 doi:10.20944/preprints201806.0137.v1

Review

Opportunities and Challenges in Data-Driven Healthcare Research

You Chen[1],*

[1] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN;
* Correspondence: you.chen@gmail.com; Tel.: +1-615-343-1939

Abstract: Health information technology has been widely used in healthcare, which has contributed a huge amount of data. Health data has four characteristics: high volume; high velocity; high variety and high value. Thus, they can be leveraged to i) discover associations between genes, diseases and drugs to implement precision medicine; ii) predict diseases and identify their corresponding causal factors to prevent or control the diseases at an earlier time; iii) learn risk factors related to clinical outcomes (e.g., patients' unplanned readmission), to improve care quality and reduce healthcare expenditure; and iv) discover care coordination patterns representing good practice in the implementation of collaborative patient-centered care. At the same time, there are major challenges existing in data-driven healthcare research, which include: i) inefficient health data exchanges across different sources; ii) learned knowledge is biased to specific institution; iii) inefficient strategies to evaluate plausibility of the learned patterns and v) incorrect interpretation and translation of the learned patterns. In this paper, we review various types of health data, discuss opportunities and challenges existing in the data-driven healthcare research, provide solutions to solve the challenges, and state the important role of the data-driven healthcare research in the establishment of smart healthcare system.

Keywords: opportunity, challenge, perspective, health data; disease prediction; clinical outcome prediction; healthcare process; data quality; quantity and quality analysis; artificial intelligence

---

1. Introduction

Health information technology (HIT) plays an important role in the healthcare system evolution, and it has had a dramatic impact on the practice of medicine. In many situations, HIT has been verified to be an effective tool to achieve high quality and safety care [14-16]. We discuss opportunities and challenges of data-driven healthcare research starting from HIT and also ending with HIT (as shown in **Figure 1**). HIT transforms data in the version of paper into electronic and hatches many novel health related information systems and services such as electronic health record systems (EHRs) [17], online health communication forums [18-19], next generation sequencing [20] and wearable devices and mobile health [21-22].

The new systems and services (e.g., EHRs) originated from HIT have contributed a huge amount of health data, which has four major characteristics [23-25]: i) high volume; it is very common to have Terabytes or Petabytes of the storage system for healthcare organizations (HCOs) to manage health data ; ii) high velocity; the health data movement is now almost real time and the update window has reduced to fractions of the seconds; iii) high variety; the health data can be stored in various formats such as database including structured and unstructured, extensible markup language, photos and short message service; and iv) high value, i.e., a patient health status can be visualized via an enhanced 360 degree of view.

High throughput of health related data provides a direct view of a person's health; however, there are many health patterns which are hidden behind the data and are not shown up in front of care providers and patients [26-27]. Thus, there is an emergent need to learn these hidden patterns.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 June 2018 doi:10.20944/preprints201806.0137.v1

Data-driven healthcare research has been proposed to achieve this goal [28-29]. In recent decades, various types of data-driven healthcare researches have been proposed, for instance, Genome-Wide Association Study (GWAS) [30-31] and Phenome-Wide Association Study (PheWAS) [32-33] have been developed to find associations between genes, diseases and drugs; drug-drug interaction studies have been implemented to detect adverse drug interactions [34]; predictive models are used to predict diseases such as Alzheimer [35] and suicide [36] at an earlier time; computational algorithms and statistical models are leveraged to identify risk factors related to patient outcomes such as unplanned readmission rates [38], mortality rates [39] and prolonged length of stay [40-41]; and healthcare process modeling aims to identify care coordination patterns representing good practice in the implementation of patient-centered care [42-43]. However, there are several major challenges existing in the data-driven healthcare research, which include but not limited to: i) inefficient health data exchange strategies; ii) biased research findings; and iii) difficulties in the evaluation, interpretation and translation of the learned patterns.

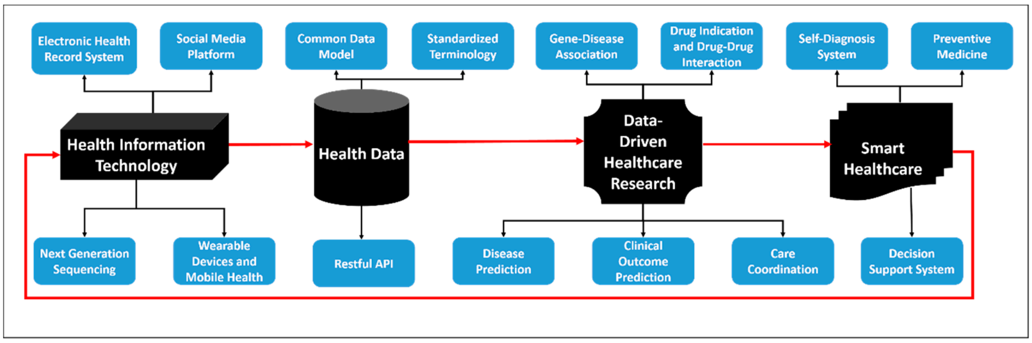

Figure 1. A circular workflow to depict the position of data-driven health care research, which originates from health information technology and finally feeds it back.

Researchers have proposed solutions to solve the aforementioned challenges. For instance, Observational Health Data Sciences and Informatics (OHDSI) built Observational Medical Outcomes Partnership (OMOP) common data model (CDM) to solve data quality and data exchange challenges [44]. Under the OMOP, data can be represented by standard terminology and transferred across HCOs via application programming interface (API) [45]. With the effort of OHDSI, data-driven healthcare research can be conducted on a large volume of data and the research findings will have a high probability not to be biased to specific HCO [45-46]. At the same time, qualitative approaches (e.g., surveys and focused group interviews) have been proposed to assess plausibility of the learned patterns and translate them into clinical practice [42, 72].

In this review, beyond showing challenges and opportunities of data-driven healthcare research, we also depict perspectives of data-driven healthcare research and its important role in building smart health care systems such as self-diagnosis systems.

2. Health Data

Health data including DNA, EHR, mobile and social median, has been generated in an accelerated way.

2.1. Genomic Data

Extraordinary progress made in genome sequencing technologies lets the generation of DNA data in a fastest and cheapest way. According to data collected by the National Human Genome Research Institute, the cost per genome is around $1,121 in July 2017 [97]. Next generation sequencing (NGS) platforms can perform sequencing of millions of small fragments of DNA in parallel and each of the three billion bases in the human genome is sequenced simultaneously, which brings down the average time of sequencing a human genome to one hour [20, 97].

2.2. EHR Data

As of 2016, over 95% of hospitals are eligible for the Medicare and Medicaid EHR Incentive Program in United States, and the health information systems they have been using include Epic Systems, Allscripts, eClinicalWorks, AthenaHealth, NextGen Healthcare, Cerner, MEDITECH (Medical Information Technology), McKesson, and Orion Health[47].

2.3. Mobile Data

By 2018, it's predicted that over 50% of smartphone users will have downloaded mobile health applications (apps), which can be grouped into two major categories: wellness and medical [48-49]. Wellness apps are typically used by patients accompanying with wearable devices, while medical apps are designed to be primarily used by physicians. Of the 100,000 mobile health apps in app stores around the world, 85% of apps are for wellness while the remaining 15% are for medical [50-51].

2.4. Social Media Data

Social media has generated huge amount of health related data. It has been recognized that over 30% of adults are likely to share information on their health, prescribed drugs, hospitals they stayed and their insurance programs in social media platforms with other patients, and doctors, [52-53]. The most popular online platforms they have been frequently accessed are WebMD, Wikipedia, health magazine websites, Facebook, YouTube, online blogs, patient communities, and Twitter [54].

3. Opportunities in Data-driven Healthcare Research

Health data contains a patient's genetics and genomics, electronic health records, daily activities, lifestyle choices and social determinants, which provides a great opportunity to implement precision medicine. Precision medicine takes into account individual variability in genes, environment variables, and lifestyle choices to design personalized disease treatment and prevention strategies [55]. For instance, if a patient has a genetic variation in gene VKORC1, which can reduce the ability of an enzyme to recycle vitamin K, and subsequently the ability of the blood to clot, then a doctor needs to prescribe a low dose of warfarin which is a medication that is used to prevent blood from clotting, also known as an anticoagulant [98]. To achieve the goal of precision medicine, researchers have done various types of data-driven healthcare researches, which can be categorized into following six major groups.

3.1. GWAS and PheWAS studies

GWAS and PheWAS have been proposed to learn associations between genes and diseases. GWAS samples a large number of genetic variants for association with a single phenotype (disease) [30] whereas PheWAS does the same procedure with many phenotypes to one gene [56]. GWAS has been studied for over decades and as of 2017, the GWAS Catalog contains 3,172 publications and 52,491 unique Single Nucleotide Polymorphism (SNP)-trait associations [31]. Researchers from Vanderbilt University have done distinguished studies on the scans of diseases for each individual gene via longitudinal EHRs [56, 100]. As of 2017, PheWAS catalog contain 1,358 EHR-derived phenotypes associated with 3,144 SNPs [32, 56].

3.2. Drug Repositioning

Drug repositioning approaches have been designed to identify and develop new therapeutic indications for existing drugs. New drug development is a costly, complex and time-consuming process. The average length of time from target discovery to approval of a new drug is about 14 years [57]. The failure rate during this process exceeds 95 percent, and the cost per successful drug can be $1 billion or more [57]. Thus, drug repositioning of approved drugs has recently gained new momentum for rapid identification and development of new therapeutics for diseases that lack

 doi:10.20944/preprints201806.0137.v1

effective drug treatment [1-3]. GWAS and PheWAS findings [30-32, 56, 58] and clinical data [59] are leveraged to discover novel indications for existing drugs. A recent study leveraged GWAS and PheWAS findings to discover new clinical targets of existing drugs [99]. They used disease-gene associations found in GWAS and PheWAS, and drug-gene associations in DrugBank to learn relations between drugs and diseases via their overlapped genes. If the learned relations between drugs and diseases are novel, which are unknown in clinical observations, then they assume these relations could provide a great opportunity to develop new functions of existing drugs. Finally, they found 744 relations between diseases and drugs (e.g., disease asthma and drug irinotecan), which have not been found in clinical observations.

3.3. Drug-Drug Interactions

Drug-drug interactions (DDIs) learning has been proposed to reduce medication errors and improve patient safety [60-63]. DDIs is a situation in which a drug affects the activity of another drug when both are administered together, and it has been one of the commonest causes of medication error. There are two typical transporter-based DDI risk evaluations: in vitro and in vivo extrapolation models [101]. University of Washington drug interaction database-Metabolism and Transport Drug Interaction Database (DIDB)-has been licensed for scientists and clinicians working in the field of DDIs since 2002 [102]. DIDB is a knowledge base which includes both in vitro and in vivo DDI data, allowing in vitro to in vivo extrapolations; at the same time, it includes DDI data coming from 291 new drug application reviews [102]. Drugbank is another major data resource which has been leveraged by researchers to explore DDIs. For instance, researchers used drugbank data and machine learning algorithms to predict DDIs and DDI induced adverse drug interactions [103].

3.4. Disease Prediction

Disease predictive models have been developed to predict diseases before their occurrences. These models usually leverage computational models and potential risk factors including biomarkers, clinical phenotypes, lifestyle behaviors or social determinants to predict diseases at an early time. For instance, mutations in the genes encoding amyloid precursor protein, presenilin 1 and presenilin 2 are responsible for early-onset autosomal dominant Alzheimer's Disease [5]; a high body mass index (BMI) and high blood cholesterol in cardiovascular diseases [4]; and socio-economic variables (e.g., income, education, or occupation) is linked to a wide range of health problems, including low birth weight, cardiovascular disease, infectious intestinal disease, hypertension, arthritis, diabetes, and cancer [6-7]. A recent study leveraged a machine learning based model and mothers' maternal data to predict neonatal encephalopathy (NE), which is a leading cause of infant mortality and long-term neurological morbidity [104]. This model can predict NE earlier than the time a child was born, which provides a great opportunity for HCOs to adopt preventative interventions to minimize the effects of distal risk factors and decrease the risk of NE.

3.5. Clinical Outcome Prediction

Outcome prediction aims to measure associations between prognostic factors and clinical outcomes. The prognostic factors include health conditions such as diseases, [40, 77]; care coordination routines such as clinical workflows [76, 78] and care team [41, 79]; environmental variables such as social determinants [80]; and healthcare payers such as health insurance programs [81]. Clinical outcome includes unplanned readmission rates [38], length of stay (LOS) in hospital [40-41], health care expenditure [73], patient satisfaction [74-75], and morbidity and mortality [36, 39]. Clinical outcome prediction can bring two major benefits. The first is it can improve efficiency of resource allocation. For instance, researchers can leverage mothers' historical health conditions (before childbirth) to predict their LOS during delivery hospitalizations [40]. HCOs can use such decision support system to estimate LOS for each patient and then conduct resource allocation accordingly. Another benefit of clinical outcome prediction is there is a big opportunity to identify causal prognostic factors related to outcome, which can potentially improve care quality (e.g., preventions of unplanned readmission) and reduce health care cost (e.g., reductions in LOS) [8-10].

3.6. Care Coordination Optimization

It has been recognized that patient centered care, which requires a transition from independent clinician working in isolation to a care team with fully interactions between each other, can improve care quality and reduce healthcare cost [105]. Communication, collaboration and care coordination between health care employees play an important role in establishing or refining patient centered care [82]. Researchers have developed various data-driven models to learn care teams or clinical workflows from EHRs [11-13, 42, 83-85]. Their results indicate that the teams and workflows learned from the data are plausible and can be interpreted by HCOs [42]. At the same time, some of researches measured associations between care team patterns and clinical outcomes to discover the team patterns representing good practice in the implementation of patient centered care [41]. Another type of study in patient centered care aims to put right care providers in place for right patients, in particular for those patients who exhibit multiple health conditions simultaneously [43]. For instance, researchers found patients with a collection of health conditions (e.g., anemia, hypogonadism, prostate cancer and bone loss) were usually co-managed by an integrated clinical workflow in a form of a bundle of care providers [43]. Compared with the traditional care strategies, which treat each of conditions independently such approach can potentially avoid replicated care (e.g., replicated tests requested by different care providers) and reduce patient visit durations (e.g., cost of time for transitions between care providers).

4. Challenges in Data-driven Healthcare Research

Health data provides strong supports to conduct the aforementioned researches to achieve the goal of precision medicine. However, there are several major challenges to utilize health data to achieve the goal, which can be categorized as follows.

4.1. Interpretation of Health Information System Utilization

There is a gap between people who design health information systems and those who use the systems. Healthcare employees who use the same health information system, may even have disparate interpretations on the system utilizations [64-65]. For instance, a patient's health information documented by a provider of one HCO may be misunderstood by providers from another HCO. Thus, health information systems maybe inappropriately utilized, and thus the data documented in such information system may be incorrectly interpreted [64].

4.2. Data Standard

Aligning data coming from different sources can provide a complete care journey for each patient, which is a necessity to implement patient centered care and value-based care [66-67]. For instance, a trauma patient may be diagnosed in disparate HCOs, such as primary care hospital, trauma center, and skilled nursing facilities. The primary care recorded his historical health information, trauma center recorded detailed surgical procedures he had received and nursing facilities recorded progresses of his post-operative recovery. Getting all information associate with the patient requires involved HCOs to coordinate with each other to ensure the information locates in each HCO could be communicated accurately. This is very important to advocate patient-centered care, which coordinates healthcare workers across disparate HCOs to make a decision for a patient's diagnosis and value-based care, which continuously monitors health status and outcome of a patient and then make a payment according to the outcome achieved. However, there are few common data models which can be served as channels to let data be communicated across HCOs. Another benefit of common data model is it can support the building of a big cohort for research purpose, and subsequently increases the power of learned knowledge [68]. For instance, if a research aims to infer causal relationship between maternal disease (e.g., depression) and neonatal disease (hypo-pituitary axis-childhood growth and development), then they need to identify subjects to get sufficient power to conduct a case-control study to test the significance of the causal relationship. In this case, it will be required to identify more subjects from various HCOs. Thus, a common data model is critically important to ensure information is accurately aligned across disparate HCOs.

4.3. Biased Finding

Most of data-driven healthcare researches have been conducted on an individual healthcare system, and thus the findings learned from such studies are biased to the specific healthcare system [104, 106-107]. For instance, performance of NE prediction models introduced in [104] are biased to the investigated patient population at Vanderbilt University Medical Center (VUMC). Majority of maternal patients admitted to VUMC are high risk and thus, findings of prediction model built on such population are biased to high risk patients. In other words, low risk patients with NE babies may not be captured by the predictive models.

Data-driven models are usually trained on unbalanced cohorts, where the number of cases is much smaller than the controls, and thus patterns learned from such models are biased, in many scenarios, resulting in always predicting the majority class (controls). For instance, NE is a rare disease and the number of cases is much smaller than the number of controls [104]. Thus, models built on such unbalanced cohorts are dominated by controls.

4.4. Interpretation of Learned Patterns

There is a gap between data-driven findings and their applications in clinical practice. Usually researchers focus on performances (e.g., predictive accuracy) of data-driven approach and seldom emphasize on interpretation and evaluation of patters learned from the data. Two typical approaches: supervised learning [69-70, 104] and unsupervised learning [41-43] exist in the data-driven healthcare research, and they both face challenges to interpret knowledge learned from data. Supervised learning is very similar with traditional hypothesis-driven case-control study, which requires experts to predefine a hypothesis and then build a cohort of cases and controls to test the hypothesis. Supervised learning requires that the cases and controls are pre-labeled by experts according to a golden standard. At the same time, it requires experts' prior knowledge to identify potential explanatory variables influencing cases or controls. Computational models are built based on the identified explanatory variables and labeled cases and controls [69-70]. This type of research aims to achieve a high accuracy of identifying cases and controls via the identified explanatory variables. However, it is hard to interpret causal relationships between explanatory and response (cases or controls) variables. For instance, a study used logistical regression to predict NE via mothers' maternal data and measure associations between NE and risk factors of the mothers [104]. The model can achieve an area under curve (AUC) of 0.87 to predict NE and identify risk factors which play the most important roles in the prediction. However, it is still hard for it to figure out causal factors leading to NE.

Unsupervised learning which does not need manual effort to build cohorts of cases and controls, automatically learns novel patterns from the data [41-43]. It is much harder to interpret patterns learned from the unsupervised than the supervised. This is because most of patterns learned from the unsupervised approaches are novel, which is difficult for experts to interpret via their domain knowledge. For instance, a study learned care teams from EHRs via an unsupervised clustering approach, and it is hard for it to evaluate if the learned care teams are plausible or if they are effective in clinical practice [42].

5. Potential Solutions

Although there are many challenges existing in data-driven healthcare research, we have witnessed big effort to solve these challenges.

5.1. Transfer Learning

To quantify differences in interpretations of health information system utilizations, researchers have done studies to measure similarities of health information system utilization behaviors [86-88]. For instance, a study investigated the transferability and stability of phenotypes learned from one health information system to another [86]. The study assumed that healthcare employees in the two HCOs (Vanderbilt University Medical Center and Northwestern University) have similar interpretations of using diagnosis codes (International Classification of Diseases, Ninth Revision, Clinical Modification) in their EHR systems. They learned phenotypes from diagnosis codes in each healthcare system, and did a cross projection of patients based on the learned phenotypes (e.g., a phenotype learned from system A used to explain patients in system B); and then they compare differences between patients projected by phenotypes learned from their own system and projected by phenotypes learned from other systems. They found that utilization behavior of standard terminology such as ICD-9 codes across disparate healthcare systems are consistent [86].

5.2. Common Data Model

To promote health data coming from disparate sources can be exchanged, OHDSI has proposed OMOP CDM, which allows for the systematic analysis of disparate observational databases via standardized terminologies [44-45]. CDM transforms data contained within disparate sources into a common format as well as a common representation (terminologies, vocabularies, coding schemes). For instance, OHDSI has developed an open source software ATLAS [71] for researchers to identify people with specific conditions, drug exposures from disparate sources. ATLAS can transform health information coming from disparate sources to a standardized observational data via CDM. At the same time, ATLAS can visualize a particular subject's health care records coming from different sources.

It is notable that data exchange between HCOs can potentially solve the problem of bias findings. This is because, CDM can let researchers to construct a big cohort to include all types of subjects (patients) from various sources. Models built on such big patient population have a potentiality to learn knowledge which are not biased to a specific healthcare system.

5.3. Under-sampling and Over-sampling

To solve the challenge of training a model on an unbalanced cohort, researchers use over-sampling (up sampling more cases to match the number of controls) or under-sampling (down sampling controls to match the number of cases) strategies to construct a balanced cohort [108-109]. For all investigated cases, under-sampling randomly selects controls whose number is the equal or close to cases, and then build balanced cohort. This process can be done many times and generate a series of cohorts. Models will be trained and validated within each cohort. Each independent model is not biased to unbalanced data. However, the main drawback of this strategy is that each model could not capture complete characteristics of controls, and thus the model has a high false positive and low positive predictive rates (many controls predicted as cases) in the practice. An alternative way to reduce high false positive rates is that given a new subject, first measuring distance between a new subject and each independent mode (e.g., average distance between the subject and all subjects involved in an independent model), and then using the model which has the smallest distance to predict the class of the new subject.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 June 2018 doi:10.20944/preprints201806.0137.v1

Over-sampling randomly samples cases to increase the number of cases to an equal number of controls. A potential problem over-sampling will approach is the overfitting. This is because, validation set may have the same cases with those in training set. An alternative way to solve overfitting is to separate validation set first, and then oversampling cases in the training set. A recent technique the synthetic minority over-sampling technique, which not only over-samples minority class, as well as under-samples the majority class, has been developed to solve drawbacks yield by both over-sampling and under-sampling [110].

5.4. Quantity and Quality Analysis

To fill the gap between patterns learned from health data and its interpretation and application in clinical practice, researchers have proposed many interpretation strategies. For instance, online survey is a popular approach aiming to recruit clinical experts to assess and interpret learned patterns [42, 72]. Usually a survey contains the learned patterns and their corresponding clinical context. Researchers send these surveys to clinical experts and ask them to assess plausibility of the patterns according to their domain knowledge [72]. For instance, a data-driven study learned care teams of a HCOs via computational models and then they invited administrative and clinical experts to assess plausibility of the learned care teams via online surveys [42]. They designed survey question for each learned care team and asked experts to determine if each learned care team satisfies their expectations. Beyond online surveys, researchers also design focus group interview to let content experts discuss and interpret the learned patterns [89].

6. Perspectives of Data-driven Healthcare Research

According to the aforementioned opportunities, challenges, and potential solutions to the challenges, data-driven healthcare research can provide big opportunities to establish clinical decision support systems or smart self-diagnosis systems.

5.1. Artificial Intelligence

Artificial intelligence has been populated in recent years [90-91]. For instance, Amazon Alexa system has been providing APIs to allow disparate types of devices communicate with their clouds [92-93]. Furthermore, they incorporate various computational models and algorithms in their clouds, which can be leveraged to automatically analyze data collected from disparate devices. In the near future, we believe smart control systems such as Alexa can connect medical devices, wearable devices, social media accounts, and shopping accounts which can monitor patients' health status (e.g., heart rate, blood pressure) and life style choices (e.g., sleeping hours, physical exercise, social behaviors, and eating habits) all the time. The computational models developed in the field of data-driven healthcare research can also be connected to clouds to provide smart clinical decisions. For instance, disease predictive models can be integrated to provide risk alerts for Parkinson disease, Alzheimer, and suicide; drug-drug interaction models can provide information for people to avoid adverse drug interactions; GWAS and PheWAS models can assist physicians in prescribing appropriate medications based on a patient's genome and phenome data; and patient aligned care team models can recommend right care teams for right patients at right time.

5.2. Smart Healthcare System

Speech recognition [94] and visualization technologies [95-96] have been progressing very fast in recent years, which provides strong support to build input (voice-assisted care) and output (360 degree of visualization of a patient health status) of a smart healthcare system. Healthcare employees can communicate with decision support systems via voices settled in smart healthcare systems, and interpret changes of patients' health status via visualized interactive graphs.

A smart healthcare system has three core components: i) inputs (e.g., smart devices, audio speech recognition, online social media and shopping account), which automatically collect patients' data; ii) computational models (e.g., data mining, deep learning), which analyze health data and discover patterns; and iii) outputs (e.g., visualization tools) to visualize the learned patterns. Data-driven healthcare research aims to conduct smart analysis on the collected data and discover valuable patterns or knowledge, which are subsequently visualized to be shown up in front of patients, physicians and HCOs.

7. Conclusions

Data-driven healthcare research plays an important role in the establishment of smart healthcare system. This paper reviews opportunities, challenges, potential solutions to challenges, and perspectives existing in data-driven healthcare research. Data-driven healthcare research originated from analysis of health data generated by HIT and its ultimate goal is to discover valuable knowledge learned from the data to feed the HIT. In other words, data-driven healthcare research both starts and ends at HIT. Although many challenges including data quality, data standards, interpretation of learned patterns and translation of the patterns into clinical practice, exist in the data-driven healthcare research, it still has a great potentiality to assist in the establishment of smart healthcare systems.

Author Contributions: You Chen performed all of the reviewing works and writing of the manuscript.

Funding: This research was supported, in part, by the National Institutes of Health under grant R00LM011933.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chong CR, Sullivan DJ Jr. New uses for old drugs. Nature. 2007; 448:645–646. doi: 10.1038/448645a.
2. Smith RB. Repositioned drugs: integrating intellectual property and regulatory strategies. Drug Discov Today Ther Strateg. 2011; 8:131. doi: 10.1016/j.ddstr.2011.06.008.
3. Padhy BM, Gupta YK. Drug repositioning: re-investigating existing drugs for new therapeutic indications. J Postgrad Med. 2011; 57:153–160. doi: 10.4103/0022-3859.81870.
4. Yusuf S, Hawken S, and Ounpuu S. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet. 2004; 364(9438):937-52.
5. Kim DH, Yeo SH, Park JM. Genetic markers for diagnosis and pathogenesis of Alzheimer's disease. Gene. 2014; 545(2):185-93.
6. Adams NL, Rose TC, Hawker J, et.al. Socioeconomic status and infectious intestinal disease in the community: a longitudinal study (IID2 study). The European Journal of Public Health. 2017; 28(1): 134-138.
7. Talala KM, Huurre TM, Laatikainen TK, Martelin TP, Ostamo AI, Prättälä RS. The contribution of psychological distress to socio-economic differences in cause-specific mortality: a population-based follow-up of 28 years. BMC Public Health. 2011; 11:138.
8. Grigioni F, Clavel MA, Vanoverschelde JL. The MIDA Mortality Risk Score: development and external validation of a prognostic model for early and late death in degenerative mitral regurgitation. European heart journal. 2017; 39(15): 1281-1291.
9. Mansoor H, Elgendy IY, Segal R, Bavry AA, Bian J. Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: A machine learning approach. Heart & Lung: The Journal of Acute and Critical Care. 2017; 46(6): 405-411.
10. Ko MC, Huang SJ, Chen CC, Chang YP, et.al. Factors predicting a home death among home palliative care recipients. Medicine. 2017; 96(41):e8210. doi: 10.1097/MD.0000000000008210.
11. Horner GN, Agboola S, Jethwani K, Tan-McGrory A, Lopez L. Designing Patient-Centered Text Messaging Interventions for Increasing Physical Activity Among Participants With Type 2 Diabetes: Qualitative Results From the Text to Move Intervention. JMIR Mhealth Uhealth. 2017; 5(4):e54.
12. Bowdoin JJ, Rodriguez-Monguio R, Puleo E, Keller D3, Roche J. The patient-centered medical home model: healthcare services utilization and cost for non-elderly adults with mental illness. Journal of Mental Health. 2017; 1-9.
13. Johnson V, Wong E, Lampman M, et.al. Comparing Patient-Centered Medical Home Implementation in Urban and Rural VHA Clinics: Results From the Patient Aligned Care Team Initiative. The Journal of ambulatory care management. 2018; 41(1): 47-57.
14. Payne P R O, Lussier Y, Foraker R E, et al. Rethinking the role and impact of health information technology: informatics as an interventional discipline. BMC medical informatics and decision making. 2016; 16(1): 40.
15. Peters T E. Transformational Impact of Health Information Technology on the Clinical Practice of Child and Adolescent Psychiatry. Child and Adolescent Psychiatric Clinics. 2017; 26(1): 55-66.
16. Sawesi S, Rashrash M, Phalakornkule K, et al. The impact of information technology on patient engagement and health behavior change: a systematic review of the literature. JMIR medical informatics. 2016; 4(1).
17. Friend TH, Jennings SJ, Levine WC. Communication Patterns in the Perioperative Environment During Epic Electronic Health Record System Implementation. Journal of medical systems. 2017; 41(2): 22.
18. Fergie G, Hilton S, Hunt K. Young adults' experiences of seeking online information about diabetes and mental health in the age of social media. Health Expectations. 2016; 19(6): 1324-1335.
19. Li X, Gray K, Verspoor K, et al. Analysing health professionals' learning interactions in online social networks: A social network analysis approach. arXiv preprint arXiv:1604.02883. 2016.

20. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016; 17(6): 333-351.

21. Wang JB, Cataldo JK, Ayala GX, et al. Mobile and Wearable Device Features that Matter in Promoting Physical Activity. Journal of mobile technology in medicine. 2016; 5(2).

22. Rodgers M, Pai V, Conroy R. Ambient and wearable sensors for human health monitoring. Active and Assisted Living: Technologies and Applications. 2016: 29.

23. Groves P, Kayyali B, Knott D, et al. The big data revolution in healthcare: Accelerating value and innovation. McKinsey Quarterly. 2013; 2: 3.

24. Feng Z, Bhat RR, Yuan X, et al. Intelligent Perioperative System: Towards Real-time Big Data Analytics in Surgery Risk Assessment. arXiv preprint arXiv:1709.10192, 2017.

25. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health information science and systems. 2014; 2(1): 3.

26. Wang Y, Kung LA, Wang WYC, et al. An integrated big data analytics-enabled transformation model: Application to health care. Information & Management. 2018; 55(1): 64-79.

27. Myers SR, Carr BG, Branas CC. Uniting Big Health Data for a National Learning Health System in the United States. JAMA pediatrics. 2016; 170(12): 1133-1134.

28. Shameer K, Badgeley M A, Miotto R, et al. Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. Briefings in bioinformatics. 2016: bbv118.

29. Rathore MM, Ahmad A, Paul A, et al. Real-time medical emergency response system: exploiting IoT and big data for public health. Journal of medical systems. 2016; 40(12): 283.

30. Marigorta UM, Denson LA, Hyams JS, et al. Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. Nature genetics. 2017.

31. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic acids research. 2017; 45(D1): D896-D901.

32. Salnikova LE, Khadzhieva MB, Kolobkov DS. Biological findings from the PheWAS catalog: focus on connective tissue-related disorders (pelvic floor dysfunction, abdominal hernia, varicose veins and hemorrhoids). Human genetics. 2016; 135(7): 779-795.

33. Claar DD, Larkin EK, Bastarache L, et al. A Phenome-Wide Association Study Identifies a Novel Asthma Risk Locus Near TERC. American journal of respiratory and critical care medicine. 2016; 193(1): 98-100.

34. Liu R, AbdulHameed MDM, Kumar K, et al. Data-driven prediction of adverse drug reactions induced by drug-drug interactions. BMC Pharmacology and Toxicology. 2017; 18(1): 44.

35. Yang J, An N, Kowall NW, et al. Data driven approaches for predictors selectors in determining Alzheimer's disease. Alzheimer's & Dementia: The Journal of the Alzheimer's Association. 2016; 12(7): P478.

36. Barak-Corren Y, Castro VM, Javitt S, et al. Predicting suicidal behavior from longitudinal electronic health records. American journal of psychiatry. 2016; 174(2): 154-162.

37. Glassman S, Carreon LY, Andersen M, et al. Predictors of Hospital Readmission and Surgical Site Infection in the United States, Denmark, and Japan: Is Risk Stratification a Universal Language?. Spine. 2017; 42(17): 1311-1315.

38. Jamei M, Nisnevich A, Wetchler E, et al. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PloS one. 2017; 12(7): e0181173.

39. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach. Academic Emergency Medicine. 2016; 23(3): 269-278.

40. Gao C, Kho A, Ivory C, Osmundson S, Malin B, Chen Y. Predicting Length of Stay for Obstetric Patients via Electronic Medical Records. Stud Health Technol Inform. 2017; 245:1019-1023.

41. Chen Y, Patel M B, McNaughton C D, et al. A Data-Driven Analysis of the Influence of Care Coordination on Trauma Outcome. Journal of the American Medical Informatics Association. 2018. https://doi.org/10.1093/jamia/ocy009.

42. Chen Y, Lorenzi N M, Sandberg W S, et al. Identifying collaborative care teams through electronic medical record utilization patterns. Journal of the American Medical Informatics Association. 2017; 24(e1): e111-e120.

43. Chen Y, Kho A N, Liebovitz D, et al. Learning Bundled Care Opportunities from Electronic Medical Records. Journal of Biomedical Informatics. 2018; 77:1-10

44. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics. 2015; 216: 574.

45. Chakrabarti S, Sen A, Huser V, et al. An Interoperable Similarity-based Cohort Identification Method Using the OMOP Common Data Model Version 5.0. Journal of Healthcare Informatics Research. 2017: 1-18.

46. Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences. 2016; 113(27): 7329-7336.

47. Centers for Medicare and Medicaid Services. EHR incentive programs: what's changed for EHR incentive programs in 2015 through 2017 (Modified Stage 2). 2017.

48. Nikou S, Bouwman H. Mobile Health and Wellness Applications: A Business Model Ontology-Based Review. International Journal of E-Business Research (IJEBR). 2017; 13(1): 1-24.

49. Albrecht UV, von Jan U. Safe, sound and desirable: development of mHealth apps under the stress of rapid life cycles. Mhealth. 2017, 3.

50. Helbostad JL, Vereijken B, Becker C, et al. Mobile health applications to promote active and healthy ageing. Sensors. 2017; 17(3): 622.

51. Tikkanen SA, Barnhouse MK. The Effects of Personal and Social Uses of Mobile Health Applications on Healthy Behaviors. Communication Studies. 2017; 68(2): 152-172.

52. Park M, Sun Y, McLaughlin M L. Social Media Propagation of Content Promoting Risky Health Behavior. Cyberpsychology, Behavior, and Social Networking. 2017; 20(5): 278-285.

53. McClellan C, Ali MM, Mutter R, et al. Using social media to monitor mental health discussions− evidence from Twitter. Journal of the American Medical Informatics Association. 2017; 24(3): 496-502.

54. Benetoli A, Chen TF, Aslani P. Consumer Health-Related Activities on Social Media: Exploratory Study. Journal of Medical Internet Research. 2017; 19(10): e352.

55. Collins FS, Varmus H. A new initiative on precision medicine. New England Journal of Medicine. 2015; 372(9): 793-795.

56. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nature biotechnology. 2013; 31(12): 1102-1111.

57. Adams CP, Brantner VV. Estimating the cost of new drug development: is it really $802 million?. Health affairs. 2006; 25(2): 420-428.

58. Sanseau P, Agarwal P, Barnes MR, et al. Use of genome-wide association studies for drug repositioning. Nature biotechnology. 2012; 30(4): 317-320.

59. Yao L, Zhang Y, Li Y, et al. Electronic health records: Implications for drug discovery. Drug discovery today. 2011; 16(13): 594-599.

60. Payne TH, Hines LE, Chan RC, et al. Recommendations to improve the usability of drug-drug interaction clinical decision support alerts. Journal of the American Medical Informatics Association. 2015; 22(6): 1243-1250.

61. Wells PS, Holbrook AM, Crowther NR, et al. Interactions of warfarin with drugs and food. Annals of internal medicine. 1994; 121(9): 676-683.

62. Iyer SV, Harpaz R, LePendu P, et al. Mining clinical text for signals of adverse drug-drug interactions. Journal of the American Medical Informatics Association. 2014: 353-362.
63. Yang CC, Yang H, Jiang L, et al. Social media mining for drug safety signal detection. Proceedings of the 2012 international workshop on Smart health and wellbeing. ACM, 2012: 33-40.
64. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. Journal of the American Medical Informatics Association. 2004; 11(2): 104-112.
65. Schneeweiss S. Learning from big health care data. New England Journal of Medicine. 2014; 370(23): 2161-2163.
66. Nguyen OK, Chan CV, Makam A, et al. Envisioning a social-health information exchange as a platform to support a patient-centered medical neighborhood: a feasibility study. Journal of general internal medicine. 2015; 30(1): 60-67
67. Adler-Milstein J, Lin SC, Jha AK. The number of health information exchange efforts is declining, leaving the viability of broad clinical data exchange uncertain. Health Affairs. 2016; 35(7): 1278-1285.
68. Gottesman O, Kuivaniemi H, Tromp G, et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. Genetics in Medicine. 2013; 15(10): 761.
69. Sheets L, Petroski GF, Zhuang Y, et al. Combining Contrast Mining with Logistic Regression To Predict Healthcare Utilization in a Managed Care Population. Applied Clinical Informatics. 2017, 8(2): 430-446.
70. Zheng T, Xie W, Xu L, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. International journal of medical informatics. 2017; 97: 120-127.
71. Knowledge Base workgroup of the Observational Health Data Sciences and Informatics (OHDSI) collaborative, Boyce R D, Voss E A, et al. Large-scale adverse effects related to treatment evidence standardization (LAERTES): an open scalable system for linking pharmacovigilance evidence sources with clinical data. Journal of biomedical semantics. 2017; 8: 1-15.
72. Chen Y, Lorenzi N, Nyemba S, et al. We work with them? Healthcare workers interpretation of organizational relations mined from electronic health records. International journal of medical informatics. 2014; 83(7): 495-506.
73. Bates DW, Saria S, Ohno-Machado L, et al. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Affairs. 2014; 33(7): 1123-1131.
74. Pantouvakis A, Bouranta N. Quality and price–impact on patient satisfaction. International journal of health care quality assurance. 2014; 27(8): 684-696.
75. Van Onsem S, Van Der Straeten C, Arnout N, et al. A new prediction model for patient satisfaction after total knee arthroplasty. The Journal of arthroplasty. 2016; 31(12): 2660-2667. e1.
76. Chen Y, Xie W, Gunter C A, et al. Inferring clinical workflow efficiency via electronic medical record utilization. AMIA Annual Symposium Proceedings. American Medical Informatics Association. 2015; 2015: 416.
77. Tian H, Chen GH, Xu Y, et al. Impact of pre-transplant disease burden on the outcome of allogeneic hematopoietic stem cell transplant in refractory and relapsed acute myeloid leukemia: a single-center study. Leukemia & lymphoma. 2015; 56(5): 1353-1361.
78. Goyal M, Jadhav AP, Bonafe A, et al. Analysis of workflow and time to treatment and the effects on outcome in endovascular treatment of acute ischemic stroke: results from the SWIFT PRIME randomized controlled trial. Radiology. 2016; 279(3): 888-897.
79. Mundt MP, Gilchrist VJ, Fleming MF, et al. Effects of primary care team social networks on quality of care and costs for patients with cardiovascular disease. The Annals of Family Medicine. 2015; 13(2): 139-148.
80. Genc G, Abboud H, Oravivattanakul S, et al. Socioeconomic status may impact functional outcome of deep brain stimulation surgery in Parkinson's disease. Neuromodulation: Technology at the Neural Interface. 2016; 19(1): 25-30.

81. Xu Y, Lu M, Li N, et al. Exploring the Impact of Health Insurance on Health Care Utilization and Outcome Using Electronic Medical Record Data. International Journal for Population Data Science. 2017; 1(1).

82. Safford MM. A New Chapter in Patient-Centered Care: Sharing the Medical Note?. Annals of internal medicine. 2018; 168(4): 298.

83. Rathert C, Wyrwich MD, Boren SA. Patient-centered care and outcomes: a systematic review of the literature. Medical Care Research and Review. 2013; 70(4): 351-379.

84. Reiss-Brennan B, Brunisholz KD, Dredge C, et al. Association of integrated team-based care with health care quality, utilization, and cost. Jama. 2016; 316(8): 826-834.

85. Yan C, Chen Y, Li B, et al. Learning Clinical Workflows to Identify Subgroups of Heart Failure Patients. AMIA Annual Symposium Proceedings. American Medical Informatics Association. 2016; 2016: 1248.

86. Chen Y, Ghosh J, Bejan C A, et al. Building bridges across electronic health record systems through inferred phenotypic topics. Journal of biomedical informatics. 2015; 55: 82-93.

87. Mehta RL, Awdishu L, Davenport A, et al. Phenotype standardization for drug-induced kidney disease. Kidney international. 2015; 88(2): 226-234.

88. Veltman JA, Vissers LELM. Standardized phenotyping enhances Mendelian disease gene identification. Nature genetics. 2015; 47(11).

89. Unertl KM, Weinger MB, Johnson KB, Lorenzi NM. Describing and modeling workflow and information flow in chronic disease care. J Am Med Inform Assoc. 2009; 16(6), 826-836.

90. Zang Y, Zhang F, Di C, et al. Advances of flexible pressure sensors toward artificial intelligence and health care applications. Materials Horizons. 2015; 2(2): 140-156.

91. Wong T Y, Bressler N M. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. JAMA. 2016; 316(22): 2366-2367.

92. Weber A. Amazon Echo: The Best User Guide to Master Amazon Echo Fast. 2016.

93. Chung H, Park J, Lee S. Digital forensic approaches for Amazon Alexa ecosystem. Digital Investigation. 2017; 22: S15-S25.

94. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. Acoustics, speech and signal processing. IEEE international conference on. 2013: 6645-6649.

95. Zhu Z. Application of Geographical Information System and Interactive Data Visualization in Healthcare Decision Making. International Journal of Big Data and Analytics in Healthcare (IJBDAH). 2016; 1(1): 49-58.

96. Shneiderman B, Plaisant C, Hesse B W. Improving healthcare with interactive visualization. Computer. 2013; 46(5): 58-66.

97. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 05/22/2018.

98. Mili FD, Allen T, Wadell PW, et al. VKORC1-1639A allele influences warfarin maintenance dosage among Blacks receiving warfarin anticoagulation: a retrospective cohort study. Future cardiology. 2018; 14(1): 15-26.

99. Yin W, Gao C, Xu Y, Li B, Ruderfer D, Chen Y. Learning Opportunities for Drug Repositioning via GWAS and PheWAS Findings. AMIA 2018 Informatics Summit. 237-246.

100. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010; 26(9): 1205-1210.

101. Prueksaritanont T, Chu X, Gibson C, et al. Drug–drug interaction studies: regulatory guidance and an industry perspective. The AAPS journal. 2013; 15(3): 629-645.

102. Yu J, Zhou Z, Tay-Sontheimer J, et al. Risk of Clinically Relevant Pharmacokinetic-based Drug-drug Interactions with Drugs Approved by the US Food and Drug Administration Between 2013 and 2016. Drug Metabolism and Disposition. 2018: dmd. 117.078691.

103. Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. Journal of the American Medical Informatics Association. 2014; 21(e2): e278-e286.

104. Li T, Gao C, Yan C, Osmundson S, Malin B, Chen Y. Predicting Neonatal Encephalopathy From Maternal Data in Electronic Medical Records. AMIA 2018 Informatics Summit. 359-368.

105. Greenwald L. Provider and Consumer Engagement in Care Coordination Models: Evidence from the US. International Journal of Integrated Care. 2015; 15(5).

106. Vijayakumar P, Nelson R G, Hanson R L, et al. HbA1c and the prediction of type 2 diabetes in children and adults. Diabetes care. 2017; 40(1): 16-21.

107. Läll K, Mägi R, Morris A, et al. Personalized risk prediction for type 2 diabetes: The potential of genetic risk scores. Genetics in Medicine. 2017, 19(3): 322.

108. Jain A, Ratnoo S, Kumar D. Addressing class imbalance problem in medical diagnosis: A genetic algorithm approach. IEEE International Conference on Information, Communication, Instrumentation and Control. 2017: 1-8.

109. García MNM, Herráez JCB, Barba MS, et al. Random Forest Based Ensemble Classifiers for Predicting Healthcare-Associated Infections in Intensive Care Units. Distributed Computing and Artificial Intelligence. 13th International Conference. Springer, Cham, 2016: 303-311.

110. Sáez JA, Luengo J, Stefanowski J, et al. SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Information Sciences. 2015; 291: 184-203.