a holistic personal lifelog where the variety across data sources is normalized and eliminated. Lifelogging generates continuous streams of data on a per-person basis, however despite the potential for real-time interactions, most of the applications for lifelogging we have seen to date do not yet operate in a real-time mode. Finally, *veracity* refers to the accuracy of data and to it sometimes being imprecise and uncertain. In the case of lifelogging, because much of our lifelog data can be derived from sensors which may be troublesome, or have issues of calibration and sensor drift. As lifelog applications will become more widespread, we can see that lifelogging does indeed have a big data application with a requirement to provide facilities to extract meaning, etc. in order to create surrogate memories based on useful and meaningful lifelogs, which is both the end goal and the big challenge for information retrieval over lifelogs (Rahimi *et al*., 2014). Table 2 (Kuo *et al*., 2014) summarizes the challenges and potential solutions of Big Data analytics.

Table 2. Big Data Analytics challenges and potential solutions

Stage	Challenges	Potential solutions
1. Aggregation	Dispersed, heterogeneous and unstructured health raw data; difficulty in sharing among different incompatible applications; networking challenge in transferring large data into or out of the cloud.	• High speed file transfer technologies • Data compression • P2P data distribution
2. Maintenance	Heavy IT burden (cost and time) in storing and maintaining large raw data for a small organization or lab; data jurisdiction issues for some projects.	• Cloud computing • Grid computing • NoSQL.
3. Integration	Challenge in integrating unstructured data; three types of challenges (functional, metadata and instance integration) in transforming and integrating large heterogeneous structured data into a suitable format.	• Structured EHR data integrations • Image integration technologies • Graph integration technologies • Unstructured clinical note integration technologies
4. Analysis	Challenges to choosing or constructing analysis models: complexity of the analysis, scale of the data, parallelization of computing model, and availability of computing resources.	• Platforms: (super, grid, cloud and heterogeneous) computing • Tools: MapReduce, Hadoop, and SAS in-memory analytics, etc. • Algorithms: distributed data mining
5. Interpretation	Result presentation and interpretation by non-technical domain experts; biases and blind spots in Big Data; ease in violating individuals' privacy if proper precautions are not taken.	• Data provenance techniques • Validating approaches • Privacy regulations

5. Future Trends in Big Data Management

In fact, the methods and results of clinical trials on the drugs we use today are still routinely and legally being withheld from doctors, researchers and patients. There is reasonable disagreement on how to interpret the trials, in which case we need full access to their methods and results, for an informed public debate in the medical academic community. This is particularly important, since there can often be shortcomings in the design of a clinical trial, which mean it is no longer a fair test of which treatment is best. Similarly, in trials described as "double blinded" – where neither doctor nor patient should be able to tell whether they're getting a placebo or the real drug – the active and placebo pills were different colors. We often choose to use treatments in medicine, knowing that they have limited benefit, and significant side-effects: but we make an informed decision, balancing the risks and benefits for ourselves. We cannot make informed decisions about which treatment is best while information about clinical trials is routinely and legally withheld from doctors, researchers, and patients (Groves et al., 2013). Anyone who stands in the way of transparency is exposing patients to avoidable harm. We need regulators, legislators, and professional bodies to demand full transparency. We also need clear audit on what information is missing, and who is withholding it. Finally, more than anything – because culture shift will be as powerful as legislation –we need to do something even more difficult (Kolker *et.al*. (2014). Pharmaceutical-industry experts, payors, and providers are now beginning to analyze big data to obtain insights. Although these efforts are still in their early stages, they could collectively help the industry address problems related to variability in healthcare quality and escalating healthcare spending. For instance, researchers can mine the data to see what treatments are most effective for conditions, identify patterns related to drug side effects or hospital readmissions, and gain other important information that can help patients and reduce costs. Fortunately, recent technologic advances in the industry have improved their ability to work with such data, even though the files are enormous and often have different database structures and technical characteristics. Many healthcare stakeholders have underinvested in information technology because of uncertain returns—although their older systems are functional, they have a limited ability to standardize and consolidate data. The nature of the healthcare industry itself also creates challenges: while there are many players, there is no way to easily share data among different providers or facilities, partly because of privacy concerns (Kofke, 2014).

The proliferation of potential evidence, patient data, and health consumer information represent opportunities for the innovative applications of analytics techniques to assist with the translation of data (in a variety of format depending on the source) to consumable knowledge. Analytics can be predictive, descriptive, or prescriptive. Evidence based medicine (EBM) is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. Literature sources include randomized controlled trials (RCT), systematic reviews, clinical guidelines, cohort studies, Quasi-Experimental studies, descriptive studies, and expert opinions. Additional sources of practice-based evidence include claims data, insurance, and other administrative hospital data (Jones, 2014).

6. Conclusion

Even within a single hospital, payor, or pharmaceutical company, important information often remains siloed within one group or department because organizations lack procedures for integrating data and communicating findings. In many cases, aggregating individual data sets into big-data algorithms is the best source for evidence, as nuances in subpopulations (such as the presence of patients with gluten allergies) may be rare enough that individual smaller data sets do not provide enough evidence to determine that statistical differences are present. The old levers for capturing value—largely cost-reduction moves, such as unit price discounts based on contracting and negotiating leverage, or elimination of redundant treatments—do not take full advantage of the insights that big data provides and thus need to be supplemented or replaced with other measures related to the new value pathways. Similarly, traditional medical-management techniques will no longer be adequate, since they pit payors and providers against each other, framing benefit plans in terms of what is and isn't covered, rather than what is and is not most effective. Finally, traditional fee-for-service payment structures must be replaced with new systems that base reimbursement on insights provided by big data—a move that is already well under way.

The growing use of Electronic Health Records (EHRs) raises issues of semantic interoperability and the quality management/improvement of large datasets derived from multiple EHRs. Improved data quality (DQ) in health organizations can improve the quality of decisions in health care. The documenting of clinical data as text in clinical notes remains a major reason for non-accessible data in EHRs. Genomics has been the cutting edge of the Big Data revolution in the life sciences, one that holds considerable (if yet-to-be-delivered) promise for enabling personalized medicine. Internet transactions and communications, cloud storage, social media and mobile devices expose more and more personal data to potential misuse. Data security, unintentional exposure or

loss of data to unauthorized parties use of the Internet, cloud computing and pooling of data all raise the data security stakes. Big Data Analytics enables all the data (medical literature, electronic health record, clinical notes, x-ray and other imaging data, insurance and claims data, and more) to be leveraged to translate data to relevant information for EBM support.

Big Data can be used to identify healthcare trends, prevent diseases, combat social inequality, and so on. Managed well data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments accountable data-handling problems, complexity and expensive or unavailable computational solutions to research problems are major issues in healthcare/biomedical research (big) data management and analysis. Development of a standardized analytic procedure will enable both an ecosystem of reusable scientific tools and workflows, and aid in this BDA endeavor, ultimately contributing to better science. Health data is different from data in other disciplines in that it includes structured EHR data, coded data, semi-structured data, unstructured data, genetic data, and other types of data. There are several potential solutions for Big Data maintenance including cloud computing, grid computing, and NoSQL/NewSQL, etc.

References

Andreu-Perez, J., Poon, C.C.Y., & Merrifield, R. D. *et al*. (2015). Big Data for health. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1193-1208.

Bottles, K., & Begoli, E. (2014). Understanding the pros and cons of Big Data Analytics. *Physician Exec*, *40*(4), 6-12.

Dendrou, C. A., Bell, J.I., & Fugger, L. (2013). Weighing in on autoimmune disease: Big Data tip the scale. *Nature Medicine*, *19,* 38-139. http://dx.doi.org/10.1038/nm.3087

Dhar, V. (2014). Big Data and predictive analytics in health care. *Mary Ann Liebert, Inc*., *2*(3), 1-4. http://dx.doi.org/10.1089/big.2014.1525

Dopazo, J. (2013). Genomics and transcriptomics in drug discovery. *Drug Discovery Today*. http://dx.doi.org/10.1016/j.drudis.2013.06.003

El-Gayar, O., & Timsina, P. (2014). Opportunities for business intelligence and Big Data Analytics in evidence based medicine. *System Sciences* (HICSS), 749-757. http://dx.doi.org/10.1109/HICSS.2014.100

Fang, R., Pouyanfar, S., Yang, Y., Chen, S.C., & Iyengar, S.S. (2016). Computational health informatics in the Big Data age: a survey. *ACM Computing Surveys (CSUR), 49*(1), 12. http://dx.doi.org/10.1145/2932707

Feldman, B., Martin, E.M., & Skotnes T. (2012). Big Data in healthcare: hype and hope. *Business Development for Digital Health*.

Groves, P., Kayyali, B., Knott, D., & Kuiken, S.V. (2013). The 'big data' revolution in healthcare. *McKinsey & Company White Paper,* Center for US Health System Reform Business Technology Office, 1-22.

Gurrin, C., Smeaton, A.F., & Doherty A.H. (2014). Lifelogging: personal Big Data, *Foundations and Trends in Information Retrieval*, *8*(1), 1-125. http://dx.doi.org/10.1561/1500000033

Han, Y., Li, L., Zhang, Y., Yuan, H., Ye, L., Zhao, J., & Duan, D.D. (2015). Phenomics of vascular disease: the systematic approach to the combination therapy. *Curr Vasc Pharmacol*, *13*(4), 433-40.

Harrison, C. (2012). 'Big data' deal for diabetes clinical trial modelling. *Nature Reviews Drug Discovery,* 1. http://dx.doi.org/10.1038/nrd3891

Jones, J.K. (2014). Identifying population data to evaluate risk, use, cost and benefit of medical products. *ASCPT Workshop: Registries and Databases in Clinical Research*, 1-37.

Klimek, P., Kautzky-Willer, A., Chmiel, A., Schiller-Fr hwirth I., & Thurner, S. (2015). Quantification of diabetes comorbidity risks across life using nation-wide big claims data. *PLoS Comput Biol, 11*(4), e1004125. http://dx.doi.org/10.1371/journal.pcbi1004125

Kofke, W.A. (2014). AUA president's panel at 2014 AUA annual meeting: genomics, Big Data, and publication pitfalls. *AUA Update,* Summer, 1-17.

Kolker, Özdemir, & Martens, L., *et al*. (2014). Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. *OMICS*, *18*(1), 10-14. http://dx.doi.org/10.1089/omi.2013.0149

Krumholz, H.M. (2014). Big Data and new knowledge in medicine: the thinking, training, and tools needed for a learning health System. *Health Aff* (Millwood), *33*(7), 1163-70. http://dx.doi.org/10.1377/hlthaff.2014.0053

Kuo, M.-H., Sahama, T., Kushniruk, A.W., Borycki, E.M., & Grunwell, D.K. (2014). Health big data analytics: current perspectives, challenges and potential solutions. *International Journal of Big Data Intelligence*, *1*(1/2), 114-126.

Lupton, D. (2014). The Commodification of Patient Opinion: The digital patient experience economy in the age of Big Data. *Sociology of Health and Illness*, *36*(6), 856-869. http://dx.doi.org/10.1111/1467-9566.12109

Ohm, P. (2015). The Underwhelming benefits of Big Data. *University of Pennsylvania Law Review Online*, *161*, 339-346.

Özdemir, V., Dove, E.S., Gürsoy, U.K., Şardaş, S., Yıldırım, A., Yılmaz, S.G., … Srivastava, S. (2015). Personalized medicine beyond genomics: alternative futures in big data—proteomics, environtome and the social proteome. *Neural Transm* (Vienna): Epub 2015 Dec 8.

Pasquale, F. (2013). Grand bargains for Big Data: The Emerging Law of Health Information. *Maryland Law Review*, *72*(3), 1-92.

Rahimi, A., Liaw, S.T., Taggart, J., Ray, P., & Yu, H. (2014). Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in electronic health records. *International Journal of Medical Informatics*, *83*, 768-778.

Tene, O., & Polonetsky, J. (2013). Big Data for all: privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property*, *11*(5), 240-272.

Viceconti, M., Hunter, P., & Hose R. (2015). Big Data, big knowledge: Big Data for personalized healthcare. *IEEE J Biomed Health Inform*, *19*(4), 1209-1215. http://dx.doi.org/10.1109/JBHI.2015.2406883

Yensen J., & Naylor, S. (2016). The Complimentary iceberg tips of Diabetes and precision medicine, 1-14. Retrieved from https://www.researchgate.net/publication/296696068_The_complementary_iceberg_tips_of_diabetes_and_precision_medicine

Zhang, W.J. (2016). Network pharmacology: A further description. *Network Pharmacology*, *1*(1), 1-14.