Review

Opportunities and Challenges in Data-Driven Healthcare Research

You Chen1,*

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.
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.

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 the 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 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 amounts 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 in 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...
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 patterns 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.
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


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


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


