Interactive Integrated Electronic Health Records

A literature review submitted

by

Shane Crinion

to

The Discipline of Bioinformatics,
School of Mathematics, Statistics & Applied Mathematics
National University *of* Ireland, Galway

in partial fulfilment of the requirements for the degree of

M.Sc. in Biomedical Genomics

April 5th 2019

Thesis Supervisor(s): Pilib Ó Broin

Abstract

Electronic Health Records (EHR) have evolved to become sophisticated predictive tools essential to precision medicine. Beyond data storage, EHRs have been used to provide clinical decision support (CDS), model disease and improve clinical efficiency using machine learning (ML) algorithms. Technical and ethical challenges such as unstandardised data and data security have hindered the development of integrated EHRs. Integration of genomic data can be used to identify novel genotype-phenotype associations and improve drug development as demonstrated through biobanks. The project aims to develop a Shiny dashboard using Synthea, a synthetic patient generator, and gene expression data to model an interactive, integrated EHR. The model will be applied to a complex disorder to identify novel disease patterns by combining clinical/molecular analytics.

Contents

List of Figures

Chapter 1

Introduction: Uses of EHRs

Electronic health records (EHR) have become a core element of high quality clinical practise and research [1, 2]. EHRs are defined by the Center of Medicaid and Medicare Services (CMS) as "digital forms of patient records that include patient information such as personal contact information, patient's medical history, allergies, test results, and treatment plan" [3]. Integration of longitudinal, clinical data provides unparalleled benefits to quality of care, cost reduction and predictive analysis of disease [4].

Beyond clinical care, EHRs are a powerful research tool and can provide comprehensive data for use in population studies [2, 5]. Complex disorders such as type 2 diabetes mellitus (T2D) require rich patient data to generate meaningful results and identify risk factors associated with progression [2]. By integrating clinical and genomic data, EHRs can be used to identify clinical epidemiology of biomarkers and gene expressions [6]. Given the rapid growth of genomic research, it is expected that digitalised genomes will soon be integrated into the EHR and will advent revolutionary discoveries in human health and disease [7].

This literature review will discuss the current clinical and research uses for EHR including cohort identification, prediction analysis and the developments in the integration of genomic data. The current challenges surround EHR used are discussed and followed by the project aims. The aims include development of an interactive dashboard for integrated genomic data to model disease patterns identifiable in combined clinical and molecular analytics.

1.1 Patient data storage

Early EHRs were implemented to facilitate health information exchange (HIE), the sharing of patient data across organisations, for patient chart data [8, 4]. Due to incentives by the Health Information Technology Act of 2009, over 80%

of acute care hospitals in the US have adopted EHRs with one billion hospital visits estimated to be documented each year [5]. The introduction of EHRs to the National Hip Fracture Database (NHFD) has reportedly decreased false death reports from 64/109 (59%) to 0 false death reports, demonstrating its potential for accuracy improvements [9]. Storing data in digitised format provides improved legibility which reduces medical errors when prescribing medication [4, 10]. The reduction of administration time brought by EHRs also increases the clinicians availability to interact with patients [11]. Digitally stored patient data has also made large scale population studies possible. [4]

1.2 Patient cohort identification

The identification of patient cohorts is invaluable for understanding of disease epidemiology, drug interactions and predicting outcome [12]. However EHRs were not developed for comparing patient which has resulted in numerous approaches for selecting patient cohorts [12]. Exclusion and inclusion by applying discrete traits can identify cohorts based on factors such as cancer stage, smoker status, infection and drug response [12, 13]. Selection of a certain phenotype (eg. cancer, drug side effect, hypertension) can be performed to identify adverse drug effects, drug repurposing or to validate a phenotype [14, 12]. This approach identified the potential application of sildenafil citrate from an anti-angina drug to an erectile dysfunction drug [15]. Also, patient demographics, clinical diagnosis codes such as International Classification of Diseases-Ninth Revision (ICD-9) codes or laboratory reports codes have been used for isolating cohorts [12, 16].

The integration of continuous health tracking (eg. fitness trackers, smartphones) can also be used to identify early signs of disease and cohort demographics [17]. Lower socio-economical classes with limited access to health care may benefit from statistics generated based on their lifestyle or cohort [18]. Clinical research limited by a lack of cohort diversity may also benefit from the integration of continuous tracking and data richness [19]. Additionally, integration of family data can be used to identify risk for disease and intervene before disease progression [1].

1.3 Clinical decision support (CDS)

High functioning EHRs have been utilised for the development of CDS systems to assist clinicians in making decisions [20]. CDS systems recommend actions to the clinician based on relevant EHR entries such as pop-up alerts, drug-dose calculators [20]. EHR-CDS is found to improve quality including patient blood pressure and diabetes testing [21]. CDS can also prevent medication errors by advising the clinician of drug interactions or patient allergies [10]. CDS is currently used with EHR to administer antibiotics, improve documentation and reduce negative patient side effects [10]. Its implementation has also increase following of clinical guidelines [20]. Costs can also be reduced through reduced unnecessary prescriptions and length of stay and the increased time with patients [10].

Chapter 2

Predictive Analysis

2.1 Disease modelling

Despite their simpler original purpose, EHRs have evolved to provide statistical prediction methods in healthcare [1]. Many common machine learning (ML) algorithms have been applied to clinical data for prediction, detection and prognosis of disease namely artificial neural networks (ANN), decision tree (DT), semi-supervised learning (SSL) and Bayesian networks (BN) [22]. Many data mining tools have also been applied to disease detection using EHR data [23]. Linear Discriminant Analysis (LDA) has been applied to identify cancer subtypes using gene expression data from microarray analysis [24]. Application of LDA to databases containing clinical and lab results for 168 primary immunodeficiency (PID) patients could classify their disease from over 250 congenital disorders with 83-94% accuracy, 60-100% sensitivity and 83-95% specificity [25]. An updated version of this model, Covariance-Regularized LDA (CRDA), includes a diagnosis-frequency vector and has been used to predict depression and anxiety with EHR data for 1000 US college students with higher accuracy and F1 score than other baseline algorithms [23].

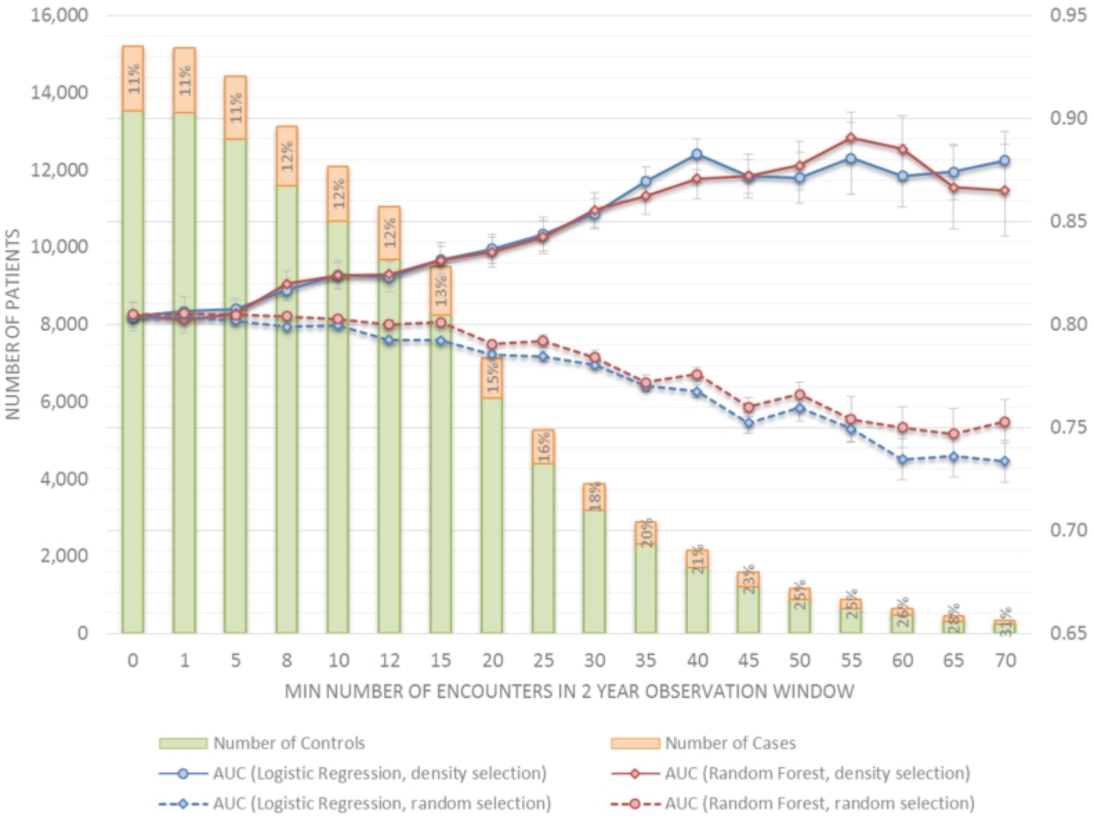

Figure 2.1: Heart failure prediction as a function of data density and size. Taken from Adkins et al. [26]

Two cases where early detection and intervention are advantageous are heart failure and suicide risk prediction [26]. A Naive Bayes Classifier (NBC) algorithm has been used to model suicide risk from 1.7 million patient records over 15 years from 2 Boston Hospitals and a limited set of ICD-9 codes. This approach could identify early risk patterns and remind the physician to contact patients at critical time (3-4 years prior to suicidal activity) [26]. ICD-9 mediated detection has also been applied to 1684 heart failure cases from Geisinger Clinic. Model predictions improved more with increased number of encounters than with improvements due to sample size (Figure 2.1). The study indicates that dense, longitudinal data is paramount to the high performance of their model [27]. Similarly a deep learning framework, Deep Patient, was developed using EHR clinical data from 704,587 patients. The disease prediction model could predict mental illness and diabetes with mean area under the curve receiver operating characteristics (AUC-ROC) of 0.6-0.863 and 0.907 respectively. These are two complex, heterogeneous disorders that can have improved diagnoses through EHR-derived data with ML algorithms. Both examples described demonstrate automated new disease pattern identification and cohort assignment [18].

ICD-9 codes could also be used to categorise rare infectious diseases with a bulk learning approach. Bulk learning analyses phenotypically similar diseases to identify the shared phenotypic components. Additionally, unique clinical subsets in critical cases are also identifiable [28]. Training was performed using 100 ICD-9 codes for infectious and parasitic diseases, demonstrating the possibility of phenotypic stratification using ICD-9 codes from EHR data [28].

2.2 Improving clinical efficiency

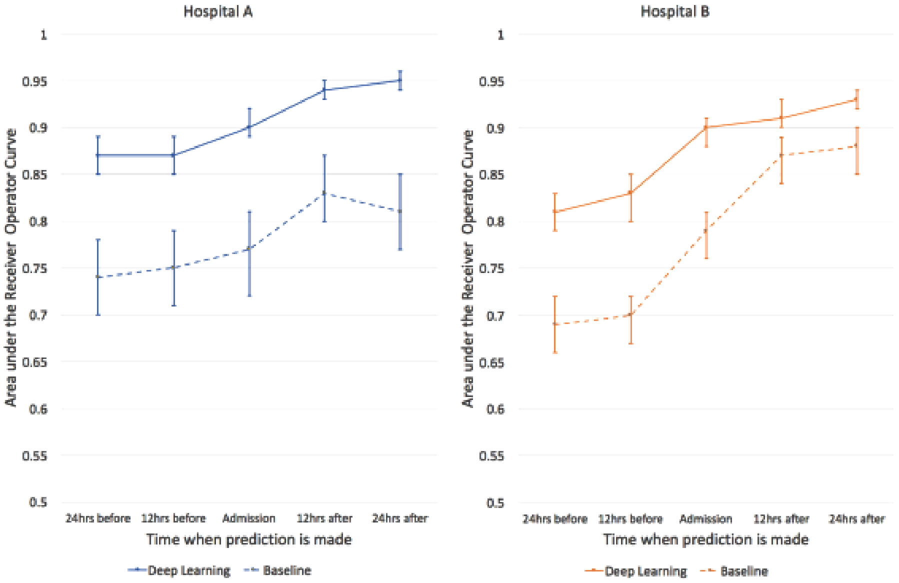

Figure 2.2: Inpatient mortality prediction indicating better results with deep learning. Taken from Rajkomar et al. [29]

Predictions made for patient mortality, ICU admissions and surveillance within hospitals can improve clinical efficiency and reduce treatment expenses [30]. Site-independent predictive modelling using deep learning models can predict medical events including in-hospital mortality, 30-day unplanned readmission and prolonged length of stay using the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) standard [29]. FHIR is a clinical standard developed for simple implementations and rigorous data exchange of EHR data [31]. The re-admission could be predicted accurately even within 30 days of initial encounter. The length of stay prediction at 24 hours after admission had an AUROC of 0.86. Discharge was predicted at admission, 24 hours after entry and at discharge with a minimum of 0.87 AUROC. Predictions using deep learning were significantly better than baseline approaches in both hospital sites (Figure 2.2).

Modified Early Warning Score (MEWS) algorithms can also accurately predict condition deterioration and plan for ICU admission. The algorithm is improved further if continuous tracking (eg. heart monitor) can also be applied [18]. MEWS and similar approaches could prevent unplanned ICU admissions and death by early recognition of warning signs [32]. Other predictive modelling algorithms include the Contrast Pattern Aided Regression Log (CPXR(Log)) logistic regression method used to accurately predict heart failure with 15.9% AUC improvement on previous models. The approach was developed in recognition of the cost reductions possible by preventing rehospitilisation [33].

Chapter 3

Challenges

3.1 Technical issues

There are many “hidden costs” that create major barriers to widespread use of EHRs including new hardware, staff training and recruitment of new employees [10, 34]. Both computationally powerful and hand-held computers are required for storage and interaction of data respectively [1]. Over 6 billion pieces of EHR data can be used by a ML algorithm, indicating the sheer amount of storage required [29]. Potential patient time is also reduced due to staff training and education [1]. New staff for installation and maintenance are required and bioinformaticians may be hired to interpret algorithms and data points used in EHR-based research [1]. All things considered, there is substantial manpower and time-consuming labour that can have immediately negative effects on the quality of clinical delivery before the benefits occur [1].

The lack of standardization has reduced the usability of EHRs for clinical research drastically [35, 29]. Only 6% of health care providers reported capability of sharing EHR data with other clinicians [35]. Interoperability is substantially limited by the specific terminology and technical specifications of each EHR entry [35]. EHR providers have even reduced data exchange capability to retain their client base [35]. Predictive statistical models requires extraction of curated variables which is extremely difficult if the test data is not standardised [29].

The extraction of EHR data for predictive modelling is hindered further by the preprocessing and cleaning required as a result of heterogeneous data [29]. Poor quality data can also result in misinformation and subsequent inappropriate treatment to the patient [1]. One solution is to select discrete variables however this can produce imprecise predictions and decrease model robustness [29, 36]. Deep learning is the most commonly used practise in machine perception and has been used for application in computer vision-to-speech recognition and NLP [29]. Deep learning does not require predictor variable specification which makes it suitable for messy unstandardised data stored in EHRs [29].

Use of the FHIR format, a flexible data structure for medical use, has been developed to simplify data inter-exchange between EHRs [29]. Another proposed solution is the Arden synthax, a HL7 recognised standard that structures patient data into self contained Medical Logic Modules (MLM). The coding is similar to natural language which improves usability and uses curly braces to improve operation transferability [37, 1, 38]. These MLMs have even contained connectors that allow contained processing of FHIR data [37].

3.2 Ethical issues

Ethical concerns exist regarding the access, interpretation and usage of clinical data [1, 39]. 34,000 patients were effect by a data breach in 2013 when a medical technician sold patient data which was accessed using EHRs [39]. There is also potential for increased medical identity theft if a patient is billed for medical services they did not receive [39, 1].

Insurance companies will require access to EHR data and the limitation of data access to authorized personnel only [39]. Using ML algorithms for “black box” predictions such as suicide risk will require careful weighing of the adverse effects against the potential benefits given the model is robust [26]. A national computer network in Singapore has been useful for data exchange however creates ambiguity regarding data ownership and informed consent [1]. In response to legal and ethical limitations of patient data, artificial patient generators such as Synthea have been developed for disease modelling and biomedical software development [40].

Despite the ethical concerns, 84 % of individuals reported support for large scale genomic data studies [36]. Privacy is a major concern for patients and this a key point to consider when improving healthcare. A potential answer to these concerns may be obligatory personal benefit in the form of accessible health information. Transparency with the patient (by personal access to EHR data) can be beneficial for patient and population health through increased awareness [41].

3.3 Integration of genomic data

A genomic EHR is widely considered as paramount to improvement of diagnosis and treatment [42, 43, 44, 26]. Merging genomic and clinical data can map genetic variants to complex heterogeneous disorders to identify novel biological pathways and drug targets [45]. The analysis of genome wide association studies (GWAS) and phenome-wide association studies (PheWAS) has been used to identify actionable genomic findings for complex disorders including cardio-metabolic traits and T2D [46, 47]. Successful integration of genomic data will likely eliminate trial and error approaches to treatment and facilitate precision medicine [46]. This potential for genomic EHRs has been recognised through the Electronic Medical Records and Genomics (eMERGE) Network initiative.

However, many of technical and ethical issues associated with EHRs are exasperated by the introduction of genomic data. Although costs are decreasing swiftly, it is still expensive to store very large gene sequence files [7]. The existing EHR structure must be redesigned to integrate genomic data successfully [45, 1]. Genomic data remains valid for the lifespan of the patient unlike other data which means greater security should also be implemented for its storage [45].

Although one study reported 84% for support large scale genomic data research, another reported that 52% expressed serious concerns about providing results to non-health-related purposes [36, 48]. In ethical terms, genetic discrimination is a possibility if life insurance underwriters are provided with genomic data on an individual's risk for potential diseases [48]. Without proper training, risk variants may be thought of as causal variants and influence treatment negatively. Access to genomic results will require data privileges to prevent medical errors [49]. Proper interpretation training is also needed to ensure that descriptive or predictive analytics are applied appropriately to gain meaningful results [11].

3.3.1 Data repositories

Biobanks are a valuable resource in personalised medicine that are used for population stratification, biomarker discovery and drug development [50]. Large scale repositories can build representative models of diseases or populations [51]. The ability to reuse patient data to investigate a range of phenotypes may be most beneficial use of biobanks [43]. Additionally, reverse genetics approaches using biobank data such as PheWAS can build rich genetic profiles by identifying phenotypes associated with a given variant [52].

The eMERGE network, the major EHR DNA biobank, contains genotyped data for 83,717 samples from 9 geographically distinct groups. eMERGE has served as a data model for development and best practise of EHRs integrated with genomic data [53]. All entries can be repurposed for use as cases/controls in any future studies and dismiss the need for new study participants [53]. The network has identified 39 million variants and improved categorisation of disease phenotypes by using phenotypic data from EHR and GWAS [53, 43]. The DiscovEHR collaborative used 50,000 participants to identify more than 4 million SNVs and combined longitudinal EHR with genomic data [18]. Gene expression data has also been integrated into EHRs to identify disease-specific gene expression [54]. A logistic regression model was used to identify reduced 5 year mortality using EHRs from 9945 breast cancer patients. The 14 case-control gene expression datasets were then used to identify 3 drug class pairs that were associated with lower mortality by their increased expression patterns [54].

3.3.2 Applications

Longitudinal EHR data has been used to improve cardiovascular disease (CVD) predictions by incorporating ML methods. These results were improved further when genetic data from the Vanderbilt BioVu biobank was integrated. Integration of ML methods and genetic data facilitated identification of age and two *MIA3* gene variants as the most reliable predictors for CVD [55]. Integration of clinical and genomic data also identified single nucleotide polymorphisms (SNP) in 9p21 associated with early myocardial infarction and CVD [36].

Gene expression signatures associated with recurrence ($p = 0.04$) and poor survival ($p = 0.004$) in hepatocellular carcinoma (HCC) could be identified when integrated with clinical and pathology data [19]. Another study using lab results from EHRs could be transmitted using FHIR standards to map 2421 lab tests to Human Phenotype Ontologies (HPO). The annotation was validated using EHR data from 15,861 respiratory complaint patients and by matching to asthma biomarkers. This also demonstrates ability to apply FHIR standards to connect medication outcomes with biomarkers [56].

The heterogeneous nature of diabetes warrants a range of measurements (eg. blood glucose, foot health, eye health, cardiovascular health) to understand the severity and symptoms of the patient [57]. These measurements may be avoided if phenotype-variation associations are identified. Many variants have been found in protein coding and non-coding regions that are associated with diabetes [46]. A T2D topilogical network has been able to identify 3 distinct clusters, of which one was also enriched for cancer malignancies [18]. These are prime examples of the new associations that can be identified by integrating clinical and genomic data.

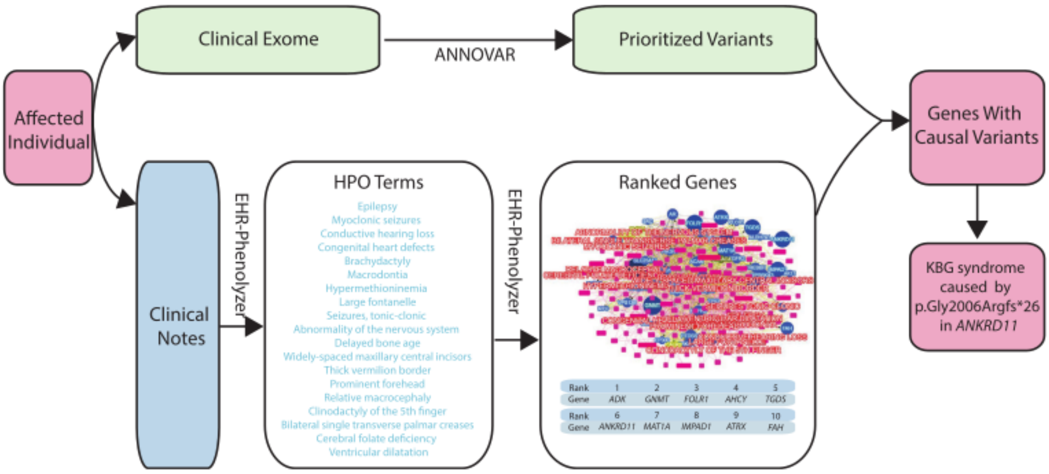

Figure 3.1: Deep Phenotyping approach to genes ranking and causal variant identification. Taken from Son et al. [58]

A Deep Phenotyping approach called EHR-Phenolyzer has been able to prioritize known causal genes for 16 of 28 monogenic disease cases based solely on their HPO concept from EHRs [58] (Figure 3.1). Two aims for the approach were (1) to prove the capability of identifying genotype-phenotype relationships and (2) to test capability of NLP to extract EHR data from records and select candidate genes. The gene-ranking ability of NLP methods was comparable to human expert ranking capability. The approach demonstrates the use of NLP phenotype extraction to link causal genes to EHR phenotypes. The method also demonstrates flexibility by using four different cohorts to ensure data exchange capability [58].

Chapter 4

Project Aims

This project involves the development of a Shiny application to analyse and interact with clinical data. Synthea will be used to model a disease cohort and perform predictive analytics. Gene expression data will then be downloaded to provide proof-of-concept for clinical and molecular analytics. The objective of the project is to develop an interactive genomic health record that can be used to obtain statistical data at a patient and cohort level.

4.1 Develop an interactive dashboard

4.1.1 R Shiny

R Shiny (https://shiny.rstudio.com/) is an R package used for web application framework and is useful for building interactive web applications for biomedical research [31]. Dashboards are easily designed and can be used on all operating systems which is profoundly important for successful integration of EHR data [11].

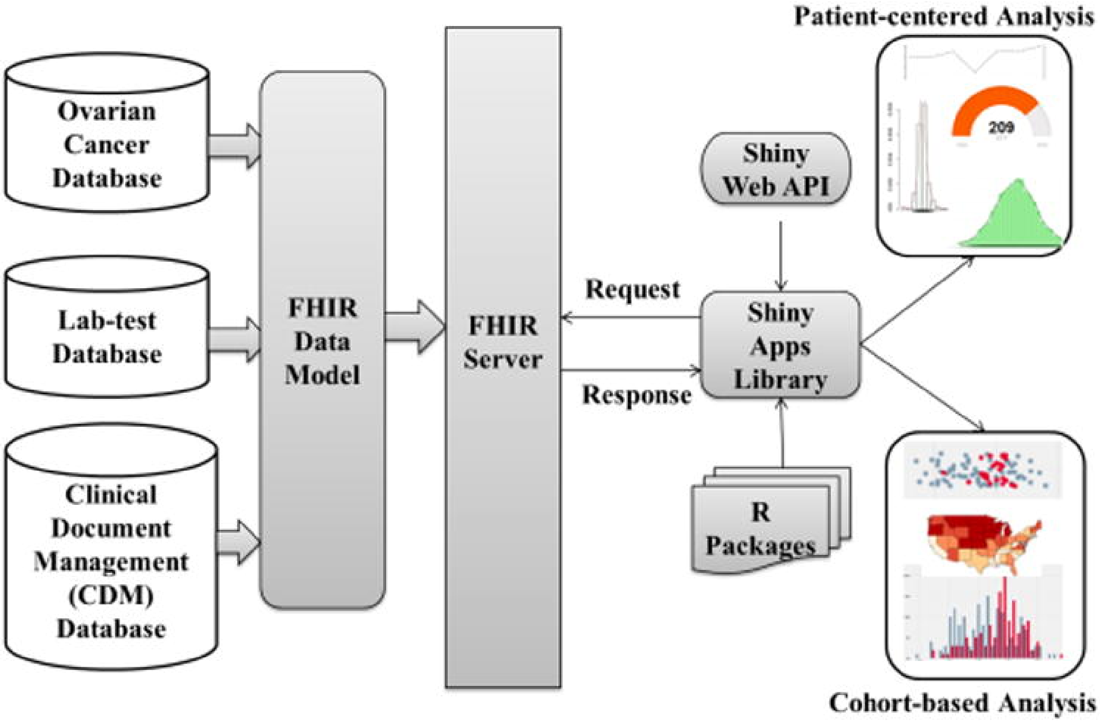

Figure 4.1: Integrated framework of clinical data using FHIR and Shiny. Taken from Hong et al. [31]

Shiny has been used with FHIR data to display interactive interfaces using ovarian-cancer datasets [31]. The framework named Shiny FHIR consisted of three modules: a FHIR server module, a workflow module for patient-centred and cohort based data analysis and an interactive user interface module connecting the Shiny web framework and FHIR server (Figure 4.1). The study demonstrates the feasibility of applying a common data exchange standard to an interactive dashboard created using Shiny [31]. EHDViz is another Shiny application that integrates EHR data with predictive models and provide visualisations for improving healthcare [59].

These examples indicate that R Shiny will be capable of displaying FHIR data and is suitable for handling clinical data.

4.1.2 Synthea

Synthea (https://synthetichealth.github.io/synthea) is an open-source package containing synthetic EHRs encoded in FHIR standard [40]. Synthea models the lifespan of patients based on the top 10 chronic conditions and reasons for medical care encounters in the US. The objective of Synthea is to address the legal and

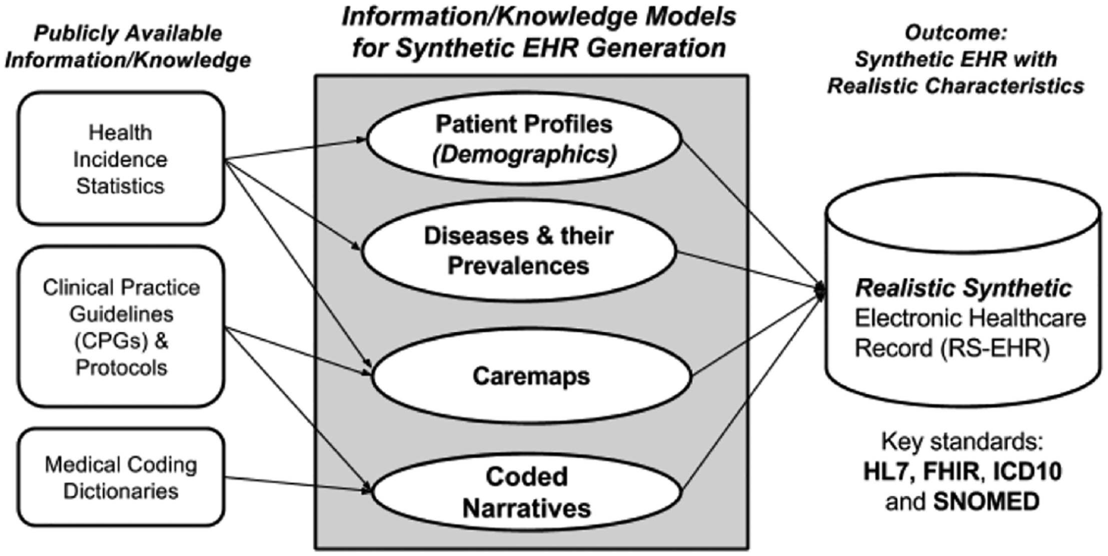
Publicly Available Information/Knowledge
Information/Knowledge Models for Synthetic EHR Generation
Outcome: Synthetic EHR with Realistic Characteristics
Health Incidence Statistics
Clinical Practice Guidelines (CPGs) & Protocols
Medical Coding Dictionaries
Patient Profiles (Demographics)
Diseases & their Prevalences
Caremaps
Coded Narratives
Realistic Synthetic Electronic Healthcare Record (RS-EHR)
Key standards: HL7, FHIR, ICD10 and SNOMED

Figure 4.2: PADASER framework for Synthea. Taken from Walonoski et al. [40]

ethical limitations that has caused lagging in health record technology [40]. The framework for Synthea is based on the Publicly Available Data Approach to the Realistic Synthetic Data (PADARSER) (Figure 4.2). The model uses publicly available health statistics as inputs to generate data from clinical workflow and disease progression. Finally, the model includes a temporal model to provide a complete profile for the patient beyond the disease of interest [40]. The longitudinal model is ideal for modelling disease progression and performing population analysis.

A previous study developed a predictor for opioid overdose using Shiny with Synthea data and ML algorithms to identify risk factors such as age, sex or address.

These studies demonstrate that Synthea can be used with the FHIR framework and for extraction of risk variants. The studies also show that Synthea data is suitable for Shiny app development when using predictive analytics. Similarly, this project will develop an application using Shiny and Synthea to identify risk factors for chronic disease.

4.2 Incorporate gene expression data

Integration of gene expression can identify disease specific gene expression and biomarkers [54]. Similarly to clinical data heterogeneity, the integration of genomic data is effected by disease polygenicity. This could be answered by the intergration of genomic data. To generate an ethical model of this, I will integrate GEO data for the synthetic diabetes patients generated above.

Gene expression data can be integrated with artificially generated patients to model patterns associated with disease.

4.2.1 Gene Expression Omnibus (GEO)

The Gene Expression Omnibus is an online public repository containing gene expression data that is publicly available for clinical research [60]. GEO accepts data of many forms and specifies criteria to allow an integrative design for large scale analysis of raw and processed data. The reusing of GEO facilitates genomic data integration and is useful in identifying gene expression to phenotype patterns. The heterogenous nature of T2D means that many patients do not respond well to certain drugs. Genetic variants associated with positive drug response may be identifiable by disease modelling using Synthea and GEO [46]. GEO has been used to study gene expression and methylation patterns in T2D patterns and identified 47 upregulated and 56 downregulated genes associated with fatty acid and glucose metabolic pathways [61]. shinyGEO is a web application that allows gene expression data analysis including differential expression analysis [62].

Gene expression data will be integrated with Synthea generated patients to model gene expression variation associated with disease. The project will provide a framework for combined clinical and molecular analytics without legal or ethical restrictions.

Chapter 5

Conclusion

The potential benefits from integration of clinical and genomic data have been recognised by the Precision Medicine Initiative, an investment of 215 million US dollars towards tailored healthcare [18]. Despite the benefits of an EHR, only 13% of respondents to a world wide study reported having an EHR in place [1]. By storing patient data on EHRs, less medical errors occur through higher quality data and CDS [58]. Predictive analysis has improved disease modelling, increased efficiency and lowered costs [1]. Challenges exist as a lack of standardisation has prevented integration of EHRs [35]. Clinical characterisations such as ICD-9 are sometimes considered insufficient for accurate diagnosis [12]. The integration of genomic data to EHRs will improve current knowledge of phenotype-variant associations and the robustness of clinical diagnosis. This will be especially useful understand the heterogeneous of complex diseases such as T2D. It can be expected that genomic EHRs will evolve with biobanks and machine learning developments as biobanks will provide the appropriate frameworks for genomic EHR implementation and ML methods to improve integration. Integrated EHRs are expected to produce significantly more meaningful results as they develop and to facilitate the widespread shift in healthcare to precision medicine.

Bibliography

[1] R. S. Evans. Electronic health records: Then, Now, and in the Future. *Yearb Med Inform*, Suppl 1:48–61, May 2016.

[2] Kipp W. Johnson, Joel T. Dudley, and Benjamin S. Glicksberg. The next generation of precision medicine: observational studies, electronic health records, biobanks and continuous monitoring. *Human molecular genetics*, 27(R1):R56–R62, April 2018.

[3] Clemens Scott Kruse, Anna Stein, Heather Thomas, and Harmander Kaur. The use of electronic health records to support population health: A systematic review of the literature. *Journal of medical systems*, 42(11):214–214, September 2018.

[4] Nir Menachemi and Taleah H. Collum. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, 4:47–55, May 2011.

[5] M. K. Ross, W. Wei, and L. Ohno-Machado. ”big data” and the electronic health record. *Yearbook of medical informatics*, 9(1):97–104, August 2014.

[6] Jonathan D. Mosley, QiPing Feng, Quinn S. Wells, Sara L. Van Driest, Christian M. Shaffer, Todd L. Edwards, Lisa Bastarache, Wei-Qi Wei, Lea K. Davis, Catherine A. McCarty, Will Thompson, Christopher G. Chute, Gail P. Jarvik, Adam S. Gordon, Melody R. Palmer, David R. Crosslin, Eric B. Larson, David S. Carrell, Iftikhar J. Kullo, Jennifer A. Pacheco, Peggy L. Peissig, Murray H. Brilliant, James G. Linneman, Bahram Namjou, Marc S. Williams, Marylyn D. Ritchie, Kenneth M. Borthwick, Shefali S. Verma, Jason H. Karnes, Scott T. Weiss, Thomas J. Wang, C. Michael Stein, Josh C. Denny, and Dan M. Roden. A study paradigm integrating prospective epidemiologic cohorts and electronic health records to identify disease biomarkers. *Nature Communications*, 9(1):3522, August 2018.

[7] Jennifer Kulynych and Henry T. Greely. Clinical genomics, big data, and electronic medical records: reconciling patient rights with research when pri-

vacy and science collide. *Journal of law and the biosciences*, 4(1):94–132, January 2017.

[8] Robert J. Carroll, Anne E. Eyler, and Joshua C. Denny. Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis. *Expert review of clinical immunology*, 11(3):329–337, March 2015.

[9] Christopher R. Gooding, Duncan Cundall-Curry, John E. Lawrence, Max E. Stewart, and Daniel M. Fountain. The use of an electronic health record system reduces errors in the national hip fracture database. *Ageing*, 48(2):285–290, March 2019.

[10] Brian Rothman, Joan C. Leonard, and Michael M. Vigoda. Future of electronic health records: Implications for decision support. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 79(6):757–768, 2012.

[11] Fadoua Khennou, Youness Idrissi Khamlichi, and Nour El Houda Chaoui. Improving the use of big data analytics within electronic health records: A case study based openehr. *Proceedings of the first international conference on intellengent computing in data sciences*, 127:60–68, January 2018.

[12] Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association*, 21(2):221–230, 2013.

[13] Jessica Watson, Brian D Nicholson, Willie Hamilton, and Sarah Price. Identifying clinical features in primary care electronic health record studies: methods for codelist development. *BMJ Open*, 7(11), 2017.

[14] Lu Wang, Scott M Damrauer, Hong Zhang, Alan X Zhang, Rui Xiao, Jason H Moore, and Jinbo Chen. Phenotype validation in electronic health records based genetic association studies. *Genetic epidemiology*, 41(8):790–800, 2017.

[15] MR Hurle, L Yang, Q Xie, DK Rajpal, P Sanseau, and P Agarwal. Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics*, 93(4):335–341, 2013.

[16] Wei-Qi Wei, Lisa A Bastarache, Robert J Carroll, Joy E Marlo, Travis J Osterman, Eric R Gamazon, Nancy J Cox, Dan M Roden, and Joshua C Denny. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PloS one*, 12(7):e0175508, 2017.

[17] Nicholas Genes, Samantha Violante, Christine Cetrangol, Linda Rogers, Eric E Schadt, and Yu-Feng Yvonne Chan. From smartphone to EHR: a case report on integrating patient-generated health data. *NPJ Digital Medicine*, 1(1):23, 2018.

[18] Venet Osmani, Li Li, Matteo Danieletto, Benjamin Glicksberg, Joel Dudley, and Oscar Mayora. Processing of electronic health records using deep learning: A review. 2013.

[19] A. Villanueva, Y. Hoshida, C. Battiston, V. Tovar, D. Sia, C. Alsinet, H. Cornella, A. Liberzon, M. Kobayashi, H. Kumada, S. N. Thung, J. Bruix, P. Newell, C. April, J. B. Fan, S. Roayaie, V. Mazzaferro, M. E. Schwartz, and J. M. Llovet. Combining clinical, pathology, and gene expression data to predict recurrence of hepatocellular carcinoma. *Gastroenterology*, 140(5):1501–1512, May 2011.

[20] Jason H. Lam and Olivia Ng. Monitoring clinical decision support in the electronic health record. *American Journal of Health-System Pharmacy*, 74(15):1130–1133, 08 2017.

[21] Rebecca G Mishuris, Jeffrey A Linder, David W Bates, and Asaf Bitton. Using electronic health record clinical decision support is associated with improved quality of care. *Am J Manag Care*, 20(10):e445–e452, 2014.

[22] Andrew J. Steele, Spiros C. Denaxas, Anoop D. Shah, Harry Hemingway, and Nicholas M. Luscombe. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PloS one*, 13:e0202344, 2018.

[23] J. Bian, L. E. Barnes, G. Chen, and H. Xiong. Early detection of diseases using electronic health records data and covariance-regularized linear discriminant analysis. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 457–460, 16-1.

[24] Desheng Huang, Yu Quan, Miao He, and Baosen Zhou. Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data. *Journal of experimental & clinical cancer research*, 28(1):149–149, December 2009.

[25] Chiharu Murata, Ana Belén Ramírez, Guadalupe Ramírez, Alonso Cruz, José Luis Morales, and Saúl Oswaldo Lugo-Reyes. Discriminant analysis to predict the clinical diagnosis of primary immunodeficiencies: a preliminary report. *Revista Alergia México*, 62(2):125–133, 2015.

[26] Daniel E. Adkins. Machine learning and electronic health records: A paradigm shift. *The American journal of psychiatry*, 174(2):93–94, February 2017.

[27] Ng Kenney, R. Steinhubl Steven, deFilippi Christopher, Dey Sanjoy, and F. Stewart Walter. Early detection of heart failure using electronic health records. *Circulation: Cardiovascular Quality and Outcomes*, 9(6):649–658, November 2016.

[28] Po-Hsiang Chiu and George Hripcsak. EHR-based phenotyping: Bulk learning and evaluation. *Journal of biomedical informatics*, 70:35–51, Jun 2017.

[29] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, May 2018.

[30] Zilma Silveira Nogueira Reis, Thais Abreu Maia, Milena Soriano Marcolino, Francisco Becerra-Posada, David Novillo-Ortiz, and Antonio Luiz Pinho Ribeiro. Is there evidence of cost benefits of electronic medical records, standards, or interoperability in hospital information systems? overview of systematic reviews. *JMIR medical informatics*, 5(3), 2017.

[31] Na Hong, Naresh Prodduturi, Chen Wang, and Guoqian Jiang. Shiny FHIR: An integrated framework leveraging Shiny R and HL7 FHIR to empower standards-based clinical data applications. *Studies in health technology and informatics*, 245:868–872, 2017.

[32] Ila D Mapp, Leslie L Davis, and Heidi Krowchuk. Prevention of unplanned intensive care unit admissions and hospital mortality by early warning systems. *Dimensions of Critical Care Nursing*, 32(6):300–309, 2013.

[33] Vahid Taslimitehrani, Guozhu Dong, Naveen L Pereira, Maryam Panahiazar, and Jyotishman Pathak. Developing EHR-driven heart failure risk prediction models using CPXR (log) with the probabilistic loss function. *Journal of biomedical informatics*, 60:260–269, 2016.

[34] Leslie Beard, Rebecca Schein, Dante Morra, Kumanan Wilson, and Jennifer Keelan. The challenges in making electronic health records accessible to patients. *Journal of the American Medical Informatics Association*, 19(1):116–120, 2011.

[35] Miriam Reisman. Ehrs: the challenge of making electronic data usable and interoperable. *Pharmacy and Therapeutics*, 42(9):572, 2017.

[36] Hall et al. Merging electronic health record data and genomics for cardiovascular research: A science advisory from the american heart association. *Circulation. Cardiovascular genetics*, 9(2):193–202, April 2016.

[37] Klaus-Peter Adlassnig, Peter Haug, and Robert A Jenders. Arden syntax: Then, now, and in the future. *Artificial intelligence in medicine*, 92:1, 2018.

[38] Matthias Samwald, Karsten Fehre, Jeroen de Bruin, and Klaus-Peter Adlassnig. The arden syntax standard for clinical decision support: Experiences and directions. *Journal of Biomedical Informatics*, 45(4):711 – 718, 2012. Translating Standards into Practice: Experiences and Lessons Learned in Biomedicine and Health Care.

[39] Fouzia F Ozair, Nayer Jamshed, Amit Sharma, and Praveen Aggarwal. Ethical issues in electronic health records: A general overview. *Perspectives in clinical research*, 6(2):73, 2015.

[40] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc*, Aug 2017.

[41] Sherry-Ann N. Brown, Hayan Jouni, and Iftikhar J. Kullo. Electronic health record access by patients as an indicator of information seeking and sharing for cardiovascular health promotion in social networks: Secondary analysis of a randomized clinical trial. *Preventive Medicine Reports*, 13:306–313, March 2019.

[42] Zornitza Stark, Lena Dolman, Teri A. Manolio, Brad Ozenberger, Sue L. Hill, Mark J. Caulfied, Yves Levy, David Glazer, Julia Wilson, Mark Lawler, Tiffany Boughtwood, Jeffrey Braithwaite, Peter Goodhand, Ewan Birney, and Kathryn N. North. Integrating genomics into healthcare: A global responsibility. *The American Journal of Human Genetics*, 104(1):13 – 20, 2019.

[43] Joshua C. Denny. Mining electronic health records in the genomics era. *PLOS Computational Biology*, 8(12):e1002823, December 2012.

[44] Meghana V. Kashyap, Michael Nolan, Marc Sprouse, Ranajit Chakraborty, Deanna Cross, Rhonda Roby, and Jamboor K. Vishwanatha. Role of genomics in eliminating health disparities. *Journal of carcinogenesis*, 14:6–6, September 2015.

[45] Joseph L. Kannry and Marc S. Williams. Integration of genomics into the electronic health record: mapping terra incognita. *Genetics in medicine: official journal of the American College of Medical Genetics*, 15(10):757–760, October 2013.

[46] James S. Floyd and Bruce M. Psaty. The application of genomics in diabetes: Barriers to discovery and implementation. *Diabetes Care*, 39(11):1858–1869, 2016.

[47] Brooke N Wolford, Cristen J Willer, and Ida Surakka. Electronic health records: the next wave of complex disease genetics. *Human molecular genetics*, 27(R1):R14–R21, 2018.

[48] Yann Joly, Hilary Burton, Bartha Maria Knoppers, Ida Ngueng Feze, Tom Dent, Nora Pashayan, Susmita Chowdhury, William Foulkes, Alison Hall, Pavel Hamet, et al. Life insurance: genomic stratification and risk classification. *European Journal of Human Genetics*, 22(5):575, 2014.

[49] Marc S. Williams. The genomic health record: Current status and vision for the future. In Reed E. Pyeritz, Bruce R. Korf, and Wayne W. Grody, editors, *Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics (Seventh Edition)*, pages 315–325. Academic Press, January 2019.

[50] Judita Kinkorová. Biobanks in the era of personalized medicine: objectives, challenges, and innovation. *EPMA Journal*, 7(1):4, 2016.

[51] Yvonne G De Souza and John S Greenspan. Biobanking past, present and future: responsibilities and benefits. *AIDS (London, England)*, 27(3):303, 2013.

[52] Joshua C Denny, Lisa Bastarache, and Dan M Roden. Phenome-wide association studies as a tool to advance precision medicine. *Annual review of genomics and human genetics*, 17:353–373, 2016.

[53] Omri Gottesman, Helena Kuivaniemi, Gerard Tromp, W Andrew Faucett, Rongling Li, Teri A Manolio, Saskia C Sanderson, Joseph Kannry, Randi Zinberg, Melissa A Basford, et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genetics in Medicine*, 15(10):761, 2013.

[54] Yen S Low, Aaron C Daugherty, Elizabeth A Schroeder, William Chen, Tina Seto, Susan Weber, Michael Lim, Trevor Hastie, Maya Mathur, Manisha Desai, et al. Synergistic drug combinations from electronic health records and gene expression. *Journal of the American Medical Informatics Association*, 24(3):565–576, 2016.

[55] Juan Zhao, QiPing Feng, Patrick Wu, Roxana A Lupu, Russell A Wilke, Quinn S Wells, Joshua C Denny, and Wei-Qi Wei. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Scientific reports*, 9(1):717, 2019.

[56] Xingmin Aaron Zhang, Amy Yates, Nicole Vasilevsky, J. P. Gourdine, Leigh C. Carmody, Daniel Danis, Marcin P. Joachimiak, Vida Ravanmehr, Emily R. Pfaff, James Champion, Kimberly Robasky, Hao Xu, Karamarie Fecho, Nephi A. Walton, Richard Zhu, Justin Ramsdill, Chris Mungall, Sebastian Kohler, Melissa A. Haendel, Clem McDonald, Daniel J. Vreeman, David B. Peden, Christopher G. Chute, and Peter N. Robinson. Semantic integration of clinical laboratory tests from electronic health records for deep phenotyping and biomarker discovery. *bioRxiv*, page 519231, January 2019.

[57] Tim Robbins, Sarah N. Lim Choi Keung, Sailesh Sankar, Harpal Randeva, and Theodoros N. Arvanitis. Diabetes and the direct secondary use of electronic health records: Using routinely collected and stored data to drive research and understanding. *Digital health*, 4:2055207618804650–2055207618804650, October 2018.

[58] Jung Hoon Son, Gangcai Xie, Chi Yuan, Lyudmila Ena, Ziran Li, Andrew Goldstein, Lulin Huang, Liwei Wang, Feichen Shen, Hongfang Liu, Karla Mehl, Emily E. Groopman, Maddalena Marasa, Krzysztof Kiryluk, Ali G. Gharavi, Wendy K. Chung, George Hripcsak, Carol Friedman, Chunhua Weng, and Kai Wang. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. *American journal of human genetics*, 103:58–73, Jul 2018.

[59] Marcus A Badgeley, Khader Shameer, Benjamin S Glicksberg, Max S Tomlinson, Matthew A Levin, Patrick J McCormick, Andrew Kasarskis, David L Reich, and Joel T Dudley. EHDViz: clinical dashboard development using open-source technologies. *BMJ open*, 6(3):e010579, 2016.

[60] Emily Clough and Tanya Barrett. The Gene Expression Omnibus database. *Methods in molecular biology (Clifton, N.J.)*, 1418:93–110, 2016.

[61] Juan Shen and Bin Zhu. Integrated analysis of the gene expression profile and dna methylation profile of obese patients with type 2 diabetes. *Molecular medicine reports*, 17(6):7636–7644, June 2018.

[62] Jasmine Dumas, Michael A Gargano, and Garrett M Dancik. shinyGEO: a web-based application for analyzing gene expression omnibus datasets. *Bioinformatics*, 32(23):3679–3681, 2016.