Briefings in Bioinformatics, 2025, **26(5)**, bbaf526
https://doi.org/10.1093/bib/bbaf526
Review

Bottlenecks in advancing and applying multiomic data integration—common data resources as rate-limiting drivers—the high-impact use case of atherosclerotic cardiovascular disease

Stephanie Bezzina Wettinger[1,2,‡], Kanita Karaduzovic-Hadziabdic[3,‡], Ritienne Attard[1], Rosienne Farrugia[1,2], Brooke N. Wolford[4], Marco Chierici[5], Giuseppe Jurman[5,6], Panagiotis Alexiou[2], José L. Peñalvo[7], Rafael S. Costa[8], José Basílio[9,10], František Sabovčik[11], Rui Vitorino[12,13], Johannes A. Schmid[14], Rajesh Shigdel[15], Baiba Vilne[16], Artemis G. Hatzigeorgiou[17], Miron Sopic[18], Yvan Devaux[19], Paolo Magni[20,21], Maria Tellez-Plaza[7,*], David P. Kreil[22,*], Aleksandra Gruca[23,*]

[1]Department of Applied Biomedical Science, Faculty of Health Sciences, University of Malta, Msida, MSD2080, Malta
[2]Centre for Molecular Medicine and Biobanking, University of Malta, Msida, MSD2080, Malta
[3]International University of Sarajevo, Hrasnicka cesta 15, 71210, Ilidza, Sarajevo, Bosnia and Herzegovina
[4]Department of Public Health and Nursing, Norwegian University of Science and Technology, Mauritz Hanssens gate 2, Trondheim, Norway
[5]Data Science for Health, Fondazione Bruno Kessler, via Sommarive 18, 38123 Trento, Italy
[6]Department of Biomedical Sciences, Humanitas University, via Rita Levi Montalcini 4, 20072, Pieve Emanuele (MI), Italy
[7]National Center for Epidemiology, Carlos III Institute of Health, calle Melchor Fernández Almagro 5, 28029 Madrid, Spain
[8]LAQV-REQUIMTE, Department of Chemistry, NOVA School of Science and Technology, NOVA University Lisbon, Campus da Caparica, 2829-516 Caparica, Portugal
[9]Institute of Pathophysiology and Allergy Research, Center of Pathophysiology, Infectiology and Immunology, Medical University of Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria
[10]INESC ID, Instituto Superior Técnico, Universidade de Lisboa, R. Alves Redol 9, 1000-029 Lisbon, Portugal
[11]Unit of Hypertension and Cardiovascular Epidemiology, Department of Cardiovascular Sciences, KU Leuven, Edward van Evenstraat 3, 3000 Leuven, Belgium
[12]Department of Medical Sciences, iBiMED, University of Aveiro, 3810-193 Aveiro, Portugal
[13]RISE-Health, Department of Surgery and Physiology, Faculty of Medicine, University of Porto, 4200-319 Porto, Portugal
[14]Institute of Vascular Biology and Thrombosis Research, Center for Physiology and Pharmacology, Medical University of Vienna, Schwarzspanierstraße 17, Physiology Building, A-1090 Vienna, Austria
[15]Department of Global Public Health and Primary Care, University of Bergen, Alrek helseklynge, blokk D, Årstadveien 175009 Bergen, Norway
[16]Bioinformatics Group, Riga Stradins University, 16 Dzirciema Street, LV-1007, Riga, Latvia
[17]Department of Computer Science and Biomedical Informatics, University of Thessaly, Papasiopoulou 2-4, 35131 Galaneika – Lamia, Greece and Hellenic Pasteur Institute Vas. Sofias Av 127, 115 21, Greece
[18]Department of Medical Biochemistry, Faculty of Pharmacy, University of Belgrade, Vojvode Stepe 450, 11 000 Belgrade, Serbia
[19]Cardiovascular Research Unit, Department of Precision Health, Luxembourg Institute of Health, 1A-B rue Edison L-1445 Strassen, Luxembourg
[20]Department of Pharmacological and Biomolecular Sciences 'Rodolfo Paoletti', Università degli Studi di Milano, via Balzaretti 9, 20133 Milan, Italy
[21]IRCCS MultiMedica, via Milanese 300, 20099 Sesto San Giovanni, Milan, Italy
[22]Bioinformatics Research, Institute of Molecular Biotechnology, Boku University Vienna, Muthgasse 18, 1190 Vienna, Austria
[23]Department of Computer Science and Networks, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

*Corresponding authors. Maria Tellez-Plaza, Department of Biomedical Sciences, Humanitas University, via Rita Levi Montalcini 4, 20072, Pieve Emanuele (MI), Italy. E-mail: m.tellez@isciii.es; David Kreil, Bioinformatics Research, Institute of Molecular Biotechnology, Boku University Vienna, Muthgasse 18, 1190 Vienna, Austria. E-mail: David.kreil@boku.ac.at; Aleksandra Gruca, Department of Computer Science and Networks, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland. E-mail: aleksandra.gruca@polsl.pl.
‡These authors contributed equally to this work.

Abstract

Despite striking successes in identifying novel biomarkers for improved patient stratification and predicting disease progression, numerous challenges remain in the effective integration and exploitation of multiomic data in biomedical applications beyond cancer, for which most bioinformatics strategies are developed and validated. That focus on cancer severely limits the effective development and advancement of algorithms in machine learning and artificial intelligence that do not suffer degraded out-of-domain performance. Generalizability and interpretability of models, however, are also required for robust insights that may translate into clinical practice. Work across different independent datasets is critical for establishing models robust towards unwanted variation in assays, protocols, and cohort populations. Disease-specific context like ethnicity, socioeconomic background, sex, lifestyle, disease phase, and tissue type also strongly affect molecular profiles. We here discuss atherosclerotic cardiovascular disease (ASCVD) as a high-impact non-cancer use case for the challenges remaining in the development and application of the latest bioinformatics approaches to multiomics data

Received: February 18, 2025. **Revised:** June 13, 2025. **Accepted:** August 31, 2025

integration. ASCVD remains the leading cause of death globally. Disease aetiology, progression, and therapy outcome depend on a complex interplay of genetic, environmental, and lifestyle factors. Integrating these diverse data types effectively remains a challenge but holds transformative potential for personalized medicine. Discovery and access to data of sufficient diversity and extent form key bottlenecks. We here compile a first comprehensive overview of key data sets in ASCVD to complement the established cancer-focused resources as a foundation for future effective development and application of state-of-the-art bioinformatics tools for multiomic data integration.

Keywords: multiomic data integration; algorithm generalizability; data diversity; common data resources; atherosclerotic cardiovascular disease (ASCDV)

Effective multiomic data introduction

While significant progress has been made in identifying novel biomarkers for better patient stratification and predicting disease progression, many challenges remain in effectively integrating and utilizing multiomic data in biomedical applications beyond cancer. Most bioinformatics strategies are developed and validated specifically for cancer that severely limits the effective development and advancement of algorithms in machine learning and artificial intelligence that do not suffer degraded out-of-domain performance. For translation into clinical practice, however, models that generalize to new cohorts are required. For robust insights, interpretable models are needed. Both issues remain open areas of research. We here chose atherosclerotic cardiovascular disease (hereafter, ASCVD) to discuss a high-impact non-cancer use case for the challenges remaining in the development and application of the latest bioinformatics approaches to multiomics data integration.

ASCVD is the leading cause of death accounting for 85% of the 17.5 million cardiovascular deaths in 2019 (WHO, Cardiovascular Diseases (CVD)) [1]. The aetiology of atherosclerosis and its acute consequences like myocardial infarction and stroke are highly multifactorial, involving a complex genetic component [2] and a less explored set of environmental/lifestyle factors (diet, physical activity, stress levels, physical environment including pollutants) [3] that influence various organs and cell types.

Recent technological advancements across various research domains have led to the era of high-throughput approaches, resulting in the collection of data at an unprecedented scale and detail, commonly known as the Big Data era [4], which has the potential to transform health outcomes as it can disentangle the complexity involved. Sources of Big Data are diverse and can include both molecular data from 'omics' technologies, lifestyle, environmental, and socioeconomic factors defined as exposome which also influence disease risk and progression. The latter can be retrieved from epidemiological studies, national registries, or electronic health records (EHRs). Omics data, in turn, constitute the genome-scale profiles of particular types of biological data, such as mRNA or protein expression or DNA methylation (*cf.* Table 1). They can characterize samples from human patients, relevant animal/cell-based models or, in the case of microbiomes, a collection of organisms. The digital transformation in cardiovascular medicine is also accelerated with the introduction of electronic medical records, telemedicine, health-related smartphone applications, or wearable sensors, which enable the rapid collection of large amounts of complementary data.

Big Data Analysis requires powerful strategies to draw useful information from all the data. Machine learning (ML)/artificial intelligence (AI), in particular, can be used to learn from data, complementing more traditional statistics-based algorithms. Irrespective of model type, generalizability of models established in one or several related studies to new cohorts remains a challenge. A hope is that by integrating diverse and complementary multimodal data, patterns and associations that are challenging to discern in single-modal data will become easier to identify robustly. This can indeed be expected due to the different noise and bias characteristics of different modalities that thus contribute different views of a complex biological state, and this is particularly critical for the heterogeneous disease clusters encompassed by ASCVD. Analysing these complementary views jointly aims to improve our understanding of the diverse processes leading to ASCVD, whether these are molecular drivers or the influence of lifestyle on cardiovascular outcomes. Translational applications will enhance patient care through the discovery of biomarkers for patient stratification and personalized therapies, while accelerating the development of novel treatments by the identification and validation of novel potential drug targets.

Notwithstanding its promise, the field struggles with a number of challenges. While natural language processing and other AI/ML algorithms provide the tools to leverage information from unstructured clinical data [5, 6], differences between EHR platforms, inclusion of studies with EHR in non-English or non-Latin alphabet text, and varying utilization of free text by physicians are problematic. The field suffers from lack of interoperability of databases, different definitions applied in databases and registries, and different laboratory techniques and research subject recruitment criteria requiring harmonization. Additionally, attention to preanalytical variables and uniform data processing procedures are critical for most -omic analyses aside from genomics [7, 8]. Prospective studies, data sharing, international consortia, and deposition of large CVD-related datasets in a common database such as the Database of Genotypes and Phenotypes (dbGaP) [9] or the European Genome-Phenome Archive [10] will help mitigate some of these challenges facilitating clinical validation. Federated data analysis initiatives such as DataShield (https://datashield.org/) [11] may help overcome data security and privacy issues that are currently hindering collaborations.

Multiomic data in ASCVD research

The revolution in the availability of genomic data driven by the reduction of sequencing costs and exponential developments in molecular technologies resulted in an unprecedented accumulation of data availability for genomic, transcriptomic, epigenomic, proteomic, metabolomic, and lipidomic measurements even in the context of ASCVD. Multiomic studies integrate those diverse biological data layers to create a comprehensive view of biological processes and disease mechanisms as each omic data type provides complementary information and a more comprehensive view of patient health status.

The heterogeneity and complexity of data generated from different omic platforms pose significant challenges in terms of data integration. Multiomic data is typically highly dimensional and, whether analysed individually or in combination, such data require analytical techniques that can accommodate high-dimensionality of analysed datasets. Most of the methods for multiomic data analysis, however, are developed and validated on cancer data, reflecting the higher awareness repositories like the GDC data portal [12], which simplify discovery and

Table 1. Overview of selected -omic technologies

-Omic layer	Description
Genomics	Data from whole-genome sequencing can be from next-generation sequencing of DNA (short reads) and third-generation sequencing (long reads) There is also sequencing of the protein-coding parts alone (exome sequencing) and microarray data which includes polymorphisms spread across the genome
Epigenomics	Focuses on the various modifications of DNA bases and chromatin accessibility on a whole-genome scale
Transcriptomics	Sequences and levels of the entire repertoire of RNA transcripts in a given cell type, tissue, or condition May include both protein-coding and non-coding RNAs In bulk transcriptomics, data from tissues with various cell populations are analysed, whereas single-cell transcriptomics are used to study effects in different cell types and states
Epitranscriptomics	Addresses the multiple chemical modifications to RNA molecules that influence their structure and function
Proteomics	Focuses on the variety of proteins produced by cells or tissues under certain conditions and on post-translational protein modifications (PTMs) such as phosphorylation
Metabolomics	Explores the variety of molecules produced by metabolic pathways, including lipids (lipidomics)
Microbiomics	Characterizes the complex microbial communities which are present in humans and impact health and disease including ASCVD

download of relevant datasets. Data from other domains, such as cardiovascular disease, exhibit different biomedical patterns and confounders, leading to critical problems in algorithm generalization and out-of-domain performance. For example, structural and compositional ASCVD data that derive from imaging of vascular cross-sections can complement omics data, in which case, optical measurements of plaque morphology and of regional cellular composition further increase the dimensionality challenges of analysis [8, 13]. In this section, we discuss specific challenges related to analysing ASCVD. We also provide a first comprehensive overview of common datasets in ASCVD to complement the established cancer-focused resources as a foundation for future effective development and application of bioinformatics methods for multiomic data analysis.

While in this review we focus on multiomic data integration from bulk samples, where omics data represent measurements averaged across many cell populations, it is worth noting the advancements in single-cell technologies and the increasing availability of single-cell data, which are discussed at length in recent technological reviews, covering transcriptomic, epigenomic, and proteomic profiles [14]. Despite the promise of additional high-resolution insights, single-cell data also bring a variety of biological and technical sources of noise, requiring sophisticated analysis tools [15].

Single-cell, single-modality analyses have also successfully been applied in the context of ASCVD, such as in recent studies advancing our understanding of atherosclerotic plaque composition and cellular heterogeneity, uncovering distinct cellular subpopulations within plaques, such as macrophage subsets (e.g. pro-inflammatory M1 and anti-inflammatory M2), smooth muscle cell phenotypes, and T-cell diversity. By analysing individual cells within plaques, researchers have identified distinct macrophage subpopulations, such as pro-inflammatory and foam cell phenotypes [16], as well as smooth muscle cell transitions contributing to plaque stability or vulnerability [17, 18]. Additionally, scRNA-Seq has revealed the role of endothelial cell activation and T-cell diversity in driving inflammation and plaque progression [19]. ScRNA-Seq technology has since been used to construct an atherosclerosis cell atlas of mice and humans, including immune cells, such as monocytes/-macrophages, dendritic cells, T cells, B cells, and natural killer cells, and non-immune cells, such as vascular smooth muscle cells, endothelial cells, pericytes, and fibroblasts [20].

Computational methods have subsequently also been developed for multiomic data integration at the single-cell level [21] that have been applied to large-scale cancer data successfully [15, 22]. The latest computational approaches have also explored multimodal generative models for single-cell data trained on related corpuses [23]. These methods are just beginning to be applied also to ASCVD [24, 25], and it will be interesting to see how such approaches translate to ASCVD cohorts more generally.

ASCVD-focused data sources

There is an increasing number of resources available when it comes to datasets for ASCVD. These include specific ASCVD datasets with data from patients with ASCVD or its effects from major studies such as the Framingham Heart Study [26], the Atherosclerosis Risk in Communities (ARIC) study [27], the Early-Onset Myocardial Infarction (EOMI) study [28], and many others. There are also plaque datasets such as Athero-Express [29] and the Biobank of Karolinska Endarterectomy (BIKE) [30]. Then, there are large population studies which are not specifically on ASCVD, but which include data on ASCVD-related phenotypes—such as the UK Biobank [31], FinnGen [32], the China Kadoorie Biobank [33], and TOPMED [34]. Finally, advancements in single-cell technologies now allow profiling at single-cell resolution. Resources such as the Athero-Express and BIKE biobanks have begun to incorporate single-cell RNA sequencing (scRNA-Seq) datasets into the context of atherosclerotic plaque biology; however, the amount of samples available is still much lower than for bulk omics data, and this limitation is exacerbated by single-cell data being noisier and thus requiring more samples for sufficiently powered statistical analyses.

A description of the major datasets, the sample sizes, and different omics data available for each one, the different sample sources, and the means of accessing the data are shown in Table 2. Though there are numerous ASCVD databases, different collections have been set up to answer different research questions, by different groups, and at different time points. The data available vary by type of sample collected (peripheral blood, local blood, plaque), the disease phenotype used to classify cases (patients undergoing endarterectomy, early-onset MI, atherosclerosis, coronary artery disease, coronary heart disease, and stroke), and the -omic technology used for data generation (GWAS versus exomes

versus genomes; bulk RNA-Seq versus scRNA-Seq; targeted versus untargeted proteomics and metabolomics). Thus, there is no standardization between the collections making integrated analysis across different datasets more complex since this variability needs to be accounted for. Furthermore, there are very few datasets for which all -omic layers are available, and for most datasets, -omic data have not been generated for the entire collection yet. This, unfortunately, limits integrative analysis within datasets due to missing layers or decreased sample sizes. A further consideration is whether the different -omic data layers were generated from the same sample or from samples collected at different time points. This information is not generally readily available, however, it is a very important consideration when analysing RNA, protein, and metabolite data. Of these data types, the transcriptome has been more widely studied and there are multiple reports highlighting the effect of preanalytical variables on RNA levels [7] highlighting an additional layer of variability that needs to be accounted for during integrative analysis.

The challenges of working with ASCVD data

Analysis of ASCVD multiomic data has a number of unique considerations that are not necessarily relevant to many other conditions.

ASCVD is a dynamic process with distinct phases and each phase has independent characteristics [35, 36], making comparisons of variables such as RNA, protein, and metabolites challenging. Prior to plaque rupture, the individual has a blood composition and a genetic background that contribute to plaque development and is representative of risk. Blood cell composition, cytokine production, and endothelial damage all contribute to the developing plaque which leads to changes in cell type composition, cytokine, and lipid content. Upon plaque rupture, the contents of the plaque are released into the bloodstream triggering atherothrombosis and platelet activation with blood clots blocking blood flow to the heart or brain. This, in turn, leads to necrosis which can result in further changes in blood composition. Thus, except for static genomic sequence data, timing of dynamic samples like blood sampling with respect to acute events (pre-acute phase, during the acute phase, post-acute phase, and timing from the event which can be stroke or myocardial infarction) is crucial when studying omics data as different times from event will result in considerable differences in relevant molecules and cell types. Additionally, it is becoming clearer that plaque development is not a uniform process but can proceed in waves of activation depending on insults that increase progression [37], and this needs to be taken account already at the stage of study design.

Testing for multiomic studies is best conducted on the same sampling. Ideally, considering that atherosclerosis is a dynamic process, sampling should be repeated at various intervals, but generally this is not possible due to the high costs involved.

There are several tissues that are relevant to ASCVD, the most important being the atheroma in arterial walls. This is generally inaccessible, though some studies on plaque or atherosclerotic tissue do exist (e.g. Athero-Express [38], the Stockholm–Tartu Atherosclerosis Reverse Network Engineering Task STARNET [39]). Most studies in humans, however, focus on systemic effects caused by changes in cell types and molecules in the circulation. Other organs will also suffer changes, particularly those associated with consequences of risk factors for ASCVD such as the pancreas, liver, and kidneys, whereas the involvement of yet other organs and tissues such as bone marrow is also being recognized as relevant [37]. Even different parts of a plaque and different plaque types have been shown to have different cellularity and protein composition which has implications for multiomic studies [40].

Compared to other conditions, systemic differences in RNA, protein, and metabolite expression might not be as severe as e.g. in cancer or sepsis. This may be further complicated by the fact that the aetiology of ASCVD is very varied and one may expect this to be reflected in the multiomic data.

Risk factors for ASCVD and even symptoms of acute events such as myocardial infarction have been shown to differ in men and women. Women tend to suffer from myocardial infarction at an older age than men, on average ~10 years older. This could be because of different effects of oestrogen and testosterone on risk for ASCVD, besides different risk factors such as response to stress, diet, and other lifestyle factors [41, 42]. It is therefore always critical that datasets record sex and present results and data by sex [43]. For women, menopausal status and day of ovarian cycle are also expected to influence multiomic data (except for genomic sequence data). Considering all these factors, the importance of good metadata describing not only sample types, tissues or cell types, methods of analysis, and year of sample collection, but also stage of disease and the timing of sampling with respect to acute events cannot be overemphasized. Ethnicity is also a highly relevant factor and should be recorded in metadata.

Furthermore, ASCVD is highly complex in which environmental and lifestyle factors have a strong influence on risk factors and are known to influence levels of -omic variables, at least in some tissues such as blood [44]. Therefore, ideally, factors such as family history, smoking status, diabetes, use of relevant medications such as statins, anthropometric measurements indicating e.g. obesity, history of previous cardiometabolic complications during gestation, ethnicity, and other relevant factors (besides of course sex and age) that can modify risk should be included in databases to allow for their inclusion in data analysis if deemed necessary [8].

An issue that affects all -omic studies is the lack of data from many populations. There is a bias in the populations studied, and the datasets that are available come mainly from developed countries, while other regions of the world are totally unrepresented. It has been recommended to include more diverse populations for omics studies [45]. Africans e.g. have greater genetic variation but are largely understudied, as are many Mediterranean populations. Additionally, environmental and lifestyle influences vary widely across the world.

Multiomic data sharing

Data sharing is crucial in biomedical sciences as collaboration is especially important for understanding complex diseases, developing new treatments, and improving patient outcomes. It is particularly important to advance research related to rare diseases, such as rare cardiac conditions [46], which requires compilation of resources from different experimental and clinical centres to allow meaningful research conclusions [47]. For example, meta-analysis of genome-wide association datasets can uncover rare cardiovascular risk factors that individual studies might miss [45].

Sharing datasets across different studies requires ensuring that all measurements are standardized and normalized which is essential for multiomic data to be comparable across different studies [48]. To ensure reproducibility and interpretability across different studies and platforms, each dataset needs to be accompanied with relevant metadata describing information about the experimental design. The metadata reporting standard guidelines help researchers on how to correctly document

Table 2. List of the major ASCVD-specific datasets including different omics data types, the sample sizes, and access information

Databases and links	Brief description/types of data	Samples	Access
Disease-specific databases			
Athero-Express/UCC-SMART • https://www.umcutrecht.nl/en/ucc-smart	*An ongoing prospective cohort study and biobank with samples from patients undergoing femoral or carotid endarterectomy. Atherosclerotic plaques, harvested during surgery, are highly phenotyped to evaluate plaque characteristics in relation to long-term cardiovascular events* • Demographics and family history • Immunohistochemical phenotyping of plaques • Cardiovascular risk factors • Whole-genome genotyping ($n = 2200$ carotid, $n = 400$ aneurysm) • RNA-Seq ($n = 700$) • scRNA-Seq ($n = 50$) • Whole-genome methylation ($n = 700$ carotid) • OLINK proteomics ($n = 688$ carotid)	Plaque samples Coronary: 4696 Carotid: ∼2400 Femoral: 1100 Aneurysm: 650	Access upon request and review regarding alignment with study objectives, ethics, and permissions. Application via the study website
Biobank of Karolinska Endarterectomy (BIKE) • https://www.omicsdi.org/dataset/geo/GSE21545 • https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-GEOD-21545	*A biobank collection of carotid plaques and peripheral blood mononuclear cells from patients undergoing endarterectomy* • DNA genotyping ($n = 610$) • RNA-Seq (plaque $n = 470$, peripheral blood $n \sim 500$, local blood $n = 230$) • scRNA-Seq (plaque $n = 4$, peripheral blood $n = 4$, local blood $n = 4$) • RNA microarrays ($n = 127$ plaques, $n = 97$ PBMC) • Proteomics; LC–MS/MS ($n = 40$ plaques) • Proteomics; 5 OLINK panels (peripheral blood $n = 700$, local blood $n = 150$) • Metabolomics (peripheral blood $n = 700$, local blood $n = 150$)	Plaque samples Carotid: 1033 Peripheral blood Local blood (at site of carotid surgery)	Access through GEO: Accession ID: E-GEOD-21545 (microarray data)
Cardiovascular disease knowledge portal • https://cvd.hugeamp.org/ • https://cvd.hugeamp.org/datasets.html	*The Cardiovascular Disease Knowledge Portal enables browsing, searching, and analysis of human genetic information linked to myocardial infarction, atrial fibrillation, and related traits* • GWAS/Exome Chip/Exomes/WGS • Exomes • Single variant association signals • Common variant gene-level associations • Related pathways • Effector gene prediction • Gene correlations • Epigenomic annotations	Varies by condition	Mostly open access; varies by dataset
Myocardial infarction 2020 GWAS: trans-ancestry • https://kp4cd.org/node/1197	*A trans-ancestry meta-analysis of genetic associations for myocardial infarction from the UK Biobank and CARDIoGRAMplusC4D cohorts* • GWAS	Cases: 61 505 Controls: 577 716	Access through UK Biobank and CARDIoGRAMplusC4D

(*Continued*)

Table 2. Continued

Databases and links	Brief description/types of data	Samples	Access
NHLBI GO-ESP: Early-Onset Myocardial Infarction (EOMI) • https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000279.v2.p1 • dbGaP Study Accession: phs000279.v2.p1	*Sequencing data of cases with extremely early-onset MI drawn from 8 cohorts: PennCATH, Cleveland Clinic Genebank, Massachusetts General Hospital Premature Coronary Artery Disease Study (MGH-PCAD), Heart Attack Risk in Puget Sound (HARPS), and Translational Research Investigating Underlying Disparities in Myocardial Infarction Patients' Health Status (TRIUMPH), the Framingham Heart Study, the Women's Health Initiative, and the Atherosclerosis Risk in Communities Study.* • Exomes	Cases: 10 024 Controls: 12 047 Cases were selected based on MI occurring in men aged ≤50 years and women aged ≤60 years	Access upon request; limited to tenure-track (or equivalent) senior investigators; requires an eRA Commons account
Atherosclerosis Risk in Communities (ARIC) Cohort • https://www.nhlbi.nih.gov/science/atherosclerosis-risk-communities-aric-study • https://aric.cscc.unc.edu/aric9/ • https://biolincc.nhlbi.nih.gov/studies/aric/ The ARIC Cohort is utilized in the following dbGaP sub-studies: • phs000090 GENEVA_ARIC • phs000223 PAGE_CALiCo_ARIC • phs000398 GO-ESP: HeartGo_ARIC • phs000668 CHARGE_ARIC • phs000860 MICORTEX • phs001536 CCDG_ARIC	*The Atherosclerosis Risk in Communities (ARIC) study started in 1987 and initially served to identify risk factors for subclinical atherosclerosis. The study included adults of black and white ethnicity between the ages of 45 and 65 who lived in four US communities. A total of 15 792 participants received an extensive examination, including medical, social, and demographic data. These participants attended 7 clinical examinations between 1987 and 2019 and participated in yearly follow-up telephone calls to assess their health status* *In the Community Surveillance Component, the same four communities were investigated to determine the community-wide occurrence of hospitalized myocardial infarction and coronary heart disease deaths in men and women aged 35–84 years. Hospitalized stroke is investigated in cohort participants only. Starting in 2006, the study conducted community surveillance of inpatient (ages ≥55 years) and outpatient heart failure (ages ≥65 years) for heart failure events beginning in 2005. Community Surveillance for non-cohorts ended in event year 2014* *As participants have aged (6000 of the original participants are now in their 80s and 90s), the study goals have shifted to focus on risk factors for heart diseases including heart attack (myocardial infarction), coronary heart disease, stroke, and heart failure. The study also aims to measure how heart disease risk factors, medical care, and health outcomes vary by race, ethnicity, sex, location, and time* • Epidemiological data • Clinical examination data • Clinical follow-up data every 3 years • SNP arrays • WGS/Exomes	~16 000 in the cohort component, followed for >35 years >400 000 in the community surveillance arm Availability of omics data varies by subproject	Access upon request through BioLINCC

(*Continued*)

Table 2. Continued

Databases and links	Brief description/types of data	Samples	Access
Centre for Common Disease Genomics [CCDG]—Cardiovascular: TexGen • https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs003010.v1.p1 • dbGaP Study Accession: phs003010.v1.p1	*A CCDG large-scale sequencing effort with a focus on early-onset heart disease in individuals from the TexGen study* • WGS	Cases: 6146 (personal or family history of CVD)	Through dbGAP authorized access: access upon request; limited to tenure-track (or equivalent) senior investigators; requires an eRA Commons account
CARDIoGRAMplusC4D • http://www.cardiogramplusc4d.org/ • https://millionhearts.hhs.gov/index.html • https://www.ebi.ac.uk/gwas/search?query=GCST90132314 • https://www.ebi.ac.uk/gwas/search?query=GCST90132315 • An interactive Manhattan plot (on RShiny): https://procardis.shinyapps.io/cadgen/	*CARDIoGRAMplusC4D (Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) together with The Coronary Artery Disease (C4D) Genetics) consortium represents a collaborative effort to combine data from multiple large-scale genetic studies to identify risk loci for coronary artery disease and myocardial infarction* *The analysis includes a meta-analysis of the CARDIoGRAMplusC4D data together with UK Biobank data and the CARDIoGRAM Exome (chip) data* • Various genotyping arrays with data imputed to the 1000 Genomes	CARDIoGRAMplusC4D Metabochip Cases: 63 746 Controls: 130 681 CARDIoGRAMplusC4D 1000 Genomes–based GWAS Cases: 60 801 (CAD) Controls: 123 504 (~55% overlap in the 2 datasets)	Summary data for the meta-analyses can be downloaded from the CARDIoGRAMplusC4D website
EPIC-CVD • https://epic.iarc.fr/research/epic-cvd/ • https://epic.iarc.fr/access/	*The cardiovascular component of the pan-European EPIC project. A case-cohort study of incident CHD and stroke aimed at identifying and assessing causality of genetic and non-genetic risk factors for CHD and stroke, derivation of predictive models of future risk, and identification of gene–environment interactions. EPIC-CVD is a prospective cohort study with participants from 23 centres in 10 European countries* • Various high-density gene arrays (on all cases and 15 000 randomly selected controls) • Biomarker data ($n = 80$; including lipids, metabolic factors, fatty acids, vitamins, and antioxidants) • Extensive questionnaire data Physical measurements	Cases: 25 000 Controls: 520 000 (data on 15 000)	Data & Biospecimen access is through application and review to determine scientific excellence, strategic priority, research strength of applicant, compliance to local ethics, and protection governance and (for data) achievement of the research goal through remote access
The Framingham Heart Study • https://www.framinghamheartstudy.org/ The Framingham Heart Study data are accessible through the following dbGAP entry: phs000153	*Prospective cohort study of 3 generations of research subjects who have been followed up to 65 years to evaluate risk factors for cardiovascular disease* • Epidemiological data • Various SNP arrays • WGS • RNA-Seq (including microRNA) • DNA methylation Metabolite profiling	~15 000 participants (6 cohorts) Initial cohort: 5209 Second-generation cohort: 5124 Third-generation cohort: 4095	Accessible upon application which needs to be approved before funding is sought. Application via the study website
STARNET—Stockholm–Tartu Atherosclerosis Reverse Networks Engineering Task • http://starnet.mssm.edu/ The data are accessible through the following dbGaP entry: phs001203	*STARNET is a genetics of RNA expression study of multiple disease-relevant tissues (blood, atherosclerotic-lesion free internal mammary artery, atherosclerotic aortic root, subcutaneous fat, visceral abdominal fat, skeletal muscle, and liver) obtained during surgery from patients with cardiovascular disease* • Exome array + imputation • RNA-Seq profiles of 7 relevant tissues	Patients: 600	Data accessible via dbGaP: access upon request; limited to tenure-track (or equivalent) senior investigators; requires an eRA Commons account

(*Continued*)

Table 2. Continued

Databases and links	Brief description/types of data	Samples	Access
Population databases			
UK Biobank • https://www.ukbiobank.ac.uk/ • https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access	*A large-scale biomedical database and resource with de-identified genetic, lifestyle, and health data from 500 000 UK participants. Includes the data used by the UK Biobank-Cardiometabolic-Consortium CHD working group which assess the use of self-reported and hospital recode data on CAD in the UK Biobank* • WGS	Cases: 10 801 (self-reported angina or other evidence of chronic CHD) Cases: 6482 (MI or revascularization) Controls: 137 914	Access requires registering with the UK Biobank, applying online and providing the required information, signing a material transfer agreement and paying a fee
The Michigan Genomics Initiative (MGI) • https://precisionhealth.umich.edu/our-research/michigangenomics/	*A biobank collection linking genotypes and electronic clinical records in Michigan Medicine patients. Participants are recruited through surgical lists or various conditions* • GWAS ($n = 80\,529$) • HLA alleles ($n = 80\,529$) • Polygenic risk scores ($n = 70\,262$) • Star alleles (pharmacogenes) ($n = 70\,262$) • WES ($n = 606$) • Targeting sequencing ($n = 964$)	~100 k consented participants	Data are available for the University of Michigan members and requires approval through the University's Institutional Review Committee External researchers can request summary statistics
TOPMED • https://topmed.nhlbi.nih.gov/ TOPMED includes the following relevant dbGaP datasets • phs001416 MESA • phs000974 FHS • phs002194 Africa6K • phs002038 TOPCHeF • phs002194 AA_CAC • phs000956 Amish • phs001368 CHS • phs001218 GeneSTAR • phs001345 GENOA • phs000964 JHS • phs001569 PROMIS • phs001215 SAFS	*The Trans-Omics for Precision Medicine (TOPMED) program integrates data from whole-genome sequencing (WGS) and other omics (e.g. metabolic profiles, epigenomics, protein and RNA expression patterns) data with molecular, behavioural, imaging, environmental, and clinical data. A primary goal of the TOPMed program is to improve scientific understanding of the fundamental biological processes that underlie heart, lung, blood, and sleep (HLBS) disorders. Data are available from multiple studies as summarized in https://topmed.nhlbi.nih.gov/* • Epidemiological data • WGS • RNA-Seq • Methylomics • Metabolomics • Proteomics	~180 k participants from more than 85 studies (different study designs)	Data for the individual studies is available through dbGaP and NHLBI BioData Catalyst

(*Continued*)

Table 2. Continued

Databases and links	Brief description/types of data	Samples	Access
China Kadoorie Biobank • https://www.ckbiobank.org/	*A large prospective study with participants, recruited from 10 geographically defined and diverse areas of China. Extensive data were collected at baseline and subsequent periodic resurveys monitored the health of participants over two decades, including the occurrence and causes of heart disease* • Epidemiological data (whole dataset) • Blood biomarker data ($n = 20\,000$) • Metabolomics ($n = 7300$) • Proteomics ($n = 1400$) • GWAS	512 000 participants	Application for data access, at a charge, is through project website. Requests are reviewed based on defined objectives, sound methodology, outputs, and timeline. Once approved, data are shared after signing of data access agreements
FinnGen • https://www.finngen.fi/en	*A large public–private partnership aiming to collect and analyse genome and health data from 500 000 Finnish biobank participants* • Axiom array—GWAS • Imputed genotypes	~500 000 participants	Embargoed for 1 year, then summary statistics made openly available through project website. Individual-level data are available upon request through the Fingenious portal
Electronic healthcare records + eMERGE • https://emerge-network.org/	*A network initiative to bring together electronic medical records and genetic data, returning validated, actionable results to clinicians* • Electronic medical records • Sequencing data of 109 genes ($N = 25\,000$) • WGS ($n = 900$) • Exome chip data ($n = 12\,865$) • Imputation dataset ($n = 105\,108$)	~105 000 participants	Access to data via collaboration agreement
Other databases and data portals			
GenomicKB • https://gkb.dcmb.med.umich.edu/	*Genomic Knowledgebase (GenomicKB) is a graph database to explore and investigate human genome, epigenome, transcriptome, and 4D nucleome with simple and efficient queries. The database uses a knowledge graph to consolidate genomic datasets and annotations from >30 consortia and portals, including 347 million genomic entities, 1.36 billion relations, and 3.9 billion entity and relation properties*	N/A	Online search engine
Genebass • https://app.genebass.org/	*A public resource of exome-based association statistics which includes 4529 phenotypes with gene-based and single variant associations using data from the UK Biobank* • Exome-based association statistics	394 841 exome datasets	Open access to merged, association statistics and metadata
HeartBioPortal • https://heartbioportal.com	*A knowledge portal featuring 44 988 human genic and intergenic regions and 6 906 292 variants across 37 cardiovascular diseases and 56 cardiometabolic quantitative traits* • GWAS • Quantitative GWAS traits • Differential expression	N/A	N/A

experimental designs, treatments, and analyses, typically through Minimum Information Guidelines like MIAME for microarray experiments [49]. These standards ensure that experiments are understandable and replicable by other scientists, with similar guidelines developed for other omics data types such as MIAPE, MIGS-MIMS, and MINSEQE [50]. Another initiative promoting transparency, accessibility, and reusability of research results is the FAIR (Findable, Accessible, Interoperable, and Reusable) open access initiative [51]. FAIR principles guide researchers in making their data and software findable, accessible, interoperable, and reusable. While these principles promote scientific collaboration and knowledge sharing, implementing them can be challenging due to fragmented data, accessibility issues, and the need for significant resources and infrastructure.

The General Data Protection Regulation (GDPR), implemented by the EU in 2018, significantly impacts multiomic data sharing and AI/ML use in healthcare by classifying genetic, biometric, and health data as sensitive. Data pseudonymization (replacing identifiable information with pseudonyms) is not applicable in the case of multiomic analysis as the personal information is inherently encoded in such datasets [52–54]. As the raw data pose a higher risk of identification compared with processed data, a potential solution is sharing aggregated data on a cohort level. While such an approach removes the risk of identifying individual patients, at the same time it results in the loss of patient-level information, which can significantly reduce its utility for research into individual differences associated with biological changes and, consequently, its utility in personalized medicine approaches [55]. Federated learning [56] offers a promising alternative by enabling collaborative analysis across multiple institutions without sharing raw data, thus protecting patient privacy while preserving the individual information needed for personalized medicine.

Approaches to data integration in ASCVD

With the increasing availability of large-scale biomedical data and modern AI/ML algorithms, great hope has been placed in the search for intricate relationships in the molecular characterization of complex diseases and, in particular, the integration of the complementary assays in multiomic datasets. Despite the high societal impact of ASCVD, studies have predominantly focused on cancer data [57, 58]. Large open repositories of cancer data [12, 59] have been instrumental in enabling the development and validation of integrated analysis methods. With the increasing collection of multiomic data for cardiovascular patients, there is a growing interest in applying the latest modern integrated analyses approaches also in this domain. Reitz *et al.* [60] discuss recent developments in multiomic analyses and network biology applied to cardiovascular diseases, with insights regarding gene regulatory networks, protein–protein interactions, signalling networks, and metabolic networks in the heart.

Several recent works have shown that the combination of multiomic data can help study the complex biology of ASCVD and already improve clinical prediction models [61]. The Framingham Heart Study, a multigenerational cohort, is a cornerstone of cardiovascular research [26]. Recent analyses within this cohort combined genome-wide association studies (GWAS), DNA methylation, and transcriptomic data to investigate subclinical atherosclerosis and myocardial infarction. This approach revealed important associations with genes such as AHRR and EXOC3, which are linked to smoking and platelet function, and FYTTD1 and PINK1, which are involved in cardiac tissue homeostasis [26]. These findings have improved predictive models for coronary atherosclerosis and myocardial infarction beyond traditional risk factors. Similarly, the Athero-Express Biobank provides both bulk and single-cell RNA sequencing data, as well as whole-genome methylation and proteomics profiles, on samples of atherosclerotic plaques obtained at endarterectomy. This has enabled the identification of unstable plaque biomarkers, including osteopontin and galectin-3. In addition, genetic analysis revealed that the rs13168867 variant in the OSMR gene was significantly associated with increased plaque vulnerability, characterized by higher intraplaque fat and lower collagen [61]. This provides a molecular basis for the assessment of future cardiovascular event risk. In parallel, the GCAT|Genomes for Life project in Catalonia provides a population-wide perspective through the integration of genomic, transcriptomic, and metabolomic data together with environmental and electronic health data. A microbiome analysis of this cohort revealed that genetic variants regulating the expression of PCSK9, APOB, and LPL modulate the plasma lipid species that mediate cardiovascular risk. This demonstrates how different omics can help uncover metabolic pathways and gene–environment interactions relevant to ASCVD pathogenesis and therapeutic prioritization [62]. Taken together, these studies illustrate the translational power of diverse modern ASCVD datasets.

While the value of complementary molecular modalities is thus increasingly recognized and the collection of matched profiles are becoming more standard, they are typically still mostly analysed separately, with results only combined in the discussion stage. While it is clear that this can already leverage some of the value of complementary profiles, it still misses the opportunity for identifying cross-modality patterns promised by a fully integrated joint analysis of complementary profiles, and this is a generic limitation, not just limited to ASCVD. With integrated methods mostly developed and validated on cancer data, however, truly integrated analyses of ASCVD data are emerging only very recently [63], and it remains to be seen which of the methods originally established on cancer data can successfully be applied to ASCVD in general.

In recent years, there have been several review papers focusing on different methods of data integration. Huang *et al.* [64] compared supervised, semi-supervised, and unsupervised methods, while others have focused on discussing early, middle, and late integration approaches reflecting when input data integration is performed relative to the actual inference process, be that AI/ML or statistical tests [65–67].

In early-stage integration (Fig. 1), individual omic data are merged first into a single integrated dataset. Analysis is then performed using factor analysis [68, 69] or other statistical or AI/ML techniques [70–72]. Even though early integration allows AI/ML models to discover interactions between different omic layers, the resulting concatenated matrix is high dimensional, complex, and requires significant processing to facilitate integration and does not consider data distribution of each omics.

Middle integration is transformation based. Transformation can either be performed independently (mixed integration) or jointly (intermediate integration) resulting in a reduced complexity of the multiomic dataset. In mixed transformation, each omic layer is transformed independently into a simpler representation followed by integrating the transformed layers into one joint transformation. The transformation can either be kernel based (mathematical functions that transform data into higher dimensional space in which linear relationships may be found within the dataset) [73, 74], graph based [75], or by using deep learning approaches [76]. In intermediate integration, data are combined in the actual inference model such as in the

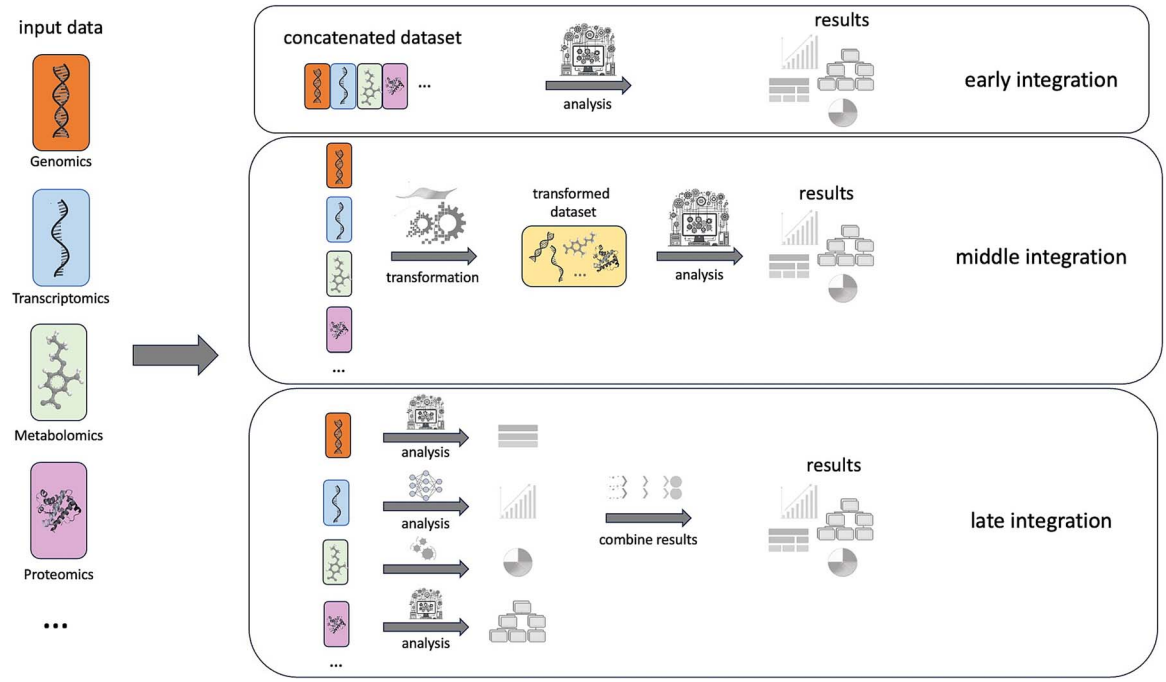

Figure 1. Early, middle, and late integration of multiomic data.

network-based approaches [77–80]; matrix factorization methods including structural learning and integrative decomposition (SLIDE) [81], non-negative matrix factorization (NMF) [82], and joint NMF [83]; and latent variable models such as multiomics factor analysis (MOFA) [69], canonical correlation analysis (CCA) [84], and sparse supervised multiple CCA (SSMCCA) [85]. These methods, in general, produce novel representations that include shared omic-specific representations common to all omics and some that are omic specific.

In late integration, outputs of individual models trained separately for different data layers are combined [86]. The main advantage of this strategy is reusing already developed models that are possibly tailored for each omics type. The obvious challenge is determining how to effectively combine individual results. Furthermore, the main disadvantage of late integration strategy is that it fails to discover interactions between different omic layers, making it less effective compared to early or middle integration approaches.

Common ML pipelines comprise multiple steps, including (1) data preprocessing, e.g. data cleaning, normalization, and optional handling of missing values; (2) optionally feature selection or dimensionality reduction; (3) linking different multiomic source variables for data integration; (4) application of ML methods depending on the task (classification, regression, clustering); and (5) analyses for an interpretation of discovered patterns and a *post hoc* evaluation of results. At each step, a wide range of methods can be employed, from frequentist to Bayesian statistics, classical ML, or recent approaches utilizing so-called deep learning. Applications of the latter to multiomic data integration have been reviewed extensively [76].

Data processing and transforms

Before applying any ML algorithm, the data need to be examined and where necessary preprocessed, including steps for quality control and mitigating technical artefacts. Specifically, for any data integration, data from different sources need to be made comparable by transformation, a process called 'normalization,' with the aim of removing differences between modalities, technologies, laboratories, or even experimental batches that are not relevant to the biomedical question investigated. This remains a non-trivial task, however, because it first depends on the biomedical question investigated (age, for instance, may be a confounder or an explicit study parameter), and it is easy to inadvertently remove true biological heterogeneity [87] which may be particularly relevant for CVD. Specifically, when control samples were included in the experimental design, probabilistic factor analysis by far outperformed other normalization approaches in removing laboratory artefacts and improving inter-laboratory concordance. It is therefore highly recommended to systematically include sufficiently many control samples in the experimental design to allow for subsequent normalization by probabilistic factor analysis (github.com/PMBio/peer) [88, 89]. Where biomedical standard controls are not naturally available in a study, even the inclusion of fixed technical controls can serve this purpose.

While earlier work has questioned the general transferability of signatures from one technology to another, seeing that microarray-derived signatures did well on RNA-Seq data but not the other way around [90], more recent work, as part of the Critical Assessment of Massive Data Analysis competition (CAMDA, www.camda.info) comparative analysis experiments, demonstrated that for modern models in some scenarios, signatures can actually successfully transfer either way (table 5 from [91]).

Recent models trained on multiple modalities suggest that combining matched modalities with a contrastive loss can also improve the separation of relevant signals from artefacts for their application to single modalities [92, 93].

Reproducibility, explainability, and interpretation

The variety of available data processing and modelling tools can affect the end-to-end reproducibility of the whole analysis process and its conclusions. In the context of translational medicine or clinical applications, regulators have sought to establish best

practice, such as in the MAQC and SEQC initiatives [88, 94–96]. A need for accountability justifies extensive benchmarking and a push towards explainable AI and interpretable ML (XAI), either through intrinsically interpretable models such as trees and ensembles of trees, or by complementing non-interpretable 'black box' models such as deep neural networks with *post hoc* methods such as SHAP [97]. Together with reproducibility, interpretability of predictive algorithms in clinical settings requiring trust can contribute to critical accountability.

Common pitfalls and potential solutions

Despite recent advances and a wide-ranging literature, multiomic data integration is still far from a solidified field with unified approaches and well-established common solutions. First, for any powerful data integration, the data need to be actually available to researchers. While in the early days, it was the cost of genome-scale assays that limited cohort sizes, nowadays clinical and organizational issues are limiting, with most extensive cohorts compiled by large national and international efforts. Even when cohorts and assay data have been compiled, however, they can be impractical to near impossible to access due to regulatory or organizational hurdles, partly reflecting the need to protect patient confidentiality and partly simply reflecting the workings of complex institutions. Even in available datasets, parts of data can be missing. Many algorithms require complete data matrices and thus various imputation techniques for handling missing data in bioinformatics have been explored in the context of multiomic imputation [98–100], including deep learning/AI for the integrative imputation of multiomic datasets [98]. Emerging techniques from AI for synthetic data generation show promise in mitigating missing or scarce data [101].

For complex and heterogeneous diseases, like ASCVD, there is a need for more advanced analytical methods focusing on a biomedical interpretation of the data [102, 103]. High sample heterogeneity, complex underlying mechanisms, and the 'curse of dimensionality' [104] from the genome-scale number of variables versus the typical cohort sizes make the identification of relevant patterns challenging. Besides collecting larger cohorts or sets of cohorts, common techniques to help identify sparse signals include model regularization and reduction of model complexity by feature selection/engineering or dimensionality reduction. While approaches like t-SNE [105] or UMAP [106] remain popular and they do display some correlations within the underlying high-dimension data, the transforms are arbitrary and do not actually preserve local or global neighbourhood structures and can thus be highly misleading [107].

While the addition of relevant datasets helps address these challenges, strong variations remain between laboratories and non-standardized protocols, from sample collection to data processing as well as sample annotation (e.g. phenotypical/clinical data). These can confound the sensitive detection of specific biomedical signals [108]. Despite international efforts to standardize data structures and semantics across the research community [109, 110], per study harmonization is still required, and so datasets need to be converted into a unified format and different semantics mapped at significant cost, and largely relying on manual curation [111]. This requires a thorough understanding of study-specific designs and data, including standard operating procedures, data collection devices, data types and formats, and content semantics [112]. Besides simple variations in column names, units of measurement, measurement devices, and granularity (e.g. per month or per day), incompatible ontologies require the greatest efforts in order to resolve e.g. inconsistent terminology of conditions, medication, or questionnaires. Some variables ideally do not require any processing, some require harmonization, while others cannot be matched and have to be omitted.

The development of tools for the identification of signatures that robustly generalize to new cohorts remains a key challenge in precision medicine [113]. In general, robust generalization can fail due to two distinct errors in measurements and algorithms: random stochastic errors, and the systemic effects of confounders or other sources of bias. Data integration holds the potential to overcome both issues by increasing the number of measurements and bringing together complementary domains with different error characteristics: Vertical data integration can combine measurements from different assays, such as gene expression and DNA methylation. Horizontal data integration can combine measurements from different cohorts, bringing together results from different populations and clinics. It is therefore puzzling that despite a plethora of new algorithms published over the years, actual successful applications of data integration methods not only remain rare but are even declining as a percentage of published studies [80]. This apparent paradox is resolved by the observation that new tools are typically introduced on a small number of datasets. Data-driven algorithms often tend to perform less effectively in scenarios different from those for which they were originally developed. This can be due to algorithms with high-dimensional parameter spaces overfitting or hyperparameters tuned to particular data characteristics and subsequently struggling with new cohorts presenting out-of-distribution data. It can be harder to identify such situations for 'black box' approaches like AI techniques such as modern deep learning algorithms. A critical first conclusion, thus, is that we need to raise the bar for empirically testing algorithms. Until advanced approaches are sufficiently tested and validated on a wider range of datasets, any recommendation of standard tools will not stand the test of time.

One approach that can increase the robustness of data-driven approaches is the early integration of prior knowledge. Specifically, a multi-partite graph-based algorithm that detects differences in functionally known pathways in the context of the disease of an individual patient has recently been shown to do well on cancer data for a large variety of data types and cohort sizes. The algorithm hierarchically fuses signals across a cohort and across multiple data types into a patient–patient network using an information theoretic similarity measure, while incorporating functional knowledge from the beginning. The algorithm performed consistently well in survival prediction and the identification of relevant molecular mechanisms even on small cohorts with 100–200 patients where other state-of-the-art algorithms failed [80]. While the algorithm was tested for six different data-type combinations and cohort sizes, all profiles were from cancer patients, and stratification was evaluated for survival analysis, complemented by discussions of implicated molecular pathways. It will thus be interesting to see how well it generalizes to other use cases and domains and, specifically, complex cardiovascular cohorts.

From Big Data to clinical practice

In this section, we will examine best practices that ensure the technically and clinically sound use of ASCVD Big Data in the real-world (RW) and underline the necessity for enhancements in the execution of integrative research which needs to incorporate increasingly recognized socioeconomic, lifestyle, and

environmental factors [114–117] into the results of analysis coming from multiomic assays.

The real-world approach enables integrative Big Data ASCVD research that is clinically valid and fair

RW data from clinical sources encompasses information drawn from various repositories linked to outcomes within a diverse population of patients, in real-life contexts. These sources include, but are not limited to, EHRs, health insurance claims, disease registries, and patient surveys. A recent review [118] assessed best practices and common difficulties regarding design, measurement, analysis, and generalizability in the setting of RW studies that evaluate safety and effectiveness of treatments in diabetes. Similarly, all these qualities are also needed for integrative Big Data studies of ASCVD to enable the translation of findings into effective and actionable clinical insights for ASCVD patient stratification and management, and in turn provide clinicians with tools oriented to real-time decision support–based integration of multiple data sources.

The most critical aspects for RW research having internal validity (i.e. the returned results are not due to chance, confounding, or other sources of bias) of the changes in the analysed factor are related to the design. The study design determines the structure of the data and the best-suited analytical models. The design of a RW study is different depending on whether the objective of the study relates to causal inference of relevant exposures on ASCVD, risk prediction, evaluation of diagnostic and prognostic tests, effectiveness of pharmacological or other interventions (possibly through emulated target trials) [119], or to inform public health policies implementation (possibly through other quasi-experimental data analysis methods) [120, 121]. Another threat to validity is misclassification of ASCVD outcomes (i.e. dependent variables) in EHRs [122]. Both measurement errors and misclassification can lead to biased results, and relevant exposures and covariates need to be identified already in the design stage. In the section 'Approaches to data integration in ASCVD', we have already stressed the importance of pre-analytic and raw-data processing.

Importantly, already at the stage of experimental design one needs to plan for the collection of information related to main potential confounders and pre-clinical and other variables that are known risk factors for outcomes, associated with exposure, and not intermediary variables in associations between exposure and outcome and which, if unbalanced, may introduce bias [7, 8]. Potential relations may be explored using approaches like variance partitioning. Multi-dimensional data visualization tools can also play a supportive role for understanding patterns in data and potential related artefacts [123, 124]. Data for men and women are best analysed separately since there are pronounced non-trivial sex-specific effects seen for atherosclerosis [125]. Where age is not a variable of interest in the study, in case–control studies, an accepted way of balancing age effects is frequency matching in 5- or 10-year age groups. When removing effects of covariates, however, caution must be taken, as a blind selection of adjustment covariates may introduce selection bias when covariates are colliders rather than confounders [126]. Overall, a carefully planned design which is suitable to the specific research question, including the *a priori* consideration of all needed information, is critical in the interpretation of complex datasets in ASCVD research. This is already pertinent when deciding on the collection times of samples: Levels of gene expression in blood in the days just after an acute event such as myocardial infarction have been shown to be very different to those in more stable states before or several months after the event [127]. At what time samples should be collected for RNA, protein, and metabolite profiling thus depends on the research questions being asked. For instance, for risk prediction the ideal is obtaining profiles a few years before events through prospective or population studies, and when this is not possible, it is important to validate findings in a prospective setting at some later point. In contrast, for identifying health complications following acute events such as heart failure after myocardial infarction or risk of second myocardial infarction, samples from right after the acute events are expected to more clearly allow the identification of biomedically relevant signals. Depending on the hypothesis examined, however, one may wait until after the changes from the acute event itself subside to study medium to longer term consequences. In summary, any experimental plan must be aware that time to acute event is a major determinant of molecular state in ASCVD.

Other issues that need to be considered from the beginning, when planning experiments, are affected by EU laws and directives prioritize health equity, ensuring fair access to resources for all. Discrimination based on sex, gender, race, ethnicity, and socioeconomic status is prohibited [128]. Health disparities incur significant direct and indirect costs [129]. Socioeconomic factors strongly correlate with CVDs [130, 131]. ASCVD prediction models vary across subgroups [132, 133], with inconsistent performance across race and gender [128, 133], potentially leading to unequal therapy distribution. Some argue that AI/ML strategies could improve care for underserved populations [134], but ensuring robustness across different contexts is crucial. Robustness, outlined in the EU's Ethics Guidelines for Trustworthy AI [135], addresses the data shift problem where models may overfit and underperform in RW research. To combat this, models require diverse training data and external validation in large independent cohorts, which ensures trust. Current ASCVD risk forecasting methods may mispredict risk for groups historically under-represented in biomedical research, unintentionally harming them. In Europe, the gold standard for 10-year ASCVD risk prediction, SCORE2 [132], and its adaptation to older ages, SCORE2-Older Person (OP) [136], do not consider socioeconomic factors [128]. They may underestimate risk for lower socioeconomic groups and overestimate for higher ones, widening health inequality gaps. In the USA, Kartoun *et al.* found disparities in risk prediction for atrial fibrillation and ASCVD across subgroups [137]. CHARGE-AF score performs well but shows higher discrimination for females and unfairness for sex and race in intermediate age groups. Evaluating risk models within subpopulations followed by recalibration is crucial for guiding clinical decisions and informing policymaking. Interestingly, the most recent Predicting Risk of Cardiovascular Disease Events (PREVENT) risk calculator developed in the context of large datasets and EHRs from 6.6 million individuals, including 46 American cohorts [138, 139], is meant to be a 'Universal Risk Calculator' and has eliminated 'race' as a predictor, and has included instead social determinants of health as measured by the social deprivation index. PREVENT includes other novel predictors such as a measure of cardiovascular-kidney-metabolic health to better capture the effects of obesity, diabetes, hyperglycaemia exposure, and kidney disease, and for high-risk patients, urine albumin-to-creatinine ratio and haemoglobin A1c. Importantly, the cardiovascular endpoint was extended to include heart failure in addition to coronary heart disease and stroke. While PREVENT has introduced substantial conceptual changes in the field of prediction, future work should test PREVENT for external validation in other global populations.

From classical to novel statistical methods to assess clinically relevant cause–effect relations and biological pathways

Classical statistical methods have become the mainstay of analyses, including smart strategies to identify the most relevant features, especially in the setting of high-dimensional omics data analysis. Pairing the Iterative Sure Independence Screening (ISIS) algorithm with shrinkage methods improves feature selection and effect estimation, such as LASSO, Minimax Concave Penalty (MCP) [140], and Smoothly Clipped Absolute Deviation (SCAD) [141] regressions [142]. Elastic net improves coefficient estimation over LASSO, but does not satisfy the oracle property [143]. MCP or SCAD penalties tend to select only one predictor. Adaptive Enet (AEnet) satisfies the oracle property and selects more than one correlated feature [143, 144]. This together with Enet, AEnet, and MSAEnet for feature selection and consistent coefficient estimation, and a bootstrap approach for empirical confidence intervals for linear, Cox, and logistic ISIS penalization methods have for instance been compiled in the SIS R package [142] (https://github.com/statcodes/SIS).

Importantly, clinically relevant studies require robust metrics for quality assessment and demonstrated added value to standard care. The choice and interpretation of metrics depend on the outcome type, such as categorical (discrimination) or time-to-event (prospective prediction). Approaches like ROC and AUC are standard for categorical outcomes, but they overlook the complex interplay of risks, benefits, and costs. Decision curve analysis (DCA) [145] evaluates each threshold based on net benefits, considering defined harms and benefits and disease prevalence in the population. DCA incorporates decision consequences into the final metric for assessing clinical value.

A critical aspect of statistical analysis is the model building strategy. In traditional epidemiology, multivariate analysis follows a construct-based model building approach. In the setting of hypothesis-based analysis, this uses a set of variables decided *a priori*, and consistent with a given causal framework, as opposed to automated variable selection methods, which have demonstrated limitations [146, 147]. In this approach, a typical strategy is considering progressively adjusted models (from simpler to more complex ones) and the impact of added variables on ASCVD risk and outcome predictions and patient stratification. An alternative to construct-based model building that has been recently brought to the cardiovascular setting is the use of AI/ML models that can be agnostically trained on vast amounts of written information (such as in EHRs) to learn patterns and relationship between words and sentences [148]. With respect to applications of AI/ML tools to the analysis of complex health-related datasets, it will be challenging but necessary to develop explainable AI algorithms. These algorithms must undergo the same regulatory and safety standards as medical products and achieve the precision of today's physicians and scientists in order to be socially accepted. In some fields, such as dermatology, this has already been achieved by automated image analysis, which often outperforms the diagnostic potential of humans [149], but for many other fields the superiority of these analysis methods has yet to be proven.

Another interesting area of evolving research interest focuses on exploiting the wealth of available genomic data to predict potential drug targets, specifically, if genetic variants have been identified using causal inference methods. For this, there are quasi-experimental methods that use genetic instrumental variants to elucidate cause–effect relationships, in particular mendelian randomization and colocalization [150], that can be applied to summarized data from genome-wide association studies. Mendelian randomization assesses, under strong assumptions, if genetic factors linked to an exposure also affect the outcome, implying causation. Colocalization investigates if two or more traits share the same genetic variants at a specific locus. While colocalization is more robust compared to Mendelian randomization, since it does not require such restricting assumptions to be fulfilled, it does not allow to estimate the causal associations and related standard errors. The two approaches, however, can provide complementary information on causal therapeutic targets [150] that can be subsequently followed up using established pharmacogenomic search engines (e.g. PharmaGKB, DGIdb, SIDER, STRING, STITCH, DrugBank, and others). Hence, there are various informative data sources and methods that can be used in Big Data integrative ASCVD research, not only to explore cause–effect relations, but also to elucidate biological implications of the research findings through bioinformatic analysis.

Going beyond these classical approaches, there have been considerable efforts at modelling mechanistic pathways to infer causal relationships [151]. Specifically, the Hipathia approach [152, 153] has recently been showcased in the COVID-19 Disease Map Challenge at the CAMDA competition, powerfully demonstrating the value of extracting disease-relevant mechanistic pathways [154]. Notably, the method has also been employed successfully in linking nanoplastic pollution to an increased risk for ischemic cardiovascular disease [155], suggesting that this approach may have high potential also in the field of ASCVD.

In summary, classical statistical methods complemented by causal inference through Mendelian randomization and bioinformatic biological pathways enrichment approaches provide powerful tools to gain new biological insights and identify potential therapeutic targets. As these methods continue to improve, they will play an increasingly important role in both research and clinical settings, advancing personalized medicine in the treatment of cardiovascular disease [156]. Finally, it has to be stated that the classical approach of measuring a few parameters for thousands of probands in clinical trials cannot be the basis of real personalized medicine, where rather measuring thousands of parameters in a single person makes sense [157] in combination with novel statistical methods and ML/AI approaches.

Seamless integration with healthcare systems

Clinically sound ASCVD Big Data research must empower clinicians, patients, and society. The final goal is not only better clinical care, but also, precision prevention medicine. For this, ASCVD Big Data analysis platforms need to be effectively integrated within the healthcare systems. The COVID-19 pandemic has further accelerated and promoted the integration of EHRs from diverse data sources. As an example, the unCoVer project [158] demonstrated effective RW research integration. Sponsored by Horizon 2020 during the COVID-19 onset, collaborators from 30+ institutions collected and analysed data on patient care across Europe. Datasets were harmonized using common data models, and made available via remote access using Opal-DataSHIELD, preserving privacy. A federated learning infrastructure (Fig. 2), compliant with GDPR, enabled online access to COVID-19 records for analysis without the data leaving the partnering institution or disclosing patient's personal information to analysts. This approach ensures secure handling of sensitive patient data including multiparty statistical analysis. For example, a study of baseline characteristics of 14 236 COVID-19 patients admitted to 6 different hospitals in Europe between 2020 and 2022 confirmed a greater proportion of cardiovascular patients amongst those recorded as

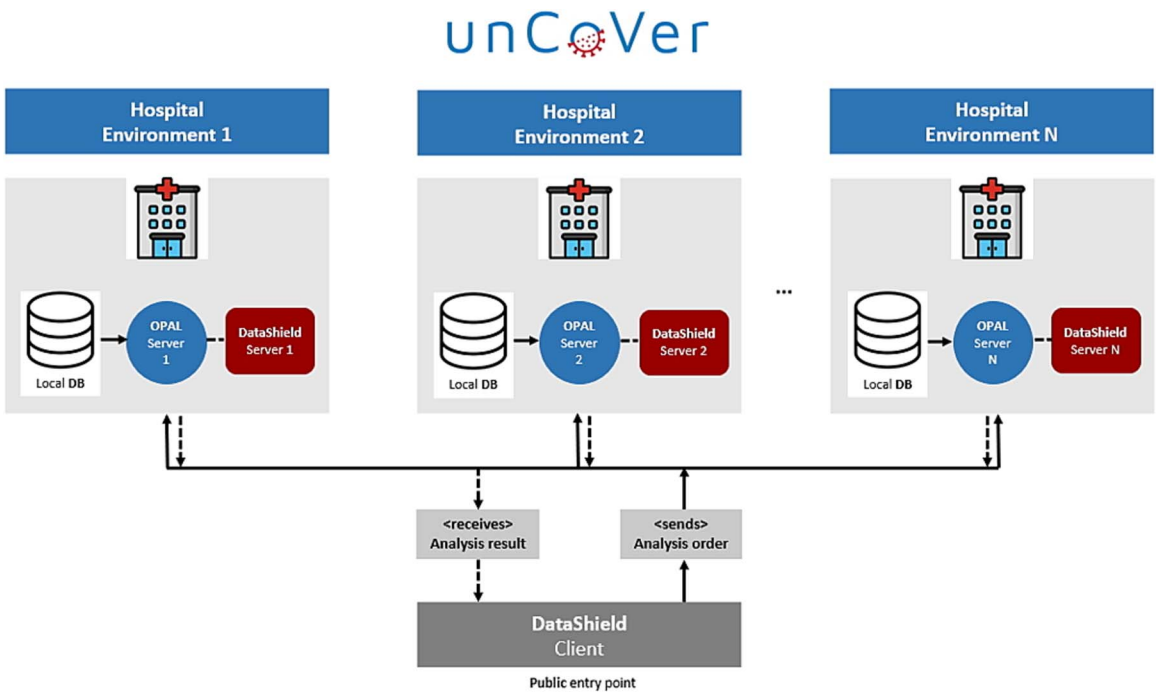

Figure 2. Secure federated computation and analysis of unCoVer data based on Opal-DataSHIELD infrastructure. unCoVer, unravelling data for rapid evidence-based response to COVID-19. From [158] with permission.

in-hospital death (9.33% versus 44.9%) [159] with a pooled age-/sex-adjusted odds ratio of 1.68 (95% CI: 1.40, 2.01) [160]. Results have been found to be consistent across all epidemic waves in Spain, after analysing EHRs from 13 974 patients and observing a higher risk of death for patients with cardiovascular-related conditions [161].

Beyond the unCoVer project, cross-centre data analysis under GDPR constraints typically follows a structured workflow: (1) harmonization of local data using common data models such as OMOP Common Data Model (OMOP-CDM) [162] or HL7 FHIR [163]; (2) implementation of secure local analysis nodes with platforms like Opal and DataSHIELD, which allow remote, privacy-preserving statistical analysis without individual-level data transfer [11, 164]; (3) coordination via federated learning infrastructure that shares only model parameters or summary statistics between centres, avoiding direct data sharing [165, 166]; and (4) governance ensured through data use agreements, institutional data protection officers, and compliance with GDPR ethical and legal standards [167]. These steps enable secure, real-world ASCVD research collaboration across institutions while maintaining patient privacy and data sovereignty.

Long COVID has now emerged as a global epidemic with persistent symptoms, including cardiovascular ones [168]. Visualization of the heart and blood vessels reveals indications of post-infectious perimyocarditis leading to ventricular failure, arterial wall inflammation, or microthrombosis in specific patient groups [169], persisting for weeks or months after SARS-CoV-2 infection resolves. In this setting, reinforcing primary/primordial cardiovascular prevention is crucial due to the increasing cardiovascular burden (both COVID related and unrelated). ASCVD Big Data research should, thus, prioritize primordial prevention interventions targeting modifiable behavioural risk factors. The growing use of smartwatches and health apps [170–172] allows individuals to input lifestyle and clinical data, potentially aiding AI/ML tools in predicting effective lifestyle changes or preventive medication for disease prevention. Users typically consent to data usage, simplifying ethics approvals. Expectations are for increased device usage, offering self-empowerment and personalized precision healthcare. Combining smartphone/smartwatch data with lifestyle and clinical factors can enhance AI algorithms for predicting synergistic lifestyle changes. This could lead to individually optimized, mild lifestyle suggestions for reaching target health values more effectively than generic advice.

Integration of EHRs and genetic data is the next step for ASCVD prediction and prevention in the clinical setting. Big Data studies demonstrate the power of polygenic risk scores (PRS) for diseases like ASCVD [173–175]. A Finnish study combined PRS with clinical risk scores, using a digital platform to communicate absolute risk to individuals [176]. After 1.5 years, higher baseline ASCVD risk correlated with positive health behaviour changes. This highlights the importance of empowering patients and clinicians with accurate risk estimates [177]. To effectively bridge multiomic insights into clinical practice, future cardiovascular risk models should be recalibrated to incorporate the PRS and other omics-derived biomarkers alongside traditional factors like age, cholesterol, and blood pressure. Embedding these enhanced risk calculators directly into EHR systems could enable real-time, automated risk assessment during clinical encounters. For example, a patient's integrated risk score—derived from both clinical and omics data—could be displayed in the physician dashboard with actionable alerts, supporting shared decision-making for preventive therapies. Furthermore, omics-informed tools must be designed with clinician usability in mind, minimizing workflow disruptions while maximizing interpretability and clinical relevance. These advancements would move precision medicine from research into routine care, especially in cardiology where risk stratification is central to long-term outcomes.

Overall, integration of multiomic Big Data, including not only genomics, but also epigenomics, transcriptomics, proteomics, and

metabolomics, can enhance precision and personalized healthcare. Social and health insurance systems must adapt to new technologies to capitalize on long-term cost savings. However, such approaches might impose an administrative burden on clinicians. According to the Phillips Global Future Health Index Report 2020, young medical professionals are driving the adoption of new technologies, but many feel unprepared to use digital data for patient care [178]. To truly impact clinical care and reduce burdens, these technologies must seamlessly integrate into existing healthcare systems, leveraging interoperable EHRs standards like Fast Healthcare Interoperability Resources (FHIR) to present data in user-friendly formats using ML. The USA leads the change, with a major proportion of healthcare providers already embracing the FHIR standard. In addition to FHIR, there are other interoperability standards for digital health such as the OMOP, OpenEHR [179], and Systematized Nomenclature of Medicine (SNOMED) [180]. The integration of automated tools for analysing medical Big Data offers the potential to alleviate the workload of healthcare professionals while enhancing the quality of care. This approach empowers clinicians to prioritize the human aspect of the patient–doctor relationship [181], resulting in a heightened standard of care.

Conclusions

Despite recent therapeutic advances, atherosclerotic cardiovascular disease still remains the leading cause of death worldwide, requiring further research and innovative strategies in order to improve prevention, diagnosis, and treatment. The establishment of high-throughput sequencing and omics technologies has enabled the generation of large datasets that capture molecular factors related to ASCVD development across different -omics dimensions. Integrative analyses of those datasets with AI/ML techniques offer unprecedented potential to identify key molecular pathways and new biomarkers to improve risk stratification and support personalized treatment. In practice, however, robust multiomic data integration remains challenging.

Technical and experimental biases already introduced during data collection complicate the integration process and require the careful application of appropriate preprocessing and normalization techniques to combine data in a meaningful way. Whenever possible, control samples should be included systematically in experimental designs to allow for more powerful normalization methods, even if only technical controls are feasible in a study. In addition, ASCVD is a very heterogeneous disease, influenced not only by genetic factors, but also important environmental, lifestyle, and socioeconomic aspects which need to be taken into account during the analysis. Recent advances on mechanistic models that highlight potential molecular causes may be of great value to the development of novel therapeutic interventions. Those approaches are complemented by black box models from AI/ML that can show superior predictive power and sensitivity in picking up subtle patterns from training data *de novo*. In this context, issues related to data privacy and ownership still hinder the exchange and collaborative use of multiomic data across different studies, which complicates such validations, especially across populations. This is largely a policy issue, which is of course affected by ethical, legal, social, and last but certainly not least, commercial considerations.

Currently, most AI/ML models are developed and validated on cancer data; therefore, generalization testing is necessary to ensure that these patterns reflect the underlying biomedical mechanisms and will robustly work also on cohorts from different times, populations, and clinics. Future research thus needs to focus on an improved validation of advanced methods. That is critical because overfitting can be hard to detect in data-limited scenarios, and this is particularly true for the high-dimensional increasingly popular 'black-box' AI techniques. This relies on the collection of benchmark datasets from diverse domains. Discovery and access to such data of sufficient diversity and extent form key bottlenecks in the advancement and effective application of bioinformatics methods for multiomic data integration. Awareness of and access to the currently available key datasets in ASCVD, as well as a concerted push by practitioners and researchers alike for increasing the data commons supporting ASCVD, is therefore critical for progress in the field.

Establishing catalogues of multiomic data repositories beyond cancer, harmonized protocols, better data-sharing practices, and greater acceptance of federated learning will accelerate the effective application and development of the latest bioinformatics methods for omics data integration. The expected improvements in disease prediction, prevention, and patient outcomes will mark a transformative shift towards precision medicine in and beyond atherosclerotic cardiovascular disease.

Key Points

- We discuss challenges and opportunities in the rapidly evolving field of machine learning and artificial intelligence for multiomic data integration.
- Key bottlenecks include discovery and access to data, as bioinformatics approaches are typically established on cancer data, with correspondingly degraded out-of-domain performance.
- We provide, for the first time, a high-value overview and details of available ASCVD datasets.
- We highlight specific characteristics of multiomic ASCVD data analysis.
- We examine best practices that ensure the technically and clinically sound use of multiomic data in the real-world applications, including considerations specific to ASCVD.

Conflict of interest: Y.D. has filed patents related to the use of RNAs for diagnostic and therapeutic purposes and is a member of the Scientific Advisory Board of the molecular diagnostic company Firalis SA.

Funding

This article is based upon work from COST Action AtheroNET, CA21153, supported by COST (European Cooperation in Science and Technology). S.B.W. and R.F. were supported by HORIZON-EIC-2022-Pathfinderchallenges-01-03 TargetMI (ID:101114924) and HORIZON-WIDERA-2022 BioGeMT (ID:101086768). R.A. was supported by HORIZON-EIC-2022-Pathfinderchallenges-01-03 TargetMI (ID:101114924). B.N.W. has received funding from the European Union's Horizon Europe Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No. 101110878. M.Ch. and G.J. were partially supported by HORIZON-MSCA-2021-SE-01-01-MSCA Staff Exchanges 2021 CardioSCOPE 101086397. P.A.'s contribution was supported by HORIZON-WIDERA-2022 BioGeMT (ID: 101086768). J.B. was supported by FCT PhD Scholarship 2020.09166.BD,

European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 951970 (OLISSIPO project), and INESC-ID Plurianual project UIDB/50021/2020 until 31.07.2023. From 01.08.2023 JB is a full-time employee of the Medical University of Vienna. F.S. has received grant from Research Foundation Flanders (1S07421N). R.V. was supported by FCT—Portuguese Foundation for Science and Technology, under iBiMED (UID 4501- Instituto de Biomedicina - Aveiro) and the Cardiovascular R&D Center—UnIC (UIDB/00051/2020 and UIDP/00051/2020), CardioNIR project – CARDIOvascular Near-InfraRed spectroscopy probing – 2021 (PTDC/EMD-EMD/3822/2021), https://doi.org/10.54499/PTDC/EMD-EMD/3822/2021. M.S. is funded by the European Union (HORIZON-MSCA-2021-PF- MAACS 101064175; HORIZON-MSCA-2021-SE-01-01 - MSCA Staff Exchanges 2021 CardioSCOPE 101086397) and the Ministry of Science, Technological Development, and Innovation, Republic of Serbia through Grant Agreement with University of Belgrade-Faculty of Pharmacy No: 451-03-47/2024-01/200161P.M. was supported in part by European Union (HORIZON-MSCA-2021-SE-01-01-MSCA Staff Exchanges 2021, CardioSCOPE 101086397) and Italian Space Agency (N. 2023-7-HH.0 CUP F13C23000050005 MicroFunExpo). M.T.P. was supported by the Spanish Funds for Research in Health Sciences, Instituto de Salud Carlos III, cofounded by European Regional Development Funds (PI22CIII/00029) and the Spanish Agency for Research (PID2019-108973RB-C21 and PID2023-147163OB-C22).

References

1. Timmis A, Aboyans V, Vardas P. *et al.* European Society of Cardiology: the 2023 atlas of cardiovascular disease statistics. *Eur Heart J* 2024;**45**:4019–62. https://doi.org/10.1093/eurheartj/ehae466
2. Aragam KG, Jiang T, Goel A. *et al.* Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat Genet* 2022;**54**:1803–15. https://doi.org/10.1038/s41588-022-01233-6
3. Alaa AM, Bolton T, Di Angelantonio E. *et al.* Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PloS One* 2019;**14**:e0213653. https://doi.org/10.1371/journal.pone.0213653
4. Oussous A, Benjelloun F-Z, Ait Lahcen A. *et al.* Big Data technologies: a survey. *J King Saud Univ Comput Inf Sci* 2018;**30**:431–48. https://doi.org/10.1016/j.jksuci.2017.06.001
5. Gobbel GT, Matheny ME, Reeves RR. *et al.* Leveraging structured and unstructured electronic health record data to detect reasons for suboptimal statin therapy use in patients with atherosclerotic cardiovascular disease. *American Journal of Preventive Cardiology* 2022;**9**:100300. https://doi.org/10.1016/j.ajpc.2021.100300
6. Witting C, Azizi Z, Gomez SE. *et al.* Natural language processing to identify reasons for sex disparity in statin prescriptions. *Am J Prev Cardiol* 2023;**14**:100496. https://doi.org/10.1016/j.ajpc.2023.100496
7. Vanhaverbeke M, Attard R, Bartekova M. *et al.* Peripheral blood RNA biomarkers for cardiovascular disease from bench to bedside: a position paper from the EU-CardioRNA COST action CA17129. *Cardiovasc Res* 2022;**118**:3183–97. https://doi.org/10.1093/cvr/cvab327
8. Sopic M, Vilne B, Gerdts E. *et al.* Multiomics tools for improved atherosclerotic cardiovascular disease management. *Trends Mol Med* 2023;**29**:983–95. https://doi.org/10.1016/j.molmed.2023.09.004
9. Tryka KA, Hao L, Sturcke A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* 2014;**42**:D975–9. https://doi.org/10.1093/nar/gkt1211
10. Lappalainen I, Almeida-King J, Kumanduri V. *et al.* The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet* 2015;**47**:692–5. https://doi.org/10.1038/ng.3312
11. *DataSHIELD: Taking the Analysis to the Data, Not the Data to the Analysis International Journal of Epidemiology*, Volume 43, December 2014, pp. 1929–44. https://doi.org/10.1093/ije/dyu188
12. Grossman RL, Heath AP, Ferretti V. *et al.* Toward a shared vision for cancer genomic data. *N Engl J Med* 2016;**375**:1109–12. https://doi.org/10.1056/NEJMp1607591
13. Khanicheh E, Qi Y, Xie A. *et al.* Molecular imaging reveals rapid reduction of endothelial activation in early atherosclerosis with apocynin independent of antioxidative properties. *Arterioscler Thromb Vasc Biol* 2013;**33**:2187–92. https://doi.org/10.1161/ATVBAHA.113.301710
14. Baysoy A, Bai Z, Satija R. *et al.* The technological landscape and applications of single-cell multi-omics. *Nat Rev Mol Cell Biol* 2023;**24**:695–713. https://doi.org/10.1038/s41580-023-00615-w
15. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 2015;**16**:133–45. https://doi.org/10.1038/nrg3833
16. Willemsen L, de Winther MP. Macrophage subsets in atherosclerosis as defined by single-cell technologies. *J Pathol* 2020;**250**:705–14. https://doi.org/10.1002/path.5392
17. Wirka RC, Wagh D, Paik DT. *et al.* Atheroprotective roles of smooth muscle cell phenotypic modulation and the TCF21 disease gene as revealed by single-cell analysis. *Nat Med* 2019;**25**:1280–9. https://doi.org/10.1038/s41591-019-0512-5
18. Cochain C, Vafadarnejad E, Arampatzi P. *et al.* Single-cell RNA-Seq reveals the transcriptional landscape and heterogeneity of aortic macrophages in murine atherosclerosis. *Circ Res* 2018;**122**:1661–74. https://doi.org/10.1161/CIRCRESAHA.117.312509
19. Fernandez DM, Rahman AH, Fernandez NF. *et al.* Single-cell immune landscape of human atherosclerotic plaques. *Nat Med* 2019;**25**:1576–88. https://doi.org/10.1038/s41591-019-0590-4
20. Li Q, Wang M, Zhang S. *et al.* Single-cell RNA sequencing in atherosclerosis: mechanism and precision medicine. *Front Pharmacol* 2022;**13**:977490. https://doi.org/10.3389/fphar.2022.977490
21. Argelaguet R, Arnol D, Bredikhin D. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:111. https://doi.org/10.1186/s13059-020-02015-1
22. Smirnov P, Przybilla MJ, Simovic-Lorenz M. *et al.* Author correction: multi-omic and single-cell profiling of chromothriptic medulloblastoma reveals genomic and transcriptomic consequences of genome instability. *Nat Commun* 2025;**16**:1085. https://doi.org/10.1038/s41467-025-56164-7
23. Yang X, Mann KK, Wu H. *et al.* scCross: a deep generative model for unifying single-cell multi-omics with seamless integration, cross-modal generation, and in silico exploration. *Genome Biol* 2024;**25**:198. https://doi.org/10.1186/s13059-024-03338-z
24. Pekayvaz K, Losert C, Knottenberg V. *et al.* Multiomic analyses uncover immunological signatures in acute and chronic coronary syndromes. *Nat Med* 2024;**30**:1696–710. https://doi.org/10.1038/s41591-024-02953-4

25. Barrero-Rodríguez R, Rodríguez JM, Naake T. *et al.* TurbOmics: a web-based platform for the analysis of metabolomics data using a multi-omics integrative approach. *bioRxiv* 2025. https://doi.org/10.1101/2025.05.09.653072
26. Møller AL, Vasan RS, Levy D. *et al.* Integrated omics analysis of coronary artery calcifications and myocardial infarction: the Framingham Heart Study. *Sci Rep* 2023;**13**:21581. https://doi.org/10.1038/s41598-023-48848-1
27. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. *Am J Epidemiol* 1989;**129**:687–702. https://doi.org/10.1093/oxfordjournals.aje.a115184
28. Do R, Stitziel NO, Won H-H. *et al.* Multiple rare alleles at LDLR and APOA5 confer risk for early-onset myocardial infarction. *Nature* 2015;**518**:102–6. https://doi.org/10.1038/nature13917
29. Verwer MC, Mekke J, Timmerman N. *et al.* Comparison of cardiovascular biomarker expression in extracellular vesicles, plasma and carotid plaque for the prediction of MACE in CEA patients. *Sci Rep* 2023;**13**:1010. https://doi.org/10.1038/s41598-023-27916-6
30. Perisic L, Aldi S, Sun Y. *et al.* Gene expression signatures, pathways and networks in carotid atherosclerosis. *J Intern Med* 2016;**279**:293–308. https://doi.org/10.1111/joim.12448
31. Bycroft C, Freeman C, Petkova D. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9. https://doi.org/10.1038/s41586-018-0579-z
32. Kurki MI, Karjalainen J, Palta P. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* 2023;**613**:508–18. https://doi.org/10.1038/s41586-022-05473-8
33. Chen Z, Chen J, Collins R. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* 2011;**40**:1652–66. https://doi.org/10.1093/ije/dyr120
34. Taliun D, Harris DN, Kessler MD. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021;**590**:290–9. https://doi.org/10.1038/s41586-021-03205-y
35. Dawson LP, Lum M, Nerleker N. *et al.* Coronary atherosclerotic plaque regression. *JACC* 2022;**79**:66–82. https://doi.org/10.1016/j.jacc.2021.10.035
36. Nayor M, Brown KJ, Vasan RS. The molecular basis of predicting atherosclerotic cardiovascular disease risk. *Circ Res* 2021;**128**:287–303. https://doi.org/10.1161/CIRCRESAHA.120.315890
37. Libby P. The changing landscape of atherosclerosis. *Nature* 2021;**592**:524–33. https://doi.org/10.1038/s41586-021-03392-8
38. Verhoeven BAN, Velema E, Schoneveld AH. *et al.* Athero-express: differential atherosclerotic plaque expression of mRNA and protein in relation to cardiovascular events and patient characteristics. Rationale and design. *Eur J Epidemiol* 2004;**19**:1127–33. https://doi.org/10.1007/s10564-004-2304-6
39. Franzén O, Ermel R, Cohain A. *et al.* Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science* 2016;**353**:827–30. https://doi.org/10.1126/science.aad6970
40. Lai Z, Wang C, Liu X. *et al.* Characterization of the proteome of stable and unstable carotid atherosclerotic plaques using data-independent acquisition mass spectrometry. *J Transl Med* 2024;**22**:247. https://doi.org/10.1186/s12967-023-04723-1
41. Poznyak AV, Sukhorukov VN, Guo S. *et al.* Sex differences define the vulnerability to atherosclerosis. *Clin Med Insights Cardiol* 2023;**17**:11795468231189044. https://doi.org/10.1177/11795468231189044
42. Dai N, Tang X, Weng X. *et al.* Sex differences in coronary inflammation and atherosclerosis phenotypes in response to imaging marker of stress-related neural activity. *Circ Cardiovasc Imaging* 2024;**17**:e016057. https://doi.org/10.1161/CIRCIMAGING.123.016057
43. Man JJ, Beckman JA, Jaffe IZ. Sex as a biological variable in atherosclerosis. *Circ Res* **126**:1297–319. https://doi.org/10.1161/CIRCRESAHA.120.315930
44. Lechner K, von Schacky C, McKenzie AL. *et al.* Lifestyle factors and high-risk atherosclerosis: pathways and mechanisms beyond traditional risk factors. *Eur J Prev Cardiol* 2020;**27**:394–406. https://doi.org/10.1177/2047487319869400
45. Leon-Mimila P, Wang J, Huertas-Vazquez A. Relevance of multi-omics studies in cardiovascular diseases. *Front Cardiovasc Med* 2019;**6**:91. https://doi.org/10.3389/fcvm.2019.00091
46. Adler A, Kirchmeier P, Reinhard J. *et al.* PhenoDis: a comprehensive database for phenotypic characterization of rare cardiac diseases. *Orphanet J Rare Dis* 2018;**13**:22. https://doi.org/10.1186/s13023-018-0765-y
47. Vlahou A, Hallinan D, Apweiler R. *et al.* Data sharing under the general data protection regulation. *Hypertension* 2021;**77**:1029–35. https://doi.org/10.1161/HYPERTENSIONAHA.120.16340
48. Chicco D, Cumbo F, Angione C. Ten quick tips for avoiding pitfalls in multi-omics data integration analyses. *PLoS Comput Biol* 2023;**19**:e1011224. https://doi.org/10.1371/journal.pcbi.1011224
49. Brazma A, Hingamp P, Quackenbush J. *et al.* Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* 2001;**29**:365–71. https://doi.org/10.1038/ng1201-365
50. Chervitz SA, Deutsch EW, Field D. *et al.* Data standards for omics data: the basis of data sharing and reuse. In: Mayer B (ed.), *Bioinformatics for Omics Data: Methods and Protocols*. Totowa, NJ: Humana Press, 2011, 31–69. doi: https://doi.org/10.1007/978-1-61779-027-0_2.
51. Wilkinson MD, Dumontier M, Aalbersberg IJJ. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018. https://doi.org/10.1038/sdata.2016.18
52. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nat Methods* 2016;**13**:251–6. https://doi.org/10.1038/nmeth.3746
53. Gürsoy G, Li T, Liu S. *et al.* Functional genomics data: privacy risk assessment and technological mitigation. *Nat Rev Genet* 2022;**23**:245–58. https://doi.org/10.1038/s41576-021-00428-7
54. Bandeira N, Deutsch EW, Kohlbacher O. *et al.* Data management of sensitive human proteomics data: current practices, recommendations, and perspectives for the future. *Mol Cell Proteomics* 2021;**20**:100071. https://doi.org/10.1016/j.mcpro.2021.100071
55. Cope H, Willis CRG, MacKay MJ. *et al.* Routine omics collection is a golden opportunity for European human research in space and analog environments. *PATTER* 2022;**3**:100550. https://doi.org/10.1016/j.patter.2022.100550
56. Yang Q, Liu Y, Chen T. *et al.* Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol* 2019;**10**:12:1–19. https://doi.org/10.1145/3298981
57. Cai Z, Poulos RC, Liu J. *et al.* Machine learning for multi-omics data integration in cancer. *iScience* 2022;**25**:103798. https://doi.org/10.1016/j.isci.2022.103798
58. Chai H, Zhou X, Zhang Z. *et al.* Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Comput Biol Med* 2021;**134**:104481. https://doi.org/10.1016/j.compbiomed.2021.104481

59. Zhang J, Bajari R, Andric D. *et al.* The International Cancer Genome Consortium Data Portal. *Nat Biotechnol* 2019;**37**:367–9. https://doi.org/10.1038/s41587-019-0055-9
60. Reitz CJ, Kuzmanov U, Gramolini AO. Multi-omic analyses and network biology in cardiovascular disease. *Proteomics* 2023;**23**:2200289. https://doi.org/10.1002/pmic.202200289
61. van Keulen D, van Koeverden I, Boltjes A. *et al.* Common variants associated with OSMR expression contribute to carotid plaque vulnerability, but not to cardiovascular disease in humans. *Front Cardiovasc Med* 2021;**8**:658915. https://doi.org/10.3389/fcvm.2021.658915
62. Obón-Santacana M, Vilardell M, Carreras A. *et al.* GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. *BMJ Open* 2018;**8**:e018324. https://doi.org/10.1136/bmjopen-2017-018324
63. Das V, Narayanan S, Zhang X. *et al.* Multi-omics data integration from patients with carotid stenosis illuminates key molecular signatures of atherosclerotic instability. *medRxiv* 2025. https://doi.org/10.1101/2025.05.12.25327328
64. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet* 2017;**8**:84. https://doi.org/10.3389/fgene.2017.00084
65. Picard M, Scott-Boyer M-P, Bodein A. *et al.* Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* 2021;**19**:3735–46. https://doi.org/10.1016/j.csbj.2021.06.030
66. Kaur P, Singh A, Chana I. Computational techniques and tools for omics data analysis: state-of-the-art, challenges, and future directions. *Arch Computat Methods Eng* 2021;**28**:4595–631. https://doi.org/10.1007/s11831-021-09547-0
67. Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform* 2018;**19**:bbw113–340. https://doi.org/10.1093/bib/bbw113
68. Stein-O'Brien GL, Arora R, Culhane AC. *et al.* Enter the matrix: factorization uncovers knowledge from omics. *Trends Genet* 2018;**34**:790–805. https://doi.org/10.1016/j.tig.2018.07.003
69. Argelaguet R, Velten B, Arnol D. *et al.* Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;**14**:e8124. https://doi.org/10.15252/msb.20178124
70. Misra BB, Langefeld CD, Olivier M. *et al.* Integrated omics: tools, advances, and future approaches. *J Mol Endocrinol* 2018;**62**:R21–R45. https://doi.org/10.1530/JME-18-0055
71. Cen W, Zhou F, Ren J. *et al.* A selective review of multi-level omics data integration using variable selection. *High-Throughput* 2019;**8**:4. https://doi.org/10.3390/ht8010004
72. Gruca A, Henzel J, Kostorz I. *et al.* MAINE: a web tool for multi-omics feature selection and rule-based data exploration. *Bioinformatics* 2022;**38**:1773–5. https://doi.org/10.1093/bioinformatics/btab862
73. Briscik M, Tazza G, Dillies M-A. *et al.* Supervised Multiple Kernel Learning approaches for multi-omics data integration. BioData Mining 2024;**17**:53. https://doi.org/10.1186/s13040-024-00406-9
74. İ. B. Aydilek, Examining effects of the support vector machines kernel types on biomedical data classification. In: *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*. Institute of Electrical and Electronics Engineers (IEEE), **2018**, pp. 1–4. https://doi.org/10.1109/IDAP.2018.8620879.
75. Valous NA, Popp F, Zörnig I. *et al.* Graph machine learning for integrated multi-omics analysis. *Br J Cancer* 2024;**131**:205–11. https://doi.org/10.1038/s41416-024-02706-7
76. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. *Brief Bioinform* 2022;**23**:bbab454. https://doi.org/10.1093/bib/bbab454
77. G. Zhou, S. Li, and J. Xia, Network-based approaches for multi-omics integration. In: Li S (ed.), *Computational Methods and Data Analysis for Metabolomics*. New York, NY: Springer US, 2020, 469–87. doi: https://doi.org/10.1007/978-1-0716-0239-3_23.
78. Wang B, Mezlini AM, Demir F. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014;**11**:333–7. https://doi.org/10.1038/nmeth.2810
79. Chierici M, Bussola N, Marcolini A. *et al.* Integrative network fusion: a multi-omics approach in molecular profiling. *Front Oncol* 2020;**10**:1065. https://doi.org/10.3389/fonc.2020.01065
80. Kańduła MM, Aldoshin AD, Singh S. *et al.* ViLoN—a multilayer network approach to data integration demonstrated for patient stratification. *Nucleic Acids Res* 2023;**51**:e6. https://doi.org/10.1093/nar/gkac988
81. Gaynanova I, Li G. Structural learning and integrative decomposition of multi-view data. *Biometrics* 2019;**75**:1121–32. https://doi.org/10.1111/biom.13108
82. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999;**401**:788–91. https://doi.org/10.1038/44565
83. Zhang S, Liu C-C, Li W. *et al.* Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res* 2012;**40**:9379–91. https://doi.org/10.1093/nar/gks725
84. Muller E, Shiryan I, Borenstein E. Multi-omic integration of microbiome data for identifying disease-associated modules. *Nat Commun* 2024;**15**:2621. https://doi.org/10.1038/s41467-024-46888-3
85. Jiang M-Z, Aguet F, Ardlie K. *et al.* Canonical correlation analysis for multi-omics: application to cross-cohort analysis. *PLoS Genet* 2023;**19**:e1010517. https://doi.org/10.1371/journal.pgen.1010517
86. I. Marín de Mas, Chapter sixteen - multiomic data integration and analysis via model-driven approaches. In: Jaumot J, Bedia C, and Tauler R (eds.), *Comprehensive Analytical Chemistry*, in Data Analysis for Omic Sciences: Methods and Applications, vol. 82. Amsterdam, The Netherlands: Elsevier, 2018, 447–76. https://doi.org/10.1016/bs.coac.2018.07.005
87. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 2017;**35**:498–507. https://doi.org/10.1016/j.tibtech.2017.02.012
88. Li S, Łabaj PP, Zumbo P. *et al.* Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotechnol* 2014;**32**:888–95. https://doi.org/10.1038/nbt.3000
89. Łabaj PP, Kreil DP. Sensitivity, specificity, and reproducibility of RNA-Seq differential expression calls. *Biol Direct* 2016;**11**:66. https://doi.org/10.1186/s13062-016-0169-7
90. Su Z, Fang H, Hong H. *et al.* An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. *Genome Biol* 2014;**15**:523. https://doi.org/10.1186/s13059-014-0523-y
91. Siriwardhana C, Datta S, Datta S. Inter-platform concordance of gene expression data for the prediction of chemical mode of action. *Biol Direct* 2016;**11**:67. https://doi.org/10.1186/s13062-016-0167-9
92. Fürst A, Rumetshofer E, Lehner J. *et al.* CLOOB: modern Hopfield networks with InfoLOOB outperform CLIP. arXiv, 2022, arXiv:2110.11316 [cs.LG]. https://doi.org/10.48550/arXiv.2110.11316
93. Radford A. *et al.* Learning transferable visual models from natural language supervision. In: Meila M, Zhang T (eds.), *Proceedings*

of the 38th International Conference on Machine Learning, vol. 139. PMLR, 2021, 8748–63. Available: https://proceedings.mlr.press/v139/radford21a.html.
94. Shi L, Campbell G, Jones WD. *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol* 2010;**28**:827–38. https://doi.org/10.1038/nbt.1665
95. Shi L, Reid LH, Jones WD. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006;**24**:1151–61. https://doi.org/10.1038/nbt1239
96. Su Z, Łabaj PP, Li S. *et al.* A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* 2014;**32**:903–14. https://doi.org/10.1038/nbt.2957
97. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Gyon I. et al. (eds.), *Advances in Neural Information Processing Systems*, vol. 30. 2017.
98. Flores JE, Claborne DM, Weller ZD. *et al.* Missing data in multi-omics integration: recent advances through artificial intelligence. *Front Artif Intell* 2023;**6**:1098308. https://doi.org/10.3389/frai.2023.1098308
99. Shahjaman M, Rahman MR, Islam T. *et al.* rMisbeta: a robust missing value imputation approach in transcriptomics and metabolomics data. *Comput Biol Med* 2021;**138**:104911. https://doi.org/10.1016/j.compbiomed.2021.104911
100. Song M, Greenbaum J, Luttrell J IV. *et al.* A review of integrative imputation for multi-omics datasets. *Front Genet* 2020;**11**:570255. https://doi.org/10.3389/fgene.2020.570255
101. Karimi-Bidhendi S, Arafati A, Cheng AL. *et al.* Fully-automated deep-learning segmentation of pediatric cardiovascular magnetic resonance of patients with complex congenital heart diseases. *J Cardiovasc Magn Reson* 2020;**22**:80. https://doi.org/10.1186/s12968-020-00678-0
102. Krittanawong C, Johnson KW, Hershman SG. *et al.* Big data, artificial intelligence, and cardiovascular precision medicine. *Expert Rev Precis Med Drug Dev* 2018;**3**:305–17. https://doi.org/10.1080/23808993.2018.1528871
103. 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. https://doi.org/10.1172/JCI129203
104. Bishop CM. *Pattern Recognition and Machine Learning. In Information Science and Statistics*. NY: Springer New York, 2006. Available: https://link.springer.com/book/9780387310732
105. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
106. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv 2020, arXiv:1802.03426v3. https://doi.org/10.48550/arXiv.1802.03426
107. Chari T, Pachter L. The specious art of single-cell genomics. *PLoS Comput Biol* 2023;**19**:e1011288. https://doi.org/10.1371/journal.pcbi.1011288
108. Leonelli S. The challenges of big data biology. *eLife* 2019;**8**:e47381. https://doi.org/10.7554/eLife.47381
109. Fortier I, Raina P, van den Heuvel ER. *et al.* Maelstrom Research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol* 2017;**46**:dyw075–105. https://doi.org/10.1093/ije/dyw075
110. Wallentin L, Gale CP, Maggioni A. *et al.* EuroHeart: European Unified Registries On Heart Care Evaluation and Randomized Trials: an ESC project to develop a new IT registry system which will encompass multiple features of cardiovascular medicine. *Eur Heart J* 2019;**40**:2745–9. https://doi.org/10.1093/eurheartj/ehz599
111. Adhikari K, Patten SB, Patel AB. *et al.* Data harmonization and data pooling from cohort studies: a practical approach for data management. *Int J Popul Data Sci* 2021;**6**:21. https://doi.org/10.23889/ijpds.v6i1.1680
112. Kumar G, Basri S, Imam AA. *et al.* Data harmonization for heterogeneous datasets: a systematic literature review. *Appl Sci* 2021;**11**:8275. https://doi.org/10.3390/app11178275
113. Petzschner FH. Practical challenges for precision medicine. *Science* 2024;**383**:149–50. https://doi.org/10.1126/science.adm9218
114. Kaufman JD, Elkind MSV, Bhatnagar A. *et al.* Guidance to reduce the cardiovascular burden of ambient air pollutants: a policy statement from the American Heart Association. *Circulation* 2020;**142**:e432–47. https://doi.org/10.1161/CIR.0000000000000930
115. Lamas GA, Bhatnagar A, Jones MR. *et al.* Contaminant metals as cardiovascular risk factors: a scientific statement from the American Heart Association. *J Am Heart Assoc* 2023;**12**:e029852. https://doi.org/10.1161/JAHA.123.029852
116. Rose JJ, Krishnan-Sarin S, Exil VJ. *et al.* Cardiopulmonary impact of electronic cigarettes and vaping products: a scientific statement from the American Heart Association. *Circulation* 2023;**148**:703–28. https://doi.org/10.1161/CIR.0000000000001160
117. Virani SS, Newby LK, Arnold SV. *et al.* 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA guideline for the management of patients with chronic coronary disease. *J Am Coll Cardiol* 2023;**82**:833–955. https://doi.org/10.1016/j.jacc.2023.04.003
118. Winterstein AG, Ehrenstein V, Brown JS. *et al.* A road map for peer review of real-world evidence studies on safety and effectiveness of treatments. *Diabetes Care* 2023;**46**:1448–54. https://doi.org/10.2337/dc22-2037
119. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016;**183**:758–64. https://doi.org/10.1093/aje/kwv254
120. Tchetgen Tchetgen EJ, Park C, Richardson DB. Universal difference-in-differences for causal inference in epidemiology. *Epidemiology* 2024;**35**:16–22. https://doi.org/10.1097/EDE.0000000000001676
121. Ruiz-Hernandez A, Navas-Acien A, Pastor-Barriuso R. *et al.* Declining exposures to lead and cadmium contribute to explaining the reduction of cardiovascular mortality in the US population, 1988–2004. *Int J Epidemiol* 2017;**46**:1903–12. https://doi.org/10.1093/ije/dyx176
122. Desai RJ, Levin R, Lin KJ. *et al.* Bias implications of outcome misclassification in observational studies evaluating association between treatments and all-cause or cardiovascular mortality using administrative claims. *J Am Heart Assoc* 2020;**9**:e016906. https://doi.org/10.1161/JAHA.120.016906
123. Dunn W Jr, Burgun A, Krebs M-O. *et al.* Exploring and visualizing multidimensional data in translational research platforms. *Brief Bioinform* 2017;**18**:bbw080–1056. https://doi.org/10.1093/bib/bbw080
124. L. Linsen, H. Hagen, and B. Hamann, Eds., *Visualization in Medicine and Life Sciences. In Mathematics and Visualization*. Berlin, Heidelberg: Springer, 2008. https://doi.org/10.1007/978-3-540-72630-2.
125. Sukhavasi K, Mocci G, Ma L. *et al.* Single-cell RNA sequencing reveals sex differences in the subcellular composition and associated gene-regulatory network activity of human carotid plaques. *Nat Cardiovasc Res* 2025;**4**:412–32. https://doi.org/10.1038/s44161-025-00628-y

126. Hernán MA, Monge S. Selection bias due to conditioning on a collider. *BMJ* 2023;**381**:p1135. https://doi.org/10.1136/bmj.p1135
127. Shi H, Zhang G, Wang J. *et al.* Studying dynamic features in myocardial infarction progression by integrating miRNA-transcription factor co-regulatory networks and time-series RNA expression data from peripheral blood mononuclear cells. *PloS One* 2016;**11**:e0158638. https://doi.org/10.1371/journal.pone.0158638
128. Varga TV. Algorithmic fairness in cardiovascular disease risk prediction: overcoming inequalities. *Open Heart* 2023;**10**:e002395. https://doi.org/10.1136/openhrt-2023-002395
129. Álvarez-Gálvez J, Jaime-Castillo AM. The impact of social expenditure on health inequalities in Europe. *Soc Sci Med* 2018;**200**:9–18. https://doi.org/10.1016/j.socscimed.2018.01.006
130. de Mestral C, Stringhini S. Socioeconomic status and cardiovascular disease: an update. *Curr Cardiol Rep* 2017;**19**:115. https://doi.org/10.1007/s11886-017-0917-z
131. Kist JM, Smit GWG, Mairuhu ATA. *et al.* Large health disparities in cardiovascular death in men and women, by ethnicity and socioeconomic status in an urban based population cohort. *eClinicalMedicine* 2021;**40**:101120. https://doi.org/10.1016/j.eclinm.2021.101120
132. SCORE2 Working Group and ESC Cardiovascular Risk Collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J* 2021;**42**:2439–54. https://doi.org/10.1093/eurheartj/ehab309
133. Kist JM, Vos RC, Mairuhu ATA. *et al.* SCORE2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the Netherlands: an external validation study. *eClinicalMedicine* 2023;**57**:101862. https://doi.org/10.1016/j.eclinm.2023.101862
134. Anderer S, Hswen Y. 'Scalable privilege'—how AI could turn data from the best medical systems into better care for all. *JAMA* 2024;**331**:459–62. https://doi.org/10.1001/jama.2023.21719
135. European Commission. Ethics guidelines for trustworthy AI. *Publications Office of the European Union, Apr* 2019.
136. SCORE2-OP working group and ESC Cardiovascular risk collaboration. SCORE2-OP risk prediction algorithms: estimating incident cardiovascular event risk in older persons in four geographical risk regions. *Eur Heart J* 2021;**42**:2455–67. https://doi.org/10.1093/eurheartj/ehab312
137. Kartoun U, Khurshid S, Kwon BC. *et al.* Prediction performance and fairness heterogeneity in cardiovascular risk models. *Sci Rep* 2022;**12**:12542. https://doi.org/10.1038/s41598-022-16615-3
138. Larkin H. What to know about PREVENT, the AHA's new cardiovascular disease risk calculator. *JAMA* 2024;**331**:277–9. https://doi.org/10.1001/jama.2023.25115
139. Khan SS, Coresh J, Pencina MJ. *et al.* Novel prediction equations for absolute risk assessment of total cardiovascular disease incorporating cardiovascular-kidney-metabolic health: a scientific statement from the American Heart Association. *Circulation* 2023;**148**:1982–2004. https://doi.org/10.1161/CIR.0000000000001191
140. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010;**38**:894–942. https://doi.org/10.1214/09-AOS729
141. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001;**96**:1348–60. https://doi.org/10.1198/016214501753382273
142. Domingo-Relloso A, Feng Y, Rodriguez-Hernandez Z. *et al.* Omics feature selection with the extended SIS R package: identification of a body mass index epigenetic multimarker in the Strong Heart Study. *Am J Epidemiol* 2024;**193**:1010–8. https://doi.org/10.1093/aje/kwae006
143. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat* 2009;**37**:1733–51. https://doi.org/10.1214/08-AOS625
144. Xiao N, Xu Q-S. Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection. *J Stat Comput Simul* 2015;**85**:3755–65. https://doi.org/10.1080/00949655.2015.1016944
145. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;**3**:18. https://doi.org/10.1186/s41512-019-0064-7
146. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J Clin Epidemiol* 2004;**57**:1138–46. https://doi.org/10.1016/j.jclinepi.2004.04.003
147. Greenland S. Modeling and variable selection in epidemiologic analysis. *Am J Public Health* 1989;**79**:340–9. https://doi.org/10.2105/AJPH.79.3.340
148. Butte AJ. Artificial intelligence—from starting pilots to scalable privilege. *JAMA Oncol* 2023;**9**:1341–2. https://doi.org/10.1001/jamaoncol.2023.2867
149. Omiye JA, Gui H, Daneshjou R. *et al.* Principles, applications, and future of artificial intelligence in dermatology. *Front Med* 2023;**10**:1278232. https://doi.org/10.3389/fmed.2023.1278232
150. Zuber V, Grinberg NF, Gill D. *et al.* Combining evidence from Mendelian randomization and colocalization: review and comparison of approaches. *Am J Hum Genet* 2022;**109**:767–82. https://doi.org/10.1016/j.ajhg.2022.04.001
151. Amadoz A, Hidalgo MR, Çubuk C. *et al.* A comparison of mechanistic signaling pathway activity analysis methods. *Brief Bioinform* 2019;**20**:1655–68. https://doi.org/10.1093/bib/bby040
152. Peña-Chilet M, Esteban-Medina M, Falco MM. *et al.* Using mechanistic models for the clinical interpretation of complex genomic variation. *Sci Rep* 2019;**9**:18937. https://doi.org/10.1038/s41598-019-55454-7
153. Hidalgo MR, Cubuk C, Amadoz A. *et al.* High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. *Oncotarget* 2016;**8**:5160–78. https://doi.org/10.18632/oncotarget.14107
154. Loucera C, Esteban-Medina M, Rian K. *et al.* Drug repurposing for COVID-19 using machine learning and mechanistic models of signal transduction circuits related to SARS-CoV-2 infection. *Sig Transduct Target Ther* 2020;**5**:290–3. https://doi.org/10.1038/s41392-020-00417-y
155. Bojic S, Falco MM, Stojkovic P. *et al.* Platform to study intracellular polystyrene nanoplastic pollution and clinical outcomes. *Stem Cells* 2020;**38**:1321–5. https://doi.org/10.1002/stem.3244
156. Levin MG, Burgess S. Mendelian randomization as a tool for cardiovascular research: a review. *JAMA Cardiol* 2024;**9**:79–89. https://doi.org/10.1001/jamacardio.2023.4115
157. Schork NJ. Personalized medicine: time for one-person trials. *Nature* 2015;**520**:609–11. https://doi.org/10.1038/520609a
158. Peñalvo JL, Mertens E, Ademović E. *et al.* Unravelling data for rapid evidence-based response to COVID-19: a summary of the unCoVer protocol. *BMJ Open* 2021;**11**:e055630. https://doi.org/10.1136/bmjopen-2021-055630
159. Peñalvo J, Mertens E, Cottam J. *et al.* Federated learning for describing COVID-19 patients and hospital outcomes:

an unCoVer analysis. *Eur J Public Health* 2022;**32**:ckac131.254. https://doi.org/10.1093/eurpub/ckac131.254
160. Mertens E, Ademovic E, Majdan M. *et al.* Associations of pre-existing comorbidities and COVID-19 in-hospital mortality: an unCoVer analyses. *Eur J Public Health* 2022;**32**:ckac130.015. https://doi.org/10.1093/eurpub/ckac130.015
161. Peláez A, Ruiz del Árbol N, Vázquez Sellán A. *et al.* Clinical characteristics and outcomes among hospitalised COVID-19 patients across epidemic waves in Spain: an unCoVer analysis. *Med Clin (Barc)* 2024;**162**:523–31. https://doi.org/10.1016/j.medcli.2023.12.030
162. Observational Health Data Sciences and Informatics. The book of OHDSI. Independently published, 2019. Available: https://ohdsi.github.io/TheBookOfOhdsi/
163. IFHIR Specification v5.0.0: R5. Accessed: Jun. 11, 2025. [Online]. Available: https://www.hl7.org/fhir/
164. Store and Document Data with Opal. Accessed: Jun. 11, 2025. [Online]. Available: https://www.obiba.org/pages/products/opal/
165. Brisimi TS, Chen R, Mela T. *et al.* Federated learning of predictive models from federated electronic health records. *Int J Med Inform* 2018;**112**:59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007
166. Rieke N, Hancox J, Li W. *et al.* The future of digital health with federated learning. *npj Digit Med* 2020;**3**:1–7. https://doi.org/10.1038/s41746-020-00323-1
167. Guidelines 03/2020 on the processing of data concerning health for the purpose of scientific research in the context of the COVID-19 outbreak | European Data Protection Board, 2020. Available: https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-032020-processing-data-concerning-health-purpose_en
168. Global Burden of Disease Long COVID Collaborators. Estimated global proportions of individuals with persistent fatigue, cognitive, and respiratory symptom clusters following symptomatic COVID-19 in 2020 and 2021. *JAMA* 2022;**328**:1604–15. https://doi.org/10.1001/jama.2022.18931
169. Gyöngyösi M, Alcaide P, Asselbergs FW. *et al.* Long COVID and the cardiovascular system—elucidating causes and cellular mechanisms in order to develop targeted diagnostic and therapeutic strategies: a joint scientific statement of the ESC working groups on cellular biology of the heart and myocardial and pericardial diseases. *Cardiovasc Res* 2023;**119**:336–56. https://doi.org/10.1093/cvr/cvac115
170. Shapiro B, Forger DB. Reducing chronic disease may just be a walk in the park. *CR Med* 2022;**3**:100874. https://doi.org/10.1016/j.xcrm.2022.100874
171. Kim J, Campbell AS, de Ávila BE-F. *et al.* Wearable biosensors for healthcare monitoring. *Nat Biotechnol* 2019;**37**:389–406. https://doi.org/10.1038/s41587-019-0045-y
172. Chen C, Ding S, Wang J. Digital health for aging populations. *Nat Med* 2023;**29**:1623–30. https://doi.org/10.1038/s41591-023-02391-8
173. Mars N, Kerminen S, Feng YA. *et al.* Genome-wide risk prediction of common diseases across ancestries in one million people. *Cell Genomics* 2022;**2**:100118. https://doi.org/10.1016/j.xgen.2022.100118
174. Khera AV, Chaffin M, Aragam KG. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018;**50**:1219–24. https://doi.org/10.1038/s41588-018-0183-z
175. Dikilitas O, Schaid DJ, Kosel ML. *et al.* Predictive utility of polygenic risk scores for coronary heart disease in three major racial and ethnic groups. *Am J Hum Genet* 2020;**106**:707–16. https://doi.org/10.1016/j.ajhg.2020.04.002
176. Widén E, Junna N, Ruotsalainen S. *et al.* How communicating polygenic and clinical risk for atherosclerotic cardiovascular disease impacts health behavior: an observational follow-up study. *Circ Genom Precis Med* 2022;**15**:e003459. https://doi.org/10.1161/CIRCGEN.121.003459
177. Jermy B, Läll K, Wolford BN. *et al.* A unified framework for estimating country-specific cumulative incidence for 18 diseases stratified by polygenic risk. *Nat Commun* 2024;**15**:5007. https://doi.org/10.1038/s41467-024-48938-2
178. Philips. *Future Health Index 2020: The Age of Opportunity: Empowering the Next Generation to Transform Healthcare* 2020. [Online]. Available: https://www.philips.com/a-w/about/news/future-health-index/reports/2020/the-age-of-opportunity/download-reports.html.
179. openEHR Conformance Guide. Accessed: Jun. 11, 2025. [Online]. Available: https://specifications.openehr.org/releases/CNF/development/guide.html
180. Overview of SNOMED CT. Accessed: Jun. 11, 2025. [Online]. Available: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html
181. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. United States: Basic Books, 2019. Available: https://www.lehmanns.de/shop/wirtschaft/44020584-9781541644632-deep-medicine

Briefings in Bioinformatics, 2025, **26(5)**, bbaf526
https://doi.org/10.1093/bib/bbaf526
Review