CHAPTER ONE

# Introduction of medical genomics and clinical informatics integration for p-Health care

**Li Tong, Hang Wu, May D. Wang, and Geoffrey Wang***
The Research Center for Biomedical Information Technology, SIAT, Shenzhen, China
*Corresponding author: e-mail address: jeff.wang@gatech.edu

## Contents

## Abstract

Achieving predictive, precise, participatory, preventive, and personalized health (abbreviated as p-Health) requires comprehensive evaluations of an individual's conditions captured by various measurement technologies. Since the 1950s, analysis of care providers' and physicians' notes and measurement data by computers to improve healthcare delivery has been termed clinical informatics. Since the 2010s, wide adoptions of Electronic Health Records (EHRs) have greatly improved clinical informatics development with fast growing pervasive wearable technologies that continuously capture the human physiological profile in-clinic (EHRs) and out-of-clinic (PHRs or Personal Health Records) to bolster mobile health (mHealth). In addition, after the Human Genome Project in the 1990s, medical genomics has emerged to capture the high-throughput molecular profile of a person. As a result, integrated data analytics is becoming one of the fast-growing areas under Biomedical Big Data to improve human healthcare outcomes. In this chapter, we first introduce the scope of data integration and review applications, data sources, and tools for clinical informatics and

*Progress in Molecular Biology and Translational Science*, Volume 190
ISSN 1877-1173
https://doi.org/10.1016/bs.pmbts.2022.05.002

medical genomics. We then describe the data integration analytics at the raw data level, feature level, and decision level with case studies, and the opportunity for research and translation using advanced artificial intelligence (AI), such as deep learning. Lastly, we summarize the opportunities in biomedical big data integration that can reshape healthcare toward p-health.

## 1. Introduction

Delivering predictive, precise, participatory, preventive, and personalized health, abbreviated as p-Health, is the primary goal of future healthcare systems that can significantly improve care quality while reducing cost. To accomplish this goal, in-clinic Electronic Health Records (EHRs), out-of-clinic Personal Health Records (PHRs),[1] and high throughput genomic data could be assessed and validated concertedly through translational data analytics pipelines consisting of six steps: data collection, data quality control, feature extraction, knowledge modeling, decision making, and action-taking. As shown in Fig. 1, this chapter will examine the challenges and opportunities in data integration analytics, from clinical informatics to genomics medicine for p-Health, at three levels—raw data, feature, and decision.

Clinical informatics was first introduced in the 1950s when the U.S. National Bureau of Standards and the U.S. Air Force used digital computers to develop expert systems (such as MYCIN and Internist-I[2]) and computerized medical records management systems. In 1959, Ledley and colleagues discussed the idea of using computational reasoning to aid medical diagnostic processes for the first time and brought out the EHR idea.[3] EHR systems improve the communication among physicians, providers, and patients, the quality of clinical decisions, and the delivery of care significantly in hospital routine practices.[4] However, clinical informatics progress was slow due to the low EHR adoption rate. In 2008, the American Hospital Association (AHA) Annual Survey reported that only 13.4% of the non-Federal acute care hospitals in the U.S. had adopted basic or comprehensive EHR systems, and only 1.6% have an EHR system with clinical decision support.[5] With the new policy, "the Health Information Technology for Economic and Clinical Health (HITECH) Act 2009",[6] in 2015, the EHR system adoption rate had increased to about 80% in the US, with 34.4% of these hospitals having EHR systems with clinical decision support[5] and, by CDC's estimate, this number reached approximately 90% in 2019.[7]

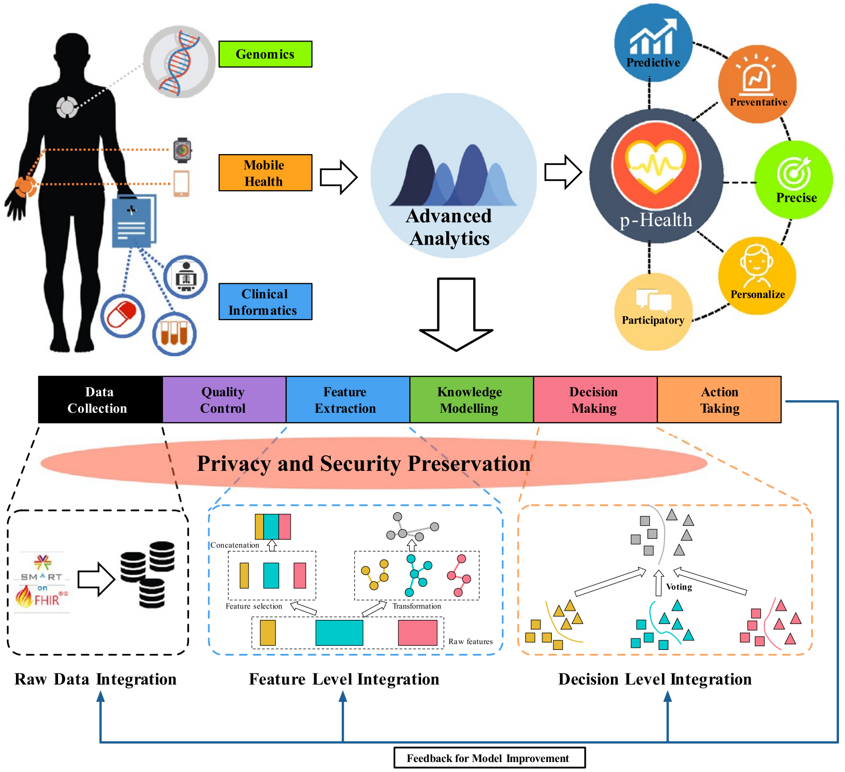


**Fig. 1** Enabling p-Health by data integration. Integrating clinical informatics with genomics and mobile health enables p-Health through advanced data analytics. The typical data analysis pipeline consists of data collection, quality control, feature extraction, knowledge modeling, decision making, and action taking. For translation to clinical applications, privacy and security preservation needs to be implemted through each step of the pipeline. Along this pipeline, data integration can be achieved at raw data level, feature level, and decision level, respectively.

The development of low-cost sensors embedded in either mobile phones or wearable devices has recently enabled out-of-clinic care data to be captured in PHRs, in addition to in-clinic EHRs. When the individual data of anthropometric measurement and patient response were collected and analyzed through the wristband-type activity sensors, the out-of-clinic PHR were conveniently analyzed and concluded in addition to the EHRs. For example, the integration of low cost sensor (PHR) data have given clinicians and patients an option to actively share lifestyle and behavior data within the EHR workflow to support clinical interventions,[8] even though the obstructive sleep apnea related outcomes were insignificant in this case.

In addition, low-cost microwave sensors with high sensitivity were also used for monitoring blood glucose levels of diabetics. The glucose levels measured are as sensitive as 70–120 mg/dL in blood with a 0.94 MHz/(mg/dL) resolution at several transmission resonances.[9] The prototype has shown the shift in AI-powered wearable signal processing microwave sensors for PHRs toward their commercialization. The continuous monitoring of a person's daily physiology and communication has enabled personalized, preventive, predictive health for early disease warning and participatory patient education, bridging the gap between in-clinic and out-of-clinic care. Thus, in this chapter, the term "EHRs" is used for both traditional in-clinic EHR and out-clinic PHR.

Since the completion of Human Genome Project in 2001,[10] applying genomics to customize clinical care (or medicine) has become possible. Genomics focuses on discovering the structure, function, and editing of genomes.[11] Based on the definition of the National Human Genome Research Institute (NHGRI), "Genomic Medicine" is an emerging medical discipline that uses the genomic information of an individual for their clinical care and health outcomes,[12] and NHGRI "Genomics" covers the study of direct information about DNA or RNA, excluding the study of downstream derived products (e.g., proteomics, metabolomics).[12] DNA sequences capture genomic variations at the single nucleotide level, e.g., single nucleotide polymorphisms (SNPs),[13] and at the chromosome level, as with structure variations (SVs).[14] RNA sequences contain genomic variations from gene expression or alternative splicing events.[15] Human diseases result from the complex interactions between genotypes and environment,[16] therefore incorporating molecular level information such as genetic variations is essential for the accomplishment of precision medicine in p-Health. The unique genetic information of individuals can reveal disease status and responses to treatment (e.g., more than 80% of rare disease are caused by genetic mutations, and genomics can play an important role in the diagnosis[17]). The genomic variations detected in DNA or RNA sequences are genotypes that complements phenotypes in genomic medicine. With the development of high-throughput next-generation sequencing (NGS) technologies, genomic data such as the whole genome of an individual, can now be sequenced for as little as $1000 in a few days.[18] As of February 2022, the NIH Genetic Testing Registry contained over 22,023 conditions with 18,741 genes genetic tests.

As shown in Fig. 1, integrating "clinical informatics" and "genomic medicine" presents challenges, including data harmonization, data quality control, and advanced data analytics to build an integrated intelligent clinical decision support system for p-Health. To overcome these challenges, data integration is performed with a six-step data analytic pipeline at three levels—raw data, feature, and decision. We will first define and then review the clinical informatics and genomic medicine in Sections 1.1 and 1.2, respectively. Then in Section 2, we discuss state-of-the-art data integration methods. In Section 3, we will address the Privacy and Security issues of the p-Health. At the end in Section 4, we conclude with the future of advanced data integration from clinical informatics to genomics for p-Health (Table 1).

**Table 1** Concept glossary for clinical informatics and genomics.

| Term | Definition | Ref. |
|---|---|---|
| Clinical Informatics | The application of informatics and information technology to deliver healthcare services. | 19 |
| Electronic Health/ Medical Records (EHRs/EMRs) | An EHR/EMR is an electronic version of a patient's medical history, that is maintained by the provider over time, and may include all of the key administrative clinical data relevant to that persons care under a particular provider, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports. | 20 |
| Health Informatics | The interdisciplinary study of the design, development, adoption, and application of IT-based innovations in healthcare services delivery, management and planning. | 21 |
| Personal Health Records (PHRs) | PHR contains health information that is managed by the individual and can be accessed and managed by authorized users in a private, secure, and confidential environment. | 1 |
| Artificial intelligence (AI) | The study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of success at some goal. | 22 |
| Machine Learning | The subfield of computer science that gives "computers the ability to learn without being explicitly programmed." | 23 |
| Deep Learning | Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. | 24 |

*Continued*

**Table 1** Concept glossary for clinical informatics and genomics.—cont'd

| Term | Definition | Ref. |
|---|---|---|
| NHGRI "Genomic Medicine" | An emerging medical discipline that uses genomic information about an individual as part of their clinical care and the health outcomes. | 12 |
| NHGRI "Genomics" | the study of direct information about DNA or RNA, excluding the study of downstream derived products (e.g., proteomics, metabolomics). | 12 |
| Genome | The complete set of DNA in an organism. | 11 |
| Transcriptome | A collection of all the gene readouts (RNA) present in a cell. | 11 |
| -Omics | The study of the –ome. For example, genomics and transcriptomics are the studies of the genome and transcriptome respectively. | 25 |
| Single Nucleotide Polymorphism (SNP) | A variation at a single position in a DNA sequence among individuals. | 13 |
| Structural variation (SV) | Genomic alterations that involve segments of DNA that are larger than 1 kb, and can be microscopic or submicroscopic. | 14 |
| Copy number variation (CNV) | Copy number variation of DNA segments ranging from kb to Mb. CNV is one type of SVs. | 26 |
| Alternative splicing | During gene expression, different combinations of splice sites can be joined to each other, resulting one gene coding for multiple proteins. | 15 |
| Next generation sequencing (NGS) | A new generation of non-Sanger-based sequencing technologies. These high throughput sequencing technologies can sequence DNA and RNA at much faster speed and lower cost compared to Sanger sequencing. | 27 |

## 1.1 Clinical informatics

Clinical informatics uses data analytics to gain new insight from individual and population health records and to improve clinical decision-making by combining data-derived knowledge and domain expert knowledge (Fig. 2). Typical data sources include the following.

*Conventional individual health records*, which is the first challenge in clinical informatics, contain various types of data, from structured billing codes to unstructured clinical notes. Throughout the hospital stay of patients, administrative information (e.g., demographics, gender information, and diagnostic codes for billing and public health reporting purposes), auxiliary clinical

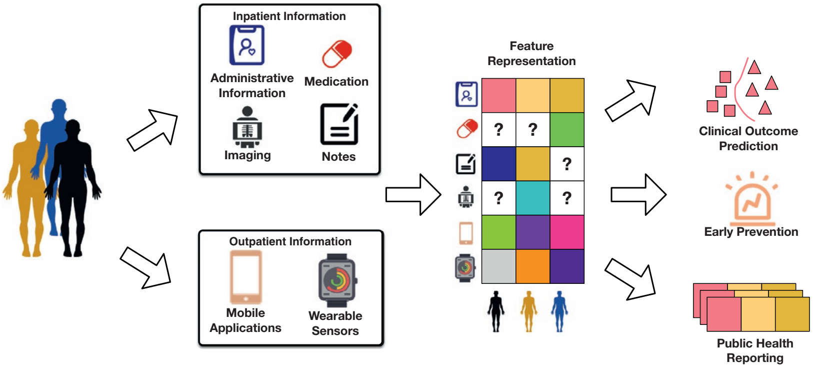


**Fig. 2** An overview of clinical informatics. In-clinic and out-of-clinic health information contribute to the primary data sources for clinical informatics. Data from different sources are combined to form the feature representation of each patient, for later clinical endpoints such as clinical outcome prediction, early prevention, and public health reporting.

data (e.g., lab tests, medical and radiological imaging, medications, genetics), continuous physiological data (e.g., bedside monitoring data in intensive care units), and unstructured clinical notes record comprehensive information about patients' clinical status and outpatient care information. The *public health data* are more abstract and often contain aggregated statistics of diseases in geographical regions and times (e.g., the National Center for Health statistics (NCHS) collects 2.6 million death certificates each year that record demographics and causes of death within the US population. Recently, billions of mobile phones and wearable sensors (e.g., FitBit) and behavior imaging have enabled daily health monitoring in *outpatient care*. Especially during the COVID-19 pandemic, the United States Centers for Medicare & Medicaid Services (CMS) and health insurance companies and mobile health service providers, have all stepped up their support for mobile health care.[28] PHRs collect data from these applications and provide a more personalized evaluation of patients. Within the context of this chapter, EHRs are considered to comprise data sources from traditional EHRs, physiological and imaging data, and out-of-clinic as "extended" EHRs to represent heterogeneous types of data existing in clinical informatics. Other data sources that contribute to health include *ontology*, such as Unified Medical Language System (UMLS) and SNOMED CT.[29] Physicians and clinical researchers use nodes to represent medical concepts and edges between nodes to encode the relationships between concepts, such as what diseases can lead to certain symptoms and what medications can help alleviate specific symptoms.

Analyzing these data has benefits to both population and individual care for p-Health. Recent progress is shown in Table 2. At the population care

**Table 2** Highlights of progress made in clinical informatics.

| Task | Data collection and quality control | Feature extraction | Knowledge modeling and decision making |
|---|---|---|---|
| *Re*–admission Prediction | Models using administrative data[44]<br>Models using real–time administrative data,[45]<br>Models adding survey or chart review data[46] | Mainly raw values with missing values imputed | Logistic regression,[47] graphical models,[48] random forests,[49] and neural networks[50] |
| Mortality Prediction/ Survival analysis | Models using EHR data,[51]<br>Models using clinical trials data[52] | Raw values regression[53] | Cox regression[53] and logistic regression, risk of mortality score[51] |
| Diagnosis Prediction | Models using EHR data[54] | Raw values and transformed values,[55]<br>Deep learning features[56,57] | Recurrent neural networks,[56] multiple layer perceptron, decision trees, nearest neighbors, generalized linear model[58] |
| Epidemics Prediction | Models using public health data[59]<br>Models incorporating social media data[60] | Mainly use raw reported values | Bayesian data analysis, text mining on social media,[60] time series mining[59] |
| Phenotyping and Risk Factors Identification | Models using EHR data,[61]<br>Models using claims data,[62]<br>Models using clinical trials data[63] | Raw values,<br>Deep learning features[57,64] | Matrix/tensor factorization,[65] Bayesian networks, unsupervised clustering[55] |
| Causal Inference | Models using simulated data,[66]<br>Models using observational data[67] | Raw data,<br>Deep learning features[68] | Nearest neighbor matching,[68] decision trees[66] and random forests,[66] panel data, propensity scores, causal networks[69] |
| Mobile Health Software Applications | Models using mobile data[70] | N/A | Apple watch, Samsung gear, Jawbone, FitBit (devices); WebMD (referencing app); AmWell, PatientIO (Telemedicine); SugarSense (Diabetes) |
| Mobile health data analytics | Wearable sensors and smartphones | Raw values[71]<br>Deep learning features[72] | Time series[73]<br>Visual analytics[74] |

level, health policymakers are interested in understanding epidemics from a statistical point of view across spatial and temporal dimensions.[30] A great amount of evidence has shown that EHRs have enabled multi-institute collaborations on data sharing, disease understanding, and, most importantly, disease dealing. The efficiency of EHRs has enabled p-Health on a personal level and has supported international collaboration in terms of developing effective vaccines and creating herd immunity. It has been one of the most efficient anti-pandemic practices in human history.[31–38] Aggregated EHRs help explore the evolution and spreading of diseases for resource allocation, which may be used to predict and prevent the outbreak of epidemics. Death certificates that contain causes of death can assist in understanding the diseases spread in the nation. At the individual care level, diagnosing patients' conditions and deciding prognosis are crucial steps. Clinical informatics can provide decision support for in-clinic physicians and care providers,[39] where clinical outcome (hospital re-admission, mortality) prediction helps prevent adverse in-clinic clinical events. Data prediction enabled through mobile sensors give early warning of medical conditions and provides physicians with out-of-clinic participatory health. EHR-associated scores have been used to predict the patients' clinical courses and outcomes and support clinical decisions. It also has been reported that respiratory support information from EHRs have been consistent with mortality prediction.[37,40] Besides prediction, grouping similar patients' records[41] and simulation[42] can suggest reference treatment options to assist clinical action-taking.

To gain more insight in clinical informatics, public datasets and tools are summarized in Table 3. Due to privacy and other health regulatory issues, only a limited number of datasets are publicly available, and comprehensive data sets may be obtained through research collaboration with specific clinical institutions. As the data available in EHRs mainly reflect phenotypes, to truly achieve personalized and precise health,[42] it is critical to expand from clinical informatics to genome medicine.[43]

## 1.2 Genomic and genomic medicine

Genomics can analyzes all genes and their interactions in each person at any point in time. Genomic medicine uses such genomic information to prevent, diagnose, and treat diseases clinically, i.e., for p-Health (Fig. 3). Human diseases result from complex interactions of genotypes and environmental factors.[16] Personalized information embedded in genomics can help physicians predict disease risk factors and patients' responses to treatments.

**Table 3** Selected databases and tools for clinical informatics.

| **Purpose** | **Data source** | **Dataset description** |
|---|---|---|
| Selected databases or projects for clinical informatics | | |
| Individual Health | Medical Information Mart for Intensive Care III (MIMIC III)[75] | Clinical, notes, waveform information for about 58,000 admissions in intensive care units |
| Individual Health | The Alzheimer's Disease Neuroimaging Initiative (ADNI)[76] | Imaging, clinical, genetics for a cohort about 2000 in an Alzheimer's disease study |
| Individual Health Records | Parkinson's Progression Markers Initiative[77] | Imaging, clinical, biomarkers for a cohort of about 800 in a Parkinson's disease study |
| Public Health | CMS 2008–2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) | About 40 million synthetic claims data |
| Public Health | Mortality data[78] | Death certificates containing demographics and causes of death for about 2 million each year in the U.S. |
| **Purpose** | **Description** | **Open source tools** |
| Selected tools for clinical informatics | | |
| Data Storage and Management | Storing health data for hospital management or personal access | FHIRBase,[79] EHRServer[80] |
| General Predictive Modeling | Building predictive pipelines using patients' health records | PARAMO[81] |
| Visual and Analytics | Visualizing health data for easy interpretation and data understanding | EHDViz[82] |
| Domain Knowledge Modeling | Representing domain knowledge as ontology for knowledge-guided analysis | UMLS[29] |
| Mobile Health | Using data collected from mobile devices for assisted clinical decision making | Open mHealth[83] |

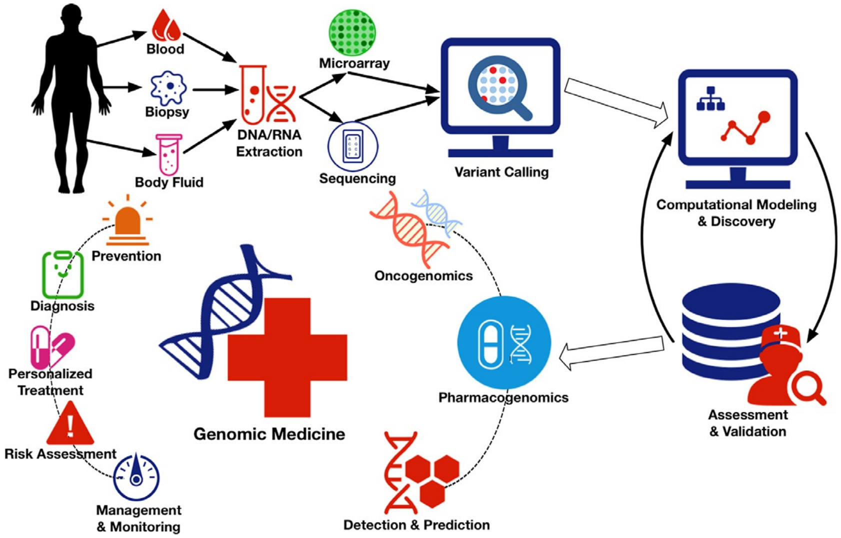


**Fig. 3** An overview of genomic medicine. DNAs/RNAs are firstly extracted from the patient's blood, biopsy, or body fluid samples, and then genotyped by either microarray or sequencing. The raw genomic data are further analyzed with bioinformatics approaches to generate a list of genomic variants. AI-based computational modeling is applied for disease related genomic variants discovery with the guidance of current knowledge bases. The identified genomic biomarkers will be evaluated and validated by domain experts and will in turn enrich the knowledge bases. Ongeogenomics, pharmacogenomics, and disease risk predication are three major components of genomic medicine, which enables disease prevention, diagnosis, personalized treatment, risk assessment, and health management and monitoring.

For example, familial mutations in tumor suppressor genes such as BRCA1 and BRCA2 increase the risk of female breast and ovarian cancers.[84] Test for these traits can prevent, diagnose, and treat breast and ovarian cancers early.

As shown in Table 4, genomic medicine has mainly been applied to three areas: *oncogenomics, pharmacogenomics, and disease risk prediction*. Oncogenomics and pharmacogenomics are the most studied examples for applying personalized genomics.[85] Disease risk prediction is especially beneficial for those diseases with genetic markers of high clinical impacts. Various cancer known to be caused by gene mutations[86] have been studied extensively with personalized genomics to identify cancer-related genetic variations for personalized diagnosis and treatment.[87] For example, Nguyen et al. have applied personal genomics to relate the outcomes for breast different cancer subtypes after breast-conserving therapy.[88] Zheng et al. have associated genetic variations with prostate cancer.[89] *Pharmacogenomics* is the study of how a person's

**Table 4** Highlights of progress made in genomic medicine.

| Task | Applications | Selected examples | Genetic variants | Computational methods and knowledge modeling |
|---|---|---|---|---|
| Oncogenomics | Identify oncogenes (understanding genetic mechanisms for cancer generation and progression) | Colorectal cancer[98]<br>Prostate cancer[89]<br>Ovarian cancer[99] | SNPs[89,99]<br>CNVs[98]<br>DNA Methylation (e.g., altered methylation of CpG islands)[98]<br>Gene expression[88] | Genome-wide association studies (GWAS)[99]<br>Cox proportional-hazards model[98]<br>Logistic-regression[89]<br>Statistic tests (Chi-square test, Student's *t*-test, log-rank test, Fisher's exact test)[88,98] |
| | Personalized treatment | Breast cancer[88]<br>Colorectal cancer[98] | | |
| Pharmacogenomics | Optimize treatments (personalized treatments, drug safety, optimal dosing) | Cystic Fibrosis[100]<br>Malignant melanoma with BRAF[90] | SNPs[90,101]<br>Gene expression[102]<br>Gene fusion[103] | GWAS[101]<br>Statistical tests[90] |
| | Drug discovery (targeted, efficient trials) | Rheumatoid arthritis[101]<br>Breast cancer with HER2[102]<br>Lung cancer with EML4–ALK fusion[103] | | |
| Disease diagnosis or risk prediction | Infectious diseases | Ebola[104] | SNPs[104,105] | GWAS[105] |
| | Rare diseases | Severe intellectual disability[106] | | |
| | Other endpoints | Alzheimer's disease[105] | | |

genome affects their response to drugs (i.e., the relationship between genetic variations and drug responses). It can guide physicians in choosing optimal drugs for patients based on their genetic variations, which minimizes adverse drug reactions and maximize effectiveness[85] for personalized care. For example, the chemotherapeutic agent Vemurafenib specifically targets melanomas with the BRAF V600E mutation, with efficacy and adverse events evaluated.[90] *Disease risks* vary with genetic variations. For example, 80% of rare diseases have genetic origins, making personalized genomics essential for early screening, diagnosis, and individualized treatment.[17] Multiple rare diseases have genetic biomarkers with sufficiently high clinical impact, and more are discovered each year. In COVID-19 drug discovery, some pharmacogenetic data were collected to repurpose old drugs for COVID-19.[37,40] In addition, in silico modeling and pharmacogenomic markers have also been utilized for these COVID-19 drug repurposing efforts. By combining genetic biomarkers using genome-wide association studies (GWAS) with environmental factors, risk factors of some diseases can be thoroughly assessed, allowing for earlier intervention and better treatment.

Table 5 lists databases and projects for large-scale genomics studies for personalized genomics. These databases or projects focus on: (1) general genetic variants discovery; (2) specific diseases; or (3) specific populations, respectively. *The first group of projects* aims to discover genetic variants and understand their functions. To identify common genetic variations and to discover novel ones, the 1000 Genomes Project,[91] followed by the 10,000 human genomes project,[92] sequenced a large number of human genomes. These projects provide a base for the clinical use of genetic variations. The encyclopedia of DNA elements project (ENCODE)[93] aims to discover the functional elements in the human genome to make sense of genomic data. UK Biobank[94] connects genetic variants with a wide range of diseases and outcomes by providing EHR (with imaging) and genomic data, which is a perfect resource for integrated data analytics. *The second group of projects* aims to better understand and treat specific diseases. These projects include the cancer genome atlas project (TCGA),[95] Alzheimer's Disease neuroimaging initiative (ADNI),[96] and Parkinson's progression markers initiative (PPMI).[77] The TCGA project focuses on utilizing cancer genomics to improve the understanding, prevention, diagnosis, and treatment of various cancers. Similarly, ADNI and PPMI projects are developed for two of the most prevalent neurodegenerative diseases, Alzheimer's disease, and Parkinson's disease, respectively. These disease-oriented databases collect multi-omics data, clinical information (EHR), and medical imaging data

**Table 5** Selected databases and tools for genomics.

| Database or project | Description | Data elements | Ref. |
|---|---|---|---|
| 1000 Genomes Project | The first project to sequence the genomes of a large number of people, which aims at finding most genetic variants with frequencies of at least 1% in the population studied. | Variant calls (VCF format). Alignments (BAM or CRAM), Raw sequence files. | 91 |
| Encyclopedia of DNA Elements Project (ENCODE) | A group of international projects aims at building a comprehensive list of functional elements in the human genome. | DNA-seq, RNA-seq, and various DNA-RNA, DNA-protein interaction assays like ChIP-seq | 93 |
| UK Biobank | A national and international health resources open for all health researchers. The database is established for improving the prevention, diagnosis, and treatment of diseases. | EMR, Questionnaires for health and lifestyle, Genetic information (whole genome sequencing, exome sequencing, genotyping), Imaging (MRI scan for 100,000 participants) | 94 |
| The Cancer Genome Atlas Project (TCGA) | A publicly available dataset with more than 2 TB of genomic data. The project aims at improving the prevention, diagnosis, and treatment of cancer by discovering the key genomic changes in 33 types of cancers. | EMR (clinical information of participants), histopathology slide images, molecular information (mRNA/mi RNA expression, protein expression, copy number, etc.) | 95 |
| Alzheimer's Disease Neuroimaging Initiative (ADNI) | An ongoing, longitudinal, multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). | EMR (demographics, clinical assessment), Genetics data (SNP), MRI image data, PET image data, Image analysis results, chemical biomarkers. | 96 |
| Parkinson's Progression Markers Initiative (PPMI) | An observational clinical study to verify progression markers in Parkinson's disease with a comprehensive set of clinical, imaging and bio-sample data. | EMR, Imaging (MRI, PET, SPECT), whole exome sequencing data, biological specimens. | 77 |

Table 5 Selected databases and tools for genomics.—cont'd

| Database or project | Description | Data elements | Ref. |
|---|---|---|---|
| Minority Health Genomic and Translational Research Bio-Repository Database (MH-GRID) | A study aims at finding the causes of severe high blood pressure and other cardio-metabolic problems in people of African ancestry | Genes, EMR (diet, sleep, body mass index, stress, access to healthy food and parks). | 97 |

**Selected tools and applications for genomics**

| Task | Description | Selected tools with ref. |
|---|---|---|
| Quality control | Identify artifacts in Microarray data or NGS data including low-quality and contaminated reads. | caCORRECT,[107] FastQC[108] |
| Read alignment | Alignment (mapping) of the short reads generated by NGS to reference sequences (genome or transcriptome) | BWA[109] |
| SNP detection | Identify SNPs from the aligned reads | GATK HaplotypeCaller[110] |
| SV detection | Identify SVs from the aligned reads | Hydra-sv[111] |
| CNV detection | Identify CNVs from the aligned reads | GATK gCNV[110] |
| Transcript quantification | Estimate gene expression levels | Cufflinks[112] |
| Normalization | Normalize systematic variations between samples | Cuffnorm[112] |
| Differential expression | Identify the differentially expressed genes in two or more conditions. | Cuffdiff[112] |
| Alternative splicing | Identify differentially spliced transcripts | Cufflinks[112] |
| Prediction and Modeling | Prediction and modeling using Omics data | OmniClassifier,[113] omniBiomarker[114] |
| Data Management | Data sharing and data integrity maintenance | ArrayWiki[115] |

(magnetic resonance imaging (MRI), positron emission tomography (PET), and pathological images) to enable data integration and a better understanding of each disease. *The third group of projects* aims to stratify medicine for specific populations. One example is the minority health genomic and translational research bio-repository database (MH-GRID).[97] Motivated by the fact that high blood pressure affects African Americans more than other racial groups, MH-GRID contains data collected from over a thousand African Americans across the US. Besides genetic data, the MH-GRID collects health-related information such as diet, sleep, body mass index, stress, and access to healthy food and parks.

Table 5 lists selected methods and tools for DNA and RNA analysis for personalized genomic medicine. DNA and RNA can first be sequenced by next-generation sequencing (NGS) platforms and then analyzed by bioinformatics approaches. Developed from the Sanger chain-termination method, NGS techniques parallelize the sequencing process and produce billions of sequences concurrently,[27] by using sequencing-by-synthesis, and have significantly lowered the cost of DNA sequencing compared to the Sanger method itself.[27] With sequences as input, the first step of genomic data analysis is sequencing data quality control to remove potential artifacts and low-quality reads. The sequenced short reads passing quality control are then aligned to the reference genome. The aligned reads are further analyzed to identify various genetic variations, e.g., SNPs, CNVs, and SVs. Based on NHGRI genomic definition presented in Section 1, genetic variation at the transcription level can be characterized by RNA-seq to obtain gene expression, alternative splicing, and gene fusion information. RNA sequencing (RNA-seq) reads are first aligned to the genome (spliced alignment) or transcriptome (un-spliced alignment) and are then quantified and normalized to yield gene expression levels. RNA-seq can capture major transcription-level regulation events like alternative splicing.[15]

The following two sections will focus on data integration, aiming at utilizing both clinical informatics and genomic medicine for improved p-Health.

## 2. State-of-the-art advanced data integration

Data integration uses multiple sources of information to provide a better understanding of a system (Fig. 4). Wong reviewed requirements and solutions to data integration and warehousing at the raw data access level in biomedicine in EnsEMBL, GenoMax, and SRS.[116] Goble et al. introduced

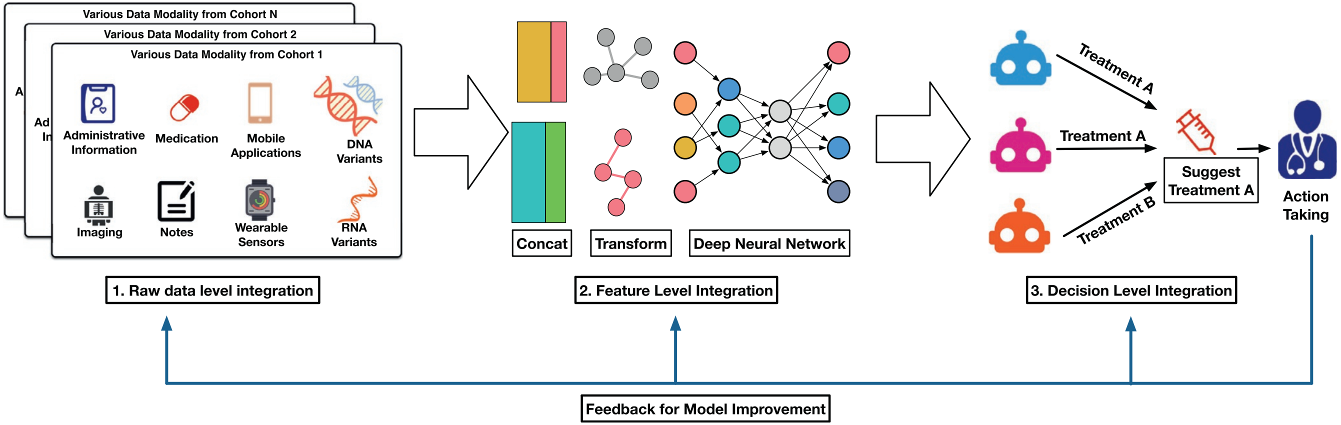

Various Data Modality from Cohort N
Various Data Modality from Cohort 2
Various Data Modality from Cohort 1
Administrative Information
Medication
Mobile Applications
DNA Variants
Imaging
Notes
Wearable Sensors
RNA Variants
1. Raw data level integration
Concat
Transform
Deep Neural Network
2. Feature Level Integration
Treatment A
Treatment A
Treatment B
Suggest Treatment A
Action Taking
3. Decision Level Integration
Feedback for Model Improvement

**Fig. 4** Summary of advanced data integration analytics. The multimodality health data can be integrated at raw data level, feature level, and decision level. At raw data level, health data from different modalities and cohorts are harmonized. At the feature level, features from different data can be integrated by concatenation, transformation, or deep learning methods. At the decision level, the outputs of multiple models are combined for decision making and further action taking by the doctors.

"a loose federation of bio-nations" to handle the heterogeneity of biomedical data sources and the ontology in data integration.[117] More recently, to obtain new insights, Gomez-Cabrero et al. described how to deal with the diversity of existing omics data types and formats of large datasets.[118] They used dynamic Bayesian networks, self-organizing maps, and network inference methods over popular data sources such as 1000 Genomes Project, ENCODE, TCGA, and ImmGen. Ritchie et al. reviewed -omic data integration by comparing meta-dimensional and multi-stage analysis.[119] Based on current work in biomedical data integration, we identify the following *four challenges* for integrating EHRs and genomic data:

The first challenge is data collection and harmonization. New technologies, such as mobile sensors and DNA sequencing, all require modality-specific techniques for data storage, visualization, and analysis by computing devices. New genomic technologies are emerging every year, ranging from microarray to third-generation sequencing, which makes genomic data heterogeneous with various formats and standards. The NGS has different platforms, such as Illumina, 454 sequencing, or SOLiD sequencing, that lead to more variation. The development of third-generation sequencing by Pacific Biosciences,[120] with a much longer read length will add further variations in genomic data. Validating the data generated by these various technologies is challenging.

The second challenge is data quality. In clinical informatics, EHR data can be unstructured and noisy, with issues such as high percentages of missing values, errors, invalid data, and outliers. Thus, data quality control processes, such as missing data imputation, data conflict resolution, and data transformation, are needed.[121] In genomic medicine, DNA sequencing quality is affected by sample contamination and sequencing errors. Thus, numerous new bioinformatics approaches (e.g., FastQC[108]) are being developed to improve the sequencing quality before downstream analysis. In addition, to establish unified standards for genomic data processing for reliable clinical applications, the US Food and Drug Administration (FDA) coordinates the microarray quality control (MAQC) project[122] and the sequencing quality control (SEQC) project[123] for a comprehensive assessment of microarray and RNA-seq, respectively.

The third challenge is to develop advanced data analytics to extract knowledge from EHR and genomic data. In clinical informatics, feature engineering constructs a meaningful set of representations for patients, including medications, lab tests, and procedures, whereas the past analytics relied on hand-crafted features or input from domain experts, which are very limited in representation and predictive power. The new opportunities are on deep learning to find representations from large data. In genomic medicine, the "curse of dimensionality" (i.e., the feature dimension is significantly larger than the patient sample size) is a well-known challenge. Genome sequencing generates millions of genomic variations such as SNPs, CNVs, gene expression levels, and alternative splicing events for each sample, while the sample size of the study may be several hundred. Directly applying traditional data analysis will result in ill-conditioned feature matrixes, and the solution is to filter out irrelevant variants use feature selection in either a supervised or unsupervised fashion. Lack of interpretability is another major challenge for EHR and genomic data analytics. Predictive models built on EHR data have shown potential in predicting the hospital length of stay and re-admission probability. However, deploying them in daily hospital practice is challenging because physicians do not fully trust the model prediction because the bases for such predications are unknown. Making sense of genomic data is also a bottleneck for translating genomic discoveries into clinical practice.[124] Current data-driven approaches are mainly based on mathematical models, statistical tests, and computational methods. Lack of biological and clinical interpretation weakens the clinical impact of the novel genetic biomarkers discovered. Experimental validation is important for translating discoveries from data mining into biological knowledge and further applications to clinical care.

The fourth challenge concerns converting knowledge into actionable decision-making. Artificial intelligence (AI)[22] started in the 1950s and has undergone a paradigm shift from symbolic logical reasoning to machine learning,[23] including the recent success of deep learning[24] in automatic representation learning. Deep learning has the potential to provide highly accurate predictive modeling, and deep reinforcement learning shows significant breakthroughs in playing video games, dialog systems, and other tasks. However, a major challenge for computer-aided decision-making in clinical informatics is that it is unethical to test different options. Moreover, it is almost impossible in practice to test two treatment options in one individual. Historical data may be used to simulate the effect of each action applied to an individual, which is termed "counterfactual inference." However, to ultimately address the problem, we need to understand the causal relationship between medical events and between treatment options and diseases, for causality beyond correlation studies.

In this section, we review state-of-the-art data integration analytics at the raw data, feature, and decision levels to enable p-Health.

## 2.1 Raw data level integration

Raw data level integration happens in the data collection and storage stage, when different modalities of data are collected from different institutions and times for the same disease. The multi-modality data from more patients can enable machine learning algorithms to achieve better predictive models.[125] How to integrate different types of data and databases is also challenging. In clinical informatics, there are 186 commercial EHR system vendors in the U.S. Market,[106] with Epic, Cerner, and MEDITECH having the highest market shares, and each of them has its specific data standard and format. As a result, multiple clinical institutions with different EHR vendors cannot exchange information directly.[107] Existing standards such as Health Level Seven International (HL7) and Clinical Document Architecture (CDA)[126] cannot support the interexchange of data from different EHRs automatically for decision support systems, while manual data conversion case-by-case is both time-consuming and prone to errors. Besides, ensuring privacy and security in the raw data integration is also crucial.[127] To integrate data, the international standard body, Health Level Seven International (HL7),[108] has made a collaborative effort to develop a new EHR standard FHIR (Fast Healthcare Interoperability Resources), and its extensions SMART-on-FHIR, to make data from different EHR interoperable. FHIR is a new emerging standard that uses a resource-centric approach (as opposed to document-centric) to specify data elements in two parts: (i) standardized healthcare data models; and (ii) a set of application programming interfaces (APIs) for interacting and modifying these data models.[128] Thus, data resources can exist as online services rather than static files so that applications can read and write these resources in real-time. SMART-on-FHIR apps are based on FHIR but work outside of the constraints of many stakeholders' existing technical and security infrastructures. Thus, opportunities in raw data level integration include extending FHIR Resources, implementing RESTful APIs, and SMART-on-FHIR apps to facilitate the raw data integration process. Examples include death reporting, health record vendor translator, and mobile health apps[129]; and SMART on FHIR Genomics.[130] In genomic medicine, currently, FHIR enables the incorporation of genomic variations into EHRs. One example for establishing standards and APIs for genomic data storing, sharing, and clinical applications is SMART on FHIR Genomics.[109] FHIR app development presents challenges and opportunities for establishing standards and APIs for genomic data storing, sharing, and clinical applications. Such uses have been reported in the medical fields of pathology, ophthalmology, NGS and meta-studies.

## 2.2 Feature level integration

The next step in the data analytics pipeline following data collection and quality control is feature extraction. The feature level integration includes both concatenation-based and model-based methods.

Directly concatenating raw features to integrate different data modalities can introduce thousands of features, but the challenges for integrating EHRs and genomics features at the feature level are that the information represented by feature vectors may have varying representation power and noise, and the computational challenges introduced by adding features together may lead to model under-fitting or non-convergence. Thus feature selection, such as L1 norm feature selection[131] and minimum-redundancy maximum-relevancy (mRMR) feature selection,[132] is needed.

Model-based feature integration often adopts an encoder-decoder framework to combine features from different modalities. For each modality, modality-specific encoders can map features of every data modality to a joint feature embedding space first. Features combined in the joint space will go through a single decoder for final reporting. Various machine learning algorithms have been researched to accomplish model-based integration. Multiple kernel learning (MKL)[133] learns a kernel transformation for each modality and then combines all kernel transformations with a weighted linear average. Canonical correlation analysis (CCA)[134] finds a new linear feature space, in which features from all modalities have maximum correlation when projected onto the new space. Probabilistic graphical model (PGM)[135] treats each feature as a random variable and performs statistical inference to obtain an integrated feature, which is also represented in the form of latent variables. These techniques are effective in several applications (Table 6). However, the disappearance of individual modality's feature learning and feature integration may result in a meaningless feature representation.

An emerging advanced AI method in feature-level integration is deep learning that can combine feature extraction and prediction, which learns a meaningful representation for high accuracy in the given task.[24] Multimodal deep learning[136] is an early work using the restricted Boltzmann machine, while CCA is also extended with a deep feedforward network in[137] for feature-level integration. In clinical research, the multi-modal analysis using deep learning improves model accuracy in medical diagnosis and imaging analysis.[138–141] More recently, EHRs, SNPs, and MRIs have been integrated with deep autoencoders for improved prediction of Alzheimer's diseases,[142] in which features from multi-modalities are first transformed

**Table 6** Selected methods and tools for biomedical data integration.

| Categories | Methods | Software or applications with ref. | Data modalities |
|---|---|---|---|
| Raw data level:<br>Ontology<br>Semantic web | Resource description framework | AlzPharm[155] | Neuron properties, pathological changes, drug responses, disease stages. |
| | Heterogeneous database system | Heterogeneous Database Integration in Biomedicine[156] | Distributed databases |
| Feature level integration:<br>Concatenation-based integration<br>Transformation-based integration | Grammatical evolution neural network | ATHENA[157] | SNPs, microarray, proteomics, sequence data, biomarkers, clinical data |
| | Bayesian network | WinBUGS[158] | SNP, mRNA (gene expression), phenotypes |
| | Multivariate Cox LASSO model | Glmpath[159] | mRNA, DNA methylation, CNV, and microRNA |
| | Feature extraction, selection, and concatenation | Anduril[160] | Gene expression, splice variant, SNP, CNV, DNA methylation, miRNA, siRNA, proliferation array |
| | Feature extraction and rule integration | Combining rule and machine learning classifiers[161] | Temporal events, rules, clinical notes |
| | Kernel-based integration | SKMsmo[162] | Gene expression, protein sequences |
| | | Clinical decision support[163] | Clinical, micro array |
| | Graph-based semi-supervised learning | Cancer clinical outcome prediction[164] | CNV, DNA methylation, Gene expression, miRNA |
| | | Protein functions prediction[165] | Co-participation in a protein complex, physical interactions, genetic interactions, Pfam domain structure, and gene expression |

**Table 6** Selected methods and tools for biomedical data integration.—cont'd

| Categories | Methods | Software or applications with ref. | Data modalities |
|---|---|---|---|
| Decision level Integration: Voting-based integration Ensemble-based integration | Majority voting/ Weighted majority voting | Predicting protein fold recognition[148] | Pseudo-amino acid composition, predicted secondary structure, hydrophobicity, normalized van der Waals volume, polarity, polarizability |
| | Bayesian networks | CONEXIC[151] | CNV, and gene expression |
| | | Breast cancer prognosis prediction[152] | Clinical and micro array |
| | Heterogeneous network | Drug-drug interactions[166] | Drug phenotypic, therapeutic, chemical, and genomic properties |

to lower dimensions with deep autoencoders and then concatenated at the intermediate feature level.[143,144]

The opportunities in advanced AI-based feature integration include constructing differentiable encoders and the interpretation of integrated features and predictive models, especially deep learning models. Differentiable encoders can extract features from both EHRs and genomic data so that the whole pipeline can be trained with backpropagation. On feature interpretation, despite the progress in building highly accurate predictive models, physicians cannot trust black-box algorithms if they cannot understand what features contribute to the final prediction and how to link features to their original physiological meanings. Perturbation analysis,[145] influence functions,[146] and visualization of neural network outputs are some pioneering works for model interpretation. Identifying important raw features that contribute to the final prediction or similar patients from historical data are two directions that may help us understand the behavior of current models.

## 2.3 Decision level integration

Decision level integration first generates multiple base models using each data modality independently as the training sets and then generates a final model by combining individual models trained in the first step.

The challenges of decision level integration are the construction of an accurate base decision model and the combination of base models that can prevent model overfitting while adding more parameters.

The conventional decision-level combination uses simple majority, the weighted majority[147] (to predict protein fold recognition[148]), and ensemble learning. Ensemble learning designs special weighting methods to address the overfitting by adjusting base models with the resampled training set (bagging[149]), sequentially increasing weights of misclassified training data (boosting[150]). For example, the random forest classifier makes decisions based on the multiple decision trees it constructed from the resampled training set. Graphical-model-based approaches, such as Bayesian networks, combine multi-omics data to better understand glioblastoma and breast cancer.[151,152]

In advanced AI methods for decision level integration, there exist three opportunities: construction of base models, increasing interaction between different data modalities, and reducing the gap between the decision and clinical action.

First, designing accurate base models is critical for final predictive accuracy. For example, in the majority voting scheme, we need to choose one or more base models for each data modality from different models, including decision tree, k nearest neighbors (kNN), support vector machine, logistic regression, and neural network. For each base model, there are also multiple parameters to be tuned. The selection of base models and parameter tuning can be tedious and also prone to overfitting, especially for small sample sizes, so inventing efficient hyper-parameter tuning algorithms such as Bayesian optimization can be an interesting direction.

Second, decision-level integration allows independent analysis of each data set, while the integration at this upper level also limits the possible interactions among different data modalities. How to use the knowledge and decision from a resource-rich modality to assist decision-making in another modality is a challenging but rewarding task. For example, we often have limited patient data in the biomedical domain. However, we have a large collection of biomedical domain knowledge, documented as research papers or medical ontologies. Transferring decisions from such knowledge or retrieving such knowledge for decision-making[153] can be helpful for physicians.

Third, bridging the huge gap in the decision-level integration can improve the final clinical action. After we design accurate predictive models with these integration methods, how to suggest viable actions (medications and procedures) for physicians remains a challenge because even though randomized trials are widely used in drug and clinical trials to determine the effect of new technology, in daily practice, physicians can only resort to observational study for such evaluations.[154] To solve the problem, it is necessary to understand the causal relationship between genomic data, medical conditions, and treatment to suggest and evaluate more reasonable actions. This causal inference modeling is a huge topic in biomedical big data analytics on its own and will not be covered in this chapter.

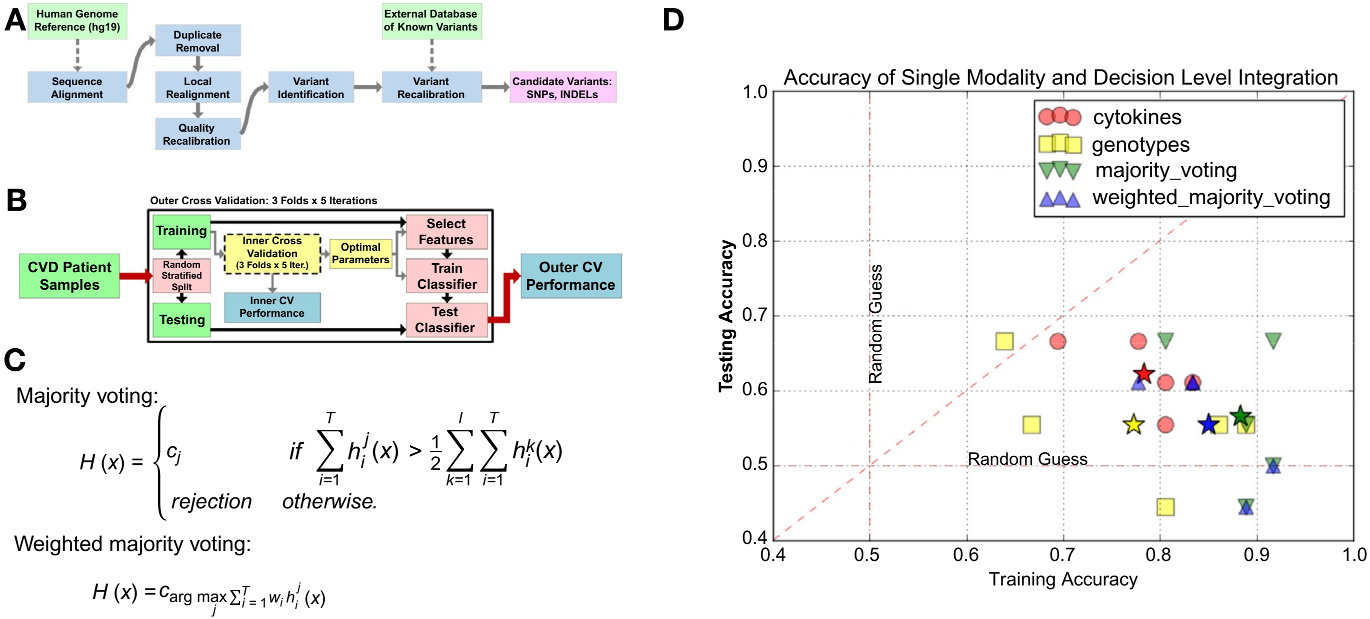
## 3. Case studies

### 3.1 Predicting length of stay after infant cardiac surgery by integrating proteomic data with genomic variations using majority voting

Immediately following newborn cardiac surgery, a profound inflammatory response may occur due to exposure of the blood to foreign materials during cardiopulmonary bypass (CPB).[167] It has been shown that the severity of systemic inflammatory responses correlates with prolonged length of stay (LOS) in either the intensive care unit (ICU) or hospital.[168] A full understanding of the mechanism for systemic inflammatory responses may help (1) identify high-risk patients and modify surgical procedures on a personalized basis to improve overall clinical outcomes; and (2) develop therapeutic strategies targeting specific biomarkers to alter clinical outcomes.

In this case study, we aim to integrate genomic data with proteomic data at the decision level for accurate prediction of clinical outcomes after infant cardiac surgery. Such study can not only facilitate biological interpretation of inflammatory response mechanisms but also provide an opportunity for biomarker translation to clinical settings. Specifically, combining postoperative serum cytokine levels and genome-wide variants, we investigate decision-level integration models that may predict the LOS of infants in the cardiac ICU after surgery.

For neonates who had undergone cardiac surgeries, blood samples were collected at admission back to the ICU immediately after the operation. We analyze 11 inflammatory markers from these postoperative blood samples. DNA of blood serum samples were sequenced at the Yerkes National Primate Research Center at Emory University using two Illumina HiSeq 1000 flowcells and 50 base-pair paired-end reads. Genomic variants, including SNPs and INDELs (i.e., insertions and deletions), were identified using The Genome Analysis Toolkit (GATK).[110] Fig. 5 summarizes the key modules and the workflow of the GATK pipeline. To reduce the number of



$$H(x) = \begin{cases} c_j & \text{if } \sum_{i=1}^{T} h_i^j(x) > \frac{1}{2} \sum_{k=1}^{l} \sum_{i=1}^{T} h_i^k(x) \\ \text{rejection} & \text{otherwise.} \end{cases}$$

$$H(x) = c_{\arg\max_j \sum_{i=1}^{T} w_i h_i^j(x)}$$

**Fig. 5** Predicting length of stay (LOS) after infant cardiac surgery by integrating proteomic data with genomic variations. (A) Genome Analysis Toolkit (GATK) pipeline for extracting genomic variations. (B) Nested cross validation to estimate prediction performance. (C) Decision level integration using majority voting and weighted majority voting, where ***T*** is the number of classifiers, ***l*** is the number of labels, ***x*** is the input data of a sample, $\boldsymbol{h_i^j}(\boldsymbol{x})$ is the probability of predicting ***x*** as label ***j*** by the ith classifier, and $\boldsymbol{w_i}$ is the weight for the ith classifier. (D) The accuracy of using single data modality and decision level integration for predicting the length of stay after infant cardiac surgery.

variants we discovered, we filtered the variants to include only variants that correspond to the inflammatory response (38 interferons and 92 interleukins (IL)) and identified 4411 variants. We binarized the LOS into long and short LOS with a threshold of 14 days. We applied nested cross-validation for prediction modeling of each data modality (Fig. 5B). We used mutual information for feature selection and k-nearest-neighbors (KNN) classifiers for prediction. We then used majority voting and weighted majority voting to combine prediction decisions of the model trained based on each data modality alone (Fig. 5C).

Fig. 5D shows the prediction results for LOS. With single data modality, accuracy using cytokines alone (test accuracy: 0.62 ± 0.042, train accuracy: 0.78 ± 0.048) outperforms accuracy using genotypes alone (test accuracy: 0.56 ± 0.070, train accuracy: 0.77 ± 0.10). After integrating cytokines and genotypes at the decision level, majority voting (test accuracy: 0.57 ± 0.089, train accuracy: 0.88 ± 0.041) and weighted majority voting (test accuracy: 0.56 ± 0.070, train accuracy: 0.85 ± 0.048) achieve similar prediction accuracies. Single data modality using cytokines alone achieves the best testing accuracy. Decision level integration improved the training accuracy but failed to improve the testing accuracy.

With decision level data integration, the training accuracy improved compared with that using a single modality alone. The improvement may be beneficial for realizing p-Health by accurate prediction of LOS after infant cardiac surgery. However, the testing accuracy with decision level integration is similar to that using genotypes alone and lower than that using cytokines alone, which may indicate overfitting when using majority voting and weighted majority voting at the decision level. The potential overfitting could be caused by the unbalanced number of features and the limited number of data modalities since we have only two data modalities with 4411 SNP genotypes while only 11 cytokines. To improve the predicting accuracy by integrating cytokines and genotypes, we may need more sophisticated data integration analytics to combine unbalanced data. In this case study, we present data integration application at the decision level for p-Health, where advanced analytics is important for making sense of multi-modality data.

## 4. Privacy and security

For translation to clinical applications, privacy and security preservation needs to be implemented through each step of the mutli-modal data integration pipeline (Fig. 1). The medical applications in mobile devices with embedded sensors provide superb services supporting healthcare interventions. Sophisticated data originating from the patients, from their medical records (EHR/PHR), and from the environments, can be used to assess users' health conditions and to generate alerts. The m-health apps often carry not only immediate illness-related data but also sensitive information, such as personal genetic condition, the results of any test and scan, medical histories, and personal information with tremendous value. If these records were not secured, the consequence could be severe, leading to financial losses and personal harm. Thus enforcing security and privacy through m-health apps on personal information is absolutely essential through data collection to predictive modeling (Fig. 1).

While these sensitive data are at stake, however, there is a lack of generally recognized security and privacy guidelines and legally binding data protection provisions to guarantee data privacy and safety. Through our in-depth medical data protection analysis of some popular m-health apps, we have evaluated the common data protection mechanisms, protection quality, and medical-related protection criteria.

In the past 20 years, the mobile "health and well-being" app (m-health app) outburst and is increasingly popular among patients due to its better health communication means with patients. Even though it is considered part of the IoT ecosystem, it has its uniqueness for entering into the patients' private space, where the goal is to achieve a flexible and low-cost healthcare service with a better clinical decision. However, along with the convenience, the improved accuracy, increased efficiency, and enhanced productivity need sharable Personal Health Records.

However, the unregulated software and hardware (mobile phone and wearables) m-health market only promises inadequate and shaggy protection of patients' personal data. Radical changes in medical data protection are needed.

Among others, the common technical privacy risks in mobile health are unencrypted traffic, embedded advertisements, and third-party analytics services. The privacy attacks could come from the internet, third-party service, Bluetooth connections, the logging system, SD card storage, and export capabilities. Potential damages include information leaks, information manipulation or information loss, and unauthorized information access from third parties.

The main security concerns of the m-health app embrace how the patient data were collected, managed, and shared, such as how to limit the information to its minimum requirement, just enough for medical services.

We have discussed the information concerning security from the following categories:

The privacy policy needs to avoid the dubious definition for protecting users from privacy issues, such as malicious or unintentional data leakage. The permission requests of the apps need to be specific with a sound reason, such as why tracking a user's location is necessary. In addition, some of the connections do not connect using HTTPS and have static code issues concerning connection components. The security and privacy attributes should be considered when transmitting health-related data, such as the transmit form and/or storage form, and how the health multimedia content is transmitted and stored (e.g., scans of x-rays images) must also be considered. These are very private medical information and are particularly important to the patients, as is the patients' geolocation information, which must be protected from third parties. A particularly important aspect of the m-Health app in China is how to use chat session transmission, as WeChat is very popular in mainland China. A lot of medical information uses WeChat to transmit sensitive medical information with exposure risk in the none-HTTPS WeChat connections, including email addresses, passwords, names, images, and health-related questions that could be leaked through this channel.

To address these potential security and privacy requirement of the m-Health app, we need to address the functional and content requirements of the m-Health app at the policy level.

## 5. Conclusion

In this chapter, we discussed state-of-the-art methods for data integration at three levels, viz., raw data, feature, and decision, for combining clinical informatics with genomics to enable evidence-based p-Health. For integration at the raw data level, the major challenge is data harmonization and data quality control. With the unified FHIR standards established by HL7, integrating different data resources with FHIR apps is the most promising solution. For integration at the feature level, the power of conventional hand-crafted features is limited in the big data era. Advanced AI methods such as deep learning can significantly increase the feature representation power with end-to-end training and are one key direction for feature-level integration. Interpretation of integrated features at the feature level is another opportunity for translating data integration to clinical practice. For integration at the decision level, the construction of base models, the lack of interactions between different data modalities, and the translation from decisions to final clinical actions are the three major challenges. Designing novel predictive models and efficient hyper-parameter tuning algorithms, enhancing interactions between different data modalities, and inferring causal relationships are opportunities for data integration at the decision level.

In conclusion, combing EHR data with clinical genomics requires advanced data integration analytics. These data integration enabled by advanced AI methods will predict mortality risk and re-admission frequency as well as provide diagnostic and prognostic information, thus reshaping p-Health care.

## References

1. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: Definitions, benefits, and strategies for overcoming barriers to adoption. *J Am Med Inform Assoc*. 2006;13(2):121–126.
2. Buchanan BG, Shortliffe EH. *Rule-Based Expert Systems*. vol. 3. Reading, MA: Addison-Wesley; 1984.
3. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis. *Science*. 1959;130 (3366):9–21.
4. Jha AK, DesRoches CM, Campbell EG, et al. Use of electronic health records in US hospitals. *N Engl J Med*. 2009;360(16):1628–1638.
5. Charles D, Gabriel M, Furukawa MF. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2013. *ONC Data Brief*. 2013;9:1–9.
6. Health UDo, Services H. *HITECH Act Enforcement Interim Final Rule*. US Department of Health Services; 2009.
7. Prevention USCfDCa. *National Electronic Health Records Survey Public Use File National Weighted Estimates. 2019*; 2019. https://www.cdc.gov/nchs/fastats/electronic-medical-records.htm. Accessed Feburary 28, 2022.
8. Kim JW, Ryu B, Cho S, et al. Impact of personal health records and wearables on health outcomes and patient response: three-arm randomized controlled trial. *JMIR Mhealth Uhealth*. 2019;7(1).
9. Omer AE, Shaker G, Safavi-Naeini S, et al. Low-cost portable microwave sensor for non-invasive monitoring of blood glucose level: novel design utilizing a four-cell CSRR hexagonal configuration. *Sci Rep-Uk*. 2020;10(1).
10. Deleted in review.
11. NHGRI. *Fact Sheets about Genetic and Genomic Science*; 2015. https://www.genome.gov/10000202/fact-sheets/. Accessed 09 November, 2015.
12. NHGRI. *National Human Genome Research Institute (NHGRI) Definition of 'Genomic Medicine'*; 2012. https://www.genome.gov/pages/about/nachgr/sept2012agendadocuments/genomic_medicine_definition_080112_rchisolm.pdf. Accessed August, 2012.
13. Nature. *Single Nucleotide Polymorphism*; 2014. https://www.nature.com/scitable.
14. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet*. 2006;7(2):85–97.
15. Matlin AJ, Clark F, Smith CWJ. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*. 2005;6(5):386–398.

16. Hunter DJ. Gene-environment interactions in human diseases. *Nat Rev Genet*. 2005;6 (4):287–298.
17. Gulbakan B, Ozgul RK, Yuzbasioglu A, Kohl M, Deigner HP, Ozguc M. Discovery of biomarkers in rare diseases: innovative approaches by predictive and personalized medicine. *EPMA J*. 2016;7:24.
18. Hayden EC. Technology: the $1,000 genome. *Nature*. 2014;507(7492):294–295.
19. AMIA.org. *Definition of Clinical Informatics*; 2017. https://www.amia.org/applications-informatics/clinical-informatics.
20. CMS.gov. *Electronic Health Records 2012*; 2017. https://www.cms.gov/Medicare/E-Health/EHealthRecords/index.html.
21. Procter RDE. Health Informatics Journal, Edinburgh, United Kingdom. Definition of health informatics [internet]. In: *Message to: Virginia van Horne*. Bethesda, MD: Content Manager HSR Information Central; 2009.
22. Russell S, Norvig P. Intelligence A. In: *A Modern Approach. Artificial Intelligence*. Vol. 25. Egnlewood Cliffs: Prentice-Hall; 1995:27.
23. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev*. 2000;44(1.2):206–226.
24. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
25. OmicsWiki. *Omes and Omics. 2013*. http://omics.org/index.php/Omes_and_Omics. Accessed 18 Aug, 2013.
26. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. 2006;444(7118):444–454.
27. Schuster SC. Next-generation sequencing transforms today's biology. *Nat Methods*. 2008;5(1):16–18.
28. Greiwe J, Nyenhuis SM. Wearable technology and how this can be implemented into clinical practice. *Curr Allergy Asthma Rep*. 2020;20(8).
29. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(Suppl 1):D267–D270.
30. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*. 2012;13(6):395–405.
31. Berdahl CT, Nguyen AT, Diniz MA, et al. Using body temperature and variables commonly available in the EHR to predict acute infection: a proof-of-concept study showing improved pretest probability estimates for acute COVID-19 infection among discharged emergency department patients. *Diagnosis (Berlin, Germany)*. 2021;8 (4):450–457.
32. Dagliati A, Malovini A, Tibollo V, Bellazzi R. Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. *Brief Bioinform*. 2021;22 (2):812–822.
33. Garry EM, Weckstein AR, Quinto K, et al. Use of an EHR to inform a claims-based algorithm to categorize inpatient COVID-19 severity. *Pharmacoepidemiol Drug Saf*. 2021;30:93.
34. Holzer KJ, Lou SS, Goss CW, et al. Impact of changes in EHR use during COVID-19 on physician trainee mental health. *Appl Clin Inform*. 2021;12(03):507–517.
35. Icten Z, Watzker A, Friedman M, Menzin J. Thrombotic and cardiovascular events among patients hospitalized with COVID-19: findings from a large EHR database. *Pharmacoepidemiol Drug Saf*. 2021;30:366–367.
36. Nault K, Grgurich P, Stempek S, Dargin J, Gray A. Effect of EHR restriction on conservation of neuromuscular blocking agents during a COVID-19 surge. *Crit Care Med*. 2021;49(1):116.
37. Osborne TF, Veigulis ZP, Arreola DM, Roosli E, Curtin CM. Automated EHR score to predict COVID-19 outcomes at US Department of Veterans Affairs. *PLoS One*. 2020;15(7).

38. Poulos J, Zhu L, Shah AD. Data gaps in electronic health record (EHR) systems: an audit of problem list completeness during the COVID-19 pandemic. *Int J Med Inform*. 2021;150.
39. Middleton B, Bloomrosen M, Dente MA, et al. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. *J Am Med Inform Assoc*. 2013;20(e1):e2–e8.
40. Yu SC, Hofford MR, Lai AM, Kollef MH, Payne PRO, Michelson AP. Respiratory support status from EHR data for adult population: classification, heuristics, and usage in predictive modeling. *J Am Med Inform Assoc JAMIA*. 2022.
41. Sun J, Wang F, Hu J, Edabollahi S. Supervised patient similarity measure of heterogeneous patient records. *ACM SIGKDD Explor Newsl*. 2012;14(1):16–24.
42. Victora CG, Habicht J-P, Bryce J. Evidence-based public health: moving beyond randomized trials. *Am J Public Health*. 2004;94(3):400–405.
43. Ashley EA. The precision medicine initiative: a new national effort. *JAMA*. 2015;313 (21):2119–2120.
44. Hammill BG, Curtis LH, Fonarow GC, et al. Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization. *Circ Cardiovasc Qual Outcomes*. 2011;4(1):60–67.
45. Billings J, Dixon J, Mijanovich T, Wennberg D. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ*. 2006;333(7563):327.
46. Krumholz HM, Chen Y-T, Wang Y, Vaccarino V, Radford MJ, Horwitz RI. Predictors of readmission among elderly survivors of admission with heart failure. *Am Heart J*. 2000;139(1):72–77.
47. Hasan O, Meltzer DO, Shaykevich SA, et al. Hospital readmission in general medicine patients: a prediction model. *J Gen Intern Med*. 2010;25(3):211–219.
48. Cai X, Perez-Concha O, Coiera E, et al. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *J Am Med Inform Assoc*. 2015;23(3):553–561.
49. Vedomske MA, Brown DE, Harrison JH. Random forests on ubiquitous data for heart failure 30-day readmissions prediction. In: *Paper presented at Machine Learning and Applications (ICMLA), 2013 12th International Conference on 2013*; 2013.
50. Ottenbacher KJ, Smith PM, Illig SB, Linn RT, Fiedler RC, Granger CV. Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke. *J Clin Epidemiol*. 2001;54(11):1159–1165.
51. Pollack MM, Ruttimann UE, Getson PR. Pediatric risk of mortality (PRISM) score. *Crit Care Med*. 1988;16(11):1110–1116.
52. Harrell F. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer; 2015.
53. Thiébaut A, Bénichou J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Stat Med*. 2004;23(24):3803–3820.
54. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med*. 2001;23(1):89–109.
55. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Informat*. 2006;2:59.
56. Lipton ZC, Kale DC, Elkan C, Wetzell R. *Learning to Diagnose with LSTM Recurrent Neural Networks*. arXiv preprint arXiv:151103677; 2015.
57. Ridgway JP, Lee A, Devlin S, Kerman J, Mayampurath A. Machine learning and clinical informatics for improving HIV care continuum outcomes. *Curr HIV/AIDS Rep*. 2021;18(3):229–236.
58. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med*. 2004;23(7):1111–1130.

59. Matsubara Y, Sakurai Y, Van Panhuis WG, Faloutsos C. FUNNEL: automatic mining of spatially coevolving epidemics. In: *Paper presented at: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2014*; 2014.
60. Lee K, Agrawal A, Choudhary A. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In: *Paper presented at: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2013*; 2013.
61. Klein TE, Chang JT, Cho MK, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. *Pharm J*. 2001;1(3):167.
62. Wang Y, Chen R, Ghosh J, et al. Rubik: knowledge guided tensor factorization and completion for health data analytics. In: *Paper presented at: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2015.
63. Sun L, Roesler J, Rösen-Wolff A, et al. CARD15 genotype and phenotype analysis in 55 pediatric patients with Crohn disease from Saxony, Germany. *J Pediatr Gastroenterol Nutr*. 2003;37(4):492–497.
64. Kale DC, Che Z, Bahadori MT, Li W, Liu Y, Wetzel R. In: *Causal Phenotype Discovery Via Deep Networks. Paper presented at: AMIA Annual Symposium Proceedings*; 2015.
65. Ho JC, Ghosh J, Steinhubl SR, et al. Limestone: high-throughput candidate phenotype generation via tensor factorization. *J Biomed Inform*. 2014;52:199–211.
66. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci*. 2016;113(27):7353–7360.
67. Chen Y, Dales R, Tang M, Krewski D. Obesity may increase the incidence of asthma in women but not in men: longitudinal observations from the Canadian National Population Health Surveys. *Am J Epidemiol*. 2002;155(3):191–197.
68. Johansson F, Shalit U, Sontag D. Learning representations for counterfactual inference. In: *Paper presented at: International Conference on Machine Learning*; 2016.
69. Heckerman D. A Bayesian approach to learning causal networks. In: *Paper presented at: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*; 1995.
70. Duncan DT, Hickson DA, Goedel WC, et al. The social context of HIV prevention and care among black men who have sex with men in three US cities: the neighborhoods and networks (N2) cohort study. *Int J Env Res Pub He*. 2019;16(11).
71. Liu Y-Y, Moreno A, Li S, Li F, Song L, Rehg JM. Learning continuous-time hidden Markov models for event data. In: *Mobile Health*. Springer; 2017:361–387.
72. Che Z, Purushotham S, Kale D, et al. Time series feature learning with applications to health care. In: *Mobile Health*. Springer; 2017:389–409.
73. Dempsey WH, Moreno A, Scott CK, et al. iSurvive: an interpretable, event-time prediction model for mHealth. In: *Paper presented at: International Conference on Machine Learning*; 2017.
74. Paweletz CP, Charboneau L, Bichsel VE, et al. Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene*. 2001;20(16):1981–1989.
75. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*. 2016;3.
76. Jack CR, Bernstein MA, Fox NC, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imaging*. 2008;27(4):685–691.
77. Marek K, Jennings D, Lasch S, et al. The Parkinson progression marker initiative (PPMI). *Prog Neurobiol*. 2011;95(4):629–635.
78. Statistics NCfH. *Instructions for Classifying Multiple Causes of Death*. NCHS Instruction Manual; 2005.
79. Team HS. FHIRbase: open source storage based on the FHIR standard ready for use in production. In: *FHIR Standard Gives Specific Directions for Exchanging Structured Medical Data and FHIRbase Is Developed to Easily Store and Retrieve Medical Data in the FHIR Format*; 2017. Accessed 08/18, 2017. Available at http://fhirbase.github.io/index.html.

80. Labs C. *EHRServer: Open Source, Service-Oriented, Clinical Data Repository*; 2017. https://cabolabs.com/en/projects.
81. Ng K, Ghoting A, Steinhubl SR, Stewart WF, Malin B, Sun JM. PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. *J Biomed Inform*. 2014;48:160–170.
82. Badgeley MA, Shameer K, Glicksberg BS, et al. EHDViz: clinical dashboard development using open-source technologies. *BMJ Open*. 2016;6(3).
83. Center T. *Open mHealth: The First and Only Open Standard for Mobile Health Data*; 2017. http://www.openmhealth.org/.
84. King MC, Marks JH, Mandell JB, Grp NYBCS. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science*. 2003;302(5645):643–646.
85. Esplin ED, Oei L, Snyder MP. Personalized sequencing and the future of medicine: discovery, diagnosis and defeat of disease. *Pharmacogenomics*. 2014;15(14):1771–1790.
86. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med*. 2004;10(8):789–799.
87. Strausberg RL, Simpson AJ, Old LJ, Riggins GJ. Oncogenomics and the development of new cancer therapies. *Nature*. 2004;429(6990):469–474.
88. Nguyen PL, Taghian AG, Katz MS, et al. Breast cancer subtype approximated by estrogen receptor, progesterone receptor, and HER-2 is associated with local and distant recurrence after breast-conserving therapy. *J Clin Oncol*. 2008;26(14):2373–2378.
89. Zheng SL, Sun JL, Wiklund F, et al. Cumulative association of five genetic variants with prostate cancer. *N Engl J Med*. 2008;358(9):910–919.
90. Chapman PB, Hauschild A, Robert C, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med*. 2011;364(26):2507–2516.
91. Altshuler D, Durbin RM, Abecasis GR, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061–1073.
92. Telenti A, Pierce LC, Biggs WH, et al. Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci U S A*. 2016;113(42):11901–11906.
93. Ecker JR. FORUM: genomics ENCODE explained. *Nature*. 2012;489(7414):52–53.
94. Allen N, Sudlow C, Downey P, et al. UK biobank: current status and what it means for epidemiology. *Health Policy Tech*. 2012;1(3):123–126.
95. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*. 2012;366(10):883–892.
96. Mueller SG, Weiner MW, Thal LJ, et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am*. 2005;15(4):869.
97. Horbal SR, Seffens W, Davis AR, et al. Associations of Apelin, Visfatin, and urinary 8-Isoprostane with severe hypertension in African Americans: the MH-GRID study. *Am J Hypertens*. 2016;29(7):814–820.
98. Liao XY, Lochhead P, Nishihara R, et al. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. *N Engl J Med*. 2012;367(17):1596–1606.
99. Pharoah PD, Tsai YY, Ramus SJ, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat Genet*. 2013;45(4):362–370. 370e361–362.
100. Accurso FJ, Rowe SM, Clancy JP, et al. Effect of VX-770 in persons with cystic fibrosis and the G551D-CFTR mutation. *N Engl J Med*. 2010;363(21):1991–2003.
101. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. 2014;506(7488):376.
102. Vogel CL, Cobleigh MA, Tripathy D, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J Clin Oncol*. 2002;20(3):719–726.
103. Soda M, Choi YL, Enomoto M, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*. 2007;448(7153):561–566.

104. Gire SK, Goba A, Andersen KG, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*. 2014;345(6202):1369–1372.
105. Harold D, Abraham R, Hollingworth P, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet*. 2009;41(10):1088–U1061.
106. de Ligt J, Willemsen MH, van Bon BW, et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med*. 2012;367(20):1921–1929.
107. Moffitt RA, Yin-Goen Q, Stokes TH, et al. caCORRECT2: improving the accuracy and reliability of microarray data in the presence of artifacts. *BMC Bioinformatics*. 2011;12:383.
108. Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 26 Apr, 2010.
109. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
110. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43 (5):491.
111. Quinlan AR, Clark RA, Sokolova S, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res*. 2010;20(5):623–635.
112. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat Protoc*. 2012;7(3):562–578.
113. Phan JH, Kothari S, Wang MD. omniClassifier: a desktop grid computing system for big data prediction modeling. *ACM BCB*. 2014;2014:514–523.
114. Phan JH, Young AN, Wang MD. omniBiomarker: a web-based application for knowledge-driven biomarker identification. *IEEE Trans Biomed Eng*. 2013;60 (12):3364–3367.
115. Stokes TH, Torrance JT, Li H, Wang MD. ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses. *BMC Bioinformatics*. 2008;9(Suppl 6):S18.
116. Wong L. Technologies for integrating biological data. *Brief Bioinform*. 2002;3 (4):389–404.
117. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform*. 2008;41(5):687–693.
118. Gomez-Cabrero D, Abugessaisa I, Maier D, et al. Data integration in the era of omics: current and future challenges. *BMC Syst Biol*. 2014;8(Suppl 2):I1.
119. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. 2015;16(2):85–97.
120. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009;323(5910):133–138.
121. Rahm E, Do HH. Data cleaning: problems and current approaches. *IEEE Data Eng Bull*. 2000;23(4):3–13.
122. Shi LM, Reid LH, Jones WD, et al. The MicroArray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*. 2006;24(9):1151–1161.
123. Consortium SM-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat Biotechnol*. 2014;32(9):903–914.
124. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat Med*. 2011;17(3):297–303.
125. Vapnik VN, Vapnik V. *Statistical Learning Theory*. Vol. 1. New York: Wiley; 1998.
126. Dolin RH, Alschuler L, Boyer S, et al. HL7 clinical document architecture, release 2. *J Am Med Inform Assoc*. 2006;13(1):30–39.

127. Clifton C, Kantarcioğlu M, Doan A, et al. Privacy-preserving data integration and sharing. In: *Paper presented at: Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*; 2004.
128. Bender D, Sartipi K. HL7 FHIR: an Agile and RESTful approach to healthcare information exchange. In: *Paper presented at: Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on 2013*; 2013.
129. Franz B, Schuler A, Kraus O. Applying FHIR in an integrated health monitoring system. *EJBI*. 2015;11(2):en56–en61.
130. Alterovitz G, Warner J, Zhang P, et al. SMART on FHIR genomics: facilitating standardized clinico-genomic apps. *J Am Med Inform Assoc*. 2015;22(6):1173–1178.
131. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. *J R Stat Soc Ser B Stat Methodol*. 2008;70(1):53–71.
132. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*. 2005;3(02):185–205.
133. Gönen M, Alpaydın E. Multiple kernel learning algorithms. *J Mach Learn Res*. 2011;12:2211–2268.
134. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. *Neural Comput*. 2004;16(12):2639–2664.
135. Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press; 2009.
136. Srivastava N, Salakhutdinov RR. Multimodal learning with deep Boltzmann machines. In: *Paper presented at: Advances in Neural Information Processing Systems*; 2012.
137. Andrew G, Arora R, Bilmes J, Livescu K. Deep canonical correlation analysis. In: *Paper presented at: International Conference on Machine Learning*; 2013.
138. Havaei M, Guizard N, Chapados N, Bengio Y. HeMIS: hetero-modal image segmentation. In: *Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention*; 2016.
139. Arya N, Saha S. Multi-modal advanced deep learning architectures for breast cancer survival prediction. *Knowl-Based Syst*. 2021;221.
140. De Silva T, Chew EY, Hotaling N, Cukras CA. Deep-learning based multi-modal retinal image registration for the longitudinal analysis of patients with age-related macular degeneration. *Biomed Opt Express*. 2021;12(1):619–636.
141. Islam KT, Wijewickrema S, O'Leary S. A deep learning based framework for the registration of three dimensional multi-modal medical images of the head. *Sci Rep-UK*. 2021;11(1).
142. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. *Sci Rep*. 2021;11(1):3254.
143. Zhang F, Li Z, Zhang B, Dua H, Wang B, Zhang X. Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease. *Neurocomputing*. 2019;361:185–195.
144. Lee G, Nho K, Kang B, et al. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Sci Rep-UK*. 2019;9.
145. Ribeiro MT, Singh S, Guestrin C. Why should i trust you? Explaining the predictions of any classifier. In: *Paper presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016.
146. Koh PW, Liang P. *Understanding Black-Box Predictions via Influence Functions. arXiv preprint arXiv:170304730*; 2017.
147. Dietterich TG. Ensemble methods in machine learning. *Multiple Classifier Syst*. 2000;1857:1–15.
148. Shen HB, Chou KC. Ensemble classifier for protein fold pattern recognition. *Bioinformatics*. 2006;22(14):1717–1722.
149. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123–140.

150. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;1189–1232.
151. Akavia UD, Litvin O, Kim J, et al. An integrated approach to uncover drivers of cancer. *Cell*. 2010;143(6):1005–1017.
152. Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*. 2006;22(14):e184–e190.
153. Rasiwasia N, Costa Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval. In: *Paper presented at: Proceedings of the 18th ACM International Conference on Multimedia*; 2010.
154. Bottou L, Peters J, Quiñonero-Candela J, et al. Counterfactual reasoning and learning systems: the example of computational advertising. *J Machine Learn Res*. 2013;14 (1):3207–3260.
155. Lam HYK, Marenco L, Clark T, et al. Research—AlzPharm: integration of neurodegeneration data using RDF. *BMC Bioinformatics*. 2007;8.
156. Sujansky W. Heterogeneous database integration in biomedicine. *J Biomed Inform*. 2001;34(4):285–298.
157. Holzinger ER, Dudek SM, Frase AT, Krauss RM, Medina MW, Ritchie MD. ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. *Pac Symp Biocomput*. 2013;385–396.
158. Fridley BL, Lund S, Jenkins GD, Wang L. A Bayesian integrative genomic model for pathway analysis of complex traits. *Genet Epidemiol*. 2012;36(4):352–359.
159. Mankoo PK, Shen R, Schultz N, Levine DA, Sander C. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One*. 2011;6(11):e24709.
160. Ovaska K, Laakso M, Haapa-Paananen S, et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med*. 2010;2 (9):65.
161. Kovačević A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *J Am Med Inform Assoc*. 2013;20(5):859–866.
162. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. *Bioinformatics*. 2004;20(16):2626–2635.
163. Daemen A, Gevaert O, De Moor B. Integration of clinical and microarray data with kernel methods. In: *Paper presented at: Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE 2007*; 2007.
164. Kim D, Shin H, Song YS, Kim JH. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform*. 2012;45(6):1191–1198.
165. Tsuda K, Shin H, Scholkopf B. Fast protein classification with multiple networks. *Bioinformatics*. 2005;21(Suppl 2):ii59–ii65.
166. Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *J Am Med Inform Assoc*. 2014;21(e2):e278–e286.
167. Brix-Christensen V. The systemic inflammatory response after cardiac surgery with cardiopulmonary bypass in children. *Acta Anaesthesiol Scand*. 2001;45(6):671–679.
168. Leclerc F, Leteurtre S, Duhamel A, et al. Cumulative influence of organ dysfunctions and septic state on mortality of critically ill children. *Am J Resp Crit Care*. 2005;171 (4):348–353.