Integrating Multi-Omics Data With EHR for Precision Medicine Using Advanced Artificial Intelligence

Li Tong, Wenqi Shi, *Graduate Student Member, IEEE*, Monica Isgut, Yishan Zhong, Peter Lais, Logan Gloster, Jimin Sun, Aniketh Swain, Felipe Giuste, and May D. Wang, *Fellow, IEEE*

(Methodological Review)

***Abstract*—With the recent advancement of novel biomedical technologies such as high-throughput sequencing and wearable devices, multi-modal biomedical data ranging from multi-omics molecular data to real-time continuous bio-signals are generated at an unprecedented speed and scale every day. For the first time, these multi-modal biomedical data are able to make precision medicine close to a reality. However, due to data volume and the complexity, making good use of these multi-modal biomedical data requires major effort. Researchers and clinicians are actively developing artificial intelligence (AI) approaches for data-driven knowledge discovery and causal inference using a variety of biomedical data modalities. These AI-based approaches have demonstrated promising results in various biomedical and healthcare applications. In this review paper, we summarize the state-of-the-art AI models for integrating multi-omics data and electronic health records (EHRs) for precision medicine. We discuss the challenges and opportunities in integrating multi-omics data with EHRs and future directions. We hope this review can inspire future research and developing in integrating multi-omics data with EHRs for precision medicine.**

***Index Terms*—Multi-omics, electronic health records, data integration, artificial intelligence, precision.**

I. Introduction

PRECISION medicine is a medical model that aims to provide customized healthcare for patients, including targeted prevention, precise diagnosis, personalized treatments, and accurate prognosis (National Research Council Committee on A Framework for Developing a New Taxonomy of Disease, 2012). Compared with the conventional one-drug-fits-all model, precision medicine aims to improve healthcare for each individual by identifying the most appropriate medical decisions and performing the most effective procedures for each person. With precision medicine in practice, optimized healthcare could be delivered to each individual while minimizing the overall healthcare cost for the whole society. In addition, it enables population-level analysis of multi-modal biomedical data to identify patient-specific differences beyond electronic health records (EHRs) in the era of Big Data.

The advancements in biotechnologies, such as next-generation sequencing (NGS), coupled with the proliferation of smart wearable devices, permit physicians to access detailed patient profiles. These range from molecular or cellular details like multi-omics data that encapsulate genetic mutations and gene expression, to objective behavior profiles chronicling daily activities and lifestyles. Furthermore, the multi-omics data enabled by techniques such as NGS can provide information about patient-level genetic variations [1]. For example, during the COVID-19 pandemic, genome sequencing was utilized to identify the subtypes of COVID-19 variants [2] and predict patients' prognoses [3], [4]. Genetic profiling can help doctors identify specific mutations in cancer diagnosis and treatment to choose the best treatment [5], which is enabled by a novel paradigm in drug development called pharmacogenomics [6]. Similarly, smart wearable devices serve as a powerful tool to keep track of personal health information for personalized diagnosis and optimized treatment. For example, diabetes monitoring devices can improve insulin dosing for diabetes patients daily [7]. In

Manuscript received 23 April 2023; revised 9 September 2023; accepted 29 September 2023. Date of publication 12 October 2023; date of current version 15 January 2024. The work of May D. Wang was supported in part by a Wallace H. Coulter Distinguished Faculty Fellowship, in part by a Petit Institute Faculty Fellowship, in part by Amazon Faculty Fellowship, and in part by Microsoft Research. *(Corresponding author: May D. Wang.)*

Li Tong and Felipe Giuste are with the Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Emory University, Atlanta, GA 30322 USA (e-mail: ltong9@gatech.edu; fgiuste@gatech.edu).

Wenqi Shi and Yishan Zhong are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: wshi83@gatech.edu; yzhong307@gatech.edu).

Monica Isgut is with the School of Biology, Georgia Institute of Technology, Atlanta, GA 30322 USA (e-mail: misgut@gatech.edu).

Peter Lais, Logan Gloster, and Aniketh Swain are with the Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30322 USA (e-mail: plais3@gatech.edu; lgloster@gatech.edu; aswain9@gatech.edu).

Jimin Sun is with the School of Computer Science, Georgia Institute of Technology, Atlanta, GA 30322 USA (e-mail: jiminsun@gatech.edu).

May D. Wang is with the Wallace H Coulter Department of Biomedical Engineering, the School of Electrical and Computer Engineering, the School of Computational Science and Engineering, the Institute of People and Technology, Winship Cancer Institute, and the Institute of Bioengineering and Bioscience, Georgia Institute of Technology and Emory University, Atlanta, GA 30322 USA (e-mail: maywang@gatech.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/RBME.2023.3324264, provided by the authors.

Digital Object Identifier 10.1109/RBME.2023.3324264

the era of biomedical data, precision medicine has made significant progress with the availability of multi-omics data and EHRs. However, making the most of these big biomedical data, especially integrating data from various modalities, remains a significant challenge to maximize the efficacy of precision medicine.

Multi-omics studies aim to understand human beings and diseases using molecular profiles at multiple levels, including the genome, transcriptome, epigenome, and metabolome. They can reflect biological activities and corresponding pathological changes at the molecular level for each individual. Thus, multi-omics data are promising to fill the gaps in understanding the disease at a personalized level in precision medicine. By analyzing these complex biological Big Data, researchers can discover novel associations between biological entities and identify biomarkers for disease diagnosis and treatment responses. However, the high-throughput multi-omics data are too large and complicated to be manually analyzed. Researchers have applied various machine learning techniques to enable computer-aided analysis of multi-omics data [7], aiding biomarkers discovery, diagnosis, and prognosis.

Electronic health records (EHRs) serve as digital version of a patient's medical records, including demographics, medical histories, vital signs, lab tests, radiology images, pathological images, diagnoses, treatment procedures, and medications. With the development of smart wearable devices, the EHRs have been extended into personal health records (PHRs) with out-of-clinic data such as the patient's daily behavior, activities, and physiological data. In precision medicine, the EHR also serves as a bridge to connect with extra patient information, such as the genetic profiles in multi-omics data. While they remain more interpretable than multi-omics data, their increasing complexity, especially with data types like extensive ECG monitoring and whole-slide imaging, have made machine learning vital for effective data management. Moreover, the variety of EHR data modalities is increasing rapidly. Thus, researchers seek help from machine learning to assist in feature extraction, feature selection, and predictive modeling.

Although multi-omics and EHR data have been widely explored for precision medicine in recent years, integrating multi-omics and EHR data is still an under-explored field. Integrating multi-omics data with EHR data can provide a more thorough patient evaluation. The patient's multi-omics data containing molecular-level features can help clinicians understand the disease and discover new genetic/cellular biomarkers. At the same time, the EHR data can contribute to clinical biomarkers discoveries and serve as quantitative phenotype targets to evaluate treatment outcomes. By combining both modalities, precision medicine can achieve more accurate subtyping of diseases and more personalized patient treatment.

In this review, we will first present the modality-specific learning in Sections II and III. We will then introduce the cross-modality integration between multi-omics data and EHR data in Section IV. We will then discuss the challenges and opportunities for this cross-modality integration in Section V. Detailed paper search strategy with PRISM is available in supplementary materials.

II. Modality-Specific Learning: Multi-Omics Data

A growing number of studies have utilized multi-omics data for various precision medicine-related machine learning tasks, including disease diagnosis, patient sub-phenotyping, and disease or mortality risk prediction. However, data from different omics modalities can be heterogeneous and noisy, with various underlying distributions and scales. Thus, the way these data are integrated significantly impacts downstream processing. This section discusses our findings from 18 recent precision medicine-based works with respect to different omics modalities, basic pre-processing steps, and multi-omics integration methods used. A summary of the reviewed papers is illustrated in **Fig. S1** supplementary materials.

A. Precision Medicine Endpoints, Diseases, and Datasets

Out of the reviewed papers, the vast majority were focused solely on one or more types of cancer. In contrast, three of the studies reviewed applied their models to data on depression [8], inflammatory bowel disease (IBD) [9], and Alzheimer's disease [10]. The reviewed papers focused primarily on cancer subtyping or subtype classification (n = 7), risk prediction or cancer survival (n = 11), drug response (n = 2), and disease diagnosis (n = 2), including some papers that had multiple clinical endpoints. **Table S1** in supplementary materials describes some of the disease-specific multi-omics datasets used. The Cancer Genome Atlas (TCGA), which was used in almost 80% (14 out of 18) of the studies, was by far the most common multi-omics dataset used. The Genomics of Drug Sensitivity in Cancer (GDSC) was the second most common dataset and was used in two studies [11] on cancer drug response prediction.

B. Omics Modalities, Pre-Processing, Quality Control, and Defining Features

The main categories of omics data used in the reviewed papers included transcriptomic (miRNA, mRNA), epigenomic (DNA methylation), copy number variation (CNV), genomic (somatic mutation), and protein expression data. These data are collected using various methods, including DNA or RNA sequencing, RNA, or single nucleotide polymorphism (SNP) microarrays, and reverse-phase protein arrays. Table I describes some of the standard omics modality-specific data collection and preprocessing steps. An essential aspect of quality control relates to the handling of missing data. Out of the papers that discussed the challenge of missing data, the most common approach involved removing features or samples with significant missing data [12], and to a lesser extent, imputing missing values with the median of the known values for each feature or using an established algorithm [12] such as k-nearest neighbor (k-NN).

In addition to basic pre-processing and quality control, another important consideration for many studies was defining input features for each omics modality. The most common approach was to map the original large-scale features from each omics dataset into genes as features. Examples of original features include DNA methylation values at specific sites

TABLE I
OMICS MODALITY PRE-PROCESSING

Omics Data Modality	Approaches/Platforms	General / Common Pre-Processing Steps
Whole genome/exome sequencing	Agilent SureSelectXT Human All Exon 50Mb bait set and Illumina HumanHT-12 v3 Expression BeadChip	BAM to FASTQ conversion BWA-based alignment to GRCh38.d1.vd1 Duplicate read/artifact flagging Indel local realignment Base quality score recalibration
mRNA-Seq	Affymetrix Human Genome U219 Array	BAM to FASTQ conversion STAR alignment to GRCh38.d1.vd1 Gene expression count using HT-Seq-Count Expression normalization using HT-Seq
scRNA-Seq	10X Genomics platforms	10X Genomics Cell Ranger pipeline: 1) Conversion to FASTQ 2) Alignment, barcoding, UMI counting 3) Aggregation of run counts and normalization of runs to same sequencing depth 4) Dimensionality reduction of feature-barcode matrices
miRNA	MegaplexTM Primer Pools and TaqMan MicroRNA Array	BAM or FASTQ alignment using BWA Read count normalization miRNA expression quantification and isoform mapping
CNV	Affymetrix SNP 6.0 array	Copy number segmentation and masking Copy number estimation
CpG Methylation	Illumina Human Methylation (HM27) and HM450 arrays	Mapping to GRCh38.d1.vd1 Type I and II probe quality screening and coordinate generation Transcript identification in GENCODE v22
Reverse Phase Protein Array	Aushon Biosystems 2470 arrayer	Slide quantification of spot intensity Relative protein level determination by standard protein curve Raw spot spatial intensity correction from controls Quality control scoring and protein loading correction

within genes or RNA transcripts derived from genes. These mapped gene-level feature matrices (samples as rows and genes as columns) would be used for downstream processing steps. For example, TCGA DNA methylation data consist of CpG island methylation values as features, and there can be many for a given gene. Takahashi et al. [13] and Chai et al. [14] found the average values for all CpG sites on each gene promoter and encoded the data as gene-specific features. Xu et al. [15] used dimensionality reduction for all CpGs on each gene to extract a single feature value. Another less common approach involved using the original large-scale features [14], [16] from each given omics data modality rather than transforming all omics datasets into *gene feature* matrices. For example, Zhang et al. [16] used approximately 350 k total features from TCGA RNA sequence, miRNA, copy number variation (CNV), and DNA methylation datasets. They scaled the features from different modalities to be in the same range. Tong et al. [17] used approximately 26 k DNA methylation features representing CpG binding sites without transforming these into gene-specific features.

C. Multi-Omics Integration Methods

Several recent reviews have defined varying categories for multi-omics integration methods. In 2019, Rappoport et al. [18] defined three methods used in cancer subtyping research studies: early integration, late integration, and middle integration. In 2020, Subramanian et al. [19] defined five multi-omics integration methods that are not mutually exclusive: similarity, correlation, network, Bayesian, multivariate, and fusion. Picard et al. [20] categorized multi-omics integration methods for machine learning applications into early, mixed, intermediate, late, and hierarchical. Given that our reviewed works focus on machine learning, we will use the categorization framework defined by Picard et al. to classify the integration methods used in the papers (see Fig. 1).

1) Dimensionality Reduction: Independently of the specific multi-omics integration methods used, most papers incorporated some form of dimensionality reduction to reduce the size of the feature space. In concordance with Picard et al., we define two categories of dimensionality reduction: a) feature extraction (establishing a learned feature embedding by transforming original features into a new feature space) and b) feature selection (removing unimportant or noisy features). The most common feature extraction methods used in the reviewed studies include principal components analysis (PCA) and autoencoders. Feature selection methods can be classified as filter-based, wrapper-based, and embedded [20]. Filter-based methods use statistical tools to estimate the feature's importance or relevance, removing the least important features from the dataset. Examples from some of the reviewed studies include maximum relevance minimum redundancy (mRMR) [21], [22], the Chi-squared test [23], and the Wilcoxon rank-sum test [24]. Wrapper-based methods involve applying a statistical learning model to different sets of features to find those that achieve the best performance, and a frequently used example in several reviewed studies was Cox proportional hazards regression. In these studies, a Cox model is trained on the input features to identify the most important features for survival prediction, and other features are removed from the dataset. Embedded methods are built into machine learning-based classifiers, such as L1 regularization. For papers using early multi-omics integration methods, dimensionality reduction was typically achieved using the concatenated multi-omics dataset. Whereas, for intermediate, mixed, late, and/or hierarchical integration, dimension reduction was performed individually for each omics modality.

2) Early Integration: A few of the reviewed papers used early integration approaches [9], [15], whereby they combined

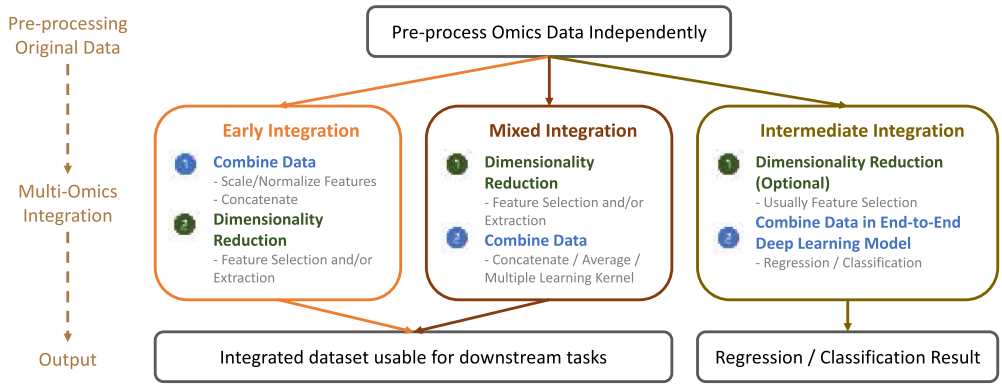

Fig. 1. Illustration of general framework for intermediate, mixed, and early integration. The figure illustrates a general flow of steps for each multi-omics integration category. After basic omics modality-specific data pre-processing, the integration step differs based on the type of framework used. Early integration involves combining the data first, usually via concatenation, and then performing dimensionality reduction which can be feature selection and/or extraction. Mixed integration involves dimensionality reduction first on each omics modality independently and combining the data using any of several approaches. Finally, intermediate integration (specifically – the deep learning-based methods) involves an optional dimensionality reduction step that is usually feature selection, followed by an end-to-end classification or regression model. Note that not all methods under each category follow the same framework. For example, some papers under mixed integration might have a further feature selection or extraction step after the combining data step. Also, the intermediate integration framework illustrates end-to-end models that integrate the omics data in a single neural network, but this is not the only process for intermediate integration.

large-scale multi-omics features into a single feature matrix after basic omics-specific pre-processing. The early integration often involved scaling or normalizing the original features. Then dimensionality reduction was performed collectively for the dataset containing all omics modalities rather than individually for each one. For example, Zhang et al. [25] concatenated pre-processed gene-level CNV features and RNA sequence features into a single matrix. They used feature extraction methods such as PCA and autoencoder to obtain a 100-feature embedding from that matrix, followed by feature selection based on the embedding using Cox regression to identify the most highly informative features for cancer survival. Another study [16] combined 350 k features from four different omics modalities into a matrix and used an autoencoder to obtain a 100-feature latent space embedding, followed by Cox regression to select survival-informative features. A similar early integration approach has also been adopted by Chai et al. [14].

3) Mixed Integration: Many of the reviewed studies used mixed integration methods [11]. After omics-specific data pre-processing, dimensionality reduction was performed individually for each separate omics modality. These processed data were combined into a single dataset via concatenation, averaging, or other methods. The most common feature extraction method was autoencoders [12], which learn new features using non-linear transformations of original features from each omics dataset, incorporating information about the distribution of each dataset. For dimensionality reduction, Tong et al. [17] first used PCA or unsupervised variance-based feature selection (UVFS) separately for each of four different omics data modalities to identify the top 100 (from PCA) or top 1000 features (from UVFS) from each modality with the greatest variance across tumor samples. Then, they used each low-dimensional omics dataset as input into its own autoencoder for further feature extraction. The authors tested two different integration methods, the best-performing one of which involved concatenating the autoencoder-derived latent space embeddings of all the omics modalities (ConcatAE). This combined dataset was then used for a downstream cancer survival prediction task. Poirion et al. [26] (DeepProg) used a similar approach, whereby they input each normalized omics dataset into its own autoencoder and then used the autoencoder-derived latent space from each omics dataset in a separate Cox proportional hazards model for feature selection to identify the features most informative of cancer survival. After feature selection, the features were concatenated into a multi-omics feature set used for downstream sub-phenotyping tasks. Takahashi et al. [13] followed a very similar overall framework. Other mixed integration methods are based on multiple kernel learning (MKL) instead of concatenation to combine omics features after pre-processing and dimensionality reduction [22], [24]. For example, SimpleMKL and SMO-MKL were used in two separate breast cancer-related studies to define a kernel for each omics modality for survival and subtype classification, respectively.

Survival Analysis Learning with Multi-Omics Neural Networks (SALMON) [27] used mixed- and hierarchical-integration approaches. Hierarchical methods use an external knowledge base for integration, such as known regulatory relationships between features in different omics modalities. SALMON leveraged external knowledge to extract features from each omics modality before combining them for downstream processing. They used a gene co-expression to integrate mRNA and miRNA omics data and then developed an eigengene matrix for each data modality, which was concatenated and used as an input dataset for downstream processing.

4) Intermediate Integration: Intermediate integration methods process omics-specific data while simultaneously considering information across the different omics modalities. A few examples of methods explicitly designed for data integration include iCluster [28], multi-factor analysis (MOFA) [29], similarity network fusion (SNF) [30], and multiple co-inertia analysis (MCIA) [31]. They were not commonly used in the reviewed studies, but the most salient trend identified for intermediate integration was the use of end-to-end deep learning models. Each omics dataset was input into its sub-network of a

classification model. Given that the omics data are integrated within the same model, information from within and between omics datasets was used to learn higher-level multi-omics features for classification.

For example, Hierarchical Integration Deep Flexible Neural Forest (HI-DFNForest) [32] was designed to use each omics data modality as input into its encoding subnetwork. The latent embeddings from the omics modalities were combined in a subsequent layer, followed by deep flexible neural forest feedforward layers to classify patients into different cancer subtypes. Lin et al. performed -omics modality-specific dimensionality reduction using the Chi-squared test to select the top 5000 features. They then input these data into a neural network called Deep Neural Networks based on Multi-Omics (DeepMO) [23], which also used an encoding sub-network for each omics modality, followed by another sub-network that combined the latent features of each modality for breast cancer subtype classification. The same general integration approach was adopted by Sharifi-Noghabi et al. [33], whereby each omics modality was input into its encoding subnetwork, followed by a concatenation of the latent space embedding from each modality and then a downstream drug response classification task in the same end-to-end network. Hassanzadeh and Wang [21] recently trained an end-to-end deep belief network model for cancer survival classification using mRNA, miRNA, and DNA methylation data. Multi-Omics Graph Convolutional Networks (MOGONET) [10] was another example of an intermediate integration method. The authors trained three omics-specific graph convolutional networks for a classification task, and each predicted its own set of labels. A tensor was set up and used as input into a View Correlation Discovery Network (VCDN) which predicted the final classification labels.

III. Modality-Specific Learning: EHR Data

EHR primarily contains two types of data: **(1) unstructured data**, such as clinical notes, medical imaging, and **(2) structured data**, such as administrative information, vital signs, laboratory results, diagnosis codes and results, prescriptions, and medications [34]. **Table S2** in supplementary materials summarizes the frequently used publicly accessible EHR datasets, including structured and unstructured EHR data.

A. Pre-Processing and Quality Control of Unstructured and Structured EHR Data

Unstructured EHR data, commonly referred to as free-text clinical notes describing a patient's condition, is the most efficient and intuitive method of clinical documentation. While this free-text format is convenient for expressing concepts and events, it makes searching, summarizing, decision support, and statistical analysis difficult because of **(1) flexible formatting, (2) atypical grammar, (3) detailed descriptions, and (4) abundant misspellings** [35]. Basic pre-processing steps include converting the numeric values to strings, converting text to the UTF-8 standard encoding format, normalizing entities, segmenting sentences, removing stop words and punctuation, and removing extraneous spaces, lines, or characters [36]. Additional challenges are associated with reusing structured EHR data, particularly **data heterogeneity, quality, and high dimensionality** [34]. Table II describes widely used EHR standards for digitally storing information related to patient admissions, conditions, care procedures/results of these procedures, and outcomes. A detailed review of exiting EHR data standard [37], [38] is available in the supplementary materials.

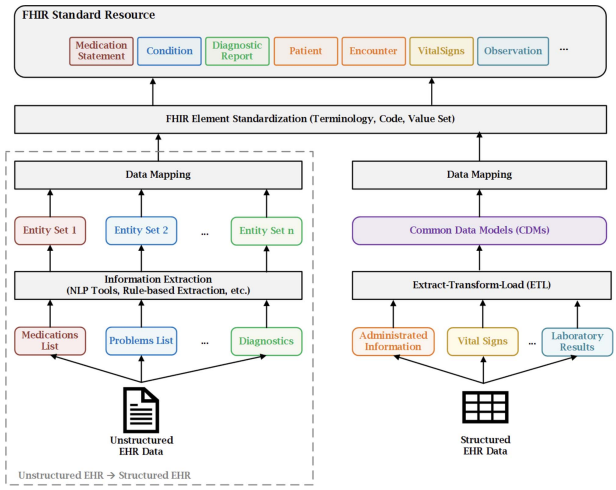

Fig. 2. Framework for integrating structured and unstructured EHR data using FHIR. This figure illustrates the framework for integrating structured and unstructured EHR data using the HL7 FHIR standard. Modeling of EHR data based on standards is essential for data interoperability and widespread use. Due to the heterogeneous approaches utilized by existing clinical systems, integrating unstructured data into a standardized data model presents unique challenges. FHIR has been extensively utilized to normalize and integrate both structured and unstructured EHR data by providing a standardized representation of a vast array of healthcare concepts.

B. EHR Data Integration

Due to the widespread utilization of EHRs in hospital systems, there are significant demands for semantic interoperability and large-scale analysis in clinical and translational research [39]. However, the lack of interoperability of EHR data between institutions complicates secondary use of EHR data, particularly in collaborative research across institutions. By representing structured EHR data with a common data model (CDM), we can facilitate large-scale data research collaboration while also enabling advanced research in precision medicine. Table III summarizes the CDMs frequently used in research when a set of structured EHR data needs to be exchanged or shared. In addition, FHIR has been widely used [37], [40] to normalize and integrate structured and unstructured EHR data by providing a standardized representation of a wide variety of healthcare concepts, as shown in Fig. 2. A detailed review of unstructured and structured EHR data integration [29], [41], [42] is available in the supplementary materials.

C. EHR Data Analysis Methods

Despite their primary purpose of operational efficiency improvement in healthcare, EHRs have been discovered to have a secondary use in clinical informatics [34]. In particular, EHRs in health information systems have been widely used for precision

TABLE II
EHR DATA STANDARDS: THESE MODELS SPECIFY STANDARDS FOR DIGITALLY STORING INFORMATION RELATED TO PATIENT ADMISSIONS, CONDITIONS, CARE PROCEDURES/RESULTS OF THESE PROCEDURES, AND OUTCOMES

Name	Description
HL7 v2	Successor to HL7 v1. Plaintext encoding of admissions, discharges, and transfers as well as exchanging orders and reports for tests and treatment. The most used healthcare interchange format. Custom software is often needed to parse and view results. Flexible but subject to variability in implementation.
HL7 v3	Successor to HL7 v2 that sought to standardize data interchange using a structured reference information model (RIM) and object-oriented ideas. Healthcare data are partitioned into four core classes with predefined relationships available between them (supplied through two additional classes).
FHIR	Based on the HL7 data-sharing standards. A RESTful protocol for transferring EHR data that returns data in a variety of machine- or human-readable formats. Built on previous iterations of HL7 and allows for web-based technologies to take advantage of these data.
CDA	The most widely adopted application of HL7 v3 is designed to be *human-readable*. Heavily incorporates XML into its organizational structure.
OpenEHR	A highly generic architecture aimed at clearly defining data semantics to be robust in the face of rapidly changing information, technology, and patient movement between healthcare systems. Developed by OpenEHR International.
xDT	A communications standard largely used in German medical centers specifying details for laboratory (LDT), accounting (ADT), treatment (BDT), and device (GDT) data transfer.
UMLS	A set of files and software that aims to promote interoperability between systems by integrating many health and biomedical vocabularies/standards. Maintained by the U.S. National Library of Medicine, a member of SNOMED International.
SNOMED CT	A designated standard for use in U.S. Federal Government systems for the electronic exchange of clinical health information and is also a required standard in interoperability specifications of the U.S. Healthcare Information Technology Standards Panel. Designed around benefiting both individuals and populations as well as facilitating evidence-based healthcare decisions.
ICD	A schema that allows for systematic classification of *diseases*, permitting the capable analysis of healthcare data and increased interoperability between health systems. Published by the World Health Organization.
CPT	A schema that allows for the systematic classification of *healthcare services* provided to patients as well as procedures. Developed and maintained by the American Medical Association.
LOINC	A schema that allows for the systematic classification of *medical laboratory tests and observations*. Publicly available at no cost. Developed and maintained by the Regenstrief Institute.
RxNorm	A collection of normalized names for clinical drugs and associated names in widely used pharmacy management and drug interaction software. Maintained by the U.S. National Library of Medicine.

TABLE III
EHR COMMON DATA MODELS (CDMS): CDMS ARE OFTEN USED IN RESEARCH WHEN THERE IS A NEED TO EXCHANGE OR SHARE A SET OF DATA FOR SOME PARTICULAR USE

Name	Description
Sentinel CDM	A relational model developed in 2008 to address the FDA's desire to create a national system capable of being used to monitor the safety of medical products that had acquired regulatory approval.
PCORNet CDM	A relational model used by the Patient-Centered Outcomes Research Institute (PCORI) to support the PCORnet Distributed Research Network (DRN). An open-source extension of the Sentinel CDM.
OMOP CDM	A person-centric way to standardize EHR information aimed at identifying populations with specific healthcare interventions, analyzing their demographics, and analyzing the effects. Adopted by the Observational Health Data Sciences and Informatics (OHDSI) Consortium.
Informatics for Integrating Biology and the Bedside (i2b2) CDM	A relational data model centered around the use of a single 'fact' table for storing patient information. Particularly suitable for representing nonstandard or local types of medical data. Used by the i2b2 tranSMART Foundation.

A data model is a representation of data typically collected about things or events and the relationships between them; a CDM is used to standardize and facilitate the exchange, pooling, sharing, or storing of data from multiple sources.

medicine tasks such as (1) **disease diagnosis** [43], [44], [45], [46], (2) **risk prediction** [38], [47], [48], [49], [50], [51], [52], [53], [54], [55], (3) **treatment effect (including drug response)** [56], [57], [58], [59], [60], [61], and **other clinical applications** [62], [63]. Until recently, most methods for evaluating huge volumes of EHR data depended on well-established machine learning and statistical techniques, including least absolute shrinkage and selection operator (LASSO), logistic regression, support vector machines (SVM), and tree-based algorithms (e.g., random forest and XGBoost). Deep learning approaches have recently shown signific success in various disciplines by effectively constructing deep hierarchical features and capturing long-range dependencies in data. An overview of machine learning-based analysis methods and a summary of the reviewed papers are illustrated in Figs. 3 and **S2**, respectively.

1) Disease Diagnosis: Modern EHR systems are densely populated with discrete medical codes that encompass every facet of patient encounters enabling computer-aided disease diagnosis and clinical decision support systems. Ensemble approaches involving multiple learning algorithms have been utilized to achieve higher predictive performance than each of the component learning algorithms alone. For example, Pang et al. [46] developed seven machine learning algorithms to predict childhood obesity after data quality control, transformation, and imputation. XGBoost outperformed other models on most standard classifier evaluation metrics, including sensitivity, precision, F1-score, accuracy, and area under the receiver operating curve (AUROC) to predict obesity. Similarly, Masino et al. [45] used standard k-fold cross-validation to automatically select features and model tuning parameters for

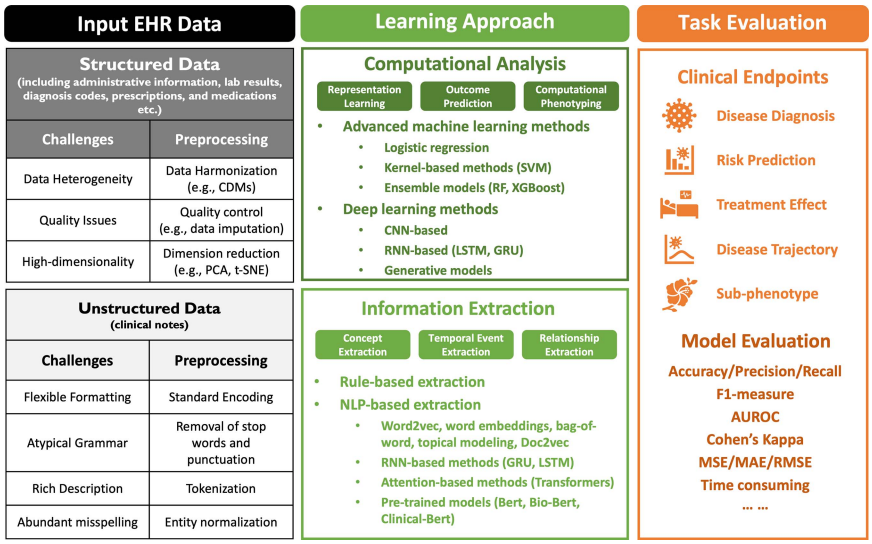

Fig. 3. Overview of machine learning-based EHR studies. The figure illustrates a general flow of steps for clinical decision support systems learning EHR data. The typical data analysis pipeline consists of data collection, quality control, feature extraction, knowledge modeling, decision making, and model evaluation.

identifying infants with sepsis using several popular machine learning models as classifiers. For deep learning techniques, the widely-used MIMIC-III dataset was adopted to train long short-term memory (LSTM) recurrent neural networks (RNNs) with attention mechanisms for daily sepsis prediction [44], thereby providing clinical decision support for a variety of clinical tasks. Additionally, the development of AI-enabled clinical decision support systems leveraging EHR data could also facilitate the diagnosis of rare diseases [43]. Researchers manually selected features that were not directly related to suspicion of acute hepatic porphyria (AHP) patients and trained several machine learning models on a small dataset of AHP patients. Potential AHP patients were identified by manually reviewing the top-ranking patients by margin distance after applying the SVM model with the best performance in cross-validation to a large EHR dataset. This study demonstrated that population-level informatics could be used to identify rare diseases, significantly enhancing the capability of identifying undiagnosed patients from a large population. Yan et al. [64] combined pathological images, which reveal intricate tissue structures, with structured data from EMRs that furnish invaluable clinical context and patient history for breast cancer classification. Specifically, they introduced a generalizable richer feature representation, as well as a fusion of high-dimensional and low-dimensional data to integrate highly heterogeneous data for improving pathological images classification accuracy. Similarly, Lu et al. [65] introduced Tumour Origin Assessment via Deep Learning (TOAD), a multi-modal learning algorithm developed to diagnose the primary tumor's origin using standard histology slides with demographic information.

2) Risk Prediction: Leveraging large-scale EHR databases, machine learning models can facilitate the successful prediction of high prevalence events. With the input feature effectively extracted from EHR data, including vital signs, lab observations, demographic information, and ICD codes, machine learning methods can achieve mortality prediction and survival analysis to improve risk adjustment [49], [50], [52]. Rasmy et al. [51] validated accurate heart failure risk prediction of RNNs in a large heterogeneous EHR data set and demonstrated the generalizability of RNN-based predictive models for heart failure onset risk. However, the utility of this method in predicting risk in smaller cohorts with rare diseases and infrequent events is unknown [38]. To test the hypothesis that machine learning techniques could aid in the identification of high-risk patients, researchers used a deep neural network to predict in-hospital mortality in patients with spontaneous coronary artery dissection. The deep neural network model had higher predictive accuracy and discriminative power than logistic regression or other machine learning models for identifying high-risk patients. Moreover, under the COVID-19 pandemic, EHR data has been explored to predict the risk of mortality [53], [55], the risk for the transition from mild to severe [47], future oxygen requirements [48], and the need for mechanical ventilation [54].

Fiterau et al. [66] introduced ShortFuse to optimize both structured covariates and time series data (e.g., wearable sensor data) during feature extraction. It integrated time series by deliberately modeling the temporal interactions and dependencies alongside structured covariates. They showed a performance boost in forecasting cartilage degeneration related to osteoarthritis and predicting surgical outcomes for patients with cerebral palsy with EHRs and sensor data. Tomašev et al. [67] integrated EHRs with acute kidney injury (AKI) monitoring data to continuously update risk predictions for potential patient deterioration. Compared with previous studies only focusing on EHRs, integrating AKI condition provided more timely alerts with ample context, ensuring healthcare professionals have enough time to intervene. Solachidis et al. [68] combined EHRs with imaging, biometric sensors, and Internet of Things devices to facilitate automatic distant monitoring of Parkinson's and Alzheimer's patients.

3) Treatment Effect and Drug Response: The increasing prevalence of EHRs offers a unique opportunity to build machine learning algorithms to identify treatment effects and drug responses for precision medicine. With the advancement of machine learning techniques, the ability to accurately predict treatment outcomes is crucial for developing personalized treatment plans and follow-up schedules. Wang et al. [61] proposed a reliable multi-objective ensemble deep learning method for predicting a high risk of treatment failure following radiotherapy in lung cancer patients based on features extracted from EHRs. To improve the prediction robustness and accuracy of individual networks, the proposed model was trained with dynamic ensemble deep learning, selected with adaptive multi-objective optimization, and evaluated with evidential reasoning fusion. The multi-objective model simultaneously evaluates sensitivity, specificity, and AUROC as objective functions, thereby overcoming the limitations of model selection based on a single objective. Another study is conducted by Chen et al. [56] to adopt multi-task deep learning to extract both outcome-predictive and treatment-specific latent representations from EHRs by jointly performing the outcome prediction and treatment category classification. Then, to estimate individual treatment effects, they predicted counterfactual outcomes by applying KNN to the learned hidden representations. For predicting drug responses, Datta et al. [57] proposed a logistic regression algorithm that unearths several known interactions from an EHR dataset of about 400000 hospitalizations to study drug-drug interactions. Mo et al. [59] collected juvenile idiopathic arthritis patients with methotrexate from EHR. They developed four machine learning models to predict drug responses, which could provide decision support for doctors to make or adjust therapeutic schemes before or after treatment. Researchers [58] also demonstrated the potential of ML to predict antibiotic resistance to bacterial infections in hospitalized patients using a model ensemble (LASSO logistic regression, XGBoost, and neural networks).

4) Other Clinical Applications: Additionally, EHR data has demonstrated its efficacy in various other clinical applications regarding precision medicine. Xu et al. [63] used a memory network-based deep learning approach LSTM to discover acute kidney injury (AKI) sub-phenotypes using structured and unstructured EHR data of patients before AKI diagnosis. Because some diseases are highly heterogeneous, identifying sub-phenotypes can aid in understanding their pathophysiology and the development of more targeted clinical interventions. In applying disease trajectories modeling, Oh et al. [62] established a generative model to extract a set of disease trajectories from EHR data. They evaluated it internally based on log-likelihood, which can be interpreted as the ability of the trajectories to explain the observed disease progression. Scheurwegs et al. [69] combined both structured EHRs and unstructured clinical notes to allocate clinical diagnostic and procedural codes, particularly ICD-9-CM, to patient stays. Their findings indicated that structured and unstructured data complement each other in predicting ICD-9-CM codes, with the late integration method harnessing this synergy most effectively. Similarly, Abhyankar et al. [70] combined structured EHR data with unstructured data (clinical notes) to identify patient cohorts, focusing on patients undergoing dialysis. The experimental results suggested that integrating structured EHR data with insights from clinical notes via straightforward queries can enrich the capability to identify specific patient cohorts. Pathak et al. [71] aimed to create a scalable informatics framework that could standardize and unify both structured and unstructured EHR data for efficient phenotype extraction. The study showcased the utility of open-source and modular resources in transforming EHR data into a standardized format, promoting its secondary use. To guide the development and evaluation of AI systems using multimodal inputs, Soenksen et al. [72] proposed a unified Holistic AI in Medicine (HAIM) framework for integrating multi-modal data. Results demonstrate that models trained using the HAIM framework consistently outperformed single-source models by margins of 6-33% across various medical demonstrations, such as chest pathology diagnoses, patient length-of-stay, and 48-hour mortality predictions. The findings highlighted the varying significance of data modalities across distinct healthcare tasks, emphasizing the importance of incorporating multi-modal data inputs for precision in results.

IV. Cross-Modality Integration Learning

Multimodal integration in the context of computing is the virtual analogy of physical sensory integration. The central nervous system and brain cortices in humans are able to integrate stimuli from multiple senses. For example, the visual perception pathway is connected to the auditory cortex [73], [74]. Researchers also uncovered the collaboration and connectivity of different brain regions [75]. The vast amount of information received from smell, sight, touch, sound, and taste interplays with each other and eventually concludes with instructions that guide human cognition and behavior. The interconnection in the human brain offers a physical basis for advanced cognitive functions that Chimpanzees cannot realize [76].

Computers are also capable of leveraging information from multiple sources. Images, audio, and texts are signals that can be combined to represent a holistic description. A model takes in the aggregated information, condenses the amount of information by dimensionality reduction, and returns key information of interest [77]. Compared to the unimodal processing system, such a design presents a synergistic effect and compensates for the shortcomings of individual modalities. The multimodal integration technique also demonstrates promising results in medical-related applications. In this section, we outline the field of radiogenomics and EHR-multi-omics, wherein common datasets and pipelines are also presented. **Table S3** summarizes the frequently used multimodal biomedical datasets that are accessible to the public, including large-scale biobanks.

A. Radiogenomics

Among multiple approaches for multimodal integration in the medical field, phenotype-genotype integration stands as one of the most common strategies (see **Fig. S3.**). The genetic root potentially elucidates the phenotype of the condition. Medical imaging data is one of the most informative sources for diagnosis among EHRs, including MRI, CT, and pathology imaging data.

Radiogenomics, or imaging genomics, is a primary research field that integrates imaging phenotypes with multi-omics.

Radiomics is a discipline that aims to uncover phenotype features in medical images that bear diagnostic and prognostic values in a quantitative rather than qualitative fashion for clinical decision support [78]. Radiomics enables holistic analysis of data tumor regions in the context of peri-tumoral tissues, largely attributed to the non-local nature of imaging. However, the lack of standardization in image acquisition and feature extraction severely limits the potential of radiographs as a modality when used alone [78]. The massive variety coupled with the usually limited data highlights the challenge of building a robust radiomic-based model.

In comparison, genomics reveals molecular-level information on the etiology and progression of conditions. Knowledge of genotypes constitutes an essential part of evidence-based disease characterization and classification [79], [80]. Nevertheless, the biases in sampling hinder the possibility of comprehensive and accurate genome data despite advanced gene-sequencing techniques. A solid tumor is usually not homogeneous, and clinicians cannot sample every region. As a result, slices of tissues that enable an unbiased understanding of the bulk environment are often infeasible. Although the continuous alterations of cell genotypes offer valuable information to understand cancer progression before and after treatments, longitudinal studies are often inconvenient and expensive. Unlike non-invasive imaging, genotype analysis would mean multiple painful and costly biopsies for patients [81].

The marriage between radiomics and genomics addresses the shortcoming of each approach. The non-invasive and highly accessible global phenotype information complements the local genotype information. Enriching data modality also opens the possibility of finding potential associations between radiographic phenotypes and genotype information. Radiogenomics increases the likelihood of discovering new biomarkers that facilitate precise stratification of patient conditions and identifying the most suitable therapeutic strategies for individuals. As such, radiogenomics is an enabling technology behind precision medicine and personalized healthcare vision [82]. In this section, the two publicly available and the most widely used datasets, The Alzheimer's Disease Neuroimaging Initiative (ADNI) and The Cancer Genome Atlas (TCGA), are described in detail with their exemplary applications.

1) Datasets: The Cancer Genome Atlas (TCGA) is a comprehensive public library encompassing detailed genomic profiles of more than 30 types of human tumors. Initiated by the National Institute of Health (NIH), TCGA gathers next-generation sequencing data from multiple centers and makes it available on Genomic Data Commons (GDC). Genetic data modalities include gene expression, miRNA expression, protein quantification, and loss of heterozygosity (LOH). Meanwhile, radiology images of the cancer cases in TCGA are available in The Cancer Imaging Archive (TCIA). TCGA has significantly propelled research on carcinogenesis across multiple cancer types. Classification of glioblastoma multiforme (GBM) genetic subtypes [83], [84], associations between genotypes and phenotypes [85], progression [86], [87], and treatment strategies [88] of GBM have been uncovered from TCGA data. Besides GBM, the TCGA dataset also enables advancements in lung cancer, leukemia, and breast cancer research.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a repository consisting of subjects classified into three groups based on their cognitive status: normal, mild cognitive impairment (MCI), and Alzheimer's disease (AD). Phenotype data modality includes MRI (18F)-fluorodeoxyglucose positron emission tomography (FDG PET), urine, serum, cerebral spinal fluid (CSF), and clinical notes [89]. The repository facilitates standardizations of longitudinal studies, discoveries of image feature extraction and analysis techniques, and correlation of radiographic phenotype and genotypes [90].

The non-small cell lung cancer (NSCLC) dataset facilitates research into the leading causes of cancer death. CT images and clinical data are available for all 211 subjects in the dataset, and 130 of the subjects also have associated genomic data [91].

2) Applications: Overview: In a radiogenomics pipeline, the materials often consist of at least radiography and genomic data, and sometimes clinical assessment results such as risk scores or treatment responses are also available. In most studies, the goal is to associate radiography features with the underlying genotypes [92]. Thus, our discussion of pipelines centers around predicting the already identified genetic biomarkers. Both radiomic and genomic data are high-dimensional, but genome-wide association studies (GWAS) are frequently used to identify genetic biomarkers of interest, greatly decreasing feature set size. In contrast, there is no standard approach to dimensionality reduction within radiomic data analysis. Such a mismatch in dimensionality highlights the challenge of radiogenomics. As a result, the efficacy of a radiogenomic analysis is governed by the radiomic feature extraction methods.

A general pipeline involves the extraction of histogram, shape, and texture features from radiographs with tools such as PyRadiomics (https://pyradiomics.readthedocs.io/en/latest/) or LiFEX (https://www.lifexsoft.org/), followed by pruning of the extracted features. Such a dimensionality reduction process is generally set as the first stage of the pipeline. In the second stage, the resulting radiographic features with lower dimensionality are then classified into the corresponding genotypes. Like the three integration approaches outlined in the multi-omics section, there are two distinct categories in radiogenomics. Differentiated by when genomic and radiographic information begins to be jointly analyzed, these two pipelines are low-dimensional integration and high-dimensional integration (see Fig. 4.).

3) Low-Dimensional Integration: In the low-dimensional integration pipeline, the genomic data is integrated after obtaining a reduced set of radiographic features. In other words, dimensionality reduction of radiographic features is achieved by solely considering the properties of the radiographs or by regression to clinical metrics such as risk scores or therapy responses. Classifications of genotypes are then performed on the remaining reduced phenotype information.

First, radiographic features are selected without consideration of either genomics or clinically relevant metrics but instead determined only by the properties of the images. Yan et al. [93] stratified infiltrative glioma patients into low and high-risk

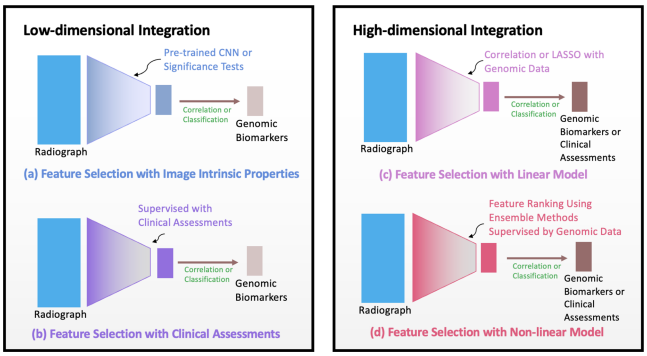

Fig. 4. Illustration of general framework for low-dimensional and high-dimensional integration in radiogenomics. Low-dimensional and high-dimensional radiogenomics integration pipelines are illustrated. (a) A pre-trained CNN model is used in dimensionality reduction, after which the reduced features are linked to genomic data. (b) The feature extraction is guided by clinical assessment data, and the most relevant features are retained to predict the genomic biomarkers. (c) Linear models such as LASSO or correlation are used to identify the most important radiographic features in the prediction of the genomic biomarkers. (d) Non-linear models such as random forests rank the importance of all features. In the high-dimensional integration pipelines, the reduced set of features can be associated with either the biomarkers or the clinical assessments.

gliomas with a ResNet-34-based model. Meanwhile, genomic studies revealed differentially expressed genes (DEGs). Then, gene set variation analysis (GSVA) quantified pathway activities in individual patients, with which the correlations to deep learning features were identified. Zhang et al. [94] developed a convolution-based network for breast tumor segmentation from MRI images. The morphological features from the segmented region achieved an AUROC of 0.69 in classifying the luminal A subtype. Kawaguchi et al. [95] innovatively include ResNet-derived and PyRadiomics-extracted features from glioma MRI scans. PCA and NMF helped dimensionality reduction. SVM, logistic regression, and random forests classified radiographic features to multiple genomic biomarkers. Besides neural networks, statistical analyses among all extracted features are also proven viable in feature selection. Huang et al. [96] studied the distributions of the features in PET scans on the ADNI dataset. Two-sample t-test, Mann-Whitney U test, and F-scores served as criteria to remove highly correlated features. SVM was used to classify the remaining features into three cognitive groups.

Second, the radiographic features predictive of clinical results are retained. Beig et al. [97] revealed a genetic basis for sex-specific phenotypes of GBM. Three groups (i.e., male, female, and combined) were constructed from MRI scans. Images in each group underwent dimensionality reduction by LASSO based on risk scores, and the remaining features were linked to DEG by the Mann-Whitney U test. Hu et al. [98] predicted the therapy response of esophageal cancer by integrating features extracted from the Xception architecture and PyGenomics handcrafted features. Among those, the top 20% predictive features for therapy response were selected. Pathway enrichment analysis was then performed with the reduced feature.

Although both approaches demonstrate promising results in genotype prediction, the second approach is more adaptive and goal oriented. Depending on the clinical endpoints, the same radiographic feature could rank vastly different in importance among all features, so the unguided dimensionality reduction in the first approach could lead to a less optimized set of reduced features. Furthermore, a technique involves using genomic data as the criteria for dimensionality reduction in the first stage. As such, radiomics and genomics are integrated, not in the second stage but in the first stage, where radiomic data remains high-dimensional.

Similar studies have explored the relationship between alterations in gene expression and tissue morphology using low-dimensional integration. They achieved this by integrating genomics data with histology and/or radiology images for various objectives: cancer classification [99], survival prognosis [100], [101], and predicting treatment responses [102], [103]. For example, recent advancements have leveraged a combination of an IDH1 mutation status and a histology profile to refine the WHO classification of diffuse glioma [104], instead of relying solely on a single one in previous studies. Similarly, Boehm et al. [105] collected a multimodal dataset of 444 patients, primarily diagnosed with late-stage high-grade serous ovarian cancer. Quantitative features crucial for prognosis were identified, such as tumor nuclear size from histopathological staining with hematoxylin and eosin, and omental texture observed on contrast-enhanced CT scans. This study revealed that these quantitative features not only held prognostic significance but also offered complementary information compared to one another and genomic features.

4) High-Dimensional Integration: In a high-dimensional integration pipeline, genotypes are used to assist dimensionality reduction of radiographic features immediately after features extraction. Radiographic features that are less predictive of genotypes are removed. Sometimes the features are further shrunk in size using PCA. After dimensionality reduction, a separate model is developed in the second stage to discover the relationship between the remaining radiographic features and the associated genotypes. The genomic-guided feature selection process in the first stage can be linear or nonlinear, differentiating the two subtypes in a high-dimensional integration pipeline.

First, linear methods such as LASSO and correlation are utilized in radiographic feature selection. Binder et al. [106] discovered a negative correlation between EGFR mutation and the overall survival rate of GBM patients. Specifically, the 2104 quantitative imaging phenomic (QIP) features from four modes of MRI brain images are reduced and pruned with SVM-based multivariate analysis. The QIP features with a low statistical significance to distinguish EGFR mutants were discarded and resulted in 17 final QIP features for robust classification of EGFR mutation status. Yi et al. [107] also adopted a similar LASSO-based dimensionality reduction method which combines CT radiographic and SNP genomic data to predict the therapy response of ovarian cancer patients. Kim et al. [108] extracted 47 imaging features from osteosarcoma ROI. The predictive value of each feature toward Ki-67 and EZRIN status was evaluated, and features with AUROC over 0.6 were retained. The remaining features were classified regarding chemotherapy response and metastasis with random forests.

Feature ranking with a non-linear model, especially random forests, represents an alternative pipeline in high-dimensional integration. Iwatate et al. [109] developed a non-invasive CT scan-based approach for identifying two biomarkers. Key features were selected from CT images. The number of features was first reduced with the Mann-Whitney U test and random forest models. An XGBoost classifier was trained on the remaining features to predict P53 and PD-L1, the two reliable biomarkers for pancreatic cancer prognosis. Hoshino et al. [110] adopted the previous pipeline [109] to predict the tumor mutation burden of patients with colorectal cancer. However, prior to the Mann-Whitney U test, correlated features with Pearson's coefficient greater than 0.9 were discarded. Saha et al. [111] developed a pipeline to identify the ER, PR, HER2, and Ki-67 breast cancer subtypes from MRI phenotypes. First, segmented images were generated via fuzzy C-means automatic segmentation, where 529 key features were selected. Then, the most significant features were evaluated through the random forests and linked to genotypes via correlation studies. A pre-screening step like the one in [110] was featured in a study by Hsu et al. [112]. The group integrated RNAseq data and Agilent microarray data as genomic data to uncover the predictive power of MRI scans to four immune subsets for glioma. At first, radiographic features that exhibited opposite correlations to the two genomic modalities were filtered out. Next, a random forests model was constructed to rank the remaining features with respect to the four genomic subsets. A decision tree then further discarded features that had weights below the overall mean weights. As a result, 9809 T1C and 9809 ADC features were reduced to 37 and 39 features, which provided over 0.6 AUC in classifying four immune subsets. Pinheiro et al. [113] classified EGFR and KRAS mutation statuses of non-small cell lung cancer (NSCLC) patients with CT images. Inter-correlated features and the least important features indicated by gradient boosting were discarded. The remaining features went through PCA and t-SNE. In the last step, XGBoost was used to classify the mutation statuses and achieved an AUC of 0.75. Similarly, several studies [92], [114], [115], [116] utilized genomics information to guide the selection of histological features or mammogram images, enhancing survival prediction for multiple cancers.

Besides random forests, other innovative ensemble learning schemes also benefit from radiographic feature selection. Nuechterlein et al. [117] extracted 35340 features from MRI scans and developed a novel dimensionality technique via repeated LASSO. The 288 most frequently recurring features were retained and further reduced to a size of 15 via PCA. The remaining features demonstrated 0.8 AUC in predicting the two Isocitrate dehydrogenase (IDH) subtypes of glioblastoma using ridge logistic regression. Haubould et al. [118] used PET-MRI data labeled with histology results to classify low-grade gliomas and high-grade gliomas. Randomized logistic regression was repeated over 200 times to extract the most contributing features. Random forests were then used to classify the underlying genotypes.

We presented two approaches to integrating genomic data with radiomic data early in the pipeline. It is worth noting that linear and non-linear approaches are not mutually exclusive, and some studies take advantage of both. For a large number of radiographic features, pre-screening with a linear model could be beneficial before building the non-linear ranking algorithm. Overall, to effectively address the dimensional disparity between radiomic and genomic data and trim the radiographic features, we identified two categories of pipelines. First, genomic data is treated as the classification or regression targets only after radiographic features reduction with a convolutional neural network or with clinical assessments. Alternatively, genomic data help to prune radiomic data from the beginning, with either a linear model or a non-linear model. The resulting small set of features is associated with the genetic biomarkers or with clinical metrics of interest. From a different perspective, the pipelines can also be differentiated by what modalities are available for radiogenomic analysis. When only radiological data and genetic data are available, and the goal is to predict the underlying genotypes with radiographs, radiographic features can be shrunk by a general approach such as pre-trained CNN where the intrinsic properties of the image determine the results, or by an approach where features less predictive of the genetic biomarkers are removed. Then, the remaining features are classified into their respective genotypes. In the latter case, the genomic data is considered twice. On the other hand, if radiological data, genomic data, and clinical assessments are available, two mirrored pipelines deserve consideration. First, radiomic features are trimmed based on clinical assessment results and then fed into a second model to predict the genotypes. Alternatively, radiomic features are trimmed based on genotypes and used to predict the clinically relevant metrics.

B. Integrating EHRs With Multi-Omics

Recent advancements in high-throughput technologies and widespread adoption of EHRs have accelerated the accumulation of multi-omics and EHRs data. Big data analytics enable the extraction of knowledge from voluminous and complex data to improve patient care quality through precision medicine. Multi-omics data has been integrated with medical records to provide polygenic risk scores in clinical practice. This section will focus on advanced data analytics methods to achieve the primary goal of personalized and precision medicine by integrating medical records with multi-omics data.

1) Biobanks: A biobank is generally defined as a systematic collection of human biological samples and associated data for research purposes [119]. For example, U.K. Biobank [120] connects genetic variants to a wide variety of diseases and outcomes through the provision of EHR (with imaging) and genomic data, making it an ideal resource for integrated data analytics. With the rapid development of biobanking, it is critical to define how such resources can facilitate the organization of massive amounts of biological data and translation of genome-based knowledge for population health benefits. To advance biomedical research and promote the development of new treatments, BBMRI-ERIC, a European research infrastructure for biobanking and biomolecular resources, was established to foster collaborations between biobanks and facilitate the exchange of biological samples between biobanks, industry, and patients. Biobanks show their

value in providing access to patient data with a higher rate of disease (especially for rare diseases), following patients longitudinally for free with time-stamped clinical encounters (i.e., disease trajectories), allowing recontact for deeper phenotyping, and translating findings into real-world practice [121].

2) Applications. Overview: This study categorizes the integration of EHR and -omics into phenotype-first studies and genotype-first studies based on the guiding information. Genotype information is the genetic material passed down through generations, while the phenotype is observable characteristics or traits of an organism. PheKB [122] is a collaborative environment for developing and validating algorithms for identifying patient characteristics within health data, enabling researchers to extract high-quality phenotype information for clinical research.

3) Phenotype-First Integration: EHRs linked to biological specimens in clinical data analytics are expected to deliver cost-effective and rapid genomic studies [121]. Compared with epidemiology studies, EHR-linked biobanks enable researchers to take a patient cohort and derive the target disease from the population instead of recruiting participants of disease cases and controls separately at the beginning of each disease investigation. After enrolling patients in a biobank, researchers can reuse their data in a variety of studies examining various diseases and treatments. Phenotype-first studies enable researchers to first extract phenotypic data to define a patient cohort with a particular disease and then deliver genetic analytics with biological follow-up.

Ritchie et al. [123] have investigated the genome-wide association study with arrhythmia risk on cardiac disease using patient cohorts identified from EHRs at five sites in the Electronic Medical Records and Genomics (eMERGE) Network. Their integration studies started with a GWAS meta-analysis to evaluate the most significant loci and then continued with a phenome-wide association study (PheWAS) to search for associated diagnoses. Their phenotype-first study, which utilized an EHR-connected biobank, provided the platform for GWAS with a comprehensive examination of the longitudinal incidence of disease associated with genetic variants. Similarly, Veturi et al. [124] proposed a unified framework in U.K. biobanks to identify genes associated with lipid traits and investigate the relationship between lipids and hundreds of other complex diseases. They began with lipid GWAS to identify 67 novel lipid-associated genes and then conducted PheWAS studies on lipid-associated genes to figure out potentially pleiotropic associations between lipids and diseases. Besides, Pena et al. [125] also proposed a phenotype-first approach to molecular diagnostic resolution in rare and undiagnosed conditions with the identification of causal genes. Existing studies show that EHR-linked biobanks could facilitate tremendous opportunities in translational research for precision medicine.

In the COVID-19 study, Su et al. [126] have integrated multi-omics data (including information on plasma proteins, metabolites, and on PBMC transcriptomic and surface-protein data, immune receptor sequences, and secreted proteins) with EHR data to investigate the immune system shift among COVID-19 patients using serial blood draws from COVID-19 patients. Their integrated analysis identifies a significant immunological change between mild and moderate infection, including increased inflammation, a drop in blood nutrients, and the emergence of novel immune cell subpopulations that intensify with disease severity.

4) Genotype-First Integration: In contrast to the classical phenotype-first approach with GWAS, genotype-first integration methods start with the gene of interest and test genes across phenome to associate specific genotypes to apparent clinical phenotypes of a complex disease or trait. In genetic epidemiological studies, genotype-first approaches aim to identify genes that contribute to disease etiology regardless of the patient's suspected diagnosis. For example, U.K. biobanks recruit population-based (i.e., disease-agnostic) cohorts without regard to disease status to investigate cross-trait genetic associations in an unbiased spectrum.

Park et al. [127] proposed a genome-first approach for associating rare variants in LMNA with diverse EHR–derived phenotypes in two medical biobanks to better understand their contribution to disease. This method first annotated rare missense variants, clustered predicted deleterious variants into a gene burden (i.e., the cumulative effects of multiple rare variants in a gene), and then performed a PheWAS of predicted variants. This genotype-first study could serve as a demonstration in successfully identifying novel ontologies for pleiotropic human genes. Similarly, Drivas et al. [128] found that common genetic variants within Mendelian ciliopathy genes could contribute to the pathogenesis of common pathology in complex diseases using U.K. biobanks. Besides, researchers also identified different ciliary sub-compartments associated with distinct sets of phenotypes to investigate primary cilium involved in common pathogenesis using genotype-first integration [129]. Existing studies demonstrated that genotype-first approaches provide a valuable framework for -omics and EHR integration of common or rare disease genetics to provide insights into the pathophysiology of human diseases [130].

V. Challenges and Opportunities

The emerging field of multimodal data integration, enhanced by multi-institutional collaboration and driven by open research data initiatives, holds significant promise for precision medicine. According to recent studies reviewed in this paper, integrating multi-omics with EHR data has improved prognosis, diagnosis, and treatment in a variety of clinical applications. However, precision medicine incorporating genomics information into daily patient care remains a challenge. Only a tiny percentage of new biomarkers are patented, and an even smaller percentage are used in real-world routine clinical practice [131]. This section will review the challenges and opportunities in integration analytics on multimodal healthcare data collection and advanced computational methods for precision medicine.

A. Challenges in Integrating Multi-Omics and EHRs

The first challenge is data collection and harmonization. Deep learning is commonly used for efficient clinical knowledge discovery and decision-making. However, since deep architectures typically have enormous parameters, proper generalization

requires a larger dataset than traditional machine learning. From this perspective, there is a compelling need to collect and annotate large medical datasets that span multiple modalities, both time- and labor-intensive. Despite the widespread use of electronic systems operated and maintained by healthcare providers to collect and store patients' medical information, there remain challenges in (1) insufficient collection of patient-reported outcomes (PROs), (2) integration of patient-centered data into EHR systems, (3) standards or resources to support harmonization of EHRs, and (4) infrastructure for reusable EHR data collection in support of continuous learning [132]. Furthermore, while techniques such as NGS have enabled researchers to sequence faster and cheaper, it has also introduced new challenges, such as rising costs associated with participant recruitment and biological sample collection and processing, as well as increasing the complexity of data management, storage, validation, analysis, and interpretation [133]. Apart from the cost of data collection, data sharing between institutions is frequently constrained by concerns about privacy and security [134]. Specifically, unforeseen privacy issues necessitated the removal of large datasets from public databases due to the possibility of personal genetic information being associated with individuals. Due to the lack of standards for different biomedical devices or health IT infrastructures, vendor-dependent data collection hinders data integration with external sources and identity matching across data sources. To enable real-world evidence-based practices and improve healthcare quality and outcomes, data harmonization across various healthcare providers with multiple modalities (i.e., EHR, multi-omics, imaging, sensor, etc.) requires joint research and department effort from all stakeholders.

The second challenge in leveraging multimodal data for precision medicine lies in population and disease heterogeneity. Trials with a more heterogeneous population are more generalizable to clinical practice and thus have a higher degree of external validity; however, they are likely to have more heterogeneous treatment effects among participants [135], [136]. Madigan et al. conducted a systematic review of observational database heterogeneity and found that 20%-40% of observational studies can shift from statistically significant in one direction to the other [137]. Additionally, due to the complexity of the body, the majority of diseases exhibit a high degree of inter-individual variability in their phenotype, which has significant clinical implications, such as the difficulty in defining objective diagnostic or therapeutic rules. In cancer studies, intra-tumor heterogeneity caused by tumor cells with unique genomes or epigenomes interacting with their surrounding microenvironment could significantly affect patient treatment [138]. Heterogeneity is critical for clinical trials and drug design since it allows for the averaging and normalizing of potential alterations, thereby enabling appropriate and personalized therapeutic targeting [139].

Data quality is another critical challenge preventing multimodal integration frameworks from improving model performance. In clinical informatics, EHR data can be unstructured and noisy with issues such as high percentages of missing values, errors, invalid data, and outliers. Mobile wearable biosensors and imaging data usually suffer from noises and artifacts. Thus, data quality control processes are needed, such as missing data imputation, conflict resolution, and transformation [140]. Additionally, interoperability between different platforms and heterogeneous data types can prevent learning models from extracting efficient knowledge or patterns from EHR data [132]. In genomic medicine, sample contaminations and sequencing errors affect the DNA sequencing quality. Platforms, protocols, and bioinformatics pipelines could introduce bias into high-throughput sequencing data. For multimodal bioinformatics, such data repositories could face challenges in maintaining accuracy and accessibility due to exponential growth, including incorrect or incomplete information, lack of annotation, duplication with minor variations, and redundancy. Thus, it is crucial to develop quality control protocols and metrics for all data modalities before performing data integration [141].

The fourth challenge is to develop advanced data analytics to mine EHR and multi-omic data for knowledge. While advances in computational techniques have made routine interrogation of genomic data possible, the size, sparsity, and high dimensionality of the resulting data continue to pose challenges for computational analysis, particularly the integration of multimodal data from multiple sources [142]. Specifically, the curse of dimensionality, which refers to a feature dimension that is significantly larger than the sample size of the patient, is a well-known issue in genomic medicine. Due to the dimensionality involved, traditional parametric statistical methods are less useful in the epistatic analysis of such exponentially growing SNPs and usually result in ill-conditioned feature matrices [143]. A key step of the sequencing processing pipeline is dimension reduction to filter out irrelevant variants using either supervised or unsupervised feature selection.

B. Opportunities and Future Directions

To alleviate these data- and model-related barriers to adopting multi-modality integration in precision medicine, in this paper, we study the interactions between multimodal clinical data and AI-enabled models to provide future directions.

1) The Collection of Multimodal Biomedical Big Data: The abundance of healthcare data available today, frequently referred to as Big Data, provides invaluable resources for new knowledge discovery with the potential to advance precision medicine [144]. Determining duplicated data is a significant challenge for large data repositories. There are two methods for avoiding potential duplications. One option is to provide everyone with a unique identifier from a central source, such as the National Institutes of Health's Global Unique Identifier [145]. Another option is to provide access to the participants in the center to create a population-based biobank (e.g., U.K. biobank) [146], similar to financial banking, and give them control over access to their health information based on their needs. Combinations of these models can also provide intriguing results. Large data sets will enable researchers to develop clinical decision support systems and personalize treatment appropriately. Using data from large populations increases the likelihood of establishing a genetic link between genotype and phenotype [147]. If reasonable privacy and security protections are in

place and databases remain transparent, the risks associated with research are manageable.

Predictive, precise, participatory, preventive, and personalized health (P4) medicine describes the primary goal of future healthcare systems to improve patient care quality [148]. In addition to existing biomedical data modalities such as EHRs, multi-omics, imaging, and mobile wearable biosensors, personal health records (PHRs) play an essential role in describing individual health conditions to achieve the goal of P4 medicine. PHRs enable the integration of EHRs with data collected from sensors or other wearable computing devices, which can be used by both healthcare providers and patients [149]. Furthermore, PHRs enable direct interaction with patients to maintain and monitor their healthcare data and make decisions that may benefit their health.

2) The Advancement in Computational Methods: In precision medicine, identifying sufficiently large and diverse datasets for training frequently becomes a significant challenge that can rarely be met within individual institutions. Furthermore, collaborations between institutions that rely on centrally shared patient data face privacy and safety issues. Despite these risks, the widespread sharing of genomic and clinical data is critical for accelerating biomedical research [150]. Federated learning [151] is a novel paradigm for data-private multi-institutional collaborations. Model-learning exploits all available data without requiring data sharing between institutions by distributing model training to data owners and aggregating their results. Clinical adoption of federated learning is expected to result in models trained on unprecedented datasets, catalyzing the transition to precision/personalized medicine.

Since deep architectures typically contain enormous parameters, proper generalization requires a larger dataset than classical machine learning. Novel deep learning-based approaches for imputation and augmentation have been developed using autoencoder and generative adversarial networks for tabular and imaging data. Compared with classical data augmentation methods, generative models get rid of simple assumptions about the underlying probability distribution of the data and achieve better performance [152]. Transfer learning is one method for circumventing the medical data scarcity problem. Domain adaptation is a subset of transfer learning that enables models to perform better when applied to multiple datasets. Domain adaptation has been developed to facilitate the transfer of knowledge from one-labeled data domain to another related but unlabeled domain. It has emerged as a promising solution for dealing with the lack of annotated training data [153]. For instance, domain adaptation combined with a generative adversarial network has been used to address missing data in diagnosing brain diseases such as Alzheimer's disease using multimodal neuroimages [154].

Health disparity refers to certain disadvantaged social groups that have persistently experienced worse healthcare treatment or higher health risks than advantaged social groups. With the progress of precision medicine, model fairness questions are asked daily, which significantly impacts biomedical experimentation design, data analysis, and healthcare decision-making. Would the health decision support system have made a different decision if the patient's sensitive attributes (e.g., race and gender) differed? When we use data to build clinical decision-support systems and deliver personalized treatment plans without considering the bias, the system will replicate the bias from the real-world data. There are primarily two types of fairness metrics: 1) group-based fairness metrics that compute the model's behavior difference between subgroups with different sensitive attributes (e.g., the difference in prediction accuracy between an advantaged and disadvantaged group) [155]; and 2) individual-based fairness metrics that assume similar individuals from opposite sensitive attribute groups make similar model decisions [156]. The definition of model fairness would assist model developers in gaining a better understanding of data and system bias to mitigate the effect of health disparities in future work.

There is a trade-off between the higher performance of complex and black-box neural networks and the lower performance of more interpretable and straightforward approaches. Explainable artificial intelligence can be adopted to foster clinical trust and accelerate the adoption of AI in clinical decision support using complex neural networks. With a better understanding of black-box systems, we will be able to develop novel methods for incorporating medical insights from healthcare professionals into the clinical decision-making process in a multimodal learning framework. Explainable AI methods, such as SHapley Additive exPlanations (SHAP) feature importance in EHR and Grad-CAM for medical imaging, have been widely applied to clinical decision support to provide model interpretations. However, making sense of genomic data is also a barrier to translating genomic discoveries into clinical practice [157]. Avsec et al. [158] proposed a Base Pair Network to reveal regulatory code by accurately predicting transcription factor binding from DNA sequences. After the high-accuracy model is trained, the output signal is traced back to the input sequences to reveal sequence motifs for model interpretation. These efforts in explainable AI can significantly benefit the adoption of AI-based approaches in clinical applications.

VI. Conclusion

Multi-modal biomedical data provides the foundation for precision medicine. The various data modalities (e.g., multi-omics data and electronic health records) enable the healthcare providers to care for the patients with a multi-scale and multi-resolution assessment. Because of the complexity and volume of these multi-modal biomedical data, AI approaches are essential for personalized diagnosis and treatment. In this review paper, we have investigated the state-of-the-art AI-based approaches for utilizing multi-omics data, electronic health records data, and cross-modality data integration. For the cross-modality data integration, we reviewed the integration of multi-omics data with medical imaging data (i.e., radiogenomics) and the integration of multi-omics data with electronic health records. We have identified four challenges in integrating multi-omics data with EHRs: 1) data collection and harmonization, 2) population and disease heterogeneity, 3) data quality, and 4) advanced data analytics methods. We present opportunities and future directions in integrating multi-omics and EHRs data for precision medicine. Novel AI-based techniques such as federated learning,

transfer learning, fairness machine learning, and explainable AI are promising to fill the gap between academic research and clinical translation. With the close collaborations during data collection, data sharing and AI-based data analytics development, the integration of multi-omics data and EHRs data will reshape the paradigm of precision medicine in the near future.

Acknowledgment

This effort was inspired by NSF workshop in pandemic preparedness workshop-III.

References

[1] Y. F. Lu et al., "Personalized medicine and human genetic diversity," *Cold Spring Harbor Perspectives Med.*, vol. 4, no. 9, Jul. 2014, Art. no. a008581, doi: 10.1101/cshperspect.a008581.

[2] I. J. Morais et al., "The global population of SARS-CoV-2 is composed of six major subtypes," *Sci. Rep.*, vol. 10, no. 1, pp. 1–9, 2020.

[3] R. Jain et al., "Host transcriptomic profiling of COVID-19 patients with mild, moderate, and severe clinical outcomes," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 153–160, 2021, doi: 10.1016/j.csbj.2020.12.016.

[4] E. Pairo-Castineira et al., "Genetic mechanisms of critical illness in COVID-19," *Nature*, vol. 591, no. 7848, pp. 92–98, Mar. 2021, doi: 10.1038/s41586-020-03065-y.

[5] E. R. Malone et al., "Molecular profiling for precision cancer therapies," *Genome Med.*, vol. 12, no. 1, Jan. 2020, Art. no. 8, doi: 10.1186/s13073-019-0703-1.

[6] J. A. Johnson, "Pharmacogenetics: Potential for individualized drug therapy through genetics," *TRENDS Genet.*, vol. 19, no. 11, pp. 660–666, 2003.

[7] M. Bersanelli et al., "Methods for the integration of multi-omics data: Mathematical aspects," *BMC Bioinf.*, vol. 17, no. Suppl 2, Jan. 2016, Art. no. 15, doi: 10.1186/s12859-015-0857-9.

[8] Y. Bhak et al., "Depression and suicide risk prediction models using blood-derived multi-omics data," *Transl. Psychiatry*, vol. 9, no. 1, pp. 1–8, 2019.

[9] Q. Huang et al., "Application of artificial intelligence modeling technology based on multi-omics in noninvasive diagnosis of inflammatory bowel disease," *J. Inflammation Res.*, vol. 14, pp. 1933–1943, 2021, doi: 10.2147/JIR.S306816.

[10] T. Wang et al., "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nature Commun.*, vol. 12, no. 1, pp. 1–13, 2021.

[11] V. Malik et al., "Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer," *BMC Genomic.*, vol. 22, no. 1, Mar. 2021, Art. no. 214, doi: 10.1186/s12864-021-07524-2.

[12] B. Ma et al., "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," *Comput. Biol. Med.*, vol. 121, 2020, Art. no. 103761.

[13] S. Takahashi et al., "Predicting deep learning based multi-omics parallel integration survival subtypes in lung cancer using reverse phase protein array data," *Biomolecules*, vol. 10, no. 10, 2020, Art. no. 1460.

[14] H. Chai et al., "Integrating multi-omics data through deep learning for accurate cancer prognosis prediction," *Comput. Biol. Med.*, vol. 134, 2021, Art. no. 104481.

[15] H. Xu et al., "Multi-Omics marker analysis enables early prediction of breast tumor progression," *Front. Genet.*, vol. 12, 2021, Art. no. 670749, doi: 10.3389/fgene.2021.670749.

[16] X. Zhang et al., "Robust prognostic subtyping of muscle-invasive bladder cancer revealed by deep learning-based multi-omics data integration," *Front. Oncol.*, vol. 11, 2021, Art. no. 689626, doi: 10.3389/fonc.2021.689626.

[17] L. Tong et al., "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, Sep. 2020, Art. no. 225, doi: 10.1186/s12911-020-01225-8.

[18] N. Rappoport and R. Shamir, "NEMO: Cancer subtyping by integration of partial multi-omic data," *Bioinformatics*, vol. 35, no. 18, pp. 3348–3356, Sep. 2019, doi: 10.1093/bioinformatics/btz058.

[19] I. Subramanian et al., "Multi-omics data integration, interpretation, and its application," *Bioinf. Biol. Insights*, vol. 14, 2020, Art. no. 1177932219899051, doi: 10.1177/1177932219899051.

[20] M. Picard et al., "Integration strategies of multi-omics data for machine learning analysis," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 3735–3746, 2021, doi: 10.1016/j.csbj.2021.06.030.

[21] H. R. Hassanzadeh and M. D. Wang, "An integrated deep network for cancer survival prediction using omics data," *Front. Big Data*, vol. 4, 2021, Art. no. 568352.

[22] Z. He et al., "Integrating somatic mutations for breast cancer survival prediction using machine learning methods," *Front. Genet.*, vol. 11, 2021, Art. no. 632901.

[23] Y. Lin et al., "Classifying breast cancer subtypes using deep neural networks based on multi-omics data," *Genes*, vol. 11, no. 8, 2020, Art. no. 888.

[24] M. Tao et al., "Classifying breast cancer subtypes using multiple kernel learning based on omics data," *Genes*, vol. 10, no. 3, 2019, Art. no. 200.

[25] L. Zhang et al., "Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma," *Front. Genet.*, vol. 9, 2018, Art. no. 477, doi: 10.3389/fgene.2018.00477.

[26] O. B. Poirion et al., "DeepProg: An ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data," *Genome Med.*, vol. 13, no. 1, pp. 1–15, 2021.

[27] Z. Huang et al., "SALMON: Survival analysis learning with multi-omics neural networks on breast cancer," *Front. Genet.*, vol. 10, 2019, Art. no. 166, doi: 10.3389/fgene.2019.00166.

[28] R. Shen et al., "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.

[29] R. Argelaguet et al., "Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets," *Mol. Syst. Biol.*, vol. 14, no. 6, Jun. 2018, Art. no. e8124, doi: 10.15252/msb.20178124.

[30] B. Wang et al., "Similarity network fusion for aggregating data types on a genomic scale," *Nature Methods*, vol. 11, no. 3, pp. 333–337, Mar. 2014, doi: 10.1038/nmeth.2810.

[31] C. Meng et al., "A multivariate approach to the integration of multi-omics datasets," *BMC Bioinf.*, vol. 15, May 2014, Art. no. 162, doi: 10.1186/1471-2105-15-162.

[32] J. Xu et al., "A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data," *BMC Bioinf.*, vol. 20, no. 1, Oct. 2019, Art. no. 527, doi: 10.1186/s12859-019-3116-7.

[33] H. Sharifi-Noghabi et al., "MOLI: Multi-omics late integration with deep neural networks for drug response prediction," *Bioinformatics*, vol. 35, no. 14, pp. i501–i509, Jul. 2019, doi: 10.1093/bioinformatics/btz318.

[34] P.-Y. Wu et al., "-Omic and electronic health record Big Data analytics for precision medicine," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 263–273, Feb. 2017, doi: 10.1109/TBME.2016.2573285.

[35] R. Leaman et al., "Challenges in clinical natural language processing for automated disorder normalization," *J. Biomed. Inform.*, vol. 57, pp. 28–37, Oct. 2015, doi: 10.1016/j.jbi.2015.07.010.

[36] A. Parthipan et al., "Predicting inadequate postoperative pain management in depressed patients: A machine learning approach," *PLoS One*, vol. 14, no. 2, 2019, Art. no. e0210575, doi: 10.1371/journal.pone.0210575.

[37] N. Hong et al., "Integrating structured and unstructured EHR data using an FHIR-based type system: A case study with medication data," *AMIA Summits Transl. Sci. Proc.*, vol. 2018, 2018, Art. no. 74.

[38] C. Krittanawong et al., "Machine learning and deep learning to predict mortality in patients with spontaneous coronary artery dissection," *Sci. Rep.*, vol. 11, no. 1, Apr. 2021, Art. no. 8992, doi: 10.1038/s41598-021-88172-0.

[39] F. Amrollahi et al., "Contextual embeddings from clinical notes improves prediction of sepsis," *AMIA Annu. Symp. Proc.*, vol. 2020, pp. 197–202, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33936391

[40] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, no. 1, pp. 1–10, 2018.

[41] B. Shickel et al., "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018, doi: 10.1109/JBHI.2017.2767063.

[42] Y. Li et al., "BEHRT: Transformer for electronic health records," *Sci. Rep.*, vol. 10, no. 1, Apr. 2020, Art. no. 7155, doi: 10.1038/s41598-020-62922-y.

[43] A. M. Cohen et al., "Detecting rare diseases in electronic health records using machine learning and knowledge engineering: Case study of acute hepatic porphyria," *PLoS One*, vol. 15, no. 7, 2020, Art. no. e0235574.
[44] D. A. Kaji et al., "An attention based deep learning model of clinical events in the intensive care unit," *PLoS One*, vol. 14, no. 2, 2019, Art. no. e0211057.
[45] A. J. Masino et al., "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," *PLoS One*, vol. 14, no. 2, 2019, Art. no. e0212665.
[46] X. Pang et al., "Prediction of early childhood obesity with machine learning and electronic health record data," *Int. J. Med. Inform.*, vol. 150, 2021, Art. no. 104454.
[47] E. Casiraghi et al., "Explainable machine learning for early assessment of COVID-19 risk prediction in emergency departments," *IEEE Access*, vol. 8, pp. 196299–196325, 2020, doi: 10.1109/ACCESS.2020.3034032.
[48] I. Dayan et al., "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature Med.*, vol. 27, no. 10, pp. 1735–1743, Oct. 2021, doi: 10.1038/s41591-021-01506-3.
[49] C. Li et al., "Predicting survival in veterans with follicular lymphoma using structured electronic health record information and machine learning," *Int. J. Environ. Res. Public Health*, vol. 18, no. 5, 2021, Art. no. 2679.
[50] T. Munkhdalai et al., "Clinical relation extraction toward drug safety surveillance using electronic health record narratives: Classical learning versus deep learning," *JMIR Public Health Surveill.*, vol. 4, no. 2, 2018, Art. no. e9361.
[51] L. Rasmy et al., "A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set," *J. Biomed. Inform.*, vol. 84, pp. 11–16, 2018.
[52] N. Sahni et al., "Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: A proof-of-concept study," *J. Gen. Intern. Med.*, vol. 33, no. 6, pp. 921–928, 2018.
[53] S. Sankaranarayanan et al., "COVID-19 mortality prediction from deep learning in a large multistate electronic health record and laboratory information system data set: Algorithm development and validation," *J. Med. Internet Res.*, vol. 23, no. 9, 2021, Art. no. e30157.
[54] S. P. Shashikumar et al., "Development and prospective validation of a deep learning algorithm for predicting need for mechanical ventilation," *Chest*, vol. 159, no. 6, pp. 2264–2273, 2021.
[55] P. D. Sottile et al., "Real-time electronic health record mortality prediction during the COVID-19 pandemic: A prospective cohort study," *J. Am. Med. Inform. Assoc.*, vol. 28, no. 11, pp. 2354–2365, Oct. 2021, doi: 10.1093/jamia/ocab100.
[56] P. Chen et al., "Deep representation learning for individualized treatment effect estimation using electronic health records," *J. Biomed. Inform.*, vol. 100, 2019, Art. no. 103303.
[57] A. Datta et al., "Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from a retrospective electronic health record (EHR) cohort," *PLoS Comput. Biol.*, vol. 17, no. 7, 2021, Art. no. e1009053.
[58] O. Lewin-Epstein et al., "Predicting antibiotic resistance in hospitalized patients by applying machine learning to electronic medical records," *Clin. Infect. Dis.*, vol. 72, no. 11, pp. e848–e855, 2021.
[59] X. Mo et al., "Early and accurate prediction of clinical response to methotrexate treatment in juvenile idiopathic arthritis using machine learning," *Front. Pharmacol.*, vol. 10, 2019, Art. no. 1155.
[60] X. Mo et al., "Early prediction of clinical response to etanercept treatment in juvenile idiopathic arthritis using machine learning," *Front. Pharmacol.*, vol. 11, 2020, Art. no. 1164.
[61] R. Wang et al., "Multi-objective ensemble deep learning using electronic health records to predict outcomes after lung cancer radiotherapy," *Phys. Med. Biol.*, vol. 64, no. 24, 2019, Art. no. 245005.
[62] W. Oh et al., "A computational method for learning disease trajectories from partially observable EHR data," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2476–2486, Jul. 2021, doi: 10.1109/JBHI.2021.3089441.
[63] Z. Xu et al., "Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks," *J. Biomed. Inform.*, vol. 102, 2020, Art. no. 103361.
[64] R. Yan et al., "Integration of multimodal data for breast cancer classification using a hybrid deep learning method," in *Proc. Intell. Comput. Theories Appl., 15th Int. Conf.*, 2019, pp. 460–469.
[65] M. Y. Lu et al., "AI-based pathology predicts origins for cancers of unknown primary," *Nature*, vol. 594, no. 7861, pp. 106–110, 2021.
[66] M. Fiterau et al., "Shortfuse: Biomedical time series representations in the presence of structured information," in *Proc. Mach. Learn. Healthcare Conf.*, 2017, pp. 59–74.
[67] N. Tomašev et al., "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, no. 7767, pp. 116–119, 2019.
[68] V. Solachidis et al., "TeNDER: Towards efficient Health systems through e-Health platforms employing multimodal monitoring," in *Proc. IEEE/ACM Conf. Connected Health: Appl., Syst. Eng. Technol.*, 2021, pp. 185–192.
[69] E. Scheurwegs et al., "Data integration of structured and unstructured sources for assigning clinical codes to patient stays," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. e1, pp. e11–e19, 2016.
[70] S. Abhyankar et al., "Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis," *J. Amer. Med. Inform. Assoc.*, vol. 21, no. 5, pp. 801–807, 2014.
[71] J. Pathak et al., "Normalization and standardization of electronic health records for high-throughput phenotyping: The SHARPn consortium," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. e2, pp. e341–e348, 2013.
[72] L. R. Soenksen et al., "Integrated multimodal artificial intelligence framework for healthcare applications," *NPJ Digit. Med.*, vol. 5, no. 1, 2022, Art. no. 149.
[73] A. Falchier et al., "Anatomical evidence of multimodal integration in primate striate cortex," *J. Neurosci.*, vol. 22, no. 13, pp. 5749–5759, Jul. 2002, doi: 10.1523/JNEUROSCI.22-13-05749.2002.
[74] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, Dec. 1976, doi: 10.1038/264746a0.
[75] J. Sepulcre et al., "Stepwise connectivity of the modal cortex reveals the multimodal organization of the human brain," *J. Neurosci.*, vol. 32, no. 31, pp. 10649–10661, Aug. 2012, doi: 10.1523/JNEUROSCI.0759-12.2012.
[76] D. J. Ardesch et al., "Evolutionary expansion of connectivity between multimodal association areas in the human brain compared with chimpanzees," *Proc. Nat. Acad. Sci.*, vol. 116, no. 14, pp. 7101–7106, 2019.
[77] A. Vrečko et al., "A computer vision integration model for a multi-modal cognitive system," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2009, pp. 3140–3147.
[78] R. J. Gillies et al., "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.
[79] L. Chin et al., "Cancer genomics: From discovery science to personalized medicine," *Nature Med*, vol. 17, no. 3, pp. 297–303, Mar. 2011, doi: 10.1038/nm.2323.
[80] S. Pepke and G. Ver Steeg, "Comprehensive discovery of subsample gene expression components by information explanation: Therapeutic implications in cancer," *BMC Med. Genomic.*, vol. 10, no. 1, Mar. 2017, Art. no. 12, doi: 10.1186/s12920-017-0245-6.
[81] V. Agarwala et al., "Real-world evidence in support of precision medicine: Clinico-genomic cancer data as a case study," *Health Affairs*, vol. 37, no. 5, pp. 765–772, 2018.
[82] S. Mitra, "Deep learning with radiogenomics towards personalized management of gliomas," *IEEE Rev. Biomed. Eng.*, vol. 16, pp. 579–593, 2023, doi: 10.1109/RBME.2021.3075500.
[83] R. Shen et al., "Integrative subtype discovery in glioblastoma using iCluster," *PLoS One*, vol. 7, no. 4, 2012, Art. no. e35236.
[84] W.-Y. Teo et al., "Relevance of a TCGA-derived glioblastoma subtype gene-classifier among patient populations," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, 2019.
[85] T. C. Steed et al., "Differential localization of glioblastoma subtype: Implications on glioblastoma pathogenesis," *Oncotarget*, vol. 7, no. 18, 2016, Art. no. 24899.
[86] S. Berendsen et al., "Adverse prognosis of glioblastoma contacting the subventricular zone: Biological correlates," *PLoS One*, vol. 14, no. 10, 2019, Art. no. e0222717.
[87] S. L. Di Jia et al., "Mining TCGA database for genes of prognostic value in glioblastoma microenvironment," *Aging*, vol. 10, no. 4, 2018, Art. no. 592.
[88] Y. T. Oh et al., "Translational validation of personalized treatment strategy based on genetic characteristics of glioblastoma," *PLoS One*, vol. 9, no. 8, 2014, Art. no. e103327.
[89] R. C. Petersen et al., "Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, Jan. 2010, doi: 10.1212/WNL.0b013e3181cb3e25.
[90] C. R. Jack Jr. et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reson. Imag.*, vol. 27, no. 4, pp. 685–691, Apr. 2008, doi: 10.1002/jmri.21049.

[91] S. Bakr et al., "A radiogenomic dataset of non-small cell lung cancer," *Sci. Data*, vol. 5, Oct. 2018, Art. no. 180202, doi: 10.1038/sdata.2018.202.

[92] J. Lipkova et al., "Artificial intelligence for multimodal data integration in oncology," *Cancer Cell*, vol. 40, no. 10, pp. 1095–1110, 2022.

[93] J. Yan et al., "Deep learning features from diffusion tensor imaging improve glioma stratification and identify risk groups with distinct molecular pathway activities," *EBioMedicine*, vol. 72, 2021, Art. no. 103583.

[94] J. Zhang et al., "Hierarchical convolutional neural networks for segmentation of breast tumors in MRI with application to radiogenomics," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 435–447, Feb. 2019.

[95] R. K. Kawaguchi et al., "Assessing versatile machine learning models for glioma radiogenomic studies across hospitals," *Cancers*, vol. 13, no. 14, 2021, Art. no. 3611.

[96] Y. Huang et al., "Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network," *Front. Neurosci.*, vol. 13, 2019, Art. no. 509.

[97] N. Beig et al., "Sexually dimorphic radiogenomic models identify distinct imaging and biological pathways that are prognostic of overall survival in glioblastoma," *Neuro-Oncol.*, vol. 23, no. 2, pp. 251–263, 2021.

[98] Y. Hu et al., "Computed tomography-based deep-learning prediction of neoadjuvant chemoradiotherapy treatment response in esophageal squamous cell carcinoma," *Radiotherapy Oncol.*, vol. 154, pp. 6–13, 2021.

[99] P. Khosravi et al., "A deep learning approach to diagnostic classification of prostate cancer using pathology–radiology fusion," *J. Magn. Reson. Imag.*, vol. 54, no. 2, pp. 462–471, 2021.

[100] R. J. Chen et al., "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Trans. Med. Imag.*, vol. 41, no. 4, pp. 757–770, Apr. 2022.

[101] R. J. Chen et al., "Pan-cancer integrative histology-genomic analysis via interpretable multimodal deep learning," *Cancer Cell*, vol. 40, no. 8, pp. 865–878, 2022.

[102] L. Feng et al., "Development and validation of a radiopathomics model to predict pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer: A multicentre observational study," *Lancet Digit. Health*, vol. 4, no. 1, pp. e8–e17, 2022.

[103] S.-J. Sammut et al., "Multi-omic machine learning predictor of breast cancer therapy response," *Nature*, vol. 601, no. 7894, pp. 623–629, 2022.

[104] D. N. Louis et al., "The 2016 world health organization classification of tumors of the central nervous system: A summary," *Acta Neuropathologica*, vol. 131, pp. 803–820, 2016.

[105] K. M. Boehm et al., "Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer," *Nature Cancer*, vol. 3, no. 6, pp. 723–733, 2022.

[106] Z. A. Binder et al., "Epidermal growth factor receptor extracellular domain mutations in glioblastoma present opportunities for clinical imaging and therapeutic development," *Cancer Cell*, vol. 34, no. 1, pp. 163–177, 2018.

[107] X. Yi et al., "Incorporating SULF1 polymorphisms in a pretreatment CT-based radiomic model for predicting platinum resistance in ovarian cancer treatment," *Biomed. Pharmacother.*, vol. 133, 2021, Art. no. 111013.

[108] B.-C. Kim et al., "Preliminary radiogenomic evidence for the prediction of metastasis and chemotherapy response in pediatric patients with osteosarcoma using 18F-FDG PET/CT, EZRIN, and KI67," *Cancers*, vol. 13, no. 11, 2021, Art. no. 2671.

[109] Y. Iwatate et al., "Radiogenomics for predicting p53 status, PD-L1 expression, and prognosis with machine learning in pancreatic cancer," *Brit. J. Cancer*, vol. 123, no. 8, pp. 1253–1261, 2020.

[110] I. Hoshino et al., "Prediction of the differences in tumor mutation burden between primary and metastatic lesions by radiogenomics," *Cancer Sci.*, vol. 113, no. 1, 2022, Art. no. 229.

[111] A. Saha et al., "A machine learning approach to radiogenomics of breast cancer: A study of 922 subjects and 529 DCE-MRI features," *Brit. J. Cancer*, vol. 119, no. 4, pp. 508–516, 2018.

[112] J. B.-K. Hsu et al., "Radiomic immunophenotyping of GSEA-assessed immunophenotypes of glioblastoma and its implications for prognosis: A feasibility study," *Cancers*, vol. 12, no. 10, 2020, Art. no. 3039.

[113] G. Pinheiro et al., "Identifying relationships between imaging phenotypes and lung cancer-related mutation status: EGFR and KRAS," *Sci. Rep.*, vol. 10, no. 1, pp. 1–9, 2020.

[114] R. J. Chen et al., "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3995–4005.

[115] W. Shao et al., "Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 99–110, Jan. 2020.

[116] A. Yala et al., "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, vol. 292, no. 1, pp. 60–66, 2019.

[117] N. Nuechterlein et al., "Radiogenomic modeling predicts survival-associated prognostic groups in glioblastoma," *Neuro-Oncol. Adv.*, vol. 3, no. 1, 2021, Art. no. vdab004.

[118] J. Haubold et al., "Non-invasive tumor decoding and phenotyping of cerebral gliomas utilizing multiparametric 18F-FET PET-MRI and MR fingerprinting," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 47, no. 6, pp. 1435–1445, 2020.

[119] S. Padmanabhan, *Handbook of Pharmacogenomics and Stratified Medicine*. New York, NY, USA: Academic, 2014.

[120] N. Allen et al., "U.K. Biobank: Current status and what it means for epidemiology," *Health Policy Technol.*, vol. 1, no. 3, pp. 123–126, 2012.

[121] E. Bowton et al., "Biobanks and electronic medical records: Enabling cost-effective research," *Sci. Transl. Med.*, vol. 6, no. 234, 2014, Art. no. 234cm3.

[122] J. C. Kirby et al., "PheKB: A catalog and workflow for creating electronic phenotype algorithms for transportability," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 6, pp. 1046–1052, 2016.

[123] M. D. Ritchie et al., "Genome-and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk," *Circulation*, vol. 127, no. 13, pp. 1377–1385, 2013.

[124] Y. Veturi et al., "A unified framework identifies new links between plasma lipids and diseases from electronic medical records across large-scale cohorts," *Nature Genet.*, vol. 53, no. 7, pp. 972–981, 2021.

[125] L. D. Pena et al., "Looking beyond the exome: A phenotype-first approach to molecular diagnostic resolution in rare and undiagnosed diseases," *Genet. Med.*, vol. 20, no. 4, pp. 464–469, 2018.

[126] Y. Su et al., "Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19," *Cell*, vol. 183, no. 6, pp. 1479–1495, 2020.

[127] J. Park et al., "A genome-first approach to aggregating rare genetic variants in LMNA for association with electronic health record phenotypes," *Genet. Med.*, vol. 22, no. 1, pp. 102–111, 2020.

[128] T. G. Drivas et al., "Mendelian pathway analysis of laboratory traits reveals distinct roles for ciliary subcompartments in common disease pathogenesis," *Amer. J. Hum. Genet.*, vol. 108, no. 3, pp. 482–501, 2021.

[129] A. Verma et al., "A phenome-wide association study of genes associated with COVID-19 severity reveals shared genetics with complex diseases in the million veteran program," *PLoS Genet.*, vol. 18, no. 4, 2022, Art. no. e1010113.

[130] J. Park et al., "Exome-wide evaluation of rare coding variants using electronic health records identifies new gene–phenotype associations," *Nature Med.*, vol. 27, no. 1, pp. 66–72, 2021.

[131] E. Drucker and K. Krapfenbauer, "Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine," *EPMA J.*, vol. 4, no. 1, pp. 1–10, 2013.

[132] R. L. Richesson et al., "Enhancing the use of EHR systems for pragmatic embedded research: Lessons from the NIH health care systems research collaboratory," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 12, pp. 2626–2640, 2021.

[133] N. Mulder et al., "Genomic research data generation, analysis and sharing–challenges in the African setting," *Data Sci. J.*, vol. 16, p. 49, 2017.

[134] A. Mammoliti et al., "Orchestrating and sharing large multimodal data for transparent and reproducible research," *Nature Commun.*, vol. 12, no. 1, pp. 1–10, 2021.

[135] S. Greenfield et al., "Heterogeneity of treatment effects: Implications for guidelines, payment, and quality assessment," *Amer. J. Med.*, vol. 120, no. 4, pp. S3–S9, 2007.

[136] Y. A. Khan et al., "Precision medicine and heterogeneity of treatment effect in therapies for ARDS," *Chest*, vol. 160, no. 5, pp. 1729–1738, 2021.

[137] D. Madigan et al., "Evaluating the impact of database heterogeneity on observational study results," *Amer. J. Epidemiol.*, vol. 178, no. 4, pp. 645–651, 2013.

[138] J. Seoane and L. De Mattos-Arruda, "The challenge of intratumour heterogeneity in precision medicine," *J. Intern. Med.*, vol. 276, no. 1, pp. 41–51, 2014.

[139] J. J. Meeks et al., "Genomic heterogeneity in bladder cancer: Challenges and possible solutions to improve outcomes," *Nature Rev. Urol.*, vol. 17, no. 5, pp. 259–270, 2020.

[140] N. G. Weiskopf et al., "A data quality assessment guideline for electronic health record data reuse," *Egems*, vol. 5, no. 1, 2017, Art. no. 14.

[141] C. A. Anderson et al., "Data quality control in genetic case-control association studies," *Nature Protoc.*, vol. 5, no. 9, pp. 1564–1573, 2010.

[142] W. Kopp et al., "Simultaneous dimensionality reduction and integration for single-cell ATAC-seq data using deep learning," *Nature Mach. Intell.*, vol. 4, no. 2, pp. 162–168, 2022.
[143] A. Chattopadhyay and T.-P. Lu, "Gene-gene interaction: The curse of dimensionality," *Ann. Transl. Med.*, vol. 7, no. 24, 2019, Art. no. 813.
[144] A. S. Panayides et al., "Radiogenomics for precision medicine with a Big Data analytics perspective," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 5, pp. 2063–2079, Sep. 2019.
[145] J. J. Cimino et al., "The national institutes of health's biomedical translational research information system (BTRIS): Design, contents, functionality and experience to date," *J. Biomed. Inform.*, vol. 52, pp. 11–27, 2014.
[146] C. Bycroft et al., "The U.K. Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.
[147] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England J. Med.*, vol. 372, no. 9, pp. 793–795, 2015.
[148] S. L. Jenkins and A. Ma'ayan, "Systems pharmacology meets predictive, preventive, personalized and participatory medicine," *Pharmacogenomics*, vol. 14, no. 2, pp. 119–122, 2013.
[149] A. Roehrs et al., "Personal health records: A systematic literature review," *J. Med. Internet Res.*, vol. 19, no. 1, 2017, Art. no. e5876.
[150] D. Balaji and S. F. Terry, "Benefits and risks of sharing genomic information," *Genet. Testing Mol. Biomarkers*, vol. 19, no. 12, pp. 648–649, 2015.
[151] M. J. Sheller et al., "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.
[152] J. Yoon et al., "Gain: Missing data imputation using generative adversarial nets," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5689–5698.
[153] A. Choudhary et al., "Advancing medical imaging informatics by deep learning-based domain adaptation," *Yearbook Med. Inform.*, vol. 29, no. 01, pp. 129–138, 2020.
[154] Y. Pan et al., "Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis," in *Proc. 21st Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2018, pp. 455–463.
[155] S. A. Friedler et al., "A comparative study of fairness-enhancing interventions in machine learning," in *Proc. Conf. Fairness, Accountability, Transparency*, 2019, pp. 329–338.
[156] C. Dwork et al., "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*, 2012, pp. 214–226.
[157] H. Zhang et al., "Translating genomic medicine to the clinic: Challenges and opportunities," *Genome Med.*, vol. 11, no. 1, pp. 1–3, 2019.
[158] Ž. Avsec et al., "Base-resolution models of transcription-factor binding reveal soft motif syntax," *Nature Genet.*, vol. 53, no. 3, pp. 354–366, 2021.