***ATVB* IN FOCUS:**
Integrative Multi-Omic Approaches in Cardiovascular Disease and Treatment

Series Editor: Joseph Loscalzo

Phenomics and Robust Multiomics Data for Cardiovascular Disease Subtyping

Enrico Maiorino, Joseph Loscalzo

ABSTRACT: The complex landscape of cardiovascular diseases encompasses a wide range of related pathologies arising from diverse molecular mechanisms and exhibiting heterogeneous phenotypes. This variety of manifestations poses significant challenges in the development of treatment strategies. The increasing availability of precise phenotypic and multiomics data of cardiovascular disease patient populations has spurred the development of a variety of computational disease subtyping techniques to identify distinct subgroups with unique underlying pathogeneses. In this review, we outline the essential components of computational approaches to select, integrate, and cluster omics and clinical data in the context of cardiovascular disease research. We delve into the challenges faced during different stages of the analysis, including feature selection and extraction, data integration, and clustering algorithms. Next, we highlight representative applications of subtyping pipelines in heart failure and coronary artery disease. Finally, we discuss the current challenges and future directions in the development of robust subtyping approaches that can be implemented in clinical workflows, ultimately contributing to the ongoing evolution of precision medicine in health care.

Key Words: algorithms ■ coronary artery disease ■ heart failure ■ multiomics ■ precision medicine

Cardiovascular diseases (CVDs) are the leading cause of global mortality, with recent estimates[1] indicating 17.8 million fatalities worldwide in 2017. CVDs are also a major cause of morbidity and disability, affecting the overall quality of life and placing a significant burden on health care systems.[2] Despite the global CVD prevalence, it has been estimated that the 2 decades between 1990 and 2012 have seen a decline in the number of cardiovascular drugs that have entered clinical trials[3] in contrast to the increase observed in other therapeutic areas such as cancer.

One of the main reasons for the decline is the low absolute rate of success of CVD drugs in clinical trials and, after approval, in practice. In recent trials, health benefits were reported for as few as 9% of the cases for simvastatin,[4] 4.5% for abciximab (compared with placebo),[5] and 2.2% for clopidogrel.[6] The limited effectiveness of existing CVD treatments can be attributed to the heterogeneity of CVD pathogenesis and its manifestations in affected patients.

Please see www.ahajournals.org/atvb/atvb-focus for all articles published in this series.
See cover image

Most CVDs exhibit different pathobiology, risk, and therapeutic response depending on a variety of factors. For example, coronary artery disease (CAD) can either be asymptomatic or present with a range of symptoms that include both chronic and acute manifestations of the disease. Heart failure (HF) is a convergent phenotype that exhibits significant heterogeneity both in terms of its clinical presentation and its etiological factors. Its main subphenotypes, HF with preserved ejection fraction (HFpEF) and HF with reduced ejection fraction (HFrEF), affect different demographics and display different comorbidities and response to therapies. Furthermore, since HF can arise from multiple pathologies, it is often classified based on the underlying causes, which can include ischemic diseases, hypertension, valvular diseases, cardiomyopathies (hypertrophic, dilated, and restrictive), and congenital heart defects.[7]

Correspondence to: Joseph Loscalzo, MD, PhD, Division of Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA. Email jloscalzo@rics.bwh.harvard.edu

This manuscript was sent to Luisa Iruela-Arsipe, Senior Consulting Editor, for review by expert referees, editorial decision, and final disposition.
For Sources of Funding and Disclosures, see page 1120.

Arterioscler Thromb Vasc Biol is available at www.ahajournals.org/journal/atvb

Nonstandard Abbreviations and Acronyms

BP	blood pressure
CAD	coronary artery disease
CVD	cardiovascular disease
DR	dimensionality reduction
EHR	electronic health record
FHS	Framingham Heart Study
HF	heart failure
HFpEF	heart failure with preserved ejection fraction
IR	individualized representation
PC	principal component

At the genetic level, numerous genetic variants have been linked to an increased CVD risk[8–13] and drug response.[14,15] Environmental exposures and lifestyle factors, such as smoking, are, of course, also atherothrombogenic, with consequent increased risk of atherosclerosis, peripheral artery disease, and abdominal aortic aneurysm.[16]

Furthermore, different CVDs are causally related, with the development of one type of CVD increasing the risk of developing others. Overall, each disease case presents a complex clinical picture that necessitates the adoption of tailored treatment strategies that can take in consideration all of its unique characteristics.[14]

PRECISION MEDICINE AND DISEASE SUBTYPING

Despite the heterogeneity of CVD manifestations, conventional therapeutic approaches often involve administering the same therapies to all patients based on clinical trial criteria that take into consideration only a small subset of the available clinical evidence about each individual.[17] This one-size-fits-all approach does not regularly take into account a person's unique genetic features, medical history, and lifestyle in tailoring therapies. Modern precision medicine approaches aim to customize treatments to the unique needs of individual patients by leveraging measurements from molecular profiling technologies (omics), laboratory diagnostics, and electronic health data.

Disease subtyping lies at the heart of this transformation (Figure, top), aiming to identify distinct groups with similar disease manifestations to devise targeted, tailored therapies.[18] Computational disease subtyping techniques harness the power of data science and machine learning, unveiling patient groups within a high-dimensional space composed of clinical features, laboratory measurements, and biological components.

To complement precision medicine approaches, the emerging network medicine paradigm views the health and disease states of an individual as a complex network of networks (interactomes) within which many biological components interact.[19–22] The nodes in these interaction networks can represent genes, proteins, or even individuals, and their connections can describe a functional relation (eg, protein binding) as well as a similarity relation (eg, patient similarity networks). Owing to its versatility, network modeling can aid at multiple stages of the computational disease subtyping workflow, ranging from the identification of subtype-specific coexpression modules,[23] to the definition of individualized molecular networks,[24] and to the integration of data across various domains to capture disease-related variability.[25]

Highlights

- Computational approaches enable precise subtyping of cardiovascular diseases for personalized treatment strategies.
- We delineate the primary categories of computational disease subtyping strategies: clinical and molecular subtyping.
- We review the basic computational workflow of disease subtyping applications, composed of a feature selection step, a data integration and representation step, and a clustering step.
- We provide an overview of selected applications in heart failure and coronary artery disease subtyping.
- We discuss current opportunities and future directions for improving computational disease subtyping and discuss translational challenges of clinical models.

In this work, we review the basic blueprint of a computational pipeline for disease subtyping, illustrating the most common methods, choices, and challenges encountered at each stage of the analysis. Next, we provide a brief overview of subtyping applications for CAD and HF. Throughout the presentation, we emphasize the importance of an integrative multiomic approach coupled with precise phenotypic data, which is itself yet another level of computational features incorporated in predictive models.

CLASSES OF DISEASE SUBTYPING APPROACHES

The diffusion of electronic health records (EHRs), advanced medical diagnostics, and sequencing technologies has created an unprecedented opportunity to characterize systematically patient populations across different domains. A variety of computational approaches have been developed to stratify patients in high-dimensional space of molecularly and clinically relevant features. At the most basic level, subtyping approaches can be distinguished as (1) clinical subtyping, based on features typically recorded in the clinical setting such as demographics, symptoms, medical history, and laboratory

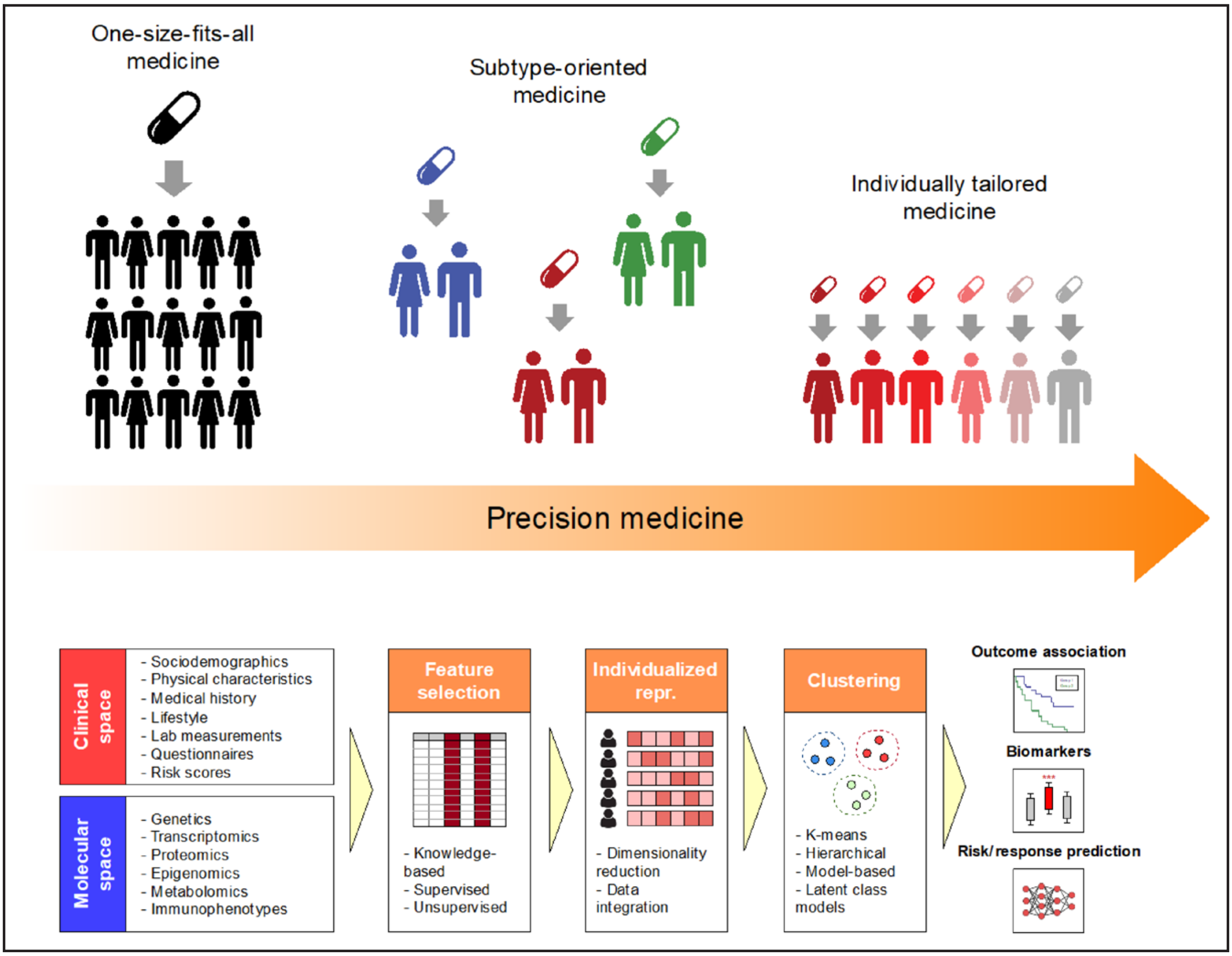

Figure. Advancing precision medicine through computational disease subtyping.

Top, Path toward precision medicine, from one-size-fits-all medicine to individually tailored medicine. **Bottom**, Basic workflow of computational disease subtyping applications. Lab indicates laboratory; and Repr., representations.

measurements and (2) molecular subtyping, based on unbiased omics assays (proteomics, transcriptomics, metabolomics, and epigenomics). Clinical subtyping aims to classify the phenotypic manifestations of the disease, which can be useful for identifying common disease progression patterns that can guide management.[26] Molecular subtyping is, instead, mainly focused on understanding the molecular processes that affect disease risk and development.[27] While not every study may fall neatly into these categories, in this review, we adhere to this primary characterization and outline the main strengths and challenges associated with each class of approaches.

CLINICAL SUBTYPING

Most CVDs have prevailing phenotypic and environmental components. The generation of detailed and extensive health information about the phenotypic condition of an individual, often referred to as deep phenotyping,[28] is an essential step in the development of precision medicine.

Cohort studies, such as the FHS (Framingham Heart Study),[29] are a primary source of precise and well-characterized phenotypic information. However, they have stringent selection criteria for patient enrollment and follow-up, and, therefore, do not always reflect the natural phenotypic variation across the general population. Following the recent diffusion of EHR platforms, observational data collected from health care institutions have become a feasible alternative for obtaining a more comprehensive pool of health information.[30]

The availability of these resources has stimulated the emergence of phenomics as the conceptual counterpart of genomics.[31,32] Under this perspective, the phenotype of an individual is described through a high-dimensional set of features, as opposed to the traditional case/control dichotomization used in most population studies. Clinical subtyping applications harness computational power to define subtypes from extensive big data sources, identifying subtle irregularities that lie beyond manual curation. In this context, advances in statistical and machine learning techniques, especially deep learning approaches applied to medical imaging data and ECG profiles,[33,34] are set to play a central role in enabling detailed phenotyping for precision medicine applications.

Nonetheless, there are multiple computational and conceptual challenges associated with clinical subtyping. Clinical variables with their diverse data types (binary, categorical, and numerical), often lacking a direct numerical representation (eg, visit notes, angiograms), are difficult to integrate in subtyping workflows and require specialized feature extraction procedures.[35]

Furthermore, as EHRs were developed primarily for billing and accounting purposes, EHR-based clinical data are noisy, incomplete, and biased.[36] Noise in data can result from reporting inaccuracies, digitization errors, or billing requirements that may not always be consonant with relevant disease features.[37] EHR data often contain missing values due to patient dropout, insufficient screening, or irregular patient–health care system interactions, which may introduce selection biases and potentially produce misleading patterns in data. To address these issues, most subtyping efforts typically involve stringent data filtering criteria and imputation operations.[38,39] Additionally, statistical and machine learning models have been introduced to account for hidden confounders and correct them when possible.[40–44]

In summary, clinical subtypes help characterize a disease's clinical presentation, differentiate severity levels, and enable a more accurate prognosis. However, the connection between phenotype and the underlying endotype is not always consistent, as different disease mechanisms may converge to similar clinical outcomes (convergent phenotypes), or, conversely, similar etiological factors may interact with individual genetics and comorbidities to produce divergent clinical outcomes (divergent phenotypes).[27,45]

MOLECULAR SUBTYPING

Molecular subtyping focuses on characterizing the biological processes underlying a specific pathobiology. The steadily decreasing costs of modern sequencing platforms have facilitated the study of tissue across multiple molecular strata (such as genomics, proteomics, and transcriptomics),[46,47] promoting the discovery of novel associations and reducing the reliance on a priori knowledge for feature selection compared with clinical subtyping. Other assays, such as flow cytometry–based immunophenotyping, provide noninvasive avenues to investigate immunologic abnormalities in many different diseases, from autoimmunity to cancer.[48]

When designing a study, selecting the appropriate tissues and omic domains for performing subtyping is a challenging task and requires careful consideration of both practical and conceptual aspects. For example, plasma and peripheral blood mononuclear cell sampling is an accessible and minimally invasive procedure that can be repeated over time, and blood can provide information about the systemic effects of the disease. However, blood-based measurements are influenced by various biological phenomena that involve different cell types and processes and, therefore, may not have sufficient sensitivity for disease variation. In contrast, while obtaining local tissue samples typically requires a biopsy or surgery of the affected area, it offers a more direct and localized view of the molecular processes associated with a specific pathophenotype. Additionally, different types of omics measurements may emphasize distinct mechanistic facets of the pathology, which could potentially be limited or partial in their scope. This issue can be mitigated by leveraging multiomics assays that profile cellular compositions across multiple biological strata. The adoption of these multimodal platforms has stimulated the development of a variety of computational techniques for multiomics data integration and variable selection.[49,50] For example, multilayer networks[51,52] have been used to represent patient-patient similarity across different omics domains[25] or to identify important disease determinants by modeling biological interactions across different molecular domains.[53] However, the integration of multiomics data comes with new challenges. Most approaches designed to model multiomics data seek shared patterns of variation across data modalities to extract insights about a phenotype. However, less attention has been devoted to detecting conflicting signals between omics that may cancel out and cause lower prediction accuracies compared with single-omics analyses.[54,55] Furthermore, different omics platforms produce data with different formats, scales, and distributions and are characterized by different sensitivities, biases, and noise levels. These discrepancies can prevent the detection of significant patterns of association involving multiple modalities.[56] Finally, most multiomics data integration efforts have focused on combining various types of molecular information, but the integration of such data with clinical information remains limited.[57] Nonetheless, the rapidly evolving research landscapes of multimodal learning[58] and network-based integration[59] hold great promise for generating robust models that can recapitulate the multifaceted nature of the cellular processes involved in CVD.

TYPICAL SUBTYPING WORKFLOW

At the computational level, subtyping translates in most cases to a clustering problem, where the subtypes to be found represent groups of patients with similar characteristics. Here, we next review the fundamental steps, choices, and caveats encountered when designing a subtyping analysis (Figure, bottom). To highlight the universal design patterns of a typical workflow, we will remain generic with respect to the specific disease being studied, occasionally providing some examples. Moreover, for brevity, we will not focus on several essential data-specific processing steps, including outlier removal, missing value imputation, and data normalization.

FEATURE SELECTION

Although current assays are becoming increasingly cost-effective, their vast dimensionality is not easily exploited due to the relatively small sample sizes they generate. For example, typical sample sizes in bulk RNA sequencing data range between the tens and hundreds of samples, while the number of detected transcripts is several orders of magnitude larger. This disparity leads to low discriminative power and overfitting (the "curse of dimensionality"[60]).

To mitigate these issues, the initial phase of a subtyping analysis involves a feature selection step, which aims to remove irrelevant or redundant variables from the analysis. Feature selection criteria can be supervised or unsupervised (Table 1). Supervised selection criteria evaluate the relevance of each feature by measuring its relation with a clinically relevant outcome of interest, such as disease progression, response to treatment, or mortality (see Reference[61] for a review). While this approach has the advantage of producing subtype classifications that are clinically relevant and interpretable, it has the limitation of identifying subtypes that may relate only to a narrow fraction of the determinants of disease variability. In contrast, unsupervised feature selection works by defining endogenous criteria for selecting features, such as having a sizable variance across the population (see Table 1 and Reference[62] for examples). Since the selection criteria derive from data, unsupervised techniques make minimal assumptions about the sources of variability to preserve in the data. However, these approaches are not designed to discern whether a strong signal of variability is clinically relevant, and the subtypes identified downstream are less interpretable and may require substantial post hoc analysis to extract insights about the disease. A challenge often encountered in feature selection for disease subtyping applications is the identification of multiple sets of features that yield similar levels of predictive performance. This phenomenon can occur due to residual redundancy of information between features, the existence of multiple competing signals in data, or a low overall signal-to-noise ratio. While these issues warrant a case-by-case assessment of their causes, general solutions include the definition of cross-validation and out-of-sample validation strategies to assess the stability and robustness of the identified feature sets. Furthermore, other studies defined ensemble feature selection schemes to find consensus sets across multiple sets of selected features.[87,88] Several studies have compared existing feature selection approaches across specific types of biomedical data, including clinical variables,[89,90] gene expression,[91] proteomics,[92] and metabolomics.[93] The selected features are the starting point for constructing individualized representations (IRs) of each subject in the study population, as explained in the next section.

GENERATION OF INDIVIDUALIZED REPRESENTATIONS

The feature selection step produces a reduced data set in which each individual in the study population is represented by a numerical vector of relevant features. However, this simple form is not always the most desirable encoding strategy for performing clustering. If the feature selection is not sufficiently conservative, the dimensionality of the vectors may remain too large, and the chosen features may include heterogeneous data types (eg, binary and categorical). For example, even a stricter filtering of gene expression data is likely to select thousands of transcripts as relevant variables. In order to mitigate the curse of dimensionality and summarize complex patterns in data more effectively, in most cases, another processing step is necessary. This step, often called feature extraction in the machine learning literature, is the process of transforming the original features in order to obtain an IR of each subject in the population. We define an IR as a mathematical object that describes implicitly or explicitly the information about a given individual contained in the original data and can be encoded by a numerical vector or even a complex object such as a network.[74] The most common approaches for building IRs for subtyping involve some combination of (1) dimensionality reduction (DR) and (2) data integration. DR methodologies aim to construct a small set of variables that carries most of the information contained in the original data. A variety of other techniques have been proposed to perform DR and can be grouped into linear techniques and nonlinear techniques. Linear techniques construct new variables as a weighted sum of the original variables according to some criteria. Principal component (PC) analysis, for example, decomposes data in a set of uncorrelated components (PCs). While the PC analysis–transformed space has the same number of curse of dimensionality features as the original space, DR is usually performed by selecting a subset of the PCs that explains most of the variability in data, discarding the rest (see Table 1 for examples and Reference[63] for a detailed review on linear techniques). Although these methods offer stability, rapid execution, and easier interpretability, they lack the ability to model nonlinear relationships that may exist among the original variables, which can result in less discriminative power. Nonlinear techniques are designed to fill this gap by trading the simplicity, stability, and computational efficiency of linear techniques with more model flexibility. In several situations, nonlinear techniques have been shown to produce

Table 1. Examples of Methodologies and Criteria Used in Subtyping Workflows

Feature selection			
Criterion type	Selection type	Examples	Basic criteria
Statistical measures	Supervised	Statistical association,[68] Gini index	Features are differentially distributed across outcome classes
Class separation	Supervised	Fisher score, CBFS,[69] ReliefF[70]	Feature values are similar within the same class and different among different outcome classes
Information theoretic	Supervised	MIM, MIFS, MRMR[71]	Features carry information about outcome class
Feature importance for prediction	Supervised	LightGBM[39]	Features are important for predicting outcome class
Correlation	Supervised	Pearson correlation, HSIC[72]	Features are correlated to outcome value
Model regularization	Supervised	LASSO, group LASSO	Features are selected in regularized regressions against outcome value
Feature variance and redundancy	Unsupervised	Variance, Pearson correlation	Features have high variance and are not too correlated with each other
Similarity preservation	Unsupervised	Laplacian score, MCFS[73]	Features preserve the similarity relations across subjects
Model regularization	Unsupervised	NDFS[74]	Features are highly discriminative of subject identity
Feature extraction and DR			
Method type	Model type	Examples	Rationale
Linear transformations	Linear	PCA, LDA, FA, NMF, GLRM[75]	Find linear combinations of original features that summarize data
Kernel based	Nonlinear	Kernel-PCA[76]	Extend linear techniques using nonlinear measures of similarity
Manifold learning	Nonlinear	t-SNE,[77] UMAP,[78] SOM[79]	Find a low-dimensional nonlinear manifold that summarizes data
Neural networks	Nonlinear	Stacked AE, Variational AE, Denoising AE[80]	Compress data with neural networks
Specialized for multiomics data integration			
Base technique	Model type	Examples	Rationale
Matrix factorization	Linear	MOFA,[81] JIVE,[82] tICA,[83] intNMF,[84] RGCCA[85]	Find linear factors that describe shared and specific variability across different modalities
Statistical modeling	Linear	iCluster[86]	Find latent clusters that summarize data across modalities
Network modeling	Network based	SNF[25]	Merge similarity matrices across different data modalities
Clustering			
Base technique	Output	Resolution parameter	Rationale
K-means clustering	Partition	Number of clusters	Find cluster centroids that minimize intracluster distances
Hierarchical clustering	Hierarchical cluster membership	Tree cut level	Aggregate points hierarchically depending on their distances
Spectral clustering	Partition	Number of clusters	Find an optimal partition in the graph of nearest neighbors of the original space
Model-based clustering	Likelihood of cluster membership	Number of latent classes	Fit a generative model to data and evaluate likelihood of membership to a cluster

For reviews of these methodologies, see References[61,62] (feature selection), References[63–66] (DR/feature extraction), and Reference[67] (clustering). AE indicates autoencoder; CBFS, clearness-based feature selection; DR, dimensionality reduction; FA, factor analysis; GLRM, generalized low-rank models; HSIC, Hilbert-Schmidt independence criterion; intNMF, integrative non-negative matrix factorization; JIVE, joint and individual variation explained; LDA, linear discriminant analysis; MCFS, min-cut–based feature-selection; MIFS, mutual information feature selection; MIM, mutual information maximization; MOFA, multiomics factor analysis; MRMR, minimal-redundancy-maximal-relevance; NDFS, non-negative discriminative feature selection; NMF, non-negative matrix factorization; PCA, principal component analysis; RGCCA, regularized generalized canonical correlation analysis; SNF, similarity network fusion; SOM, self-organizing maps; t-SNE, t-distributed stochastic neighbor embedding; tICA, tensorial independent component analysis; and UMAP, uniform manifold approximation and projection.

representations that perform better in downstream analyses such as classification, clustering, and data visualization[94] (see Table 1 for examples and Reference[64] for a comprehensive overview). However, they are more prone to overfitting data in low sample size settings, that is, they may construct variables that include the noise in data and have low generalizability.

In situations where subtyping is performed across multiple domains, for example, multiomics, it is often beneficial to use methodologies designed to integrate multiple data modes. Integrating multiple data types in a single IR before performing the analysis is often referred to as early integration or early fusion. The counterpart of early integration is late integration, where separate analyses are performed for each data type (eg, clustering), and the results are merged downstream.[95] A middle ground is termed intermediate integration where the integration is performed implicitly by a joint model in the main analysis. The added advantage of performing early or intermediate integrative DR in this way is that one can produce

IRs that encode the shared and domain-specific signals of each data type in a single description. Most of these approaches are designed for multiomics data integration and incorporate extensions of traditional techniques such as factor analysis (see Table 1 for examples and References[65],[66] for an overview).

Nonetheless, most of the IRs based on multiomics synthesize the overall molecular state of an individual as a function of the separate concentrations of different molecules, disregarding their functional context and interactions. One promising direction of research is to integrate diverse types of molecular interaction data (eg, protein-protein interactions and gene regulatory interactions) to construct individualized networks that capture the unique cellular interactions occurring in each individual.[24,96] Another network methodology, similarity network fusion,[25] integrates the pairwise similarities between individuals across multiple data types to construct a merged patient-patient similarity network. After constructing the IRs, pairwise distances can be computed using various measures (eg, Euclidean distance for vector-based IRs or graph edit distance for network-based IRs) and then used as input for the clustering process detailed in the next section.

CLUSTERING AND BIOMARKER IDENTIFICATION

Once the IRs have been generated for each subject in the study population, the next step is to identify groups of individuals who are similar in the IR space. More precisely, the objective is to assign every individual in the population to a subtype in such a way that pairs of individuals within the same subtype are more similar than pairs of individuals of different subtypes. This approach is an unsupervised clustering task and is the core of the analysis. Cluster analysis is a vast field of research, and hundreds of different algorithms have been proposed to address different situations with varying performances. A famous theoretical result in machine learning, the no-free-lunch theorem,[97] states that no optimization algorithm can perform consistently better than all other algorithms in all possible situations. In practice, this means that there is no universal algorithmic silver bullet for solving a clustering problem, that is, the choice of the clustering algorithm for a specific application has to be assessed on a case-by-case basis. While a comprehensive overview of clustering algorithms is beyond the scope of this review (see Reference[67] for an overview), there are 4 major classes of algorithms that are most commonly used in subtyping applications, namely, K-means clustering, spectral clustering, hierarchical clustering, and model-based clustering.

All clustering algorithms require specific parameter choices that are application-dependent. To obtain the optimal partition, multiple parameter configurations are tested and evaluated through several quality measures. Common measures include the compactness and separation of the found clusters (eg, silhouette width, Dunn index, and Davies-Bouldin index[98]), the stability of the partition to noise and resampling,[99] or the complexity of the clustering model (eg, Bayesian information criterion[100]). To improve cluster robustness, clustering algorithms can be executed with multiple parameterizations, a practice referred to as consensus clustering, the outputs of which are then aggregated to identify a consensus partition that averages all of the individual solutions. Furthermore, several solutions have been developed for situations where cluster boundaries are not well-defined and a point can belong to multiple clusters simultaneously. This setting, called soft clustering,[101] is particularly useful in applications where a subtype is composed of multiple independent mechanisms or in the presence of overlapping subphenotypes.[102] However, soft clustering algorithms come with a higher computational cost owing to the combinatorial complexity of soft partitions, and the resulting partial cluster assignments may prove more challenging to interpret.

Once the optimal partition has been found, a post hoc statistical analysis is performed to discover relevant cluster biomarkers[103] or associate the found subtypes with relevant scientific outcomes such as mortality and hospitalization.[38,103,104] Some caution must be exercised in interpreting the *P* values resulting from the statistical analysis since the clustering operation forces the separation of data into groups, causing artificial *P*-value inflation.[27] As a possible solution, post hoc statistical testing that accounts for clustering structure has been proposed in some specific contexts.[105] However, a general pragmatic approach is to consider the *P* values as a descriptive measure of difference instead of evidence of true statistical significance. A final avenue for validating clustering results and demonstrating generalizability is to replicate the results on a different cohort, showing that the original classification yields distinct subtypes in the validation data set.

CVD SUBTYPING

Owing to the complex pathobiology of most CVDs, the CVD subtyping literature is sparse and heterogeneous. In many cases, the authors follow different workflows that may not incorporate the basic steps described above. Here, we provide a nonexhaustive overview of several recent studies that have as their main objective the identification of different subtypes of a CVD. The salient features of these and other studies are summarized in Table 2.

HF With Preserved Ejection Fraction

HFpEF accounts for approximately half of the total HF prevalence in the population.[117] However, as opposed to HFrEF with reduced ejection fraction, HFpEF has proven to be unresponsive or weakly responsive to

Table 2. Selected CVD References Subtyping Studies

ID	PMID	Year	Disease	Subtyping type	Feature selection	Individualized representation	Main data type	Clustering
Shah et al[38]	25398313	2014	HFpEF	Clinical	Correlation thresholding	Feature vectors	Demo, Phys, Lab	Model-based clustering
Kao et al[106]	26250359	2015	HFpEF	Clinical	Knowledge based	Feature vectors	Demo, Comorb, Lab	Latent class analysis
Segar et al[107]	31637815	2019	HFpEF	Clinical	Correlation thresholding	Feature vectors	Demo, Clin, Lab	Model-based clustering
Cohen et al[103]	31926856	2020	HFpEF	Clinical	Knowledge based	Feature vectors	Demo, MedHx, Comorb	Latent class analysis
Hedman et al[108]	31911501	2020	HFpEF	Clinical	Only continuous variables, variable clustering	Feature vectors	Clin, Lab	Model-based clustering
Woolley et al[109]	33651430	2021	HFpEF	Molecular	Knowledge based	PCA	Proteomics	Hierarchical clustering
Wu et al[110]	33868594	2021	HFpEF	Molecular	Genome wide	SNF (implicit)	mRNA/miRNA expr, methylation	Spectral clustering
Wosiak and Zakrzewska[70]	NA	2018	CAD	Clinical	RCA, CFS, ReliefF	Feature vectors	Demo, Phys, Lab	K means, Gaussian mixtures
Peng et al[111]	30805932	2019	CAD	Molecular	NA	Feature vectors	mRNA expression	Consensus clustering
Flores et al[104]	34845917	2021	CAD	Clinical	Knowledge based	Generalized low-rank modeling	Demo, MedHx, Env, Lab, SNPs	K means
Ding et al[112]	35733129	2022	CAD	Molecular	Knowledge based	Feature vectors	mRNA expression	NMF/consensus/HC
Guo et al[113]	28266630	2017	CAD	Clinical	NA	Feature vectors	BP	K means
Ding et al[39]	36105873	2022	AIS	Clinical	Supervised association	Feature vectors	Demo, Lab, Comorb, Lab	Gaussian mixture model
Cho et al[114]	32762883	2019	CVDs	Clinical	Knowledge based	TDA based	Echocardiographic measurements	TDA based
Palou-Marquez et al[68]	33836805	2021	CVDs	Molecular	Association with CVD outcome	MOFA	mRNA expr, methylation	Association with CVD events
Verdonschot et al[115]	33156912	2020	DCM	Clinical	Correlation thresholding	Factor analysis on mixed data	Demo, Phys, Lab	Hierarchical clustering
Maron et al[24]	33558530	2021	HCM	Molecular	Feature value thresholding	Individualized networks	mRNA expr	Individualized analysis
Tromp et al[116]	29584721	2018	HF	Clinical	Knowledge based	Feature vectors	Comorbidities	Latent class analysis

AIS indicates acute ischemic stroke; CAD, coronary artery disease; CFS, correlation-based feature selection; Clin, clinical measurements; Comorb, comorbidities; CVD, cardiovascular disease; DCM, dilated cardiomyopathy; Demo, demographics; Env, environmental exposures; HC, hierarchical clustering; HCM, hypertrophic cardiomyopathy; HF, heart failure; HFpEF, heart failure with preserved ejection fraction; ID, Identifier; Lab, laboratory measures; MedHx, medical history; MOFA, multiomics factor analysis; NA, not available; NMF, nonnegative matrix factorization; PCA, principal component analysis; Phys, physical examination findings; PMID, Pubmed ID; RCA, reversed correlation algorithm; SNF, similarity network fusion; SNP, single-nucleotide polymorphism; and TDA, topological data analysis.

therapy[118,119] (until very recently[120]). The far greater phenotypic heterogeneity of HFpEF cases compared with HF with reduced ejection fraction has led to the hypothesis that HFpEF may be caused by a complex combination of pathobiological processes and risk factors.[121] Finding different disease subtypes with clear pathogenesis, therefore, has the potential to identify groups of individuals who are more likely to respond to specific therapies.

In the seminal study of Shah et al,[38] the authors proposed a subtyping approach that integrates various continuous clinical features, including ECG and echocardiographic data, to identify clinical subtypes of HFpEF. To generate compact IRs, they performed unsupervised feature selection by choosing only the most informative features in groups of highly correlated variables (>0.6). The IRs were then clustered through model-based clustering, popularized by the R package mclust,[122] which allows one to impose a fixed covariance structure among patient variables to produce nonspherical clusters. They selected the partition that best summarized data with the lowest number of clusters via the Bayesian information criterion, finding 3 overall clusters (phenogroups) that are characterized by distinct clinical characteristics and increasing prevalence of hospitalization and death. The authors replicated their findings on a validation cohort, showing that their proposed phenogroup characterization has prognostic relevance.

Other studies have built upon the workflow of Shah et al. To model categorical features such as sex, Kao et al[106] used latent class analysis. Segar et al overcame one of the main model limitations of both Shah et al and Kao et al, that is, the handling of exclusively homogeneous variable types, by using a model-based clustering algorithm capable of modeling heterogeneous feature types.[123] This methodology can, therefore, include more complete information in the clustering and produce richer subtypes.

While most of these studies are focused on clinical phenotypes, several of them perform a post hoc differential analysis of select protein biomarkers detected, for example, by conventional immunoassays.[103,108] In addition to the phenotypic differences, the identified phenogroups highlight significant differences in biomarker concentrations, suggesting different pathobiological foundations between groups. Other studies explored a molecular characterization of HFpEF subtypes. Woolley et al[109] proposed the first proteomic subtyping of HFpEF cases by building IRs with PC analysis on the original space of protein concentrations and applying hierarchical clustering to define 4 molecular subtypes. Wu et al[110] proposed a variant of similarity network fusion, called ne-SNF (network enhancement similarity network fusion), to integrate mRNA expression, miRNA expression, and DNA methylation, thereby identifying multiomic subtypes. The clusters, found via spectral clustering, were compared with the clusters obtained with the same procedure applied to single-omics data sets. In most cases, ne-SNF–integrated subtypes yielded the most significant differences in survival profiles between different clusters, supporting the conclusion that multiomics data integration is crucial for identifying comprehensive subtypes with clinical relevance. Multiomics assessments of large CVD cohorts, however, are still rare, and the sample sizes are typically modest owing to the costs required to perform them. Further studies are needed to delineate more precise and robust molecular subtypes and identify clear mechanistic insights on HFpEF pathogenesis, but this early works appear quite promising as a path toward detecting meaningful signals of difference among well-defined cohorts.

Coronary Artery Disease

The risk of CAD is affected by genetic[124,125] and nongenetic[126,127] factors. The clinical presentation of patients with CAD manifests as well-recognized phenotypic heterogeneity, which has stimulated the development of a variety of clinical subtyping approaches. Flores et al[104] identified clinical subtypes of CAD within the GenePAD study cohort, which includes phenotypic and genetic biomarkers of individuals diagnosed with the disease. To handle clinical features of different types (qualitative, categorical, and ordinal), IRs were evaluated via generalized low-rank modeling—a generalization of PC analysis capable of modeling heterogeneous data types.[128] By using internal validation measures of cluster separability, compactness, and stability, they identified 4 phenotypically distinct subtypes of CAD whose differences were not detectable through conventional risk assessment. To relate CAD variability with essential hypertension and blood pressure (BP) patterns, Guo et al[113] assessed a temporal series of BP data from ambulatory BP monitoring devices in hypertensive patients with and without CAD. Among the identified subtypes, they found that hypertensive patients with nocturnal systolic BP rise register the highest prevalence of CAD, indicating that short-term temporal variation of BP—an often-overlooked feature in subtyping studies—may carry significant information on CAD risk independent of the mere presence of documented hypertension.

From the molecular perspective, Peng et al[111] integrated multiple gene expression data sets for molecular CAD subtyping at the transcriptomic level. The 3 subtypes found through consensus K-means clustering were characterized by age-independent differences of CAD extent (measured by the Duke prognostic CAD index[129]), indicating that transcriptomic subtyping may reveal different mechanistic determinants across CAD cases. Ding et al[112] restricted their focus to the potential relationship between CAD development and ferroptosis—a recently discovered biological process of iron-dependent (redox mediated) cell death that has a crucial role in CVD pathobiology.[130] Under the hypothesis that differential activation of ferroptosis pathways may delineate different CAD endotypes, the authors aggregated external sources to build a list of ferroptosis-related genes and construct the IRs based on their mRNA expression levels. They then applied nonnegative matrix factorization and found 2 molecular subtypes with significant differences in their associated Duke CAD index and age, suggesting that ferroptosis-related expression may be a potential biomarker of clinical relevance. It, however, remains a matter for future investigation as to how the ferroptosis-specific subtypes relate to the systemic endotypes and clinical manifestations of CAD. Overall, unsupervised clustering techniques have highlighted significant heterogeneity in CAD features, both at the phenotypic and the mechanistic levels. However, further studies are needed to assess agreement among current classifications and to determine the mechanistic relationships connecting molecular and clinical subtypes.

OUTLOOK

Many of the conceptual and technical challenges of disease subtyping arise from the lack of a clear and broadly accepted ground truth to serve as a reference point. Therefore, advances in computational disease subtyping approaches are contingent on the ever-evolving definition of what is a useful subtype classification. Several studies have discussed what constitutes a well-defined subtype, with most proposals calling for disease classes that are uniform in terms of mechanistic processes, prognosis, and treatability.[27,131–133] Nonetheless, current subtype discovery approaches can satisfy only a subset of these criteria. While molecular subtypes exhibit a stronger association with the disease's underlying causative biology, clinical approaches prioritize prognosis and treatability, resulting in inconsistent classifications.[102] A successful convergence of subtype definitions will depend upon the generation of more abundant and precise data encompassing both the molecular and clinical facets of the disease. Single-cell sequencing technologies are revolutionizing CVD research by offering unprecedented insights into the cellular diversity and molecular mechanisms that underlie disease progression.[134] For example, recent studies have leveraged these new resources to profile the transcriptomic profiles of adult human cardiomyocytes[135] and investigate their role in HF.[136] Furthermore, innovative spatial transcriptomics techniques have opened up new, promising avenues for understanding the cellular microenvironments and intercellular interactions forming in healthy and diseased hearts.[135,137] Concurrently, large-scale coordination of multiple hospitals and clinics will be necessary to combine EHR-based clinical data with shared nomenclature, formats, and protocols.[138]

From the computational standpoint, more powerful approaches will be required to leverage this increasingly vast deluge of data. Nevertheless, current machine learning and statistical models frequently lack interpretability owing to their high complexity—an issue that could be exacerbated as data size, dimensionality, and heterogeneity continue to grow. As in clinical settings, the explicability of model prediction is as important as its accuracy; it will be critical to push for the development of interpretable machine learning models and diagnostic criteria for characterizing their output.[139] In this vein, popular measures such as the Shapley Additive Explanations have been adapted to the subtyping context to characterize better the generated patient clusters[140] or even to discover new clusters.[141]

Alongside technical advancements, implementing these models into clinical practice poses several challenges. First, significant issues arise when the patient population considered for training the model is not an accurate representation of the target population. For example, mismatches in the population country or health system require careful model recalibration.[142,143] Second, many computational subtype classifications require the measurement of a multitude of patient features that are seldom available in clinical settings. Therefore, the implementation of parsimonious procedures that can be executed with reduced feature sets and that can be integrated into existing clinical workflows will require close collaboration among clinicians, data scientists, and information technology professionals. Third, there are regulatory and ethical considerations for the use of clinical models in clinical practice, especially in regard to the presence of biases and fairness issues in the model.[144]

In summary, the advancement of precision medicine relies on the synergistic interactions among data science, biomedical research, and clinical practice. By harnessing the power of computational approaches in a more targeted and patient-centric manner, we can ultimately enhance the diagnosis, prognosis, and treatment of CVD, paving the way for a new era in personalized health care.

ARTICLE INFORMATION

Received December 12, 2022; accepted May 10, 2023.

Affiliations

Channing Division of Network Medicine (E.M., J.L.) and Division of Cardiovascular Medicine, Department of Medicine (J.L.), Brigham and Women's Hospital, Harvard Medical School, Boston, MA.

Acknowledgments

The authors wish to thank Stephanie C. Tribuna for expert secretarial assistance.

Sources of Funding

This work was supported, in part, by the National Institutes of Health (NIH) grants HL155107, HL155096, HG007692, and HL119145; by American Heart Association grants D700382 and AHA957729; and by European Union grant 101057619 to J. Loscalzo and by NIH grant K01HL166705 to E. Maiorino.

Disclosures

None.

REFERENCES

1. Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, Abbastabar H, Abd-Allah F, Abdela J, Abdelalim A. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2018;392:1736–1788. doi: 10.1016/S0140-6736(18)32203-7
2. Luengo-Fernandez R, Leal J, Gray A, Petersen S, Rayner M. Cost of cardiovascular diseases in the United Kingdom. *Heart*. 2006;92:1384–1389. doi: 10.1136/hrt.2005.072173
3. Hwang TJ, Lauffenburger JC, Franklin JM, Kesselheim AS. Temporal trends and factors associated with cardiovascular drug development, 1990 to 2012. *JACC Basic Transl Sci*. 2016;1:301–308. doi: 10.1016/j.jacbts.2016.03.012
4. Mukherjee D, Topol EJ. Pharmacogenomics in cardiovascular diseases. *Prog Cardiovasc Dis*. 2002;44:479–498. doi: 10.1053/pcad.2002.123467
5. Investigators EPIC. Use of a monoclonal antibody directed against the platelet glycoprotein IIb/IIIa receptor in high-risk coronary angioplasty. *N Engl J Med*. 1994;330:956–961. doi: 10.1056/NEJM199404073301402
6. Clopidogrel in Unstable Angina to Prevent Recurrent Events Trial Investigators. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. *N Engl J Med*. 2001;345:494–502. doi: 10.1056/NEJMoa010746
7. Ziaeian B, Fonarow GC. Epidemiology and aetiology of heart failure. *Nat Rev Cardiol*. 2016;13:368–378. doi: 10.1038/nrcardio.2016.25
8. MacRae CA, Roden DM, Loscalzo J. The future of cardiovascular therapeutics. *Circulation*. 2016;133:2610–2617. doi: 10.1161/CIRCULATIONAHA.116.023555
9. Dai X, Wiernek S, Evans JP, Runge MS. Genetics of coronary artery disease and myocardial infarction. *World J Cardiol*. 2016;8:1–23. doi: 10.4330/wjc.v8.i1.1
10. Koyama S, Ito K, Terao C, Akiyama M, Horikoshi M, Momozawa Y, Matsunaga H, Ieki H, Ozaki K, Onouchi Y, et al. Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic

risk loci for coronary artery disease. *Nat Genet*. 2020;52:1169–1177. doi: 10.1038/s41588-020-0705-3

11. Arvanitis M, Tampakakis E, Zhang Y, Wang W, Auton A, Dutta D, Glavaris S, Keramati A, Chatterjee N, Chi NC. Genome-wide association and multi-omic analyses reveal ACTN2 as a gene linked to heart failure. *Nat Commun*. 2020;11:1–12. doi: 10.1038/s41467-020-14843-7
12. Dehghan A, Bis JC, White CC, Smith AV, Morrison AC, Cupples LA, Trompet S, Chasman DI, Lumley T, Volker U, et al. Genome-wide association study for incident myocardial infarction and coronary heart disease in prospective cohort studies: the CHARGE consortium. *PLoS One*. 2016;11:e0144997. doi: 10.1371/journal.pone.0144997
13. Dichgans M, Pulit SL, Rosand J. Stroke genetics: discovery, biology, and clinical applications. *Lancet Neurol*. 2019;18:587–599. doi: 10.1016/S1474-4422(19)30043-2
14. Leopold JA, Loscalzo J. Emerging role of precision medicine in cardiovascular disease. *Circ Res*. 2018;122:1302–1315. doi: 10.1161/CIRCRESAHA.117.310782
15. Dean L, Kane M. Clopidogrel therapy and CYP2C19 genotype. March 8, 2012 [updated December 1, 2022]. In: Pratt VM, Scott SA, Pirmohamed M, et al, eds. *Medical Genetics Summaries* [Internet]. Bethesda, MD: National Center for Biotechnology Information (US); 2012. https://www.ncbi.nlm.nih.gov/books/NBK84114/
16. Malakar AK, Choudhury D, Halder B, Paul P, Uddin A, Chakraborty S. A review on coronary artery disease, its risk factors, and therapeutics. *J Cell Physiol*. 2019;234:16812–16823. doi: 10.1002/jcp.28350
17. Antman EM, Loscalzo J. Precision medicine in cardiology. *Nat Rev Cardiol*. 2016;13:591–602. doi: 10.1038/nrcardio.2016.101
18. Saria S, Goldenberg A. Subtyping: what it is and its role in precision medicine. *IEEE Intelligent Systems*. 2015;30:70–75. doi: 10.1109/MIS.2015.60
19. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet*. 2011;12:56–68. doi: 10.1038/nrg2918
20. Barabási A-L. *Network Medicine—From Obesity to the "Diseasome"*. In: Mass Medical Soc; 2007:404–407.
21. Chan SY, Loscalzo J. The emerging paradigm of network medicine in the study of human disease. *Circ Res*. 2012;111:359–374. doi: 10.1161/CIRCRESAHA.111.258541
22. Sonawane AR, Aikawa E, Aikawa M. Connections for matters of the heart: network medicine in cardiovascular diseases. *Front Cardiovasc Med*. 2022;9:873582. doi: 10.3389/fcvm.2022.873582
23. Sun P, Wu Y, Yin C, Jiang H, Xu Y, Sun H. Molecular subtyping of cancer based on distinguishing co-expression modules and machine learning. *Front Genet*. 2022;13:866005. doi: 10.3389/fgene.2022.866005
24. Maron BA, Wang RS, Shevtsov S, Drakos SG, Arons E, Wever-Pinzon O, Huggins GS, Samokhin AO, Oldham WM, Aguib Y, et al. Individualized interactomes for network-based precision medicine in hypertrophic cardiomyopathy with implications for other clinical pathophenotypes. *Nat Commun*. 2021;12:873. doi: 10.1038/s41467-021-21146-y
25. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11:333–337. doi: 10.1038/nmeth.2810
26. Chen DP, Weber SC, Constantinou PS, Ferris TA, Lowe HJ, Butte AJ. Clinical arrays of laboratory measures, or "clinarrays," built from an electronic health record enable disease subtyping by severity. Paper/Poster presented at: AMIA Annual Symposium Proceedings; 2007.
27. Dahl A, Zaitlen N. Genetic influences on disease subtypes. *Annu Rev Genomics Hum Genet*. 2020;21:413–435. doi: 10.1146/annurev-genom-120319-095026
28. Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat*. 2012;33:777–780. doi: 10.1002/humu.22080
29. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet*. 2014;383:999–1008. doi: 10.1016/S0140-6736(13)61752-3
30. Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet*. 2011;12:417–428. doi: 10.1038/nrg2999
31. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. *Nat Rev Genet*. 2010;11:855–866. doi: 10.1038/nrg2897
32. Kapur S, MacRae CA. Deep phenotyping in cardiovascular disease. *Curr Treatment Options Cardiovasc Med*. 2021;23:1–9. doi: 10.1007/s11936-020-00881-3
33. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med*. 2019;25:70–74. doi: 10.1038/s41591-018-0240-2
34. Mannil M, Eberhard M, von Spiczak J, Heindel W, Alkadhi H, Baessler B. Artificial intelligence and texture analysis in cardiac imaging. *Curr Cardiol Rep*. 2020;22:131. doi: 10.1007/s11886-020-01402-1
35. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, Rosand B, Li Y, Zhang M, Chang D, et al. Neural natural language processing for unstructured data in electronic health records: a review. *Computer Sci Rev*. 2022;46:100511. doi: 10.1016/j.cosrev.2022.100511
36. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. 2013;20:117–121. doi: 10.1136/amiajnl-2012-001145
37. Bower JK, Patel S, Rudy JE, Felix AS. Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise. *Curr Epidemiol Rep*. 2017;4:346–352. doi: 10.1007/s40471-017-0130-z
38. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, Bonow RO, Huang CC, Deo RC. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*. 2015;131:269–279. doi: 10.1161/CIRCULATIONAHA.114.010637
39. Ding L, Mane R, Wu Z, Jiang Y, Meng X, Jing J, Ou W, Wang X, Liu Y, Lin J, et al. Data-driven clustering approach to identify novel phenotypes using multiple biomarkers in acute ischaemic stroke: a retrospective, multicentre cohort study. *EClinicalMedicine*. 2022;53:101639. doi: 10.1016/j.eclinm.2022.101639
40. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2018;25:1419–1428. doi: 10.1093/jamia/ocy068
41. Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. *J Am Med Inform Assoc*. 2018;25:289–294. doi: 10.1093/jamia/ocx110
42. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. *J Biomed Inform*. 2014;51:24–34. doi: 10.1016/j.jbi.2014.03.016
43. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMS (Wash DC)*. 2013;1:1035. doi: 10.13063/2327-9214.1035
44. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380:1347–1358. doi: 10.1056/nejmra1814259
45. Gill EE, Smith ML, Gibson KM, Morishita KA, Lee AHY, Falsafi R, Graham J, Foell D, Benseler SM, Ross CJ, et al; PedVas Initiative Investigators. Different disease endotypes in phenotypically similar vasculitides affecting small-to-medium sized blood vessels. *Front Immunol*. 2021;12:638571. doi: 10.3389/fimmu.2021.638571
46. Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet*. 2018;19:299–310. doi: 10.1038/nrg.2018.4
47. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. 2017;18:83. doi: 10.1186/s13059-017-1215-1
48. Bleesing JJ, Fleisher TA. Immunophenotyping. Paper/Poster Presented at: Seminars in Hematology; 2001.
49. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. 2015;16:85–97. doi: 10.1038/nrg3868
50. Wu C, Zhou F, Ren J, Li X, Jiang Y, Ma S. A selective review of multi-level omics data integration using variable selection. *High Throughput*. 2019;8:4. doi: 10.3390/ht8010004
51. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinf*. 2016;17:15. doi: 10.1186/s12859-015-0857-9
52. Aleta A, Moreno Y. Multilayer networks in a nutshell. *Annu Rev Condens Matter Phys*. 2019;10:45–62. doi: 10.1146/annurev-conmatphys-031218-013259
53. Liu X, Maiorino E, Halu A, Glass K, Prasad RB, Loscalzo J, Gao J, Sharma A. Robustness and lethality in multilayer biological molecular networks. *Nat Commun*. 2020;11:6043. doi: 10.1038/s41467-020-19841-3
54. Joshi A, Rienks M, Theofilatos K, Mayr M. Systems biology in cardiovascular disease: a multiomics approach. *Nat Rev Cardiol*. 2021;18:313–330. doi: 10.1038/s41569-020-00477-1
55. Duan R, Gao L, Gao Y, Hu Y, Xu H, Huang M, Song K, Wang H, Dong Y, Jiang C, et al. Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLoS Comput Biol*. 2021;17:e1009224. doi: 10.1371/journal.pcbi.1009224
56. Tarazona S, Arzalluz-Luque A, Conesa A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nat Comput Sci*. 2021;1:395–402. doi: 10.1038/s43588-021-00086-z

57. Lopez de Maturana E, Alonso L, Alarcon P, Martin-Antoniano IA, Pineda S, Piorno L, Calle ML, Malats N. Challenges in the integration of omics and non-omics data. *Genes (Basel)*. 2019;10:238. doi: 10.3390/genes10030238
58. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. *Brief Bioinform*. 2022;23:bbab454. doi: 10.1093/bib/bbab454
59. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform*. 2018;19:1370–1381. doi: 10.1093/bib/bbx066
60. Teschendorff AE. Avoiding common pitfalls in machine learning omic data science. *Nat Mater*. 2019;18:422–427. doi: 10.1038/s41563-018-0241-z
61. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: a data perspective. *ACM Comput Surveys (CSUR)*. 2017;50:1–45. doi: 10.1145/3136625
62. Solorio-Fernández S, Carrasco-Ochoa JA, Martínez-Trinidad JF. A review of unsupervised feature selection methods. *Artif Intell Rev*. 2020;53:907–948. doi: 10.1007/s10462-019-09682-y
63. Cunningham JP, Ghahramani Z. Linear dimensionality reduction: survey, insights, and generalizations. *J Machine Learn Res*. 2015;16:2859–2900.
64. Espadoto M, Martins RM, Kerren A, Hirata NS, Telea AC. Toward a quantitative survey of dimension reduction techniques. *IEEE Trans Vis Comput Graph*. 2019;27:2153–2173. doi: 10.1109/TVCG.2019.2944182
65. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, Baudot A. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat Commun*. 2021;12:124. doi: 10.1038/s41467-020-20430-7
66. Vahabi N, Michailidis G. Unsupervised multi-omics data integration methods: a comprehensive review. *Front Genet*. 2022;13:854752. doi: 10.3389/fgene.2022.854752
67. Hennig C, Meila M. Cluster analysis: an overview. *Handbook of cluster analysis*. 2015:1–20. doi: 10.1201/b19706
68. Palou-Marquez G, Subirana I, Nonell L, Fernandez-Sanles A, Elosua R. DNA methylation and gene expression integration in cardiovascular disease. *Clin Epigenetics*. 2021;13:75. doi: 10.1186/s13148-021-01064-y
69. Seo M, Oh S. CBFS: high performance feature selection algorithm based on feature clearness. *PLoS One*. 2012;7:e40419. doi: 10.1371/journal.pone.0040419
70. Wosiak A, Zakrzewska D. Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis. *Complexity*. 2018. doi: 10.1155/2018/2520706
71. Gao W, Hu L, Zhang P. Class-specific mutual information variation for feature selection. *Pattern Recognit*. 2018;79:328–339. doi: 10.1016/j.patcog.2018.02.020
72. Song L, Smola A, Gretton A, Borgwardt KM, Bedo J. Supervised feature selection via dependence estimation. Paper/Poster Presented at: Proceedings of the 24th International Conference on Machine Learning; 2007.
73. Cai D, Zhang C, He X. Unsupervised feature selection for multi-cluster data. Paper/Poster Presented at: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2010.
74. Li Z, Yang Y, Liu J, Zhou X, Lu H. Unsupervised feature selection using nonnegative spectral analysis. Paper/Poster Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; 2012.
75. Schuler A, Liu V, Wan J, Callahan A, Udell M, Stark DE, Shah NH. Discovering patient phenotypes using generalized low rank models. Paper/Poster Presented at: Biocomputing 2016: Proceedings of the Pacific Symposium; 2016.
76. Schölkopf B, Smola A, Müller KR. Kernel principal component analysis. Paper/Poster Presented at: International Conference on Artificial Neural Networks; 1997.
77. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579–2605.
78. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2019;37:38–44. doi: 10.1038/nbt.4314
79. Kohonen T. *Self-Organizing Maps*. Springer Science & Business Media; 2012.
80. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016.
81. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*. 2018;14:e8124. doi: 10.15252/msb.20178124
82. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Ann Appl Stat*. 2013;7:523–542. doi: 10.1214/12-AOAS597
83. Teschendorff AE, Jing H, Paul DS, Virta J, Nordhausen K. Tensorial blind source separation for improved analysis of multi-omic data. *Genome Biol*. 2018;19:76. doi: 10.1186/s13059-018-1455-8
84. Chalise P, Fridley BL. Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. *PLoS One*. 2017;12:e0176278. doi: 10.1371/journal.pone.0176278
85. Tenenhaus A, Philippe C, Guillemot V, Le Cao KA, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. *Biostatistics*. 2014;15:569–583. doi: 10.1093/biostatistics/kxu001
86. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, Ladanyi M, Sander C. Integrative subtype discovery in glioblastoma using iCluster. *PLoS One*. 2012;7:e35236. doi: 10.1371/journal.pone.0035236
87. Ben Brahim A, Limam M. Ensemble feature selection for high dimensional data: a new method and a comparative study. *Adv Data Anal Classification*. 2018;12:937–952. doi: 10.1007/s11634-017-0285-y
88. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. Paper/Poster Presented at: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II 19; 2008.
89. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *J Clin Epidemiol*. 2016;71:76–85. doi: 10.1016/j.jclinepi.2015.10.002
90. Shilaskar S, Ghatol A. Feature selection for medical diagnosis: evaluation for cardiovascular diseases. *Expert Syst Appl*. 2013;40:4146–4153. doi: 10.1016/j.eswa.2013.01.032
91. Ang JC, Mirzal A, Haron H, Hamed HN. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Trans Comput Biol Bioinform*. 2016;13:971–989. doi: 10.1109/TCBB.2015.2478454
92. Lualdi M, Fasano M. Statistical analysis of proteomics data: a review on feature selection. *J Proteomics*. 2019;198:18–26. doi: 10.1016/j.jprot.2018.12.004
93. Antonelli J, Claggett BL, Henglin M, Kim A, Ovsak G, Kim N, Deng K, Rao K, Tyagi O, Watrous JD, et al. Statistical workflow for feature selection in human metabolomics data. *Metabolites*. 2019;9:143. doi: 10.3390/metabo9070143
94. Siwek K, Osowski S. Autoencoder versus PCA in face recognition. Paper/Poster Presented at: 2017 18th International Conference on Computational Problems of Electrical Engineering (CPEE); 2017.
95. Picard M, Scott-Boyer MP, Bodein A, Perin O, Droit A. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J*. 2021;19:3735–3746. doi: 10.1016/j.csbj.2021.06.030
96. Nakazawa MA, Tamada Y, Tanaka Y, Ikeguchi M, Higashihara K, Okuno Y. Novel cancer subtyping method based on patient-specific gene regulatory network. *Sci Rep*. 2021;11:23653. doi: 10.1038/s41598-021-02394-w
97. Adam SP, Alexandropoulos S-AN, Pardalos PM, Vrahatis MN. No free lunch theorem: a review. *Approximation and Optimization*. 2019:57–82. doi: 10.1007/978-3-030-12767-1_5
98. Liu Y, Li Z, Xiong H, Gao X, Wu J. Understanding of internal clustering validation measures. Paper/Poster Presented at: 2010 IEEE International Conference on Data Mining; 2010.
99. Von Luxburg U. Clustering stability: an overview. *Foundations and Trends in Machine Learning*. 2010;2:235–274. doi: 10.1561/2200000008
100. Chen SS, Gopalakrishnan PS. Clustering via the Bayesian information criterion with applications in speech recognition. Paper/Poster Presented at: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181); 1998.
101. Ferraro MB, Giordani P. Soft clustering. *Wiley Interdiscip Rev Comput Stat*. 2020;12:e1480. doi: 10.1002/wics.1480
102. Khera AV, Kathiresan S. Is coronary atherosclerosis one disease or many? Setting realistic expectations for precision medicine. *Circulation*. 2017;135:1005–1007. doi: 10.1161/CIRCULATIONAHA.116.026479
103. Cohen JB, Schrauben SJ, Zhao L, Basso MD, Cvijic ME, Li Z, Yarde M, Wang Z, Bhattacharya PT, Chirinos DA, et al. Clinical phenogroups in heart failure with preserved ejection fraction: detailed phenotypes, prognosis, and response to spironolactone. *JACC Heart Fail*. 2020;8:172–184. doi: 10.1016/j.jchf.2019.09.009
104. Flores AM, Schuler A, Eberhard AV, Olin JW, Cooke JP, Leeper NJ, Shah NH, Ross EG. Unsupervised learning for automated detection of coronary artery disease subgroups. *J Am Heart Assoc*. 2021;10:e021976. doi: 10.1161/JAHA.121.021976

105. Zhang JM, Kamath GM, Tse DN. Valid post-clustering differential analysis for single-cell RNA-Seq. *Cell Syst.* 2019;9:383–392.e6. doi: 10.1016/j.cels.2019.07.012
106. Kao DP, Lewsey JD, Anand IS, Massie BM, Zile MR, Carson PE, McKelvie RS, Komajda M, McMurray JJ, Lindenfeld J. Characterization of subgroups of heart failure patients with preserved ejection fraction with possible implications for prognosis and treatment response. *Eur J Heart Fail.* 2015;17:925–935. doi: 10.1002/ejhf.327
107. Segar MW, Patel KV, Ayers C, Basit M, Tang WHW, Willett D, Berry J, Grodin JL, Pandey A. Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis. *Eur J Heart Fail.* 2020;22:148–158. doi: 10.1002/ejhf.1621
108. Hedman AK, Hage C, Sharma A, Brosnan MJ, Buckbinder L, Gan LM, Shah SJ, Linde CM, Donal E, Daubert JC, et al. Identification of novel pheno-groups in heart failure with preserved ejection fraction using machine learning. *Heart.* 2020;106:342–349. doi: 10.1136/heartjnl-2019-315481
109. Woolley RJ, Ceelen D, Ouwerkerk W, Tromp J, Figarska SM, Anker SD, Dickstein K, Filippatos G, Zannad F, Metra M, et al. Machine learning based on biomarker profiles identifies distinct subgroups of heart failure with preserved ejection fraction. *Eur J Heart Fail.* 2021;23:983–991. doi: 10.1002/ejhf.2144
110. Wu Y, Wang H, Li Z, Cheng J, Fang R, Cao H, Cui Y. Subtypes identification on heart failure with preserved ejection fraction via network enhancement fusion using multi-omics data. *Comput Struct Biotechnol J.* 2021;19:1567–1578. doi: 10.1016/j.csbj.2021.03.010
111. Peng XY, Wang Y, Hu H, Zhang XJ, Li Q. Identification of the molecular subgroups in coronary artery disease by gene expression profiles. *J Cell Physiol.* 2019;234:16540–16548. doi: 10.1002/jcp.28324
112. Ding L, Long F, An D, Liu J, Zhang G. Construction and validation of molecular subtypes of coronary artery disease based on ferroptosis-related genes. *BMC Cardiovasc Disord.* 2022;22:283. doi: 10.1186/s12872-022-02719-1
113. Guo Q, Lu X, Gao Y, Zhang J, Yan B, Su D, Song A, Zhao X, Wang G. Cluster analysis: a new approach for identification of underlying risk factors for coronary artery disease in essential hypertensive patients. *Sci Rep.* 2017;7:43965. doi: 10.1038/srep43965
114. Cho JS, Shrestha S, Kagiyama N, Hu L, Ghaffar YA, Casaclang-Verzosa G, Zeb I, Sengupta PP. A network-based “phenomics” approach for discovering patient subtypes from high-throughput cardiac imaging data. *JACC Cardiovasc Imaging.* 2020;13:1655–1670. doi: 10.1016/j.jcmg.2020.02.008
115. Verdonschot JAJ, Merlo M, Dominguez F, Wang P, Henkens M, Adriaens ME, Hazebroek MR, Mase M, Escobar LE, Cobas-Paz R, et al. Phenotypic clustering of dilated cardiomyopathy patients highlights important pathophysiological differences. *Eur Heart J.* 2021;42:162–174. doi: 10.1093/eurheartj/ehaa841
116. Tromp J, Tay WT, Ouwerkerk W, Teng T-HK, Yap J, MacDonald MR, Leineweber K, McMurray JJ, Zile MR, Anand IS. Multimorbidity in patients with heart failure from 11 Asian regions: a prospective cohort study using the ASIAN-HF registry. *PLoS Med.* 2018;15:e1002541. doi: 10.1371/journal.pmed.1002583
117. Borlaug BA, Paulus WJ. Heart failure with preserved ejection fraction: pathophysiology, diagnosis, and treatment. *Eur Heart J.* 2011;32:670–679. doi: 10.1093/eurheartj/ehq426
118. Massie BM, Carson PE, McMurray JJ, Komajda M, McKelvie R, Zile MR, Anderson S, Donovan M, Iverson E, Staiger C, et al; I-PRESERVE Investigators. Irbesartan in patients with heart failure and preserved ejection fraction. *N Engl J Med.* 2008;359:2456–2467. doi: 10.1056/NEJMoa0805450
119. Yusuf S, Pfeffer MA, Swedberg K, Granger CB, Held P, McMurray JJ, Michelson EL, Olofsson B, Ostergren J; CHARM Investigators and Committees. Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-Preserved trial. *Lancet.* 2003;362:777–781. doi: 10.1016/S0140-6736(03)14285-7
120. Solomon SD, McMurray JJV, Claggett B, de Boer RA, DeMets D, Hernandez AF, Inzucchi SE, Kosiborod MN, Lam CSP, Martinez F, et al; DELIVER Trial Committees and Investigators. Dapagliflozin in heart failure with mildly reduced or preserved ejection fraction. *N Engl J Med.* 2022;387:1089–1098. doi: 10.1056/NEJMoa2206286
121. Adamczak DM, Oduah MT, Kiebalo T, Nartowicz S, Beben M, Pochylski M, Ciepluch A, Gwizdala A, Lesiak M, Straburzynska-Migaj E. Heart failure with preserved ejection fraction-a concise review. *Curr Cardiol Rep.* 2020;22:82. doi: 10.1007/s11886-020-01349-3
122. Fraley C, Raftery A, Wehrens R. Incremental model-based clustering for large datasets with small clusters. *J Comput Graphical Stat.* 2005;14:529–546. doi: 10.1198/106186005x59603
123. Marbac M, Biernacki C, Vandewalle V. Model-based clustering of Gaussian copulas for mixed data. *Commun Stat - Theory Methods.* 2017;46:11635–11656. doi: 10.1080/03610926.2016.1277753
124. McPherson R, Tybjaerg-Hansen A. Genetics of coronary artery disease. *Circ Res.* 2016;118:564–578. doi: 10.1161/CIRCRESAHA.115.306566
125. Ozaki K, Tanaka T. Molecular genetics of coronary artery disease. *J Hum Genet.* 2016;61:71–77. doi: 10.1038/jhg.2015.70
126. Wilson PW. Established risk factors and coronary artery disease: the Framingham Study. *Am J Hypertens.* 1994;7:7S–12S. doi: 10.1093/ajh/7.7.7s
127. Mack M, Gopal A. Epidemiology, traditional and novel risk factors in coronary artery disease. *Heart Fail Clin.* 2016;12:1–10. doi: 10.1016/j.hfc.2015.08.002
128. Boyd S, Zadeh R, Horn C, Udell M. Generalized low rank models. *Foundations and Trends® in Machine Learning.* 2016;9:1–118. doi: 10.1561/2200000055
129. Felker GM, Shaw LK, O’Connor CM. A standardized definition of ischemic cardiomyopathy for use in clinical research. *J Am Coll Cardiol.* 2002;39:210–218. doi: 10.1016/s0735-1097(01)01738-7
130. Chen Z, Yan Y, Qi C, Liu J, Li L, Wang J. The role of ferroptosis in cardiovascular disease and its therapeutic significance. *Front Cardiovasc Med.* 2021;8:733229. doi: 10.3389/fcvm.2021.733229
131. Agusti A, Bel E, Thomas M, Vogelmeier C, Brusselle G, Holgate S, Humbert M, Jones P, Gibson PG, Vestbo J, et al. Treatable traits: toward precision medicine of chronic airway diseases. *Eur Respir J.* 2016;47:410–419. doi: 10.1183/13993003.01359-2015
132. Boland MR, Hripcsak G, Shen Y, Chung WK, Weng C. Defining a comprehensive verotype using electronic health records for personalized medicine. *J Am Med Inform Assoc.* 2013;20:e232–e238. doi: 10.1136/amiajnl-2013-001932
133. Castaldi PJ, Boueiz A, Yun J, Estepar RSJ, Ross JC, Washko G, Cho MH, Hersh CP, Kinney GL, Young KA, et al; COPDGene Investigators. Machine learning characterization of COPD subtypes: insights from the COPDGene study. *Chest.* 2020;157:1147–1157. doi: 10.1016/j.chest.2019.11.039
134. Miranda AMA, Janbandhu V, Maatz H, Kanemaru K, Cranley J, Teichmann SA, Hubner N, Schneider MD, Harvey RP, Noseda M. Single-cell transcriptomics for the assessment of cardiac disease. *Nat Rev Cardiol.* 2023;20:289–308. doi: 10.1038/s41569-022-00805-7
135. Litvinukova M, Talavera-Lopez C, Maatz H, Reichart D, Worth CL, Lindberg EL, Kanda M, Polanski K, Heinig M, Lee M, et al. Cells of the adult human heart. *Nature.* 2020;588:466–472. doi: 10.1038/s41586-020-2797-4
136. Koenig AL, Shchukina I, Amrute J, Andhey PS, Zaitsev K, Lai L, Bajpai G, Bredemeyer A, Smith G, Jones C, et al. Single-cell transcriptomics reveals cell-type-specific diversification in human heart failure. *Nat Cardiovasc Res.* 2022;1:263–280. doi: 10.1038/s44161-022-00028-6
137. Kuppe C, Ramirez Flores RO, Li Z, Hayat S, Levinson RT, Liao X, Hannani MT, Tanevski J, Wunnemann F, Nagai JS, et al. Spatial multi-omic map of human myocardial infarction. *Nature.* 2022;608:766–777. doi: 10.1038/s41586-022-05060-x
138. Mandl KD, Kohane IS. Federalist principles for healthcare data networks. *Nat Biotechnol.* 2015;33:360–363. doi: 10.1038/nbt.3180
139. Reddy S. Explainability and artificial intelligence in medicine. *Lancet Digit Health.* 2022;4:e214–e215. doi: 10.1016/S2589-7500(22)00029-2
140. Su C, Hou Y, Xu J, Brendel M, Zhu Y, Henchcliffe C, Cheng F, Wang F. Comprehensively modeling heterogeneous symptom progression for Parkinson’s disease subtyping. *medRxiv.* 2022. doi: 10.1101-2021.07.18.21260732. PPR: 372957
141. Schulz MA, Chapman-Rounds M, Verma M, Bzdok D, Georgatzis K. Inferring disease subtypes from clusters in explanation space. *Sci Rep.* 2020;10:12900. doi: 10.1038/s41598-020-68858-7
142. Damen JA, Pajouheshnia R, Heus P, Moons KGM, Reitsma JB, Scholten R, Hooft L, Debray TPA. Performance of the Framingham risk models and pooled cohort equations for predicting 10-year risk of cardiovascular disease: a systematic review and meta-analysis. *BMC Med.* 2019;17:109. doi: 10.1186/s12916-019-1340-7
143. Joseph P, Yusuf S, Lee SF, Ibrahim Q, Teo K, Rangarajan S, Gupta R, Rosengren A, Lear SA, Avezum A, et al; PURE Investigators. Prognostic validation of a non-laboratory and a laboratory based cardiovascular disease risk score in multiple regions of the world. *Heart.* 2018;104:581–587. doi: 10.1136/heartjnl-2017-311609
144. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, Zhao J, Snowdon JL. Precision medicine, AI, and the future of personalized health care. *Clin Transl Sci.* 2021;14:86–93. doi: 10.1111/cts.12884