Title

Cardiovascular informatics: building a bridge to data harmony.

Journal

Cardioscience, 118(3)

Authors

Caufield, John
Sigdel, Dibakar
Fu, John
et al.

Publication Date

2022-02-21

DOI

10.1093/cvr/cvab067

Peer reviewed

ESC European Society of Cardiology

Cardiovascular Research (2022) **118**, 732–745
doi:10.1093/cvr/cvab067

REVIEW

Cardiovascular informatics: building a bridge to data harmony

John Harry Caufield [1,2]*, Dibakar Sigdel [1,2], John Fu [1], Howard Choi[1], Vladimir Guevara-Gonzalez[1], Ding Wang[2], and Peipei Ping[1,2,3,4,5]

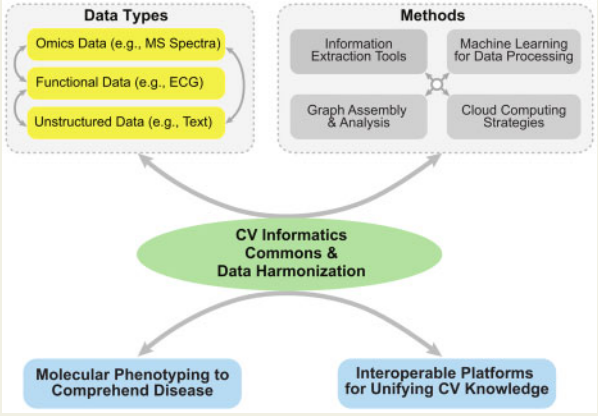
[1]NHLBI Integrated Cardiovascular Data Science Training Program at University of California, Los Angeles (UCLA), Suite 1-609, MRL Building, 675 Charles E. Young Dr. South, Los Angeles, CA 90095-1760, USA; [2]Department of Physiology, UCLA School of Medicine, Suite 1-609, MRL Building, 675 Charles E. Young Dr. South, Los Angeles, CA 90095-1760, USA; [3]Department of Medicine (Cardiology), UCLA School of Medicine, Suite 1-609, MRL Building, 675 Charles E. Young Dr. South, Los Angeles, CA 90095-1760, USA; [4]UCLA Bioinformatics Interdepartmental Program and Medical Informatics Home Area, Suite 1-609, MRL Building, 675 Charles E. Young Dr. South, Los Angeles, CA 90095-1760, USA and [5]Scalable Analytics Institute (ScAi), UCLA School of Engineering, Los Angeles, CA 90095, USA

Received 1 September 2020; editorial decision 28 February 2021; accepted 3 March 2021; online publish-ahead-of-print 5 March 2021

Abstract

The search for new strategies for better understanding cardiovascular (CV) disease is a constant one, spanning multitudinous types of observations and studies. A comprehensive characterization of each disease state and its biomolecular underpinnings relies upon insights gleaned from extensive information collection of various types of data. Researchers and clinicians in CV biomedicine repeatedly face questions regarding which types of data may best answer their questions, how to integrate information from multiple datasets of various types, and how to adapt emerging advances in machine learning and/or artificial intelligence to their needs in data processing. Frequently lauded as a field with great practical and translational potential, the interface between biomedical informatics and CV medicine is challenged with staggeringly massive datasets. Successful application of computational approaches to decode these complex and gigantic amounts of information becomes an essential step toward realizing the desired benefits. In this review, we examine recent efforts to adapt informatics strategies to CV biomedical research: automated information extraction and unification of multifaceted -omics data. We discuss how and why this interdisciplinary space of CV Informatics is particularly relevant to and supportive of current experimental and clinical research. We describe in detail how open data sources and methods can drive discovery while demanding few initial resources, an advantage afforded by widespread availability of cloud computing-driven platforms. Subsequently, we provide examples of how interoperable computational systems facilitate exploration of data from multiple sources, including both consistently formatted structured data and unstructured data. Taken together, these approaches for achieving data harmony enable molecular phenotyping of CV diseases and unification of CV knowledge.

Graphical Abstract



Keywords Informatics • Data science • Machine learning • Open data • Cloud computing

*Corresponding author. Tel: 1 (310) 267-5624, Email: jcaufield@mednet.ucla.edu.

1. Introduction

Answering biomedical questions through large-scale data analyses has been explored for much of the past century[1] and beyond. For cardiovascular (CV) disease research in particular, ever-growing sources of multi-level -omics data and observational data hold promise as new reservoirs of mechanistic evidence. Bridging the knowledge gap between genotype and phenotype has become a major challenge and a tantalizing goal. The high incidence and impact of CV disease, still the leading disease-related cause of mortality worldwide,[2] suggests more precise investigations and processing of data are necessary. Existing approaches in this area have been helpful but likely insufficient, and therefore, despite a mounting collection of information in the field (see *Figure 1*), connecting numerous observations and heterogeneous datasets remains imposing. Modern informatics research creates new avenues and new answers with the application of ever more powerful computational approaches.

Transformative progress driven by computation is now real and rapidly attainable. *In silico* methods can now transform previously intractable data into biomedically meaningful findings.[3] These advances are paired with efficient techniques for data acquisition and integration. A pressing need for cross-validated, generalizable, and standardized predictive models remains,[4] as does a need for a 'universal language' of clinical data interoperability.[5] In the meantime, new projects yield new tools, data, and infrastructure, collectively populating the CV informatics commons.

CV biology and medicine stand to massively benefit from recent computational advances. We may estimate the extent of research activity in this space by measuring funding allocated to machine learning (ML) and computational approaches through the NIH NHLBI. Using a previously defined search strategy,[6] we find that >$295 million in support has been allocated to 540 different projects since 2017, all involving some variety of ML methodology. In 2019, when NHLBI had an overall budget of roughly $3.5 billion,[7] funding for these ML projects comprised ~3%.

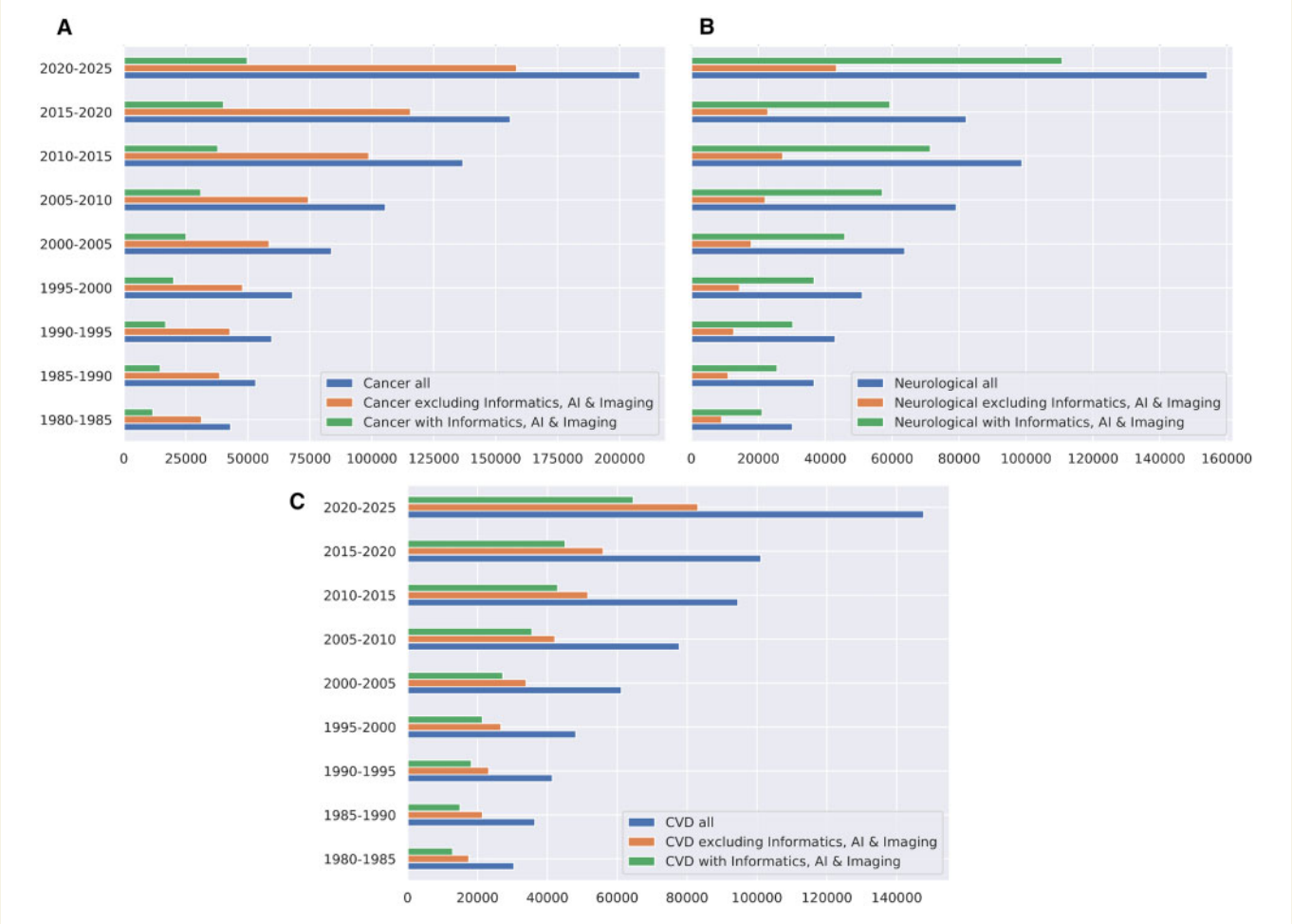


Figure 1 Published reports on major disease categories. Counts of citations on PubMed from 1980 to 2018, per 5-year period, for each of three disease-related categories, determined by Medical Subject Heading (MeSH) term assignment: (*A*) cancer (i.e. 'Neoplasms'), (*B*) neurological disease (i.e. 'Neurological Diseases'), and (*C*) cardiovascular diseases (CVD). Further categorization is determined by MeSH term assignment for each of the disease categories with and without the terms 'Informatics', 'Artificial Intelligence', or 'Diagnostic Imaging', including any of their child MeSH terms (e.g. 'Machine Learning' or 'Deep Learning'). Citation counts for 2020–25 are projections.

Though small in comparison to the full size of this funding source, just 700 projects incorporating ML received $334 million in NHLBI support for the entire period between 1985 and 2016. We can therefore observe that the past several years have witnessed adoption of ML in particular as a practical (and crucially, fundable) component of CV research.

In the following sections, we review recent advancements in the application of informatics and ML approaches to CV biology and medicine. As image analysis and computational modeling in CV contexts are covered comprehensively by other recent reviews,[8–11] we instead focus on strategies concerning particularly large or unwieldy data. Learning from large, unstructured data collections poses particular challenges: most research methods are designed to work with structured data of known formats and discrete values, yet many data sources contain unpredictable contents with variable systems of organization. Understanding and organizing unstructured data, particularly in CV biomedicine, requires extensive knowledge of the domain itself.[12] Similarly, -omics datasets vary in size and structure, demanding intuitive, domain-sensitive methods to compare them effectively.[13] The goal of CV informatics is therefore to create tools and resources for enhancing knowledge at scales far beyond any single researcher or clinician's capabilities while emulating the human capacity to integrate disparate pieces of knowledge.

2. Resources, methods, and philosophies supporting cardiovascular informatics

2.1 Why is cardiovascular informatics relevant now?

CV informatics is the specific application of computational and data-driven methods to studies of CV biology and disease. Informative gene and protein expression studies yield massive, high-dimensional data,[14] as do investigations into epigenetics,[15] requiring automated means for their management and analysis. Meanwhile, clinical researchers seek to make sense of vastly multifaceted observational data.[16] Each study poses unique challenges in how to manage, analyze, and validate results. Researchers face yet further challenges in sharing their data, allowing others to replicate their findings, and comparing observations against those from other domains (e.g. does a pathway of interest have known roles in cancer or neurological disease?). CV informatics focuses on solving analysis challenges imposed by the sheer size, complexity, or intractability of these data types. Its methods may also assist with identifying connections that seasoned investigators may otherwise overlook.

Even cursory analyses of biomedical data are increasingly commonplace in CV research. Out of the >3.3 million entries on PubMed concerning CV phenomena or diseases (as of December 2020), >3000 also involve applications of ML and >62 000 concern informatics approaches. The adoption of computational approaches in CV research (*Figure 1C*) is particularly striking over the past 5–10 years and is paralleled by growth in publications in cancer (*Figure 1A*) and nervous system disease studies (*Figure 1B*) (Shameer *et al.*[4] made a similar comparison for cardiology and ML in 2018, expecting near logarithmic growth). ML is a driving force in CV informatics, as evidenced by >300% growth in CV ML papers since 2015. Reports on informatics approaches are not limited to bioinformatics or medical informatics journals: in an analysis of just over 150 000 PubMed citations published since 2010 and concerning CV topics and informatics (including ML and diagnostic imaging), we find that no single journal contributes >2% of the total (~3000 citations). Each of the top 20 journals by publication count in this set have a CV focus (e.g. *JACC*) with the exception of *PLoS One*.

2.2 New -omics and multi-omics approaches are data rich

Innovations in comprehensive, high-throughput molecular biology have been a boon for research in CV informatics. Two specific groups of approaches have made particularly impactful contributions to this field in recent years: methods exploring previously underexplored biological scales and those integrating multiple data types. The first category includes epigenomic approaches such as broad chromatin accessibility studies and single-cell approaches (e.g. single-cell RNA-seq). Consideration of the chromatin landscape may challenge assumptions about the development of metabolic disease and heart failure[17] while RNA-seq of individual CV progenitor cells permits focused views into the development of biologically necessary structures.[18,19] Integrated, multi-omics approaches are also of particular interest to research as they may provide multiple sources of evidence for biomarker discovery, e.g. for heart failure,[20] arterial ischemic stroke,[21] or calcific aortic valve disease.[22] In practice, a reliable biomarker of CV disease may incorporate a full panel of signals from multiple high-throughput biomolecular, physiological, imaging, and even behavioral diagnostics, but integrating varied data meaningfully requires novel computational approaches.[23] New methods for producing large-scale data, whether from single cells or from multiple scales of multi-omic observations, require purpose-built analysis strategies to yield meaningful conclusions from their data.

Newly created -omics datasets span the various molecular domains of the body, joining genomes, with epigenomes, transcriptomes, proteomes, metabolomes, and beyond.[23] Most tools used for multidimensional data integration fall within one of the following five categories: clustering/dimensionality reduction-based methodologies, predictive modeling approaches, pairwise integration, network-based methodologies, and composite approaches.[23] The selection of a proper integration technique involves the consideration of data-driven statistical patterns and biological interpretability, but the relative consideration of these two aspects is dependent on the investigator's specific applications. In some cases, data may even contain spatial or temporal properties. We have found that a specific clustering method, deep convolutional embedded clustering (DCEC), is an effective way to join multi-omics data in a time series.[24] This method can cluster individual molecules from proteomes and metabolomes using the visual similarity of their temporal trends. As compared to other conventional clustering approaches, DCEC identifies more clusters relevant to Reactome pathways in a mammalian heart proteome, demonstrating its capacity as an effective approach for integrative analysis of temporal multi-omics data. Further exploration of this and other novel computational approaches will assist in the process of selecting appropriate integration techniques for CV omics data.

2.3 The advent of open data and AI democratization

The concepts of open source methods, open data, and biomedical AI have been enthusiastically adopted in biomedical data science research and data driven applications.[25–28] We define 'open source' as freely available, redistributable, and modifiable software and methods.[29] Data that are similarly 'open' are therefore highly compatible with open source as it is freely accessible and redistributable with few limitations.[30,31] AI, whether it works through open source methods/data or not, may be

defined as algorithmic simulation of human thought and decision-making processes.[30] Under ideal conditions, AI methods accomplish otherwise manual tasks with computational efficiency and precision while requiring minimal human oversight.[32] Adoption of novel computational methods in CV research may lead to previously unexpected avenues of impact, e.g. an algorithm developed for identification of heart disease biomarkers may rapidly stimulate research in other areas.

CV informatics is made feasible through the integration of open data, open source tools, and modern AI approaches. Impactful computational innovation can now happen far more rapidly and inexpensively (and, with proper technical considerations, reproducibily[33]) than in previous years.[34] AI and ML approaches are frequently distributed through open source code frameworks and made available through open distribution platforms (e.g. Github; see section 'Availability of open source methods' below). Infrastructure for finding open data has also improved, e.g. Google Dataset Search[35] now supports searches across >25 million datasets, including >100 open CV data resources. Cross-database search platforms (such as DataMed,[36] which indexes >3900 CV-relevant datasets) also enhance the findability of open data. We view the collective open code, open data, and common platforms as a *CV informatics commons*. Evaluating the components of the commons for their applicability to CV research questions, however, requires very specific data unlikely to be publicly available.[37] Open data complement observational data to support finding biologically meaningful connections.

2.4 General-purpose biomedical data sources relevant to cardiovascular disease studies

Publicly accessible knowledgebases (KBs) are premier examples of open data. A selection of biomolecular and disease KBs, along with example contents, is provided in *Table 1*. Some KBs, such as UniprotKB,[38] contain curated collections of entries linked to literature citations, while resources such as dbGaP,[39] are primarily intended as indices of large biomolecular and/or clinical datasets. Still other resources, e.g. Disease Ontology[40] or SNOMED CT,[41] support clearly delineated definitions of concepts and relationships among them. These data resources are generally open and accessible, with a few exceptions concerning licencing and secondary applications. This contrasts with closed, less easily searchable data (e.g. access to data from large-scale clinical trials may require a fee and/or license, depending on the data desired). We also make the distinction between knowledgebases (*Table* 1A) and metadata collections (*Table* 1B), the latter primarily serving to index sets of data resources rather than individual data points.

Table 1 Select biomolecular knowledgebases and metadata collections

Data Type	KBs	Example
A. Knowledgebases		
Drugs	DrugBank,[34] DrugCentral,[35] PharmGKB,[45] R epoDB,[48] RxNorm[49]	The small molecule N-(6-Aminohexyl)-5-Chloro-1-Naphthalenesulfonamide (DrugBank ID *DB04513*) has *Troponin I, cardiac muscle* as a target
Enzymes	BRENDA[41]	5 substrates for human troponin I, including *EC 3.4.22.16* (cathepsin H) in BRENDA
Diseases	Disease Ontology,[39] ICD-10/11,[37, 38] OMIM[44]	*Cardiomyopathy* has code *I42* in ICD-10-CM
Genetic variants	DisGeNET[42]	121 genetic variants associated with *familial idiopathic cardiomyopathy* in DisGeNet
Metabolites	Human Metabolome Database[43]	*Troponin I, cardiac muscle* is linked to the metabolite calcium, HMDB ID *HMDB0000464*
Protein–protein interactions	IntAct[44]	21 protein interactors for human *troponin I* in IntAct
Post-translational modifications	iPTMNet[45]	29 modification sites on *Troponin I, cardiac muscle* in iPTMNet
Clinical procedures	LOINC[46]	The diagnostic process of *Hypertrophic cardiomyopathy gene targeted mutation analysis in Blood or Tissue by Sequencing* has LOINC code 81860-9
Model organisms and their genetics	Model Organism Databases (in the Alliance of Genome Resources)[47]	The genotype of mouse strain *C57BL/6J-b2b904.1Clo* (*MGI : 5431511*) is associated with dilated cardiomyopathy
Biochemical pathways	Reactome,[48] WikiPathways[49]	*Troponin I* participates in 4 reactions in the *Muscle contraction (Homo sapiens)* pathway as per Reactome
Biomedical terminology and text	NCBI Disease Corpus,[43] SNOMED CT,[50] VASC[54]	*Familial cardiomyopathy (disorder)* is in SNOMED CT (SCTID: 35728003)
Proteins	UniProtKB[37]	*Troponin I, cardiac muscle* has accession code *P19429* in UniProtKB
B. Metadata collections		
Clinical trials	Clinicaltrials.gov[50]	24 studies concerning *Cardiomyopathy, Familial*
Genotype and phenotype studies	dbGaP[38]	26 phenotype datasets involving *cardiomyopathy* in dbGaP
Proteomes	ProteomeXchange (Peptide Atlas, PRIDE, MassIVE)[51]	73 proteome datasets with 'heart' or 'cardiac' in their titles on ProteomeXchange
Biomolecular and omics data	TOPMed[52]	36 TOPMed study datasets are available through dbGaP

A selection of KBs and metadata collections relevant to specific data types are provided here.

Methods may now be developed for real-world, public results before application to more focused datasets such as privately accessible electronic health records (EHRs).

Irrespective of their contents, KBs are most interoperable and accessible when paired with an application programming interface, or API. APIs define the set of commands allowing one system to interact with another, e.g. the specific command for a mass spectrometry data analysis platform to look up protein identifiers from UniProt. For the CV researcher, APIs are the way to obtain massive quantities of KB data reliably without requiring knowledge of the KB's internal structure or technical operation. Rapid access to all participants in the Reactome pathway (and UniProt accessions, where applicable), for example is available through just a single API call (specifically, the address https://reactome.org/ContentService/data/participants/R-HSA-5576891). For a clinical investigator, APIs are the bridge between an EHR and an existing clinical cohorts resource (e.g. Clinicaltrials.gov). APIs have become an integral part of the biomedical informatics landscape.

2.5 Opportunities and challenges in adopting open data and code

Infrastructure to support freely accessible distribution of data and computational methods is a cornerstone of CV informatics. A code repository made publicly available on the GitHub version control platform (or similar services, including Bitbucket[55] and GitLab[56]) serves as an up-to-date, openly available version of software and methods documentation. Such resources may serve as the basis for new work or components of larger projects, including as easily deployable components hosted through cloud computing platforms. As of July 2020, projects and code specifically concerning CV topics were common on GitHub (*Table* 2).

Though repositories and code shared on GitHub are plentiful, the vast majority are not indexed in biomedicine-specific manners or with consistent vocabularies (e.g. Medical Subject Headings or MeSH). Such an organizational framework is key to locating relevant digital resources: e.g. a search for 'Heart Diseases', will return >923 000 potentially relevant documents (*Table* 2A). For example, seeking 'Cardiac Imaging Techniques' yields just one repository, whereas the child MeSH terms (e.g. 'Angiocardiography' or 'Echocardiography') return >40 related projects (*Table* 2B). These resources would have remained hidden and unappreciated if investigators did not know specific keywords required. We would hope to locate projects and code *about* 'Heart Diseases' rather than every document containing the phrase itself. Similarly, a search for the term 'Cardiac Imaging Techniques' would ideally return all relevant objects concerning *any* specific imaging techniques used with the heart, beyond this generic phrase alone. Findability and accessibility

Table 2 Open computational methods and related resources available through GitHub

Search term	Repository	Code	Repository languages	Code languages
A. Repositories and code involving cardiovascular diseases				
Cardiovascular diseases	2056 (216)	1 801 877 (318 101)	Jupyter Notebook (82), R (24), Python (18)	Text (96 880), JSON (58 448), HTML (36 884)
Cardiovascular abnormalities	2	95 794	Jupyter Notebook (1)	Text (24 561), CSV (18 685), JSON (14 430)
Cardiovascular infections	4	151 377	CSS (2)	JSON (38 171), Text (36 994), CSV (24 269)
Heart diseases	1806	923 661	Jupyter Notebook (853), Python (321), R (120)	Text (339 113), JSON (131 006), HTML (106 944)
Pregnancy complications, cardiovascular	0	56 522	NA	Text (19 952), JSON (7387), HTML (6957)
Vascular diseases	28	256 422	Jupyter Notebook (8), Python (8), HTML (3)	Text (79 191), JSON (56 656), CSV (29 415)
B. Repositories and code involving cardiovascular imaging				
Diagnostic imaging	158 (109)	643 404 (404 974)	Python (25), Jupyter Notebook (15), Dockerfile (8)	Text (129 333), JSON (60 187), XML (44 811)
Cardiac imaging techniques	1	74 131	NA	Text (38 724), JSON (8496), XML (4964)
Angiocardiography	0	6422	NA	CSV (3961), Text (1091), JSON (337)
Cardiac-gated imaging techniques	0	100	NA	XML (24), Text (22), JSON (13)
Coronary angiography	8	45 988	Python (3), MATLAB (2), SAS (1)	Text (13 506), XML (7421), CSV (5681)
Echocardiography	38	84 874	Jupyter Notebook (7), Python (6), Java (4)	Text (30 994), XML (12 909), JSON (8901)
Myocardial perfusion imaging	2	21 125	MATLAB (1), Python (1)	CSV (4903), Text (4706), JSON (2662)
Radionuclide ventriculography	0	5790	NA	CSV (3843), Text (416), JSON (342)

Counts are approximate as of July 2020. Estimates have been provided where exact counts are not available. Search terms correspond to MeSH headings; parent heading (e.g. 'Cardiovascular Diseases') counts include counts of the child headings shown (e.g. 'Heart Diseases') and results for the heading itself, shown in parentheses. Repository Languages and Code Languages include the top three coding languages or formats of the Repository or Code count, as indexed by Github, respectively.

of open-source methods appropriate for CV informatics tasks thus remain challenging.

Though repository creators tag their resources with self-selected key words rather than those corresponding to a standard structure of concepts and technologies, without tags corresponding to a controlled vocabulary or ontology, this metadata is of limited interoperability. In addition, most searches starting with MeSH terms, as *Table 2* illustrates, find either a small set of repositories in varied programming languages or numerous documents in general text or high-level data formats (e.g. CSV or JSON). These results provide no further aggregation of metadata properties such as value types (e.g. as is tracked by a database such as dbGaP). As with inconsistent tagging, the limitation on this front may be a lack of broadly accepted and adhered-to standards and of GitHub's general purpose: this infrastructure is not specifically designed for biomedical informatics use cases. A lack of hierarchical topic-based indexing limits findability of open source resources.

2.6 Cloud computing infrastructure supporting cardiovascular informatics

Given the existence of open data and methods, where and how can computation be performed? Purchasing access to distributed resources is a clear and popular option. These services generally provide infrastructure (IaaS), a full platform (PaaS), software (SaaS), and/or data (DaaS) as a service.[57,58] IaaS includes virtual computational resources (e.g. Amazon Elastic Compute Cloud, EC2; or Google Cloud's Compute Engine) while PaaS offers access to complex hardware and software platforms (e.g. Amazon Software Development Kit or Google App Engine). SaaS assists creation of computational pipelines by leasing or providing cost-per-use access to established software applications, e.g. health record software. DaaS provides access to managed datasets, such as the 100 public datasets provided through Amazon Web Services.[59]

Performing -omics data analysis or information extraction in CV research is increasing the domain of cloud computing as numerous users may desire access to data and processes. Tools implemented through common platforms like the American Heart Association Precision Medicine Platform[60] can alleviate technical barriers to entry by providing purpose-build, readily usable infrastructure. Researchers can access and work with -omics data through HeartBioPortal,[61] or when working with private data, can run analyses on local computing clusters running the Galaxy platform.[62]

3. Emerging informatics approaches in structured and unstructured cardiovascular data analysis

3.1 Recent accomplishments in cardiovascular multi-omics data integration

Continual analyses of genomes, proteomes, and other -omes can reveal novel associations, particularly after integrating different data sources. Schlotter *et al.*[22] performed such a study to explore regulatory networks in calcific aortic valve disease, while Lalowski *et al.*[63] used multi-omics to identify pathways crucial to regeneration of the fetal heart. Our group has also proposed a statistical method for handling multi-omics data concerning cardiac remodeling.[24] These results and methods collectively highlight an ongoing informatics challenge: large datasets require extensive data normalization efforts and careful design to ensure resulting models are broadly applicable. One strategy for unifying multiple observation types is network medicine, a set of methods with profound implications for CV disease prognosis, diagnosis, and therapy.[64] Such approaches enable searching complex, interconnected collections of biomedical entities (e.g. proteins, genes, drugs, and pathways) to investigate the pathophysiology of CV disease, drug discovery (or, in some cases, repurposing), and understanding the molecular phenotypes of rare diseases.[65–68] Considering each patient's personalized network of biomedical entities and interactions opens further opportunities to learn from complex network analysis.[5,64,69] In one recent application, researchers constructed a network of CV diseases and non-coding RNAs, assisting in the identification of miRNA biomarkers.[70] Integrated multi-omics may soon become the default in CV research rather than a convenient path toward novel insights.

Linking phenotype with clinical observations requires extensive data engineering but can yield improved interoperability and reusability of data resources. The final product may be a polygenic risk score (PRS), though scope varies: Khera *et al.*[71] developed PRSs for coronary artery disease and atrial fibrillation while Cirulli *et al.*[72] compared thousands of phenotypes using >70 000 exome sequences. An observational study by Elliot *et al.*[73] found that incorporation of polygenic risk scores could moderately assist risk stratification with regard to coronary artery disease. Jamal *et al.*[74] used publicly-available KBs to build models for predicting adverse CV drug reactions: in a comparison of >500 models for 36 different adverse reactions (e.g. tachycardia), the researchers predicted novel reactions, such as those for the anti-malarial drug mefloquine.

3.2 Adaptation of text mining and information extraction tools

Publicly available biomedical manuscripts contain abundant experimental and clinical observations. Though writing and reading individual papers has long been the cornerstone of scholarly communication, informatics approaches enable comprehensive analyses of voluminous document collections. Working with biomedical text in this manner requires rendering unstructured data sources searchable. Recent research from the NIH National Library of Medicine has resulted in the tools PubTator Central[75] and LitSense,[76] both of which extend traditional literature search to sentence-level context. These new approaches enable curious researchers to reduce the trial-and-error necessary in usual phrase or keyword-based search,[77,78] and instead enable comparison of results for specific claims such as 'cardiac arrest is treated with hypothermia' (for reference, LitSense finds >60 instances of this statement).

Our group has found that text mining strategies can drive creation of novel insights. Working with data mining engineers, we applied a novel Context-aware Semantic Online Analytical Processing platform (CaseOLAP) to an exploration of how proteins of the extracellular matrix relate to subtypes of CV disease in the biomedical literature.[79] We have empowered text mining approaches to accomplish information extraction, or IE: in this set of processes, unstructured text is translated into structured data, concepts, and relationships. The methods to do so may be manual or automated. We have developed standardized methods for IE in CV biomedicine through identification of metadata from clinical case reports.[80,81]

Information extraction need not be completely unguided: it can benefit from existing approaches to enforcing structure on observations. Structured reporting can help to ensure that clinical observations fit a standardized template *and* are consistently interpretable through text

mining.[82] A study based on cardiac magnetic resonance imaging (such as the reporting described by Johnson *et al.*,[83] for example) may be driven by a text mining system for comparing findings extracted from standardized diagnostic reports. Checklists for the contents of observational studies, such as that described by the STROBE Statement,[84] can also assist in pre-defining the types of content in each document. Data mining becomes far more efficient and accurate with the expectation of consistency in each report.

3.3 Graph-based data integration and exploration

Assembling heterogeneous observations in a way conducive to highlighting novel, mechanistically-relevant relationships is a non-trivial task but well-suited to graph methods. Much like traditional databases, these approaches assign each data point or concept to a unique record, yet the basic unit of each graph is a relationship between at least two records. Focusing on two proteins with potential contributions to disease risk, we may rapidly identify relevant pathways and drug therapies with potential for repurposing. A rigorous search of the full graph encompassing a wide variety of data types allows quantification of the strength of each relationship as well as the basis for new relationship predictions. This approach offers particular potential in combination with data curated from health records, such that individual patients are represented as graphs of their medical histories *and* relevant experimental observations, all in a format compatible with segments of other clinical cohorts.[85]

4. Informatics approaches in clinical research and investigations of cardiovascular disease phenotypes

4.1 Development of large clinical data sources

The goal of a 'learning health system' in which the contents of EHRs seamlessly flow between health systems, government agencies, researchers, and epidemiologists[86] appears more attainable each year. Worldwide efforts to implement and learn from EHRs vary but are constantly improving, from a Swiss law mandating interoperable EHR adoption as of April 2017[87] to major efforts toward establishing interoperable health information systems in China.[88] About 86% of US physicians were using electronic medical records or EHR systems by 2017.[89] Though these data resources are, in theory, readily available for addressing investigations into CV disease, they are by no means the only sources of clinical observations.

Studies of patient presentation features have leveraged large-scale longitudinal studies not originally conceived as foundations for modern informatics strategies. Both the Framingham Heart Study[90] and the Cardiovascular Health Study[91] (initiated in 1948 and in 1989, respectively, both well before the modern resurgence of ML) remain valuable resources. The Multi-Ethnic Study of Atherosclerosis (MESA), initiated in 2000, serves as another source of longitudinal clinical observations.[92] These datasets offer the benefit of providing both demographic and clinical features along with more focused research observations, with recent study cycles including genetic analyses. One characteristic challenge in working with these long-running studies is their population bias. With the exception of MESA, the largely white cohorts in these studies have limited their efficacy in the development of new predictive models for non-white patient populations.[93]

4.2 Data mining, deep representation, and information extraction

The data associated with biomedical documents include both document contents and their *metadata*. Metadata can result from the process of assembling a document (e.g. the period of time described within a case report) and from description of its contents (e.g. the report describes a case of heart failure in a male patient with Kawasaki disease). This latter type of concept-level metadata is commonly described through keywords and controlled vocabularies, such as MeSH.[94] Clinical concepts and phenotypes are popularly represented through coding systems such as International Classification of Diseases (ICD).[95] The newly released ICD-11[96] system has reframed how it categorizes CV conditions, particularly regarding cerebrovascular disorders.[97,98] Though some efforts, such as the Cardiovascular Disease Ontology (CVDO),[99] have been produced to enable CV-focused metadata creation, such resources remain rare. We have made efforts to define the general types of metadata in clinical case reports, with the objective of supporting both manual and automated information extraction from documents describing clinical cases.[80]

One of the central challenges in biomedical information extraction is the enforcement of a consistent structure upon otherwise noisy, loosely arranged data. Any single pathological phenotype or diagnosis may be expressed through text or numerical data in a variety of ways. A case of 'heart failure with preserved ejection fraction' may be accurately described as 'HFpEF', heart failure with an ejection fraction above 50%, or even cardiac dysfunction with no explicit 'HFpEF' diagnosis. Nevertheless, we must fit these similar events into a consistent data model, or a set of categories and rules defining the objects or events in our data and how they relate to each other. Traditionally a job for relational databases where one object corresponds to one database entry—often visualized as a table or spreadsheet—models now also fit data into graph structures. The resulting heterogeneous graphs are known as knowledge graphs[85,100] (KGs), or collections of relationships between different types of concepts and entities. One such graph-based project combines >2 million relationships with the goal of identifying candidates for drug repurposing.[101]

Categorizing and processing clinical reports can transform these massive record collections into rich data sources. Digesting cardiology reports in particular has been explored for >40 years, with an early approach proposed by Gabrieli and Merrill in 1980.[102] More recent efforts allow cardiology reports to be automatically distinguished from those of other specialties[103] and can identify patients with trileaflet aortic stenosis and coronary artery disease from free-text echocardiography and cardiac catheterization reports with much higher predictive power than billing codes alone.[104] Another system for processing echocardiography reports, 'EchoInfer', can consistently extract quantitative and qualitative values describing CV structure and function.[105]

Mining the observations described within text data is most informative when merged with large-scale experimental observations. We present a general framework for data mining in *Figure 2*, showing how a processing pipeline can incorporate methods for text processing (i.e. named entity recognition and relationship extraction) and those for data integration (here, in the form of a graph or network). We note areas where cloud computing infrastructure offers particular benefits; training models is an ideal use case, for example as resources may be shut off when training is

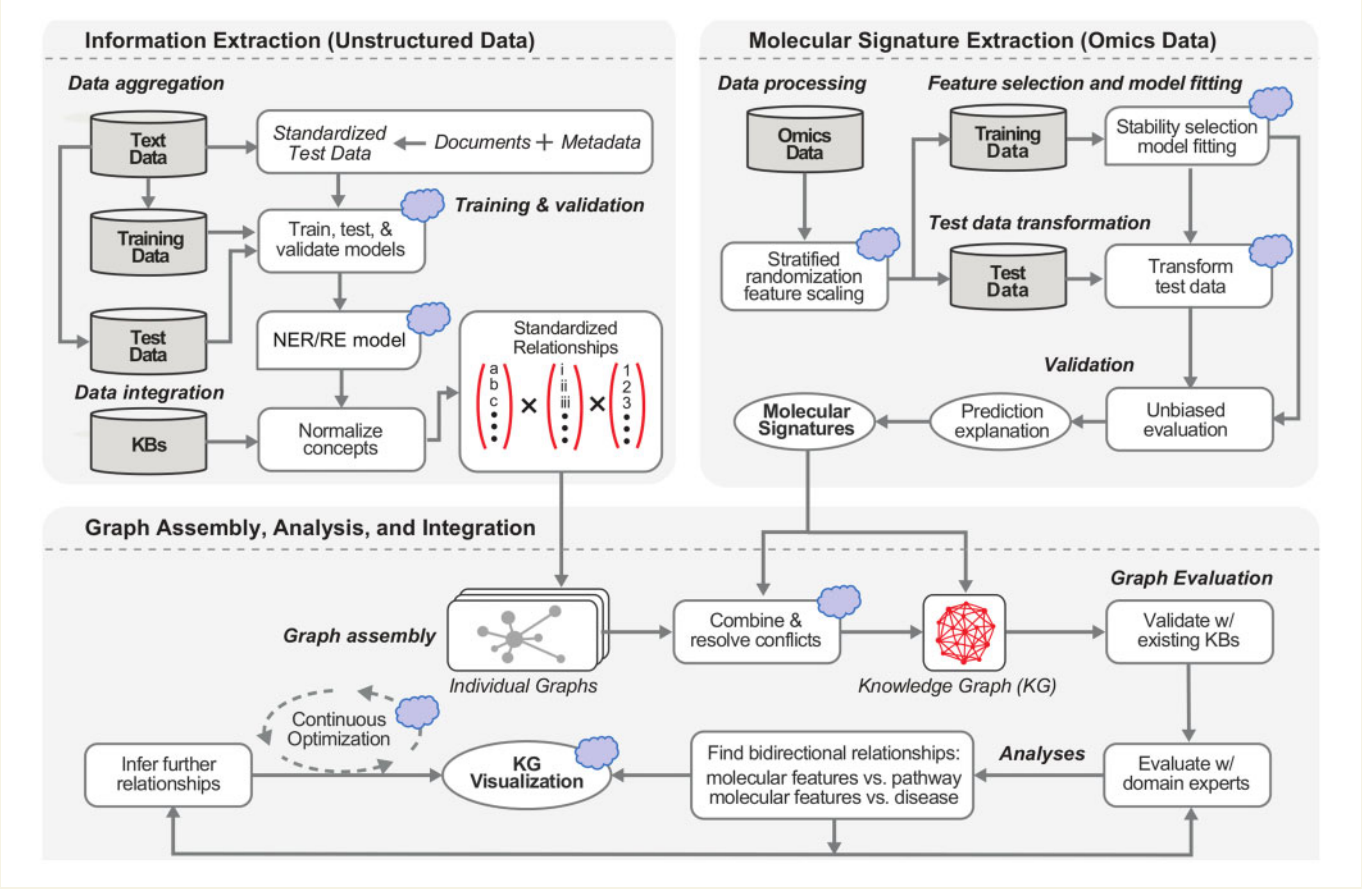


Figure 2 An example workflow for biomedical data mining and integration. A process for combining unstructured data (e.g. contents of biomedical text documents), structured data from -omics studies, and KB contents requires three components: information extraction, molecular signature extraction, and assembly of data into a unified structure. Here, the unified structure is a graph, referred to as a Knowledge Graph or KG. After the preparation of computational models for named entity recognition (NER) and relationship extraction (RE), these models are applied to a set of standardized text documents to yield sets of standardized relationships (i.e. one for each document). Molecular signature extraction from omics data follows a similar procedure: training and test data are used to assemble a model capable of identifying molecular signatures of one or more experimental conditions. All signatures and relationships are combined into a single KG. The KG must be evaluated by domain experts, but each round of evaluation and identification of specific biomedical relationships is used to improve and optimize the KG. Cloud icons represent tasks ideal for implementation through cloud computing (e.g. training, testing, and validation of text processing modules is computationally intensive and therefore a good candidate for accomplishing through cloud infrastructure).

complete. A complete end-to-end pipeline may be implemented in the cloud in practice. Regardless of the computational infrastructure, the resulting collection of concepts and relationships may then be assembled into a KG, with each part resembling that in *Figure 3*.

Application of data mining to biomedical text continues to face challenges in broader application. A survey of methods for processing pathology records found that insufficient validation and the absence of shared datasets limited the extent to which methods could be compared, though most manuscripts reported high accuracy.[106] Adapting successful approaches from one healthcare setting or clinical domain poses challenges in resolving data structures (in some cases, even minor differences in EHR software can cause data to vary structurally).[107] Specific resources supporting development of text mining tools for phenotype-level analysis may help, e.g. the PhenoCHF set contains 300 annotated discharge summaries for congestive heart failure patients.[108] Disease-focused resources such as these allow methods to be adapted to a condition's unique context and presentation features. Isolating specific values defining the features of a large dataset (i.e. a *representation*) may also be a viable long-term strategy for aggregating large quantities of clinical observations. Recent work has prepared representations of EHR contents from hundreds of thousands[109] (and in one case, >1 million) patients.[110]

4.3 Machine learning-driven models for disease risk prediction and identification

Intensive algorithmic studies of numerous CV disease presentations yield powerful predictive models. A recent example can be seen in a project by Williams *et al.*[111] concerning connections between plasma protein expression and 11 different phenotypes used as health indicators, including metrics such as lean body mass and values such as primary CV event risk. Protein measurements from >16 800 patients support accurate inference of a subset of phenotypes, primarily those concerning body fat. In a much larger study, an analysis of >3.3 million percutaneous coronary intervention procedures found that a combination of ML approaches could predict periprocedural bleeding with high accuracy.[112] A study of

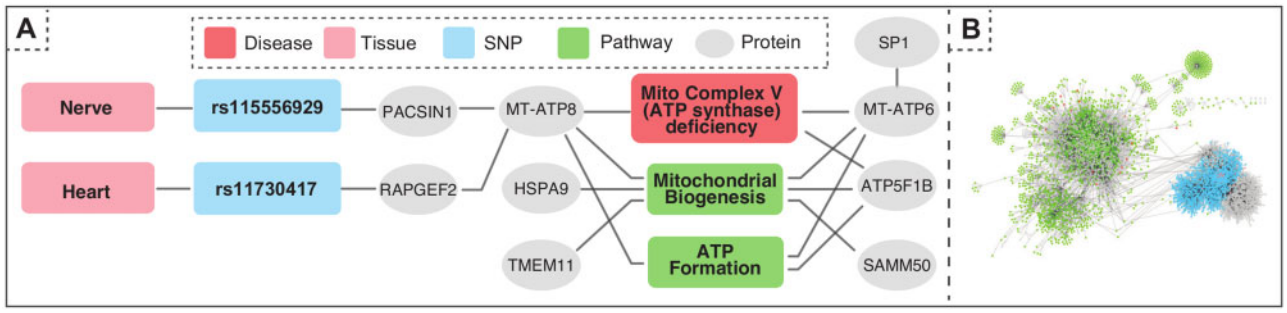


Figure 3 Segments of a knowledge graph containing relationships in metabolic and CV disease. (*A*) An example is shown using a knowledge graph to illustrate five types of concepts (colored shapes) along with their relationships (lines denoting their relationships). A knowledge graph is constructed through information extraction from biomedical literature available in PubMed and from established knowledgebases (e.g. protein–protein interactions from IntAct and tissue-specific SNPs from the GTEx database). Relationships are data type specific, e.g. a protein connected to a pathway (here, from the Reactome knowledgebase) indicates the protein participates in the pathway. (*B*) A more complex KG, consisting of >5700 aggregated concepts and >21 000 relationships with respect to SNPs, proteins, and pathways.

>24 970 adverse CV incidents was used to build a neural network-driven model capable of accurately predicting over 7% (~350) more incidents than established methods (i.e. American College of Cardiology evaluation guidelines).[113] Work conducted through the TOPCAT (Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Agonist) trial specifically investigated HFpEF, finding that a random forest model could anticipate mortality and hospitalization with fairly reliable accuracy but likely requires further adaptation for use outside the trial cohort.[114] Some model-driven efforts seek to identify and discover CV risk factors: the RFMiner framework,[115] a phenome-wide association study reported by Hyppönen *et al.*,[116] and a project by Pickhardt *et al.*[117] (in this case, with metrics originally from abdominal CT scans not intended as CV diagnostics) have all found associations and potential biomarkers for adverse CV events.

Integrating numerous data sources can empower models with collections of metrics that may have limited statistical value when considered alone. Samad *et al.*[118] assembled models of patient outcome based on integrated ECG results and EHRs from >171 000 patients, obtaining solid accuracy in its survival estimations from just 10 features. A 2016 analysis by Motwani *et al.*[119] examined how validating a mortality model with a 5-year follow-up could work for coronary artery disease cases. In this case, the ML-driven model was more accurate than baseline metrics, including using coronary computed tomographic angiography alone. A separate model assembled by Commandeur *et al.*[120] was used to estimate likelihood of myocardial infarction and cardiac death. The model is capable of more predictive ability than any single metric. Work by Suchard *et al.*[121] sought to compare the efficacy and safety of drug therapies for hypertension. They then developed hazard ratios for several drug classes based on claims and EHRs from >4.9 million patients across multiple countries, finding support for use of thiazide diuretics. Though compelling and informative, the results of each of these efforts will require extensive validation before translation to clinical practice.

Text mining and NLP methods are practical components of the CV informatics toolbox and have been successfully employed for estimating disease risk. A novel system developed by Patterson *et al.*[122] analyzes text descriptions from echocardiogram and radiology reports from the Veterans Aging Cohort Study to extract measurements most relevant to disease. The system appears to effectively extract values for concepts including ejection fraction or mitral valve regurgitation. A project by Diller *et al.*[123] applied a deep learning model to >40 000 EHRs to categorize adult congenital heart disease and pulmonary hypertension cases by diagnosis, New York Heart Association class, potential treatments, and other outputs. Their models suggest deep learning on multifaceted data is a viable strategy for these tasks, e.g. accuracy of their diagnosis classification model exceeded 91%, as compared to the 96% accuracy achieved by cardiologists performing classification manually. NLP has also been applied to such diverse applications as comparisons of CV disease presentations between dental and medical records,[124] triaging transient ischemic attack cases,[125] and collaborative efforts between computers and medical domain experts.[126] Additional projects, including those predicting hypertension and arterial disease risk, have been systematically reviewed by Sheikhalishahi *et al.*[127]

5. Major challenges and limitations of informatics approaches

Application of informatics approaches to any area of biomedicine presents major challenges on technical and conceptual fronts. As a comprehensive assessment of these issues is sufficient material for its own review (or, indeed, its own field of study) we list both obstacles and limitations as we see most pertinent to applications of them to CV research. Modern CV research requires integration of data from disparate sources: multiple -omics data types, unstructured data, imaging, observational results, and signal data may play roles, among numerous others. No automated way yet exists to reliably ensure that all data formats are interoperable, and even a single method (e.g. mass spectrometry used for proteomics) must use an array of data formats,[128] depending on the laboratory or its data processing pipeline. Similarly, informatics approaches are largely limited by their interoperability; successful informatic applications require users to gain considerable technical experience (much of them were rarely offered through training in molecular and CV biology). Implementing and testing ML-driven methods also requires understanding and/or knowledge of programming. Long-term longitudinal studies present an especially thorny issue: a model trained on data today may be obsolete in a few years in the future and its

foundational infrastructure may not be available for the duration of the entire study.

AI or ML methods are clearly not universally optimal choices for investigating or predicting CV phenomena. Regression may remain a perfectly acceptable model for some situations, as Loring *et al.*[129] found with predicting atrial fibrillation outcomes. Li *et al.*[130] found for predicting CV disease risk in a large cohort, and Frizzell *et al.*[131] observed for heart failure hospital readmission prediction. Increasingly complicated computational models are also accompanied by commensurate challenges in interpretability, potentially without meaningful benefits in accuracy. Minimally interpretable models may yield meaningless patterns with few opportunities to derive their provenance.[132] Model interpretability remains an open research area, including with predictive models for hypertension.[133] ML methods should therefore be considered one set of tools among many in a researcher's tool chest, to be deployed when appropriate and tested against alternatives. Intensive computational approaches may support data processing tasks contributing to a rigorous analysis, as shown in *Figure 2*, rather than serving as complete predictive models.

The value of any informatics method depends upon the quality of its data source and the machine readability of the data format; CV informatics faces a major challenge in this regard. Data processing to minimize the likelihood of misleading conclusions due to a lack of statistical power or the presence of confounding factors is a critical prerequisite. Noisy, complicated observational datasets rich in confounding variables are prohibitive when a naive computational method (or, at times, a new investigator) arrives at interpretations as causal.[12] Algorithmic approaches capable of finding hidden connections in data may also find spurious correlations. Investigations toward using wearable heart sensor data to infer disease risk found that patient age, sex, and medication usage were all potential confounders,[134] for example. In some settings, data may lack impactful factors: large epidemiological studies based on medical records may be unexpectedly impacted by difficult-to-quantify socioeconomic factors[135] or underlying differences between control and test cohorts.[136] Unfortunately, there are no available solutions to these challenges at the present time. Nevertheless, as an increasing number of our data analyses are automated, domain knowledge and expert guidance remain indispensable in evaluating results of all informatic approaches.

The onset of intelligent data processing machines has brought along with it a pressing question of how new knowledge can be maximized whilst maintaining patient data privacy and integrity. Groups intending to use these novel sources of biomedical data analysis must walk a fine line between respecting a patient's desire to oversee access to their personal information and the inclination to incorporate these data in generating new discoveries.[137] Just as patients worry that these data repositories storing their data can become the target of an attack and leave sensitive information exposed, those in possession of the data face reputational losses and violations if this were to occur. The decreased accessibility of this data for researchers presents the opportunity for centralized, controlled data analysis platforms (e.g. the AHA Precision Medicine Platform[60]), and the foundation of the CV informatics commons. Regulation of AI in the United States and in the EU depends on whether the AI system is classified as a medical device in either territory (i.e. by the US Federal Food, Drug, and Cosmetic Act or the EU Medical Device Regulation and the Regulation on in vitro diagnostic medical devices, respectively) as well as by the US Cybersecurity and Infrastructure Security Act of 2018 and the EU Cybersecurity Act of 2019. There is, at this time, no one-size-fits-all solution to biomedical data privacy nor a regulation covering all instances of data protection.

Pressing questions remain about the future of CV informatics. Who should benefit from advances? CV disease incidence is not consistently distributed across populations.[138,139] At times, research metrics may be out of alignment with project goals: in the course of building a risk prediction model for atherosclerotic disease, Pfohl *et al.*[140] concluded that maximizing model performance is incompatible with the goal of developing a 'fair' model, i.e. a model equivalently effective across cohorts of varying demographic features. Computational approaches designed to be applicable to large populations—or based on data from representative cohorts—must take these considerations into account or risk exacerbation of disease correlated with cultural, ethnic, social, and economic factors.[141]

6. Concluding remarks

Widespread adoption of established, standardized biomedical indexing systems by open source repositories could rapidly provide a foundation for the CV informatics commons. This scenario represents an ideal opportunity for the application of interoperable metadata standards. One clear example of standardized metadata is employed by Figshare: this data-sharing repository includes a controlled system of categories along with user-generated keywords. In addition, items in Figshare are frequently linked to published, peer-reviewed manuscripts, enhancing their findability. Second, collaboration among repository builders and biomedical scientists would offer significant benefits to ensure real-world use cases being faithfully represented in the open source. For example, Synapse was developed as an open source platform specific for biomedical research needs, e.g. indexing by each research community.[142] Similarly, the Reactome KB indexes its entries with multiple, interoperable identifiers corresponding to cellular processes and pathways, integrated and facilitated by a common interface.[49] Conveniently, any creator of a new repository can contribute to better standardization today by indexing their projects with multiple MeSH terms or terms from other ontologies (see *Table 1*) as keywords. Finally, detailed documentation accompanying a project can increase its findability (or simply allow you to find its components and data in the future).[143] Future efforts in metadata creation and standardization therefore stand to noticeably improve the FAIRness of resources central to CV informatics.

We see CV informatics as both a necessary confluence of interdisciplinary strategies and the means by which CV investigators can access previously inconceivable reaches of disease-relevant knowledge. Much of the pressure to implement advanced computational methods arises directly from the increasing size and complexity of biomolecular data: a single -omics screen holds a wealth of information, unobtainable without comprehensive, automated analysis. Meanwhile, the much-celebrated advent of EHRs promises to increase efficiency and reduce the chance of errors, yet these documents do not inherently support development of reasoning or narrative.[144] While no single ML or data processing approach will effectively extract meaning from noisy data across all of biomedicine,[145] adoption of open data, open science, and cloud computing practices enables a CV informatics commons to empower a global CV research community. Future advancements in this field must address the unique benefits of both human *and* computational thought processes. Together, we will illuminate new areas of the multi-dimensional, systemic phenomena driving CV disease.

Data availability

The data underlying this article are available in the article.

Acknowledgement

The authors thank Samir Akre for assistance with conceptual development of data exploration workflows.

Conflict of interest: none declared.

Funding

This work was supported by the National Heart, Lung and Blood Institute at the National Institutes of Health [awards R35 HL135772, T32 HL139450, R01 HL146739] and the UCLA Laubisch Endowment to Peipei Ping.

References

1. Greene JA, Lea AS. Digital futures past—the long arc of big data in medicine. *N Engl J Med* 2019;**381**:480–485.
2. Roth GA, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, Abbastabar H, Abd-Allah F, Abdela J, Abdelalim A, Abdollahpour I, Abdulkader RS, Abebe HT, Abebe M, Abebe Z, Abejie AN, Abera SF, Abil OZ, Abraha HN, Abrham AR, Abu-Raddad LJ, Accrombessi MMK, Acharya D, Adamu AA, Adebayo OM, Adedoyin RA, Adekanmbi V, Adetokunboh OO, Adhena BM, Adib MG, Admasie A, Afshin A, Agarwal G, Agesa KM, Agrawal A, Agrawal S, Ahmadi A, Ahmadi M, Ahmed MB, Ahmed S, Aichour AN, Aichour I, Aichour MTE, Akbari ME, Akinyemi RO, Akseer N, Al-Aly Z, Al-Eyadhy A, Al-Raddadi RM, Alahdab F, Alam K, Alam T, Alebel A, Alene KA, Alijanzadeh M, Alizadeh-Navaei R, Aljunid SM, Alkerwi A, Alla F, Allebeck P, Alonso J, Altirkawi K, Alvis-Guzman N, Amare AT, Aminde LN, Amini E, Ammar W, Amoako YA, Anber NH, Andrei CL, Androudi S, Animut MD, Anjomshoa M, Ansari H, Ansha MG, Antonio CAT, Anwari P, Aremu O, Ärnlöv J, Arora A, Arora M, Artaman A, Aryal KK, Asayesh H, Asfaw ET, Ataro Z, Atique S, Atre SR, Ausloos M, Avokpaho EFGA, Awasthi A, Quintanilla BPA, Ayele Y, Ayer R, Azzopardi PS, Babazadeh A, Bacha U, Badali H, Badawi A, Bali AG, Ballesteros KE, Banach M, Banerjee K, Bannick MS, Banoub JAM, Barboza MA, Barker-Collo SL, Bärnighausen TW, Barquera S, Barrero LH, Bassat Q, Basu S, Baune BT, Baynes HW, Bazargan-Hejazi S, Bedi N, Beghi E, Behzadifar M, Behzadifar M, Béjot Y, Bekele BB, Belachew AB, Belay E, Belay YA, Bell ML, Bello AK, Bennett DA, Bensenor IM, Berman AE, Bernabe E, Bernstein RS, Bertolacci GJ, Beuran M, Beyranvand T, Bhalla A, Bhattarai S, Bhaumik S, Bhutta ZA, Biadgo B, Biehl MH, Bijani A, Bikbov B, Bilano V, Bililign N, Bin Sayeed MS, Bisanzio D, Biswas T, Blacker BF, Basara BB, Borschmann R, Bosetti C, Bozorgmehr K, Brady OJ, Brant LC, Brayne C, Brazinova A, Breitborde NJK, Brenner H, Briant PS, Britton G, Brugha T, Busse R, Butt ZA, Callender CSKH, Campos-Nonato IR, Campuzano Rincon JC, Cano J, Car M, Cárdenas R, Carreras G, Carrero JJ, Carter A, Carvalho F, Castañeda-Orjuela CA, Castillo Rivas J, Castle CD, Castro C, Castro F, Catalá-López F, Cerin E, Chaiah Y, Chang J-C, Charlson FJ, Chaturvedi P, Chiang PP-C, Chimed-Ochir O, Chisumpa VH, Chitheer A, Chowdhury R, Christensen H, Christopher DJ, Chung S-C, Cicuttini FM, Ciobanu LG, Cirillo M, Cohen AJ, Cooper LT, Cortesi PA, Cortinovis M, Cousin E, Cowie BC, Criqui MH, Cromwell EA, Crowe CS, Crump JA, Cunningham M, Daba AK, Dadi AF, Dandona L, Dandona R, Dang AK, Dargan PI, Daryani A, Das SK, Gupta RD, Neves JD, Dasa TT, Dash AP, Davis AC, Davis Weaver N, Davitoiu DV, Davletov K, De La Hoz FP, De Neve J-W, Degefa MG, Degenhardt L, Degfie TT, Deiparine S, Demoz GT, Demtsu BB, Denova-Gutiérrez E, Deribe K, Dervenis N, Des Jarlais DC, Dessie GA, Dey S, Dharmaratne SD, Dicker D, Dinberu MT, Ding EL, Dirac MA, Djalalinia S, Dokova K, Doku DT, Donnelly CA, Dorsey ER, Doshi PP, Douwes-Schultz D, Doyle KE, Driscoll TR, Dubey M, Dubljanin E, Duken EE, Duncan BB, Duraes AR, Ebrahimi H, Ebrahimpour S, Edessa D, Edvardsson D, Eggen AE, El Bcheraoui C, El Sayed Zaki M, El-Khatib Z, Elkout H, Ellingsen CL, Endres M, Endries AY, Er B, Erskine HE, Eshrati B, Eskandarieh S, Esmaeili R, Esteghamati A, Fakhar M, Fakhim H, Faramarzi M, Fareed M, Farhadi F, Farinha C. S E S, Faro A, Farvid MS, Farzadfar F, Farzaei MH, Feigin VL, Feigl AB, Fentahun N, Fereshtehnejad S-M, Fernandes E, Fernandes JC, Ferrari AJ, Feyissa GT, Filip I, Finegold S, Fischer F, Fitzmaurice C, Foigt NA, Foreman KJ, Fornari C, Frank TD, Fukumoto T, Fuller JE, Fullman N, Fürst T, Furtado JM, Futran ND, Gallus S, Garcia-Basteiro AL, Garcia-Gordillo MA, Gardner WM, Gebre AK, Gebrehiwot TT, Gebremedhin AT, Gebremichael B, Gebremichael TG, Gelano TF, Geleijnse JM, Genova-Maleras R, Geramo YCD, Gething PW, Gezae KE, Ghadami MR, Ghadimi R, Ghasemi Falavarjani K, Ghasemi-Kasman M, Ghimire M, Gibney KB, Gill PS, Gill TK, Gillum RF, Ginawi IA, Giroud M, Giussani G, Goenka S, Goldberg EM, Goli S, Gómez-Dantés H, Gona PN, Gopalani SV, Gorman TM, Goto A, Goulart AC, Gnedovskaya EV, Grada A, Grosso G, Gugnani HC, Guimaraes ALS, Guo Y, Gupta PC, Gupta R, Gupta R, Gupta T, Gutiérrez RA, Gyawali B, Haagsma JA, Hafezi-Nejad N, Hagos TB. Hailegiyorgis TT. 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.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;**25**:44–56.
4. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? *Heart* 2018;**104**:1156–1164.
5. Leopold JA, Maron BA, Loscalzo J. The application of big data to cardiovascular disease: paths to precision medicine. *J Clin Invest* 2020;**130**:29–38.
6. Annapureddy AR, Angraal S, Caraballo C, Grimshaw A, Huang C, Mortazavi BJ, Krumholz HM. The National Institutes of Health funding for clinical research applying machine learning techniques in 2017. *Npj Digit Med* 2020;**3**:13.
7. Congressional Research Service. National Institutes of Health (NIH) Funding: FY1994-FY2020. 2020 January. Report No.: R43341.
8. Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, Pandey M, Maliakal G, van Rosendael AR, Beecy AN, Berman DS, Leipsic J, Nieman K, Andreini D, Pontone G, Schoepf UJ, Shaw LJ, Chang H-J, Narula J, Bax JJ, Guan Y, Min JK. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J* 2019;**40**:1975–1986.
9. Dey D, Slomka PJ, Leeson P, Comaniciu D, Shrestha S, Sengupta PP, Marwick TH. Artificial intelligence in cardiovascular imaging. *J Am Coll Cardiol* 2019;**73**:1317–1335.
10. Niederer SA, Lumens J, Trayanova NA. Computational models in cardiology. *Nat Rev Cardiol* 2019;**16**:100–111.
11. Henglin M, Stein G, Hushcha PV, Snoek J, Wiltschko AB, Cheng S. Machine learning approaches in cardiovascular imaging. *Circ Cardiovasc Imaging* 2017;**10**:e005614.
12. Kagiyama N, Shrestha S, Farjo PD, Sengupta PP. Artificial intelligence: practical primer for clinical research in cardiovascular disease. *J Am Heart Assoc* 2019;**8**:e012788.
13. Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. *Yearb Med Inform* 2014;**9**:14–20.
14. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* 2008;**8**:37–49.
15. Harst P. V D, Windt L. D, Chambers JC. Translational perspective on epigenetics in cardiovascular disease. *J Am Coll Cardiol* 2017;**70**:590–606.
16. Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, Thiel GJM, van, Cronin M, Brobert G, Vardas P, Anker SD, Grobbee DE, Denaxas S; Innovative Medicines Initiative 2nd programme, Big Data for Better Outcomes, BigData@Heart Consortium of 20 academic and industry partners including ESC. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *Eur Heart J* 2018;**39**:1481–1495.
17. Kimball TH, Vondriska TM. Metabolism, epigenetics, and causal inference in heart failure. *Trends Endocrinol Metab* 2019;S1043276019302346.
18. Lescroart F, Wang X, Lin X, Swedlund B, Gargouri S, Sànchez-Dànes A, Moignard V, Dubois C, Paulissen C, Kinston S, Göttgens B, Blanpain C. Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* 2018;**359**:1177–1181.
19. Paik DT, Tian L, Lee J, Sayed N, Chen IY, Rhee S, Rhee J-W, Kim Y, Wirka RC, Buikema JW, Wu SM, Red-Horse K, Quertermous T, Wu JC. Large-scale single-cell RNA-seq reveals molecular signatures of heterogeneous populations of human induced pluripotent stem cell-derived endothelial cells. *Circ Res* 2018;**123**:443–450.
20. Meder B, Haas J, Sedaghat-Hamedani F, Kayvanpour E, Frese K, Lai A, Nietsch R, Scheiner C, Mester S, Bordalo DM, Amr A, Dietrich C, Pils D, Siede D, Hund H, Bauer A, Holzer DB, Ruhparwar A, Mueller-Hennessen M, Weichenhan D, Plass C, Weis T, Backs J, Wuerstle M, Keller A, Katus HA, Posch AE. Epigenome-wide association study identifies cardiac gene patterning and a novel class of biomarkers for heart failure. *Circulation* 2017;**136**:1528–1544.
21. Goldenberg NA, Everett AD, Graham D, Bernard TJ, Nowak-Göttl U. Proteomic and other mass spectrometry based "omics" biomarker discovery and validation in pediatric venous thromboembolism and arterial ischemic stroke: current state, unmet needs, and future directions. *Proteomics Clin Appl* 2014;**8**:828–836.
22. Schlotter F, Halu A, Goto S, Blaser MC, Body SC, Lee LH, Higashi H, DeLaughter DM, Hutcheson JD, Vyas P, Pham T, Rogers MA, Sharma A, Seidman CE, Loscalzo J, Seidman JG, Aikawa M, Singh SA, Aikawa E. Spatiotemporal multi-omics mapping generates a molecular atlas of the aortic valve and reveals networks driving disease. *Circulation* 2018;**138**:377–393.
23. Arneson D, Shu L, Tsai B, Barrere-Cain R, Sun C, Yang X. Multidimensional integrative genomics approaches to dissecting cardiovascular disease. *Front Cardiovasc Med* 2017;**4**: 8.
24. Chung NC, Mirza B, Choi H, Wang J, Wang D, Ping P, Wang W. Unsupervised classification of multi-omics data during cardiac remodeling using deep learning. *Methods* 2019;**166**:66–73.
25. O'Boyle NM, Guha R, Willighagen EL, Adams SE, Alvarsson J, Bradley J-C, Filippov IV, Hanson RM, Hanwell MD, Hutchison GR, James CA, Jeliazkova N, Lang AS, Langner KM, Lonie DC, Lowe DM, Pansanel J, Pavlov D, Spjuth O, Steinbeck C, Tenderholt AL, Theisen KJ, Murray-Rust P. Open Data, Open Source and Open Standards in chemistry: the Blue Obelisk five years on. *J Cheminform* 2011;**3**:37.
26. Levin N, Leonelli S, Weckowska D, Castle D, Dupré J. How do scientists define openness? Exploring the relationship between open science policies and research practice. *Bull Sci Technol Soc* 2016;**36**:128–141.

27. Shaikh AR, Butte AJ, Schully SD, Dalton WS, Khoury MJ, Hesse BW. Collaborative biomedicine in the age of big data: the case of cancer. *J Med Internet Res* 2014;**16**: e101.
28. McKiernan EC, Bourne PE, Brown CT, Buck S, Kenall A, Lin J, McDougall D, Nosek BA, Ram K, Soderberg CK, Spies JR, Thaney K, Updegrove A, Woo KH, Yarkoni T. How open science helps researchers succeed. *eLife* 2016;**5**:e16800.
29. Gacek C, Arief B. The many meanings of open source. *IEEE Softw* 2004;**21**:34–40.
30. Paton C, Kobayashi S. An open science approach to artificial intelligence in healthcare: a contribution from the international medical informatics association open source working group. *Yearb Med Inform* 2019;**28**:47–051.
31. Murray-Rust P. Open data in science. *Ser Rev* 2008;**34**:52–64.
32. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017;**69**: S36–S40.
33. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. *JAMA* 2020;**323**:305–306.
34. Payne PRO, Lussier Y, Foraker RE, Embi PJ. Rethinking the role and impact of health information technology: informatics as an interventional discipline. *BMC Med Inform Decis Mak* 2016;**16**:40.
35. Google. Google Dataset Search. https://datasetsearch.research.google.com/ (20 February 2021).
36. Ohno-Machado L, Sansone S-A, Alter G, Fore I, Grethe J, Xu H, Gonzalez-Beltran A, Rocca-Serra P, Gururaj AE, Bell E, Soysal E, Zong N, Kim H. Finding useful data across multiple biomedical data repositories using DataMed. *Nat Genet* 2017;**49**: 816–819.
37. Allen B, Agarwal S, Kalpathy-Cramer J, Dreyer K. Democratizing AI. *J Am Coll Radiol* 2019;**16**:961–963.
38. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**:D158–D169.
39. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007;**39**:1181–1186.
40. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, Parkinson H, Schriml LM. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 2015;**43**:D1071–8.
41. Millar J. The need for a global language—SNOMED CT introduction. *Stud Health Technol Inform* 2016;**225**:683–685.
42. Schomburg I, Jeske L, Ulbrich M, Placzek S, Chang A, Schomburg D. The BRENDA enzyme information system—from a database to an expert system. *J Biotechnol* 2017;**261**:194–206.
43. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2017;**45**:D833–D839.
44. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, Sajed T, Johnson D, Li C, Karu N, Sayeeda Z, Lo E, Assempour N, Berjanskii M, Singhal S, Arndt D, Liang Y, Badran H, Grant J, Serra-Cayuela A, Liu Y, Mandal R, Neveu V, Pon A, Knox C, Wilson M, Manach C, Scalbert A. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* 2018;**46**:D608–D617.
45. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 2014;**42**:D358–D363.
46. Ross KE, Huang H, Ren J, Arighi CN, Li G, Tudor CO, Lv M, Lee J-Y, Chen S-C, Vijay-Shanker K, Wu CH. iPTMnet: integrative bioinformatics for studying PTM networks. *Methods Mol Biol* 2017; 333–353.
47. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J, Williams W, Case J, Maloney P. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem* 2003;**49**: 624–633.
48. Ruzicka L, Howe DG, Ramachandran S, Toro S, Van Slyke CE, Bradford YM, Eagle A, Fashena D, Frazer K, Kalita P, Mani P, Martin R, Moxon ST, Paddock H, Pich C, Schaper K, Shao X, Singer A, Westerfield M. The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources. *Nucleic Acids Res* 2019;**47**:D867–D873.
49. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. *Nucleic Acids Res* 2019:gkz1031.
50. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* 2018;**46**:D661–D667.
51. McCray AT. Better access to information about clinical trials. *Ann Intern Med* 2000; **133**:609–614.
52. Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, Campbell DS, Bernal-Llinares M, Okuda S, Kawano S, Moritz RL, Carver JJ, Wang M, Ishihama Y, Bandeira N, Hermjakob H, Vizcaíno JA. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* 2017;**45**:D1100–D1106.
53. Brody JA, Morrison AC, Bis JC, O'Connell JR, Brown MR, Huffman JE, Ames DC, Carroll A, Conomos MP, Gabriel S, Gibbs RA, Gogarten SM, Gupta N, Jaquish CE, Johnson AD, Lewis JP, Liu X, Manning AK, Papanicolaou GJ, Pitsillides AN, Rice KM, Salerno W, Sitlani CM, Smith NL, Heckbert SR, Laurie CC, Mitchell BD, Vasan RS, Rich SS, Rotter JI, Wilson JG, Boerwinkle E, Psaty BM, Cupples LA; CHARGE Analysis and Bioinformatics Working Group. Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat Genet* 2017; **49**:1560–1563.
54. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, Pitsillides AN, LeFaive J, Lee S-B, Tian X, Browning BL, Das S, Emde A-K, Clarke WE, Loesch DP, Shetty AC, Blackwell TW, Smith AV, Wong Q, Liu X, Conomos MP, Bobo DM, Aguet F, Albert C, Alonso A, Ardlie KG, Arking DE, Aslibekyan S, Auer PL, Barnard J, Barr RG, Barwick L, Becker LC, Beer RL, Benjamin EJ, Bielak LF, Blangero J, Boehnke M, Bowden DW, Brody JA, Burchard EG, Cade BE, Casella JF, Chalazan B, Chasman DI, Chen Y-DI, Cho MH, Choi SH, Chung MK, Clish CB, Correa A, Curran JE, Custer B, Darbar D, Daya M, de Andrade M, DeMeo DL, Dutcher SK, Ellinor PT, Emery LS, Eng C, Fatkin D, Fingerlin T, Forer L, Fornage M, Franceschini N, Fuchsberger C, Fullerton SM, Germer S, Gladwin MT, Gottlieb DJ, Guo X, Hall ME, He J, Heard-Costa NL, Heckbert SR, Irvin MR, Johnsen JM, Johnson AD, Kaplan R, Kardia SLR, Kelly T, Kelly S, Kenny EE, Kiel DP, Klemmer R, Konkle BA, Kooperberg C, Köttgen A, Lange LA, Lasky-Su J, Levy D, Lin X, Lin K-H, Liu C, Loos RJF, Garman L, Gerszten R, Lubitz SA, Lunetta KL, Mak ACY, Manichaikul A, Manning AK, Mathias RA, McManus DD, McGarvey ST, Meigs JB, Meyers DA, Mikulla JL, Minear MA, Mitchell BD, Mohanty S, Montasser ME, Montgomery C, Morrison AC, Murabito JM, Natale A, Natarajan P, Nelson SC, North KE, O'Connell JR, Palmer ND, Pankratz N, Peloso GM, Peyser PA, Pleiness J, Post WS, Psaty BM, Rao DC, Redline S, Reiner AP, Roden D, Rotter JI, Ruczinski I, Sarnowski C, Schoenherr S, Schwartz DA, Seo J-S, Seshadri S, Sheehan VA, Sheu WH, Shoemaker MB, Smith NL, Smith JA, Sotoodehnia N, Stilp AM, Tang W, Taylor KD, Telen M, Thornton TA, Tracy RP, Van Den Berg DJ, Vasan RS, Viaud-Martinez KA, Vrieze S, Weeks DE, Weir BS, Weiss ST, Weng L-C, Willer CJ, Zhang Y, Zhao X, Arnett DK, Ashley-Koch AE, Barnes KC, Boerwinkle E, Gabriel S, Gibbs R, Rice KM, Rich SS, Silverman EK, Qasba P, Gan W, Papanicolaou GJ, Nickerson DA, Browning SR, Zody MC, Zöllner S, Wilson JG, Cupples LA, Laurie CC, Jaquish CE, Hernandez RD, O'Connor TD, Abecasis GR; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021;**590**:290–299.
55. Atlassian. Bitbucket; 2020. https://bitbucket.org/ (6 March 2020, date last accessed).
56. GitLab. GitLab; 2020. https://about.gitlab.com/ (21 February 2020, date last accessed).
57. Kuo AM-H. Opportunities and challenges of cloud computing to improve health care services. *J Med Internet Res* 2011;**13**:e67.
58. Sobeslav V, Maresova P, Krejcar O, Franca TCC, Kuca K. Use of cloud computing in biomedicine. *J Biomol Struct Dyn* 2016;1–10.
59. Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. *Biol Direct* 2012;**7**:43.
60. Kass-Hout TA, Stevens LM, Hall JL. American Heart Association precision medicine platform. *Circulation* 2018;**137**:647–649.
61. Khomtchouk BB, Vand KA, Koehler WC, Tran D-T, Middlebrook K, Sudhakaran S, Nelson CS, Gozani O, Assimes TL. HeartBioPortal: an internet-of-omics for human cardiovascular disease data. *Circ Genomic Precis Med* 2019; **12**:baaa115.
62. Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Käll L, Lehtiö J, Lukasse P, Moerland PD, Griffin TJ. Multi-omic data analysis using Galaxy. *Nat Biotechnol* 2015;**33**:137–139.
63. Lalowski MM, Björk S, Finckenberg P, Soliymani R, Tarkia M, Calza G, Blokhina D, Tulokas S, Kankainen M, Lakkisto P, Baumann M, Kankuri E, Mervaala E. Characterizing the key metabolic pathways of the neonatal mouse heart using a quantitative combinatorial omics approach. *Front Physiol* 2018;**9**:365.
64. Chan SY, Loscalzo J. The emerging paradigm of network medicine in the study of human disease. *Circ Res* 2012;**111**:359–374.
65. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JLY, Toufighi K, Mostafavi S, Prinz J, St. Onge RP, VanderSluis B, Makhnevych T, Vizeacoumar FJ, Alizadeh S, Bahr S, Brost RL, Chen Y, Cokol M, Deshpande R, Li Z, Lin Z-Y, Liang W, Marback M, Paw J, San Luis B-J, Shuteriqi E, Tong AHY, van Dyk N. The genetic landscape of a cell. *Science* 2010;**327**:425–431.
66. Lusis AJ, Weiss JN. Cardiovascular networks: systems-based approaches to cardiovascular disease. *Circulation* 2010;**121**:157–170.
67. Hossain ME, Uddin S, Khan A, Moni MA. A framework to understand the progression of cardiovascular disease for type 2 diabetes mellitus patients using a network approach. *Int J Environ Res Public Health* 2020;**17**:596.

68. Xiong Y, Ruan L, Guo M, Tang C, Kong X, Zhu Y, Wang W. Predicting disease-related associations by heterogeneous network embedding. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018. pp. 548–555.
69. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, Schneider R, Bagos PG. Using graph theory to analyze biological networks. *BioData Min* 2011;**4**:10.
70. Wu R, Lin Y, Liu X, Zhan C, He H, Shi M, Jiang Z, Shen B. Phenotype–genotype network construction and characterization: a case study of cardiovascular diseases and associated non-coding RNAs. *Database* 2020;**2020**:baz147.
71. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018;**50**:1219–1224.
72. Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, Fath DM, Sandoval E, Isaksson M, Schlauch KA, Grzymski JJ, Lu JT, Washington NL. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat Commun* 2020;**11**:542.
73. Elliott J, Bodinier B, Bond TA, Chadeau-Hyam M, Evangelou E, Moons KGM, Dehghan A, Muller DC, Elliott P, Tzoulaki I. Predictive accuracy of a polygenic risk score–enhanced prediction model vs a clinical risk score for coronary artery disease. *JAMA* 2020;**323**:636–645.
74. Jamal S, Ali W, Nagpal P, Grover S, Grover A. Computational models for the prediction of adverse cardiovascular drug reactions. *J Transl Med* 2019;**17**:171.
75. Wei C-H, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* 2019;**47**: W587–W593.
76. Allot A, Chen Q, Kim S, Vera Alvarez R, Comeau DC, Wilbur WJ, Lu Z. LitSense: making sense of biomedical literature at sentence level. *Nucleic Acids Res* 2019;**47**: W594–W599.
77. Penning de Vries BBL, van Smeden M, Rosendaal FR, Groenwold RHH. Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice. *J Clin Epidemiol* 2020;**121**:55–61.
78. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev* 2015;**4**:5.
79. Liem DA, Murali S, Sigdel D, Shi Y, Wang X, Shen J, Choi H, Caufield JH, Wang W, Ping P, Han J. Phrase mining of textual data to analyze extracellular matrix protein patterns across cardiovascular disease. *Am J Physiol Heart Circ Physiol* 2018;**315**: H910–H924.
80. Caufield JH, Liem DA, Garlid AO, Zhou Y, Watson K, Bui AAT, Wang W, Ping P. A metadata extraction approach for clinical case reports to enable advanced understanding of biomedical concepts. *J Vis Exp* 2018;e58392.
81. Caufield JH, Zhou Y, Garlid AO, Setty SP, Liem DA, Cao Q, Lee JM, Murali S, Spendlove S, Wang W, Zhang L, Sun Y, Bui A, Hermjakob H, Watson KE, Ping P. A reference set of curated biomedical data and metadata from clinical case reports. *Sci Data* 2018;**5**:180258.
82. Altman DG. Making research articles fit for purpose: structured reporting of key methods and findings. *Trials* 2015;**16**:53.
83. Johnson EM, Gage KL, Feuerlein S, Jeong D. Cardiac magnetic resonance for the evaluation of suspected cardiac thrombus: conventional and emerging techniques. *J Vis Exp* 2019;**148**:e58808.
84. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M; for the STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med* 2007;**4**:e297.
85. Ping P, Watson K, Han J, Bui A. Individualized knowledge graph. *Circ Res* 2017;**120**: 1078–1080.
86. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med* 2010;**2**:57cm29.
87. De Pietro C, Francetic I. E-health in Switzerland: the laborious adoption of the federal law on electronic health records (EHR) and health information exchange (HIE) networks. *Health Policy* 2018;**122**:69–74.
88. Zhang H, Han BT, Tang Z. Constructing a nationwide interoperable health information system in China: the case study of Sichuan Province. *Health Policy Technol* 2017; **6**:142–151.
89. Myrick K, Ogburn D, Ward B. *Percentage of Office-Based Physicians using any Electronic Health Record (EHR)/Electronic Medical Record (EMR) System and Physicians That Have a Certified EHR/EMR System, by U.S. State: National Electronic Health Records Survey, 2017*. National Center for Health Statistics; 2019.
90. 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.
91. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, O'Leary DH, Psaty B, Rautaharju P, Tracy RP, Weiler PG. The cardiovascular health study: design and rationale. *Ann Epidemiol* 1991;**1**:263–276.
92. Bild DE, Detrano R, Peterson D, Guerci A, Liu K, Shahar E, Ouyang P, Jackson S, Saad MF. Ethnic differences in coronary calcification: the Multi-Ethnic Study of Atherosclerosis (MESA). *Circulation* 2005;**111**:1313–1320.
93. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med* 2018;**378**:981–983.
94. U.S. National Library of Medicine. Medical Subject Headings (MeSH) Fact Sheet. https://www.nlm.nih.gov/mesh/ (30 March 2018, date last accessed).
95. CDC/National Center for Health Statistics. ICD-10-CM Official Guidelines for Coding and Reporting; 2017.
96. The Lancet. ICD-11. *Lancet* 2019;**393**:2275.
97. Feigin V, Norrving B, Sudlow CLM, Sacco RL. Updated criteria for population-based stroke and transient ischemic attack incidence studies for the 21st century. *Stroke* 2018;**49**:2248–2255.
98. Franklin RCG, Béland MJ, Colan SD, Walters HL, Aiello VD, Anderson RH, Bailliard F, Boris JR, Cohen MS, Gaynor JW, Guleserian KJ, Houyel L, Jacobs ML, Juraszek AL, Krogmann ON, Kurosawa H, Lopez L, Maruszewski BJ, St. Louis JD, Seslar SP, Srivastava S, Stellin G, Tchervenkov CI, Weinberg PM, Jacobs JP. Nomenclature for congenital and paediatric cardiac disease: the International Paediatric and Congenital Cardiac Code (IPCCC) and the Eleventh Iteration of the International Classification of Diseases (ICD-11). *Cardiol Young* 2017;**27**:1872–1938.
99. Barton A, Rosier A, Burgun A, Ethier J-F. The cardiovascular disease ontology. In: Garbacz P, Kutz O eds. Formal Ontology in Information Systems. Amsterdam, Netherlands: IOS Press; 2014. pp. 409–414.
100. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from. *Sci Rep* 2017;**7**:5994.
101. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, Green A, Khankhanian P, Baranzini SE. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* 2017;**6**:e26726.
102. Gabrieli ER, Merrill FD. Computerizing a cardiology practice: condensing narrative text. *Proc Symp Comput Appl Med Care* 1980;**2**:841–843.
103. Weng W-H, Wagholikar KB, McCray AT, Szolovits P, Chueh HC. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med Inform Decis Mak* 2017;**17**:155.
104. Small AM, Kiss DH, Zlatsin Y, Birtwell DL, Williams H, Guerraty MA, Han Y, Anwaruddin S, Holmes JH, Chirinos JA, Wilensky RL, Giri J, Rader DJ. Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease. *J Biomed Inform* 2017;**72**: 77–84.
105. Nath C, Albaghdadi MS, Jonnalagadda SR. A natural language processing tool for large-scale data extraction from echocardiography reports. *PLoS One* 2016;**11**: e0153749.
106. Burger G, Abu-Hanna A, N de K, Cornet R. Natural language processing in pathology: a scoping review. *J Clin Pathol* 2016;**69**:949–955.
107. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, Crockett SD, Gourevitch RA, Dean KM, Mehrotra A. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc*; 2017;**24**:986–991.
108. Alnazzawi N, Thompson P, Batista-Navarro R, Ananiadou S. Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. *BMC Med Inform Decis Mak* 2015;**15**:S3.
109. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016;**6**: 26094.
110. Liu J, Zhang Z, Razavian N. Deep EHR: chronic disease prediction using medical notes. In: *Proceedings of the 3rd Machine Learning for Healthcare Conference, PMLR*; 2018. pp. 440–464.
111. Williams SA, Kivimaki M, Langenberg C, Hingorani AD, Casas JP, Bouchard C, Jonasson C, Sarzynski MA, Shipley MJ, Alexander L, Ash J, Bauer T, Chadwick J, Datta G, DeLisle RK, Hagar Y, Hinterberg M, Ostroff R, Weiss S, Ganz P, Wareham NJ. Plasma protein patterns as comprehensive indicators of health. *Nat Med* 2019; **25**:1851–1857.
112. Mortazavi BJ, Bucholz EM, Desai NR, Huang C, Curtis JP, Masoudi FA, Shaw RE, Negahban SN, Krumholz HM. Comparison of machine learning methods with national cardiovascular data registry models for prediction of risk of bleeding after percutaneous coronary intervention. *JAMA Netw Open* 2019;**2**:e196835.
113. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017;**12**:e0174944.
114. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, Jacoby DL, Masoudi FA, Spertus JA, Krumholz HM. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC Heart Fail* 2020;**8**:12–21.
115. Xiao Y, Fang R. RFMiner: risk factors discovery and mining for preventive cardiovascular health. In: *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE; 2017. pp. 278–279.
116. Hyppönen E, Mulugeta A, Zhou A, Santhanakrishnan VK. A data-driven approach for studying the role of body mass in multiple diseases: a phenome-wide registry-based case-control study in the UK Biobank. *Lancet Digit Health* 2019;**1**:e116–e126.
117. Pickhardt PJ, Graffy PM, Zea R, Lee SJ, Liu J, Sandfort V, Summers RM. Automated CT biomarkers for opportunistic prediction of future cardiovascular events and mortality in an asymptomatic screening population: a retrospective cohort study. *Lancet Digit Health* 2020;**2**:e192–e200.

118. Samad MD, Ulloa A, Wehner GJ, Jing L, Hartzel D, Good CW, Williams BA, Haggerty CM, Fornwalt BK. Predicting survival from large echocardiography and electronic health record datasets. *JACC Cardiovasc Imaging* 2019;**12**:681–689.
119. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, Andreini D, Budoff MJ, Cademartiri F, Callister TQ, Chang H-J, Chinnaiyan K, Chow BJW, Cury RC, Delago A, Gomez M, Gransar H, Hadamitzky M, Hausleiter J, Hindoyan N, Feuchtner G, Kaufmann PA, Kim Y-J, Leipsic J, Lin FY, Maffei E, Marques H, Pontone G, Raff G, Rubinshtein R. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J* 2016;ehw188.
120. Commandeur F, Slomka PJ, Goeller M, Chen X, Cadet S, Razipour A, McElhinney P, Gransar H, Cantu S, Miller RJH, Rozanski A, Achenbach S, Tamarappoo BK, Berman DS, Dey D. Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study. *Cardiovasc Res* 2019;cvz321.
121. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, Reich CG, Duke J, Madigan D, Hripcsak G, Ryan PB. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet* 2019;**394**:1816–1826.
122. Patterson OV, Freiberg MS, Skanderson M, J. Fodeh S, Brandt CA, DuVall SL. Unlocking echocardiogram measurements for heart disease research through natural language processing. *BMC Cardiovasc Disord* 2017;**17**:151.
123. Diller G-P, Kempny A, Babu-Narayan SV, Henrichs M, Brida M, Uebing A, Lammers AE, Baumgartner H, Li W, Wort SJ, Dimopoulos K, Gatzoulis MA. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. *Eur Heart J* 2019;**40**:1069–1077.
124. Patel J, Mowery D, Krishnan A, Thyvalikakath T. Assessing information congruence of documented cardiovascular disease between electronic dental and medical records. *AMIA Annu Symp Proc AMIA Proc* 2018;**2018**:1442–1450.
125. Bacchi S, Oakden-Rayner L, Zerner T, Kleinig T, Patel S, Jannes J. Deep learning natural language processing successfully predicts the cerebrovascular cause of transient ischemic attack-like presentations. *Stroke* 2019;**50**:758–760.
126. Gennatas ED, Friedman JH, Ungar LH, Pirracchio R, Eaton E, Reichmann LG, Interian Y, Luna JM, Simone CB, Auerbach A, Delgado E, Mj van der L, Solberg TD, Valdes G. Expert-augmented machine learning. *Proc Natl Acad Sci USA* 2020;**117**:4571–4577.
127. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Med Inform* 2019;**7**:e12239.
128. Martens L, Vizcaíno JA. A golden age for working with public proteomics data. *Trends Biochem Sci* 2017;**42**:333–341.
129. Loring Z, Mehrotra S, Piccini JP, Camm J, Carlson D, Fonarow GC, Fox KAA, Peterson ED, Pieper K, Kakkar AK. Machine learning does not improve upon traditional regression in predicting outcomes in atrial fibrillation: an analysis of the ORBIT-AF and GARFIELD-AF registries. *Europace* 2020;**22**:1635–1644.
130. Li Y, Sperrin M, Ashcroft DM, Staa T. V. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. *BMJ* 2020;m3919.
131. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, Bhatt DL, Fonarow GC, Laskey WK. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol* 2017;**2**:204–209.
132. Peek N, Holmes JH, Sun J. Technical challenges for big data in biomedicine and health: data sources, infrastructure, and analytics. *Yearb Med Inform* 2014;**23**:42–47.
133. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak* 2019;**19**:146.
134. Ballinger B, Hsieh J, Singh A, Sohoni N, Wang J, Tison G, Marcus G, Sanchez J, Maguire C, Olgin J, Pletcher M. DeepHeart: semi-supervised sequence learning for cardiovascular risk prediction. *Proc AAAI Conf Artif Intell* 2018;**32**:1.
135. Glicksberg BS, Li L, Badgeley MA, Shameer K, Kosoy R, Beckmann ND, Pho N, Hakenberg J, Ma M, Ayers KL, Hoffman GE, Dan Li S, Schadt EE, Patel CJ, Chen R, Dudley JT. Comparative analyses of population-scale phenomic data in electronic medical records reveal race-specific disease networks. *Bioinformatics* 2016;**32**:i101–i110.
136. Alageel S, Gulliford MC. Health checks and cardiovascular risk factor values over six years' follow-up: matched cohort study using electronic health records in England. *PLoS Med* 2019;**16**:e1002863.
137. Malin B, Goodman K; Section Editors for the IMIA Yearbook Special Section. Between access and privacy: challenges in sharing health data. *Yearb Med Inform* 2018;**27**:55–059.
138. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Sorlie P, Stone NJ, Wilson PWF. 2013 ACC/AHA Guideline on the assessment of cardiovascular risk. *J Am Coll Cardiol* 2014;**63**:2935–2959.
139. Rosengren A, Smyth A, Rangarajan S, Ramasundarahettige C, Bangdiwala SI, AlHabib KF, Avezum A, Bengtsson Boström K, Chifamba J, Gulec S, Gupta R, Igumbor EU, Iqbal R, Ismail N, Joseph P, Kaur M, Khatib R, Kruger IM, Lamelas P, Lanas F, Lear SA, Li W, Wang C, Quiang D, Wang Y, Lopez-Jaramillo P, Mohammadifard N, Mohan V, Mony PK, Poirier P, Srilatha S, Szuba A, Teo K, Wielgosz A, Yeates KE, Yusoff K, Yusuf R, Yusufali AH, Attaei MW, McKee M, Yusuf S. Socioeconomic status and risk of cardiovascular disease in 20 low-income, middle-income, and high-income countries: the Prospective Urban Rural Epidemiologic (PURE) study. *Lancet Glob Health* 2019;**7**:e748–e760.
140. Pfohl S, Marafino B, Coulet A, Rodriguez F, Palaniappan L, Shah NH. Creating fair models of atherosclerotic cardiovascular disease risk. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society—AIES '19*. Honolulu, HI, USA: ACM Press; 2019. pp. 271–278.
141. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;**169**:866–872.
142. Sage Bionetworks. Synapse. https://www.synapse.org/ (2 April 2020, date last accessed).
143. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. Good enough practices in scientific computing. *PLoS Comput Biol* 2017;**13**:e1005510. Ouellette F, ed.
144. Cimino JJ. Putting the "why" in "EHR": capturing and coding clinical cognition. *J Am Med Inform Assoc* 2019;**26**:1379–1384.
145. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;**319**:1317–1318.