Taylor & Francis
Taylor & Francis Group

Expert Review of Precision Medicine and Drug Development

Personalized medicine in drug development and clinical practice

ISSN: (Print) 2380-8993 (Online) Journal homepage: http://www.tandfonline.com/loi/tepm20

Big data, artificial intelligence, and cardiovascular precision medicine

Chayakrit Krittanawong, Kipp W. Johnson, Steven G. Hershman & W.H. Wilson Tang

To cite this article: Chayakrit Krittanawong, Kipp W. Johnson, Steven G. Hershman & W.H. Wilson Tang (2018) Big data, artificial intelligence, and cardiovascular precision medicine, Expert Review of Precision Medicine and Drug Development, 3:5, 305-317, DOI: 10.1080/23808993.2018.1528871

To link to this article: https://doi.org/10.1080/23808993.2018.1528871

Accepted author version posted online: 26 Sep 2018.
Published online: 10 Oct 2018.

EXPERT REVIEW OF PRECISION MEDICINE AND DRUG DEVELOPMENT
2018, VOL. 3, NO. 5, 305–317
https://doi.org/10.1080/23808993.2018.1528871

Taylor & Francis
Taylor & Francis Group

REVIEW

Check for updates

Big data, artificial intelligence, and cardiovascular precision medicine

Chayakrit Krittanawong[a], Kipp W. Johnson[b], Steven G. Hershman[c,d] and W.H. Wilson Tang[e,f,g]

[a]Department of Internal Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [b]Institute for Next Generation Healthcare, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA; [c]Department of Medicine, Stanford University, Stanford, CA, USA; [d]Division of Cardiovascular Medicine, Department of Medicine, Stanford University, Stanford, CA, USA; [e]Department of Cardiovascular Medicine, Heart and Vascular Institute, Cleveland Clinic, Cleveland, OH, USA; [f]Department of Cellular and Molecular Medicine, Lerner Research Institute, Cleveland, OH, USA; [g]Center for Clinical Genomics, Cleveland Clinic, Cleveland, OH, USA

ABSTRACT

Introduction: Cardiovascular diseases (CVDs) are chronic, heterogeneous diseases which are generally classified according to clinical presentation. However, the arrival of big data and analytical methods presents an opportunity to better understand these disease entities.

Areas covered: This review article highlights: (1) the potential of a big data approaches with emerging technology to explore the heterogeneity of CVDs; (2) current challenges of a big data approach; and (3) the future of precision cardiovascular medicine.

Expert commentary: Overall, most of the current data utilizing big data techniques remain largely descriptive and retrospective. Precision medicine, or N-of-1, approaches have not yet allowed for consistent interpretation since there is no 'standard' of how to best apply treatment approaches in a field where evidence-based medicine is based largely on randomized controlled trials. The risk score and biomarker-based approaches have been utilized with some 'validation' studies, but more in-depth biomarkers (i.e. *pharmacogenomic biomarkers*) have failed to demonstrate incremental benefits. Exploring novel CVD phenotypes by integrating existing medical variables, multi-omics, lifestyle, and environmental data using artificial intelligence is vitally important and may allow us to digitize future clinical trials, potentially leading to novel therapies.

ARTICLE HISTORY
Received 29 July 2018
Accepted 24 September 2018

KEYWORDS
Big data; cardiovascular precision medicine; precision medicine; big data approach; omics

1. Heterogeneous cardiovascular diseases

Cardiovascular diseases (CVDs) are chronic, heterogeneous diseases that have generally been identified and categorized into phenotypes according to their clinical presentation. However, due to the complexity of chronic CVDs, it is likely that multiple independent etiologies manifest similarly in the clinic. This ultimately results in differing responses to standardized treatment regimens, which are derived from broad disease characterizations. Understanding the reasons for these differences presents an avenue through which to improve patient care. Although the heterogeneous pathophysiology of CVDs has been extensively studied, the emergence of new analytical methods drawn from the statistical and computer science communities presents a powerful tool for better understanding. CVDs are associated with multiple phenotypes that result from genetics, metabolomics, environmental, and behavioral or lifestyle perturbations [1,2]. Hypertension, atrial fibrillation (AF), heart failure with preserved ejection fraction (HFpEF), Takotsubo syndrome, Cardiorenal syndrome, and spontaneous coronary artery dissection are known to be heterogeneous in their etiology and pathophysiology, and different phenotypes may respond to treatment in different ways [3–7]. Most clinical research studies are based on current clinical diagnosis and known validated parameters to investigate endpoints or outcomes. However, many parameters are not well-validated, and there are some emerging variables or combinations of variables that could potentially be used as guided parameters for prognosis and treatment in order to replace older metrics [8–10]. The diagnostic criteria of diastolic dysfunction or HFpEF, for example, are not well-defined, and the guidelines have varied over time [8,11]. Recent studies have demonstrated that an artificial intelligence (AI) method involving high-dimensional unsupervised clustering may have the potential to classify heterogeneous clinical CV conditions more accurately than current diagnostic criteria [6,12].

2. Big data and precision medicine: where we are

The zeitgeist of the information age may be the use of so-called 'big data' to analyze, interpret, and alter the human condition. Biomedical science, and cardiovascular medicine, in particular, is at the forefront of this movement. Central components of the use of big data are effective strategies for the challenges of storing, managing, and analyzing a multitude of large of datasets. The term 'big data,' used in modern-day scientific communities, medical literature, and at scientific conferences, is frequently referred to as the 5 Vs (volume, velocity, variety, veracity, and valorization), which cannot be analyzed or interpreted using traditional data processing methods [13]. However, the definition of big data is still

CONTACT Chayakrit Krittanawong ✉ Chayakrit.Krittanawong@mountsinai.org Department of Internal Medicine, Icahn School of Medicine at Mount Sinai, 1000 10th Ave, New York, NY 10019

tenuous and not well-established. Datasets do not necessarily need to be a large number of observations, but they may be considered 'big data' due to the potential of the data in the context of innovation, how meaningful it is, if it is multidimensional, and how its value will increase over time [14]. Examples of big data include datasets combining human gut microbiome sequencing, genomics, metabolomics, proteomics, transcriptomics, social media data, and data from standardized electronic health records (EHRs) or precision medicine platforms (e.g. AHA Precision Medicine Platforms or the UCSF Precision Medicine Platform) [15,16]. Several decades of translational, epidemiological, and clinical multiethnic studies of CVDs have been found to be largely inconsistent. With emerging analytic technology, a big data approach would attempt to classify heterogeneous CVDs that could facilitate precision CV medicine [17]. To date, many curated and uncurated medical and environmental databases are freely available to the public which could be used for data analysis. Tables 1–3 demonstrate both known variables (i.e. clinical variables, genetics or multi-omics variables) and potential latent variables, including environmental factors (i.e. media consumption, transportation use, restaurant selection, or illicit drugs use), epidemiological factors (i.e. Google Flu Trends) may be explored in CVDs. Some particularly exciting resources for precision medicine are the so-called 'biobanks.' These are mass collections of biomedical specimens which may be linked to retrospective EHRs in order to facilitate a wide variety of retrospective analyses [18]. Well-curated biobanks like Mount Sinai's BioMe, Vanderbilt's BioVU, Northwestern's NUgene, Penn Medicine's BioBank, Stanford Cardiovascular Institute's Biobank (SCVI) and GenePool, or more recently the massive UK BioBank (n = 500,000 patients) are exciting opportunities for biomedical discovery in precision medicine, and they can be accessed by various innovative actors, public and private, throughout the world. However, drawbacks for this research are the often limiting data usage agreement policies for these resources, which in some cases (i.e. Mount Sinai's BioMe), only allow use by faculty members from the participating institutions. As such, much of the research potential from these important biobanks are siloed away, unable to fulfill their great potential. A novel method of collecting big data is using mobile health apps. Studies like MyHeart Counts [19], Health eHeart [20], MyGene Rank [21], and the Apple Heart Study [22] have used the app store as a recruitment tool and iOS applications for data collection; using such an approach, it is not uncommon to recruit as many as ~10^5 participants. Many such studies are designed to have an open data portal accessible to qualified researchers [23–25]. Other study apps, like VascTrac, are applied to patients populated in a clinical setting [26]. In contrast, resources containing uncurated or unprocessed big data are much harder to use, but the application of big data into clinical decision-making using emerging techniques drawn from the field of AI, machine learning (ML), or deep learning (DL) has the potential to transform the current practice of cardiovascular health (CVH) into precision medicine [17,27,28]. Big data analysis using AI allows us to classify heterogeneous CVDs into more precise phenotypes of CVD, leading to personalized, targeted therapy [29]. To date, big data holds great promise for solutions in CV research in various aspects. First, big data can be used to allow integration of EHR, multi-omic data, gut microbiome sequencing, diet consumption diaries, physical activity information, sleep habit information from wearable technology, and emotional sentiments from social media posts to determine the multidimensional associations between these factors [30,31]. Second, the relationships between variables from big data tend to show nonlinear relationships, which require an advanced tool like AI for sophisticated analysis. However, the main limitation of a big data approach is the heterogeneity of multiple databases (i.e. different ICD code versions, different diagnostic criteria, different laboratories, and different software vendors) [32,33]. Therefore, the harmonization of data, particularly from different databases, is needed before performing an analysis and creating an automated prediction model for CVH recommendations for individuals. In conclusion, a big data approach to the study of heterogeneous CVD is currently challenging but appears promising. Thus, future AHA/ACC/ESC guidelines may be needed to take a big data approach into account.

3. Data processing step

In general, there are several steps required to apply big data to cardiovascular medicine (Figure 1). First, and most importantly, the discovery of datasets pertinent to the task at hand is required. This may include searching the wide variety of databases that are already available (Tables 1–3). De-identification is a crucial step for data privacy to protect patient information according to the HIPAA Privacy Rule, although this should generally be performed before the data is released [34]. Nonetheless, researchers re-using data have an obligation to maintain the confidentiality of any patient records they may analyze and to take appropriate steps to safeguard their data. Second, synchronization between different databases can generate new insights of disease pathogenesis, particularly heterogeneous diseases [35]. There are many data warehouse management tools that can be used to assist with database integration such as Google's visualizer [36], Galaxy [37], Spark SQL [38], Amazon Redshift [39], BIME Analytics [40], and Google BigQuerry [41]. However, there are certain limitations. First, the integration between different databases, particularly those including clinical variables and lifestyle variables, is still a limitation because of the heterogeneity in any number of variables which may be shared among those databases. For example, participant IDs (or even participants) are usually not shared across different freely available resources – in many cases, this makes patient-level analyses impossible. Second, these datasets have generally not been designed to work well together in the context of file format, columns/rows, transformation, or distribution. Third, some databases such as toxicology or metagenomics are designed primarily for the experts in those fields using specific terminology which may be hard to explore or combine without publicly available resources such as wiki-style websites. Fourth, data imputation is a quality control step that can be applied to improve data quality and accuracy after analysis [35,42,43]. Fifth, data modeling is a common term used in ML [44]. It is a model that needs to be generated. In general, the implementation of

Table 1. Examples of Omics database.

Omics database	Type of data	Details	Number of samples	Link
Global Biobank Engine	Phenotypes, variants, genetics, HLA alleles	A web-based tool that enables the exploration of the relationship between genotype and phenotype	500,000 individuals	biobankengine.stanford.edu
Trans-Omics for Precision Medicine (TOPMed)	Omics data – RNA, gene, and metabolite	RNA, gene, and metabolite profiles from individuals who participated in the NHLBI-funded Multi-Ethnic Study of Atherosclerosis (MESA)	Over 90,000 genomes sequences and over 30,000 whole genome sequences in dbGAP	https://www.nhlbi.nih.gov/news/2016/toward-precision-medicine-first-whole-genomes-topmed-now-available-study
BioMe	EHR-linked bio and data repository in New York City	Epidemiologic, molecular, genomic, environment, and lifestyle	32,000 participants	http://icahn.mssm.edu/research/ipm/programs/biome-biobank
Merck Molecular Activity Challenge	The training and test datasets for machine learning practice	Molecule ID, Molecular descriptors and features	15 biological activity data sets	https://github.com/RuwanT/merck
The Human Metabolome Database (HMDB)	Metabolite and protein sequences	(1) Chemical data, (2) clinical data, and (3) molecular biology/biochemistry data	114,099 metabolite entries and 5702 protein sequences	http://www.hmdb.ca/
UK biobanks	Whole genome sequencing, exome sequencing, and genotyping	Genome, exome, online questionnaires (diet, cognitive function, work history and digestive health), EHR, images	500,000 people aged between 40 and 69 years in 2006–2010	http://www.ukbiobank.ac.uk/
Genomics England	Genome sequencing	Genome sequence data, obtained from samples of blood, tissue, and saliva	100,000 genomes and 70,000 patients and family	https://www.genomicsengland.co.uk/the-100000-genomes-project/data/current-research/
UK10K	DNA sequencing	DNA sequence at an order of magnitude deeper than the 1000 Genomes Project for Europe by carrying out genome-wide sequencing of 4000 samples from the TwinsUK and ALSPAC cohorts	Whole genome cohorts (4000), neurodevelopment sample sets (up to 3000 whole exomes), obesity sample sets (2000 whole exomes), and rare diseases sample sets (1000 whole exomes)	http://www.uk10k.org/
PubChem	Chemistry	Chemical structures, identifiers, chemical, physical properties, biological activities, patents, health, safety and toxicity data	95,414,874 compounds, 250,188,056 substances, 1,252,883 bioassays, and 236,181,958 bioActivities	pubchem.ncbi.nlm.nih.gov
MetaCyc	Metabolism	Both primary and secondary metabolism, associated metabolites, reactions, enzymes, and genes	2642 pathways from 2941 different organisms	metacyc.org
Molecular Transducers of Physical Activity (MoTrPAC)	Omics during exercise	$170M NIH Consortium on impact of activity on molecular health	TBD (There is no public data yet)	https://www.motrpac.org/
Chemical Entities of Biological Interest (ChEBI)	Chemistry	'Small' chemical compounds IntEnz, KEGG COMPOUND, PDBeChem, ChEMBL	46,477 fully curated entries, each of which is classified within the ontology and assigned multiple annotations	www.ebi.ac.uk/chebi/
Protein Data Bank (PDB)	Protein	3D shapes of proteins, nucleic acids, and complex assemblies	44,165 distinct protein sequences, 38,467 structures of human sequences, and 10,027 nucleic acid containing structures	www.rcsb.org
The Universal Protein Resource (UniProt)	Proteome and proteins	Functional information on proteins and proteome	Peptide sequences from 172,997 human with 557,713 reviewed and 116,030,110 unreviewed proteins	http://www.uniprot.org/
GenBank	CoreNucleotide (the main collection), dbEST (expressed sequence tags), and dbGSS (genome survey sequences)	DNA sequences	DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI	www.ncbi.nlm.nih.gov/genbank/
The Toxin and Toxin Target Database (T3DB)	Toxin	Mechanisms of toxicity and target proteins for each toxin, detailed toxin data, pollutants, pesticides, drugs, and food toxins	3670 common toxins and environmental pollutants	http://www.t3db.ca/
SMPDB (The Small Molecule Pathway Database)	Small molecule	Small molecule pathways	30,000 human metabolic and disease pathways	http://smpdb.ca/

(Continued)

Table 1. (Continued).

Omics database	Type of data	Details	Number of samples	Link
The Golm Metabolome Database (GMD)	Metabolomics	A repository of sum formula with source tagged annotations for properties such as InChI strings, CAS numbers, IUPAC names, synonyms, cross references or KEGG Pathway names	2.1 million unique sum formula from more than 150 public available databases	http://gmd.mpimp-golm.mpg.de/
BRENDA	Enzymes, organism, pathway, reaction	Comprehensive enzyme database	7341 different enzymes	www.brenda-enzymes.org
MassBank	Mass spectra of metabolites	High-resolution mass spectra of metabolites	605 electron-ionization mass spectrometry (EI-MS), 137 fast atom bombardment MS, and 9276 electrospray ionization (ESI)-MS (n) data of 2337 authentic compounds of metabolites, 11,545 EI-MS and 834 other-MS data of 10,286 volatile natural and synthetic compounds, and 3045 ESI-MS [2] data of 679 synthetic drugs	massbank.eu/MassBank/
BioCyc	Metabolic pathways	Metabolic pathways and operons	13,075 Pathway/genome databases	biocyc.org
NHLBI Exome Sequencing Project	Exome sequencing data	Gene name (HUGO, upper or lower case), gene ID (from NCBI Entrez Gene), chromosomal location, dbSNP rs ID to study genetic contributions to the risk of several heart, lung, and blood phenotypes	>7000 individuals	http://evs.gs.washington.edu/EVS/
Ensembl Genomes	Genomic data	Bacteria, protists, fungi, plants, and invertebrate metazoan genome-scale data	44,048 bacteria, 189 protists, 811 fungi, 45 plants, and 68 Metazoa	http://ensemblgenomes.org/info/data
UCSC Genome Browser	Genomic data	CRISPR/Cas9 trac, gene Interactions, refSeq Genes track and GTEx Gene Track	180 assemblies and over 100 species	genome.ucsc.edu/cgi-bin/hgGateway
Human Microbiome Project	Microbiome data	The collection of all the microorganisms living in association with the human body. These communities consist of a variety of microorganisms including eukaryotes, archaea, bacteria and viruses.	86,843 files, 30,688 samples (the microbial communities from 300 healthy individuals, across several different sites on the human body: nasal passages, oral cavity, skin, gastrointestinal tract, and urogenital tract)	hmpdacc.org
MicrobiomeDB	Microbiome data	Geographic environmental features, 16S rRNA genes, and antibiotic exposures	13,565 samples	http://microbiomedb.org/mbio/
EBI Metagenomics	Metagenomics	All genomes present in any given environment without the need for prior individual identification	129,051 data sets, 17,545 metagenomes and 1727 metatranscriptomes	www.ebi.ac.uk/metagenomics/
Phytozome	Genomic data	All gene sets in Phytozome have been annotated with KOG, KEGG, ENZYME, Pathway and the InterPro family of protein	Phytozome hosts 93 assembled and annotated genomes, from 82 Viridiplantae species	phytozome.jgi.doe.gov/pz/portal.html
UniProt Metagenomic and Environmental Sequences (UniMES)	Metagenomic and environmental data	Metagenomic and environmental data (the amino acid sequence, protein name or description, taxonomic data and citation information)	171,510 human, 83,587 mouse, and 59,676 zebrafish	www.uniprot.org/help/unimes
The HBT (Human Brain Transcriptome)	Genome-wide, exon-level transcriptome	A total of 16 brain regions were sampled: the cerebellar cortex, mediodorsal nucleus of the thalamus, striatum, amygdala, hippocampus, and 11 areas of the neocortex. Genome-wide genotyping data for 2.5 million markers	Over 1340 tissue samples sampled from both hemispheres of postmortem human brains	http://hbatlas.org/
1000 Genomes Project	Whole-genome sequencing	A comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations	84.4 million variants from 2504 individuals	http://www.internationalgenome.org/

(Continued)

Table 1. (Continued).

Omics database	Type of data	Details	Number of samples	Link
Greengenes	Small-subunit rRNA gene (SSU)	Archaeal and bacterial 16S SSU rDNA sequences online full-length small-subunit rRNA gene (SSU) database	90,000 public 16S small-subunit rRNA gene sequences	http://greengenes.lbl.gov
H-Invitational Database (H-InvDB)	Human genes and transcripts	Curated annotations of human genes and transcripts that include gene structures, alternative splicing variants, non-coding functional RNAs, protein functions, functional domains, subcellular localizations, metabolic pathways, protein 3D structure, genetic polymorphisms (SNPs, indels, and microsatellite repeats), relation with diseases, gene expression profiling, and molecular evolutionary features, protein–protein interactions (PPIs) and gene families/groups.	120,558 human mRNAs extracted from the International Nucleotide Sequence Databases (INSD), in addition to 54 978 human FLcDNAs	http://www.h-invitational.jp/

Table 2. Examples of clinical and environmental/lifestyle database.

Database	Type of data	Details	Numbers of samples	Link
Nationwide Inpatient Sample (NIS)	Clinical	ICD-9-CM, demographic, expected payment source, total charges, discharge status, length of stay, severity and comorbidity	NIS collects annual data on 7–8 million hospital stays, reflecting all discharges from around 1000 hospitals	https://www.hcup-us.ahrq.gov/db/nation/nis/nisdbdocumentation.jsp
Nationwide Readmissions Database (NRD)	Clinical	Diagnosis, procedure, patient demographics, expected payment source, costs associated with readmissions, reasons for readmissions, impact of health policy changes	Discharge data from 27 geographically dispersed States	https://www.hcup-us.ahrq.gov/db/nation/nrd/nrddbdocumentation.jsp
Nationwide Emergency Department Sample (NEDS)	Clinical	ICD-9-CM, demographics, expected payment source, total ED charges, total hospital charges, hospital characteristics	Discharge data for ED visits from 953 hospitals located in 34 States and the District of Columbia	https://www.hcup-us.ahrq.gov/db/nation/neds/nedsdbdocumentation.jsp
Women's Health Initiative	Clinical	2 major parts: a Clinical Trial and an Observational Study from heart disease, breast and colorectal cancer, and osteoporosis in postmenopausal women	Clinical trial (68,132 women) and observational study (93,676 women) from women aged 50–79 between 1993 and 1998	https://www.whi.org/researchers/SitePages/Get%20Involved.aspx
Multi-Ethnic Study of Atherosclerosis (MESA)	Clinical	Multi-Ethnic Study from Columbia University, Johns Hopkins University, Northwestern University, UCLA, University of Minnesota, and Wake Forest University	6814 men and women	www.mesa-nhlbi.org
Atherosclerosis Risk in Communities (ARIC)	Clinical	Cardiovascular risk factors, medical care, and disease by race, gender, location, and date	470,000 men and women (aged 35–84 years)	http://www2.cscc.unc.edu/aric/opportunities_for_new_investigators
Sleep Heart Health Study (SHHS)	EEG, EKG, and polysomnograms	Multi-cohort study focused on sleep-disordered breathing and cardiovascular outcome	5804 adults (aged 40 and older)	sleepdata.org/datasets/shhs
Coronary Artery Risk Development in Young Adults (CARDIA)	Clinical	From 4 centers: Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA	5115 black and white men and women (aged 18–30 years)	www.cardia.dopm.uab.edu

(*Continued*)

Table 2. (Continued).

Database	Type of data	Details	Numbers of samples	Link
Jackson Heart Study (JHS)	Clinical	Clinical variables, labs, imaging, interview, and physical activity	5306 African-American residents living in the Jackson, MS, metropolitan area of Hinds, Madison, and Rankin Counties	www.jacksonheartstudy.org
Cardiovascular Health Study (CHS)	Clinical	Extensive initial physical and laboratory evaluations to identify cardiovascular risk factors, such as high blood pressure, high cholesterol, and pre-diabetes; subclinical disease (e.g. carotid artery atherosclerosis, left ventricular enlargement, and transient ischemia)	5888 men and women aged 65 or older in four U.S. communities – Sacramento, CA; Hagerstown, MD; Winston-Salem, NC; and Pittsburgh, PA	chs-nhlbi.org
Twitter	Social media	Curate Tweets and 3 Twitter API platforms standard and premium (free) but enterprise (paid)	Over 900 million existing Twitter accounts	https://developer.twitter.com/en/products/products-overview
IBM Watson (blog, facebook pages, Twitter, news)	Millions of data and social media sources	Several analytical packages (regular, plus, and professional), and several types of data (blog, facebook pages, Twitter, news)	Upto 10,000,000 rows per dataset and upto 500 columns per dataset	www.ibm.com/us-en/marketplace/watson-analytics
PhysioBank	Digital recordings of physiologic signals and related data	Clinical, waveforms, EKGs, RR interval, oxygen saturation variability, gait and balance data	Over 75 databases 100,000 samples	www.physionet.org
MIMIC, MIMIC-II, MIMIC-III	Clinical	Demographics, vital sign measurements, laboratory test results, procedures, medications, nurse and physician notes, imaging reports, and out-of-hospital mortality	30,000–60,000 admissions of patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012	mimic.physionet.org
National Health and Nutrition Examination Survey (NHANES)	Nutrition	Demographic, dietary, questionnaire	39,695 persons for NHANES-III, 27,801 persons for NHANES-II, and 32,000 persons for NHANES-I	wwwn.cdc.gov/nchs/nhanes/Default.aspx
The NHANES National Youth Fitness Survey (NNYFS)	Physical activity and fitness levels	Demographic, dietary, questionnaire, physical activity monitor, aerobic fitness – maximal and submaximal exercise test, and muscle strength	1640 children and adolescents aged 3–15	www.cdc.gov/nchs/nnyfs/index.htm
YouTube-8M	Video	Lifestyle	6.1 Million Video IDs, 2.6 billion audio/visual features, and 3,862 Classes	research.google.com/youtube8m/
UCF101 dataset	Video	Lifestyle	13,320 videos	http://crcv.ucf.edu/data/UCF101.php
UCF-Sports	Video	Lifestyle collected from various sports which are typically featured on broadcast television channels such as the BBC and ESPN	150 sequences with the resolution of 720 × 480	http://crcv.ucf.edu/data/UCF_Sports_Action.php
J-HMDB	Video	Collected from movies or the Internet	5100 clips of 51 different human actions	http://jhmdb.is.tue.mpg.de/
THUMOS 2015 dataset	Video	Lifestyle	430 h of video data and 45 million frames	http://www.thumos.info/home.html
DAVIS 16 and 17	Video	Lifestyle	50 sequences, 3455 annotated frames	davischallenge.org
Sports-1M	Video	Lifestyle	1,133,158 video URLs which have been annotated automatically with 487 labels	github.com/gtoderici/sports-1m-dataset/blob/wiki/ProjectHome.md
TRECVID MED dataset	Several types of video datasets (i.e. IACC.1.A-C, YFCC100M, HAVIC)	Data from a small number of known professional sources – broadcast news organizations, TV program producers, and surveillance systems	Several categories of video dataset (depends on year)	www-nlpir.nist.gov/projects/trecvid/trecvid.data.html
Uber 2B trip data	Text, Lifestyle	Lifestyle	North America, Central & South America, Europe, Africa, South Asia, Australia & New Zealand	movement.uber.com
Yelp Open Dataset	Text, Lifestyle	JSON and SQL datasets	5,200,000 reviews, 174,000businesses, 200,000 pictures, 11 metropolitan areas	www.yelp.com/dataset
Quora Question Pairs	Text, Lifestyle	Questions in Quora competition is to predict which of the provided pairs of questions contain two questions with the same meaning	N/A	www.kaggle.com/c/quora-question-pairs/data
Google Audioset	Audio	A hierarchical graph of event categories, covering a wide range of human and animal sounds, musical instruments and genres, and common everyday environmental sounds	632 audio event classes and a collection of 2,084,320 human-labeled 10-s sound clips drawn from YouTube videos	research.google.com/audioset/dataset/index.html

(*Continued*)

Table 2. (Continued).

Database	Type of data	Details	Numbers of samples	Link
NYC Taxi dataset	Taxi in New York City	Data containing information on our various indicators, trip counts, crash history, etc., and also raw trip data from a variety of sources	Millions of trip records from both yellow medallion taxis and green street hail livery	http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
OpenFDA	Date, drugs, events	Drugs, devices, and foods and subcategories (i.e. adverse events, enforcement reports, classification, registration, labelling)	8,733,422 drug adverse event reports, 65,523 food adverse event reports, and 7,353,142 device adverse event reports	open.fda.gov/tools/downloads/
SEER Research Data	Epidemiologic	Cancer incidence data from population-based cancer registries	10,050,814 cases (9,099,524 malignant cases and 9,776,139 cases)	https://seer.cancer.gov/seertrack/data/request/
UNSD Environmental Indicators	Environment	NOx emissions, SO_2 emissions, CO_2 emissions, CH_4 and N_2O emissions, Climatological disasters, Hydrological disasters, and Inland Water Resources	Environmental data (air pollution, climate changes, greenhouse gases) from 183 countries	unstats.un.org/unsd/envstats/qindicators.cshtml
DrugBank	Drug	More than 200 data fields with half of the information being devoted to drug data and the other half devoted to drug target or protein data	11,203 drug entries including 2,562 approved small molecule drugs, 966 approved biotech (protein/peptide) drugs, 121 nutraceuticals and over 5183 experimental drugs	www.drugbank.ca
The Toxin and Toxin Target Database (T3DB)	Toxin	Mechanisms of toxicity and target proteins for each toxin detailed toxin data with comprehensive toxin target information pollutants, pesticides, drugs, and food toxins	3670 common toxins and environmental pollutants	http://www.t3db.ca/
FooDB	Food, nutrients	Food, compounds, nutrients, contents detailed compositional, biochemical and physiological information structure, chemical class, its physico-chemical data, its food source(s), its color, its aroma, its taste, its physiological effect, presumptive health effects (from published studies), and concentrations in various foods	28,000 food components and food additives	http://foodb.ca/
PhysioNet	Electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiology (EKG), and oxygen saturation (SaO2)	Large collections of recorded physiologic signals (PhysioBank) and related open-source software (PhysioToolkit)	1985 subjects from MGH sleep laboratory for the diagnosis of sleep disorders (from PhysioNet Cardiology Challenge 2018)	www.physionet.org
UCI Machine Learning Repository	Machine learning dataset	Machine learning	436 data sets	archive.ics.uci.edu/ml/index.php
Open Images Dataset V4	Machine learning dataset	Machine learning (a validation set (41,620 images), and a test set (125,436 images)	15,440,132 boxes and 30,113,078 image-level labels	github.com/openimages/dataset
The National Survey on Drug Use and Health (NSDUH)	Survey data	Tobacco, alcohol, and drug use, mental health and other health-related issues in the United States	70,000 people	nsduhweb.rti.org
Google Flu Trends and Google Dengue Trends	Text	Flu Trends since 2008	50 million of the most common search queries in the United States	https://www.google.org/flutrends/about/
Cardiac MRI dataset	Images	Cardiac MRI	33 subjects and 7980 images (20 frames and 8–15 slices along the long axis)	http://www.cse.yorku.ca/~mridataset/
The CardioVascular Research Grid (CVRG)	Clinical, gene, and protein expression	Multiscale data sets (SNP, mRNA expression, protein expression, imaging, ECG, clinical data) from Canine Heart Atlas, Mouse Hearts, In-Vivo Human Heart CT Image Data	Multiple variables (clinical, gene, and protein expression) from 15 canine hearts	http://cvrgrid.org/
Influenza Research Database (IRD)	Epidemiology	Strain, segment and protein sequence data surveillance sample information	5621 structural and functional sequence features in influenza proteins	www.fludb.org
Risk-Adjusted Inpatient Mortality Rates and Hospital Ratings for California Hospitals, 2012	Clinical	Risk-adjusted mortality rates, quality ratings, and number of deaths and cases for 6 medical conditions treated (acute stroke, acute myocardial infarction, heart failure, gastrointestinal hemorrhage, hip fracture and pneumonia) in California hospitals for 2012	Depends on conditions and procedures (from 300 to 64,000)	data.chhs.ca.gov/dataset/california-hospital-inpatient-mortality-rates-and-quality-ratings
Community Health Status Indicators (CHSI)	Clinical	Community health (e.g. obesity, heart disease, cancer)	Over 200 measures for each of the 3,141 United States counties	www.cdc.gov/ophss/csels/dphid/CHSI.html

Table 3. Examples of public data search.

Omics data search	Type of data	Details	Numbers of samples	Link
GWAS Catalog	Articles summarizing GWAS and SNP-trait associations	Human genome wide association studies (GWAS) and association results	3395 publications and 62,174 unique SNP-trait associations	https://www.ebi.ac.uk/gwas/
GWAS Central	Articles	Comprises all known SNPs and other variants, allele and genotype frequency data, plus genetic association significance findings from public databases such as dbSNP and the DBGV	1605 studies (2,935,163 unique dbSNP markers)	www.gwascentral.org
KEGG pathway	Metabolism, molecular interactions, reactions and relations, environmental information processing, and cellular processes	Gene/protein (KEGG GENES) Reaction (KEGG REACTION) Drug (KEGG DRUG)	2706 entries for pathway diagrams 110,018 entries in 24 complete genomes and 12 partial genomes 5,645 entries in the COMPOUND section	https://www.genome.jp/kegg/pathway.html
ExAC Browser (Beta) | Exome Aggregation Consortium	Exome sequencing data	Harmonize exome sequencing data from a variety of large-scale sequencing projects	60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies	http://exac.broadinstitute.org/
Global Biobank Engine	Genotypes and phenotypes	Biobank explorer is currently seeded with data from UKBB allowing exploration between genotypes and phenotypes	392,292 participants from the UKBB	http://gbe.stanford.edu
gnomAD browser beta | genome Aggregation Database	Exome and genome sequencing data	Exome and genome sequencing data from a variety of large-scale sequencing projects	123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals sequenced	http://gnomad.broadinstitute.org
Gene Expression Omnibus (GEO)	Gene and functional genomics data	Freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data	4,348 DataSet	www.ncbi.nlm.nih.gov/geo/
Sequence Read Archive (SRA)	Nucleotide Sequence	High-throughput sequencing data and is part of the International Nucleotide Sequence Database Collaboration (INSDC) that includes at the NCBI Sequence Read Archive (SRA), the European Bioinformatics Institute (EBI), and the DNA Database of Japan (DDBJ)	>500 billion reads consisting of 60 trillion base pairs	www.ncbi.nlm.nih.gov/sra
The database of Genotypes and Phenotypes (dbGaP)	Genotype and phenotype	Phenotype data, association (GWAS) data, summary level analysis data, SRA (Short Read Archive) data, reference alignment (BAM) data, VCF (Variant Call Format) data, expression data, imputed genotype data, image data	Over 100,000 individuals	www.ncbi.nlm.nih.gov/gap
The Phenotype-Genotype Integrator (PheGenI)	Genome-wide association study (GWAS) catalog data with several databases	Merges NHGRI genome-wide association study (GWAS) catalog data with several databases housed at the National Center for Biotechnology Information (NCBI), including Gene, dbGaP, OMIM, eQTL, and dbSNP	66,063 association records (54,282 from dbGaP and 11,781 from the NHGRI GWAS catalog)	www.ncbi.nlm.nih.gov/gap/phegeni
Healthmap.org	News, twitter	Infectious Disease Outbreaks	An automated process, updating realtime, the system monitors, organizes, integrates, filters, visualizes and disseminates online information about emerging diseases	http://www.healthmap.org
CDC WONDER	Epidemiologic	Mortality (deaths), cancer incidence, HIV and AIDS, tuberculosis, vaccinations, natality (births), census data	20 collections of public-use data for U.S. births, deaths, cancer diagnoses, tuberculosis cases, vaccinations, environmental exposures, and population estimates	wonder.cdc.gov
SEER*Explorer	Cancer statistics	Gender, race, calendar year, age, and for a selected number of cancer sites, by stage and histology	308,745,538 patients	seer.cancer.gov/explorer/
USDA National Nutrient Database	Nutrition	Different types of foods and nutrients % fat and % lean and types of serving methods	7793 different foods and nutrients	ndb.nal.usda.gov/ndb/
Nutrition, Physical Activity, and Obesity: Data, Trends and Maps	Graph, tables	Obesity, breastfeeding, physical activity, other health behaviors and related environmental and policy data	Either nationally or by state in the US	https://www.cdc.gov/nccdphp/dnpao/data-trends-maps/index.html
TOXMAP	Graph, tables	NCI SEER cancer and disease mortality data, Canadian National Pollutant Release Inventory (NPRI) data U.S. commercial nuclear power plants, and Coal power plant data from the EPA Clean Air Markets Program	Either nationally or by state in the US	https://toxmap.nlm.nih.gov/toxmap/news/2018/06/new-version-of-toxmap-now-available.html

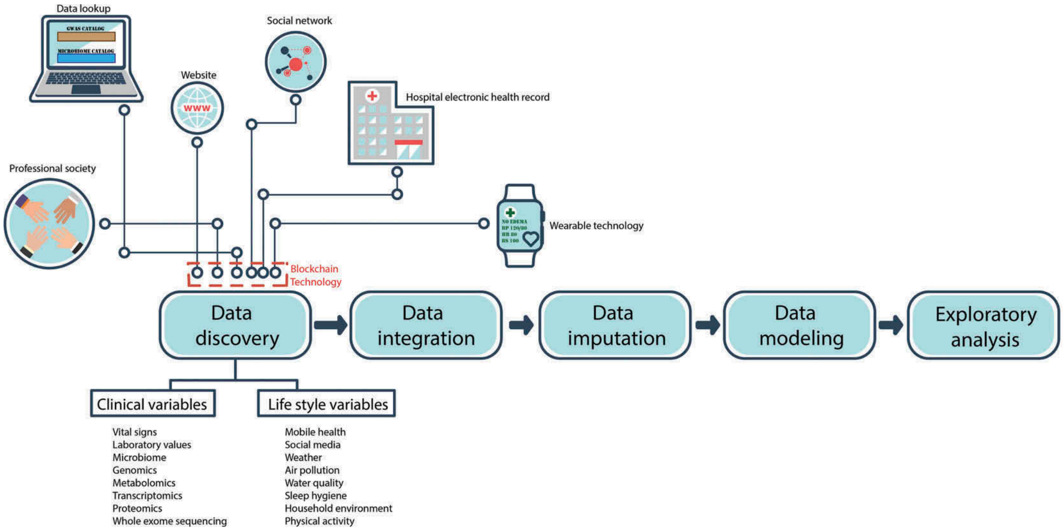

Figure 1. Big data process flow for cardiovascular medicine.

existing models (algorithms) is commonly used, as it is much easier and sufficient algorithms already exist which may be applied to important problems. Finally, an exploratory analysis is based on data-driven hypotheses rather than investigator-driven hypothesis [45]. For example, there have been papers showing clustering of phenotypes (phenomapping) [6], there are papers using systems biology methods to look at distinct endophenotypes [46], and there are also papers dissecting out response predictors with patterns [47].

4. Current challenges

It is important to delineate some of the challenges of implementing a big data approach in cardiovascular medicine. First, integrating big data into clinical trials is challenging because clinical trials are usually designed under ideal conditions, among select patients, and monitored by highly qualified physicians [48]. In order to perform analysis using big data with traditional statistical methods could be difficult. Smart clinical trials that are guided by AI to recruit patients (e.g. Deep 6 AI), do dynamic matching (e.g. SYNERGY-AI; NCT03452774), or to do direct targeted therapy are also promising [49]. Second, heterogeneity and disparities of different datasets can be challenging to utilize. Third, latent variables might have been ignored in those heterogeneous diseases in previous studies. Briefly, latent or unknown variables can be categorized into hidden medical variables and lifestyle variables. Hidden medical variables could act as new parameters to characterize accurate myocardial function, novel serum metabolites, or new parameters for subclinical arteriosclerosis [9,10]. HFpEF, for example, could potentially be subcategorized into more mechanistically and molecularly homogenous, discrete genotypes, phenotypes, and etiologies [6,11]. Lifestyle variables are often quite novel because most studies have not included high-definition lifestyle variables in their analyses [50]. However, integrating deeply phenotyped lifestyle factors into medical records can be difficult because of data privacy and the lack of publically available application programming interfaces for consumer devices to interact with EHRs [51]. Lifestyle variables may include dietary intake [52], physical activity [30], sleep hygiene [53], air pollution [54], ergonomics [55], income [56], domestic violence [57], working hours [58], and workplace wellness [59]. To date, most recent research has been collected on lifestyle variables mainly by questionnaires or interviews, leading to recall or social desirability biases [60]. Advancement of wearable technology could be used to track real-time activity and integrate those hidden variables into a person's medical history. For example, the etiologies of HF readmission are heterogeneous and perhaps related to medication compliance and dietary habits [61]. Integrating lifestyle variables could potentially track the main problems with real-world variables rather than tracking them inside of a hospital and preventing recall biases from patient histories [60,62]. However, there remains a need to collect better and more consistent data from wearable devices – most consumer devices are not approved by the FDA for clinical monitoring of patients, and this may be a limitation in some cases. In addition, wearable devices have a number of validation issues, and it is unclear if they motivate long-term behavioral change [63,64]. For example, in a BEAT-HF trial, a combination of remote patient monitoring with care transition management did not reduce 180-day all-cause readmission after hospitalization for HF [65]. Fourth, data quality, data inconsistency, data instability, and validation of big data are also barriers, and therefore the imputation of big data is critical [66]. More data, more entropy, and more heterogeneity result in lower-quality databases [67]. Therefore, the pre-analytic process of big data needs to be assessed and imputed systematically. For example, though the methodology of reducing heterogeneity in meta-analysis is not yet perfect, it can reduce significant biases [68]. Fifth, some other limitations of a big data approach are heterogeneity of multiple databases (i.e., different ICD code versions, different diagnostic criteria, different laboratories, and different software vendors) [13,14]. Hence, synchronizing existing data to generate meaningful analysis can be very challenging. Sixth, although de-identification seems to be a solution in big data research, studies have shown that re-identification can be done in various ways. For example, anonymous genetic data stores could be unmasked by matching their data to a sample of their DNA [69] or matching social networks for information that might yield insights into the genetic basis for complex human traits [70]. Seventh, to date there has been little evidence to suggest that DNA testing has little or no impact in motivating behavior change [71]. Therefore, the genomic information, or GWAS, impacting long-term behavior change may still need hand curation [72]. In addition, distinguishing signals from noise in Omics data and software validation are required [73]. For example, using different types of software (i.e. PLINK, QCTOOL, Vcftools, BOLTs, or EPACTS) may reflect different results. Lastly, another important challenge in the use of big data in cardiovascular medicine is the ascertainment of causality from observational and retrospective studies. Most AI and ML methods do not explicitly utilize a framework to model causality. Consider the humorous case of age-related gray hair and CVD. The presence of both gray hair, wrinkles, baldness, and CVD are highly correlated [74–76]. However, if we were to pursue this strong association in an attempt to design therapies (e.g. hair dyes or wrinkle cream), we would be wholly unsuccessful in preventing CVDs. This is an important limitation that all big-data analyses must account for – however, there do exist emerging methods to perform causal inference from observational datasets, such as the parametric G formula [77]. We recently completed one application of the parametric G formula, in which we used retrospective EHR data to demonstrate the relative correctness of a clinical trial for hypertension that had been called into question [78]. However, EHR data also has some limitations, such as the accuracy of ICD 9 codes [79–81].

5. Implementation of big data in clinical practice

Several resources are still the main starting points for any big data search in cardiovascular medicine. The utilization of these datasets could facilitate precision CV medicine. The integration of the Internet of Things, social media, Omics and big data technologies, and AI could create a new concept of smart health, integrating real-world variables into hospital-related variables, and leading to improved quality of patient care and hospital workflow [82–85]. Today, with the help of the Internet, there are many types of websites providing either datasets for public use or data search (Tables 1–3). The implementation of big data analytics that links these databases together is crucial. However, there may be some barriers or restrictions. Academic institutions usually have many resources and can provide their own biobank (i.e. the Mayo Clinic Biobank, Cleveland Clinic's Biorepository, SCVI Biobank, Mount Sinai's BioMe, Vanderbilt's BioVU, or Northwestern's NUgene). Most biobanks are designed so they can be accessed by various innovative actors, public and private, throughout the world. Integration of these biobanks in ongoing research is worth exploring. Training in bioinformatics or coordinating with data scientists is also important [86]. In addition, using online community support for data analysis such as Github, Stack Overflow, Kaggle, and Biostars is increasingly recognized and utilized in the medical community. Previous research has acknowledged many confounders in clinical research; however, none of them have mentioned real-world lifestyle factors such as seafood/cereal/coffee consumption, watching movies, playing video games, or personal hygiene. These real-world factors could potentially be confounders in CVD burdens, for example, HF readmission, recurrent AF, labile INR, statin sensitivity, or stent thrombosis. These integrations can increase dimensional research into new translation research by including real-world environmental factors.

6. Expert commentary

Though many of the technical issues for a big data approach remain to be solved, the potential for big data analysis to improve cardiovascular quality of care and patient outcome is tremendous. To date, the key findings from previous studies in this field are inconclusive. For example, strong evidence that the attempt to change behavior using either wearables or genomic information is lacking. The ultimate goal of big data analysis is to unify heterogeneous databases into homogenous databases using advanced computational power, such as AI. In addition, we believe that big analysis using AI will advance clinical trials in the context of recruiting patients, distributing drugs randomly and fairly between two arms, assisting drug delivery, and predicting outcomes of trials in advance. However, the biggest challenge is to combine heterogeneous variables from various datasets and implement these into clinical practice. In addition, there are candidate genes, novel biomarkers, and parameters emerging every day, which makes it almost impossible for current guidelines to remain current. Moreover, decision-making using these novel profiles without guidelines can be challenging and may face ethical dilemmas. Future studies should integrate big data analysis to better explore the robustness of novel CVD phenotypes and smart clinical trial design for targeted therapy. Targeting components of the CVD phenotypes such as specific genes, specific metabolites, and the specific gut microbiome in CVD may prove to be valuable. This phenotype-based classification system could be helpful for the identification of new biomarkers and potential targeted therapies, and it may lead to the development of tailored/customized future clinical trials.

7. Five-year view

In the realm of the big data era, genetic polymorphisms, plasma metabolomics, and proteomics may help to identify new biomarkers and potential novel therapeutic targets for CVH. We hope and believe that these tools will soon emerge as best practices in day-to-day clinical medicine. The next step is to create on-demand predictive analytics in clinical practice using the results of a big data approach, which shows great promise in cardiovascular medicine. In clinical practice, the implementation of sophisticated analytics tools with 'omic' data, the human microbiome, physical activity, environmental factors, and lifestyle factors might help identify novel phenotypes of CVD patients. Today, genetic risk scores are starting to stratify patients based on risk before the disease presents [87,88]. A big data approach could potentially transform medicine into a more personalized approach using sophisticated algorithms generated from a combination of real-world factors and medical variables to calculate the risk and benefits of CVH-related behaviors in individuals. For example, taking into account a persons patterns of dietary intake, medication compliance, and daily life activities using wearable technology, storing this data in a secure system (i.e. cloud or blockchain), and transferring it to an EHR could generate a predictive analysis with prompt recommendations in regards to maximum fruit intake and minimal carbohydrate intake for individuals in their discharge summary. The results of this type of analysis would be transferred to primary care physicians, collected in wearable technology with warning messages, and could appear in a patient's history in the EHR system. This proposed model could potentially be a modifiable factor to weigh CVD risk and benefit based on individuals.

Key issues

- A phenotype-based classification using multi-omics, lifestyle, and environmental data with new analytical methods and high computational power could potentially transform future clinical trials.
- Data cleaning and data imputation are keys to unlocking big data analysis.
- The data, so far, on both wearables and genomic information evoking long-term behavior change is negative or, at best, neutral.
- Biobanks and curated public databases may play an important role in big data analysis.
- Although there are many limitations to the proposed approach that have already been clearly tested, there is tremendous potential for big data analysis to improve cardiovascular quality of care and patient outcome.

Funding

This paper was not funded.

Declaration of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

References

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Gaye B, Tafflet M, Arveiler D, et al. Ideal cardiovascular health and incident cardiovascular disease: heterogeneity across event subtypes and mediating effect of blood biomarkers: the PRIME study. J Am Heart Assoc. 2017 Oct 17;6(10).
2. Jose PO, Frank AT, Kapphahn KI, et al. Cardiovascular disease mortality in Asian Americans. J Am Coll Cardiol. 2014;64:2486–2494.
3. Gordon RD. Heterogeneous hypertension. Nat Genet. 1995;11:6–9.
4. Darbar D, Herron KJ, Ballew JD, et al. Familial atrial fibrillation is a genetically heterogeneous disorder. J Am Coll Cardiol. 2003;41:2185–2192.
5. Inohara T, Shrader P, Pieper K, et al. Association of atrial fibrillation clinical phenotypes with treatment patterns and outcomes: a multicenter registry study. JAMA cardiology. 2018;3:54–63.
6. Shah SJ, Katz DH, Selvaraj S, et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation. 2015;131:269–279.
7. Krittanawong C, Bomback AS, Baber U, et al. Future direction for using artificial intelligence to predict and manage hypertension. Curr Hypertens Rep. 2018;20:75.
8. Balaney B, Medvedofsky D, Mediratta A, et al. Invasive validation of the echocardiographic assessment of left ventricular filling pressures using the 2016 diastolic guidelines: head-to-head comparison with the 2009 guidelines. J Am Soc Echocardiography: Official Publication Am Soc Echocardiography. 2018;31:79–88.
9. Pislaru C, Alashry MM, Thaden JJ, et al. Intrinsic wave propagation of myocardial stretch, a new tool to evaluate myocardial stiffness: a pilot study in patients with aortic stenosis and mitral regurgitation. J Am Soc Echocardiography: Official Publication Am Soc Echocardiography. 2017;30:1070–1080.
10. Laaksonen R, Ekroos K, Sysi-Aho M, et al. Plasma ceramides predict cardiovascular death in patients with stable coronary artery disease and acute coronary syndromes beyond LDL-cholesterol. Eur Heart J. 2016;37:1967–1976.
11. Krittanawong C, Kukin ML. Current management and future directions of heart failure with preserved ejection fraction: a contemporary review. Curr Treat Options Cardiovasc Med. 2018;20:28.
12. Guo Q, Lu X, Gao Y, et al. Cluster analysis: a new approach for identification of underlying risk factors for coronary artery disease in essential hypertensive patients. Sci Rep. 2017;7:43965.
13. Bellazzi R. Big data and biomedical informatics: a challenging opportunity. Yearb Med Inform. 2014;9:8–13.
14. Scruggs SB, Watson K, Su AI, et al. Harnessing the heart of big data. Circ Res. 2015;116:1115–1119.
15. Kass-Hout TA, Stevens LM, Hall JL. American Heart Association precision medicine platform. Circulation. 2018;137:647–649.
16. Gourraud P-A, Henry R, Cree BAC, et al. Precision medicine in chronic disease management: the MS bioscreen. Ann Neurol. 2014;76:633–642.
17. Krittanawong C, Zhang H, Wang Z, et al. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol. 2017;69:2657–2664.

    •• This is a useful review about artificial intelligence in cardiovascular medicine.
18. Glicksberg BS, Johnson KW, Dudley JT. The next generation of precision medicine: observational studies, electronic health records, biobanks and continuous monitoring. Hum Mol Genet. 2018;27:R56–r62.
19. McConnell MV, Shcherbina A, Pavlovic A, et al. Feasibility of obtaining measures of lifestyle from a smartphone app: the MyHeart counts cardiovascular health study. JAMA cardiology. 2017;2:67–76.

    • This study provides an example of a potential smartphone application study in cardiovascular health.
20. Guo X, Vittinghoff E, Olgin JE, et al. Volunteer participation in the health eHeart study: a comparison with the US population. Sci Rep. 2017;7:1956.
21. Muse ED, Wineinger NE, Schrader B et al. Moving beyond clinical risk scores with a mobile app for the genomic risk of coronary artery disease. bioRxiv. 2017.
22. [cited 2018 Oct 6]. Access online at https://med.stanford.edu/appleheartstudy.html.
23. Bot BM, Suver C, Neto EC, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data. 2016;3:160011.

24. Chan Y-FY, Bot BM, Zweig M, et al. The asthma mobile health study, smartphone data collected using ResearchKit. Sci Data. 2018;5:180096.
25. Webster DE, Suver C, Doerr M, et al. The Mole Mapper study, mobile phone skin imaging and melanoma risk data collected using ResearchKit. Sci Data. 2017;4:170005.
26. Ata R, Gandhi N, Rasmussen H, et al. IP225 VascTrac: a study of peripheral artery disease via smartphones to improve remote disease monitoring and postoperative surveillance. J Vasc Surg. 2017;65:115S–116s.
27. Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial intelligence in cardiology. J Am Coll Cardiol. 2018;71:2668–2679.
28. Krittanawong C, Tunhasiriwet A, Zhang H, et al. Deep learning with unsupervised feature in echocardiographic imaging. J Am Coll Cardiol. 2017;69:2100–2101.
29. Shameer K, Johnson KW, Glicksberg BS, et al. Machine learning in cardiovascular medicine: are we there yet? Heart. 2018;104:1156–1164.
30. Krittanawong C, Aydar M, Kitai T. Pokémon Go: digital health interventions to reduce cardiovascular risk. Cardiol Young. 2017;27:1625–1626.
31. Ding MQ, Chen L, Cooper GF, et al. Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. Mol Cancer Res 2017.
32. Anwar S, Negishi K, Borowszki A, et al. Comparison of two-dimensional strain analysis using vendor-independent and vendor-specific software in adult and pediatric patients. JRSM Cardiovasc Disease. 2017;6:2048004017712862.
33. O'Malley KJ, Cook KF, Price MD, et al. Measuring diagnoses: ICD code accuracy. Health Serv Res. 2005;40:1620–1639.
34. Standards for privacy of individually identifiable health information. Office of the Assistant Secretary for Planning and Evaluation, DHHS. Proposed rule. Federal register 1999;64:59918–60065.
35. Verma SS, de Andrade M, Tromp G, et al. Imputation and quality control steps for combining multiple genome-wide datasets. Front Genet. 2014;5:370.
36. Hendler J. Data integration for heterogenous datasets. Big Data. 2014;2:205–215.
37. Blankenberg D, Coraor N, Von Kuster G, et al. Integrating diverse databases into an unified analysis framework: a galaxy approach. J Bioll Databases Curation 2011. 2011: bar011.
38. Shkapsky A, Yang M, Interlandi M, et al. Big data analytics with datalog queries on spark. Proceedings ACM-Sigmod International Conference on Management of Data. San Francisco, CA, USA. 2016;2016:1135–1149.
39. [cited 2018 Oct 6].Amazon AWS http://aws.amazon.com/.
40. Forbes A The future of BIME. 2018
41. Pan C, McInnes G, Deflaux N, et al. Cloud-based interactive analytics for terabytes of genomic variants data. Bioinformatics. 2017;33:3709–3715.
42. Coleman JR, Euesden J, Patel H, et al. Quality control, imputation and analysis of genome-wide genotyping data from the illumina HumanCoreExome microarray. Brief Funct Genomics. 2016;15:298–304.
43. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284.
44. Luo G, Stone BL. Automating construction of machine learning models with clinical big data: proposal rationale and methods. JMIR Res Protoc. 2017 Aug 29;6(8):e175.
45. Naik AW, Kangas JD, Sullivan DP, et al. Active machine learning-driven experimentation to determine compound effects on protein patterns. Elife. 2016;5:e10047.
46. Eppinga RN, Hagemeijer Y, Burgess S. Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. Nat Genet. 2016 Dec;48(12):1557-1563. doi: 10.1038/ng.3708.
47. Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. Comput Meth Prog Bio. 2016;130:54–64.
48. Mayo CS, Matuszak MM, Schipper MJ, Jolly S, Hayman JA, Ten Haken RK. Big data in designing clinical trials: opportunities and challenges. Front Oncol. 2017;7:187.
49. Say LEAFC Goodbye to clinical trials that don't teach. 2018.
50. Assi N, Thomas DC, Leitzmann M, et al. Are metabolic signatures mediating the relationship between lifestyle factors and hepatocellular carcinoma risk? Results from a nested case-control study in EPIC. Cancer epidemiol Biomarkers Prevention. 2018;27:531–540.
51. Filkins BL, Kim JY, Roberts B, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res. 2016;8:1560–1580.
52. Krittanawong C, Tunhasiriwet A, Zhang H, et al. Is white rice consumption a risk for metabolic and cardiovascular outcomes? A systematic review and meta-analysis. Heart Asia. 2017;9:e010909.
53. Krittanawong C, Tunhasiriwet A, Wang Z, et al. Association between short and long sleep durations and cardiovascular outcomes: a systematic review and meta-analysis. Eur Heart J Acute Cardiovasc Care. 2017;2048872617741733.
54. Hartiala J, Breton CV, Tang WH, et al. Ambient air pollution is associated with the severity of coronary atherosclerosis and incident myocardial infarction in patients undergoing elective cardiac evaluation. J Am Heart Assoc. 2016 Jul 28;5(8).
55. Djindjic N, Jovanovic J, Djindjic B, et al. Associations between the occupational stress index and hypertension, type 2 diabetes mellitus, and lipid disorders in middle-aged men and women. Ann Occup Hyg. 2012;56:1051–1062.
56. Orth-Gomer K, Deter HC, Grun AS, et al. Socioeconomic factors in coronary artery disease - results from the SPIRR-CAD study. J Psychosom Res. 2018;105:125–131.
57. Mason SM, Wright RJ, Hibert EN, et al. Intimate partner violence and incidence of hypertension in women. Ann Epidemiol. 2012;22:562–567.
58. Kivimaki M, Jokela M, Nyberg ST et al. Long working hours and risk of coronary heart disease and stroke: a systematic review and meta-analysis of published and unpublished data for 603,838 individuals. Lancet (London, England) 2015;386:1739–1746.
59. Ryu H, Jung J, Cho J, et al. Program development and effectiveness of workplace health promotion program for preventing metabolic syndrome among office workers. Int J Environ Res Public Health. 2017 Aug 4;14(8).
60. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. 2016;9:211–217.
61. Retrum JH, Boggs J, Hersh A, et al. Patient-identified factors related to heart failure readmissions. Circ Cardiovasc Quality Outcomes. 2013;6:171–177.
62. Larsson SC, Tektonidis TG, Gigante B, et al. Healthy lifestyle and risk of heart failure: results from 2 prospective cohort studies. Circ Heart Fail. 2016;9:e002855.
63. Murakami H, Kawakami R, Nakae S, et al. Accuracy of wearable devices for estimating total energy expenditure: comparison with metabolic chamber and doubly labeled water method. JAMA Intern Med. 2016;176:702–703.
64. Jakicic JM, Davis KK, Rogers RJ, et al. Effect of wearable technology combined with a lifestyle intervention on long-term weight loss: the idea randomized clinical trial. Jama. 2016;316:1161–1171.
65. Ong MK, Romano PS, Edgington S, et al. Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure: the better effectiveness after transition–heart failure (beat-hf) randomized clinical trial. JAMA Intern Med. 2016;176:310–318.

    - **This study provides evidence of the association between wearable devices and long-term behavioral change.**

66. Dinov ID. Methodological challenges and analytic opportunities for modeling and interpreting big healthcare data. Gigascience. 2016;5:12.
67. Coakley MF, Leerkes MR, Barnett J, et al. Unlocking the power of big data at the National Institutes of Health. Big Data. 2013;1:183–186.
68. Egger M, Smith GD, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315:629.
69. Gymrek M, McGuire AL, Golan D, et al. Identifying personal genomes by surname inference. Science. 2013;339:321–324.
70. Hayden EC. The genome hacker. Nature. 2013;497:172.

71. Hollands GJ, French DP, Griffin SJ et al. The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-analysis. BMJ. 2016 Mar 15;352:i1102.
72. Presley CJ, Tang D, Soulos PR, et al. Association of broad-based genomic sequencing with survival among patients with advanced non–small cell lung cancer in the community oncology setting. Jama. 2018;320:469–477.
73. Saracci R. Epidemiology in wonderland: big data and precision medicine. Eur J Epidemiol. 2018;33:245–257.
74. Schnohr P, Lange P, Nyboe J, et al. Gray hair, baldness, and wrinkles in relation to myocardial infarction: the Copenhagen City Heart Study. Am Heart J. 1995;130:1003–1010.
75. Lesko SM, Rosenberg L, Shapiro S. A case-control study of baldness in relation to myocardial infarction in men. Jama. 1993;269:998–1003.
76. Ford ES, Freedman DS, Byers T. Baldness and ischemic heart disease in a national sample of men. Am J Epidemiol. 1996;143:651–657.
77. Lin SH, Young J, Logan R, et al. Parametric mediational g-formula approach to mediation analysis with time-varying exposures, mediators, and confounders. Epidemiology. 2017;28:266–274.
78. Johnson KW, Glicksberg BS, Hodos RA, et al. Causal inference on electronic health records to assess blood pressure treatment targets: an application of the parametric g formula. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. Fairmont Orchid, Hawaii, Puako, HI. 2018;23:180–191.
79. Ahmad FS, Chan C, Rosenman MB, et al. Validity of cardiovascular data from electronic sources: the multi-ethnic study of atherosclerosis and HealthLNK. Circulation. 2017;136:1207–1216.
80. Krittanawong C, Kumar A, Virk HUH, et al. Trends in incidence, characteristics, and in-hospital outcomes of patients presenting with spontaneous coronary artery dissection (from a national population-based cohort study between 2004 and 2015). Am J Cardiol. In press.
81. [cited 2018 Oct 6]. https://www.federalregister.gov/d/2018-15390 Aoa.
82. Talboom JS, Huentelman MJ. Big data collision: the internet of things, wearable devices and genomics in the study of neurological traits and disease. Hum Mol Genet. 2018;27:R35–r39.
83. Kang M, Park E, Cho BH, et al. Recent patient health monitoring platforms incorporating internet of things-enabled smart devices. Int Neurourol J. 2018;22:S76–82.
84. Ozdemir V, Hekim N. Birth of industry 5.0: making sense of big data with artificial intelligence, "The internet of things" and next-generation technology policy. Omics: J Integr Biol. 2018;22:65–76.
85. Dey N, Ashour AS. Medical cyber-physical systems: a survey. 2018;42. p. 74.
86. Krittanawong C. Future physicians in the era of precision cardiovascular medicine. Circulation. 2017;136:1572–1574.
87. Muse ED, Wineinger NE, Spencer EG, et al. Validation of a genetic risk score for atrial fibrillation: a prospective multicenter cohort study. PLoS Med. 2018;15:e1002525.
88. Knowles JW, Ashley EA. Cardiovascular disease: the rise of the genetic risk score. PLoS Med. 2018;15:e1002546.