ESC European Society of Cardiology

European Heart Journal (2018) **39**, 1481–1495
doi:10.1093/eurheartj/ehx487

CLINICAL REVIEW
Prevention and epidemiology

Big data from electronic health records for early and late translational cardiovascular research: challenges and potential

Harry Hemingway[1,2]*, Folkert W. Asselbergs[1,2,3], John Danesh[4], Richard Dobson[1,2,5], Nikolaos Maniadakis[6], Aldo Maggioni[6], Ghislaine J.M. van Thiel[3], Maureen Cronin[7], Gunnar Brobert[8], Panos Vardas[6], Stefan D. Anker[9,10], Diederick E. Grobbee[11], and Spiros Denaxas[1,2]; On behalf of the Innovative Medicines Initiative 2nd programme, Big Data for Better Outcomes, BigData@Heart Consortium of 20 academic and industry partners including ESC†

[1]Research Department of Clinical Epidemiology, The Farr Institute of Health Informatics Research, University College London, 222 Euston Road, London NW1 2DA, UK; [2]The National Institute for Health Research, Biomedical Research Centre, University College London Hospitals NHS Foundation Trust, University College London, 222 Euston Road, London NW1 2DA, UK; [3]Department of Cardiology, University Medical Center Utrecht, Heidelberglaan 100, Utrecht 3584 CX, The Netherlands; [4]MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Worts Causeway, Cambridge CB1 8RN, UK; [5]NIHR Biomedical Research Centre for Mental Health (IOP), King's College London, De Crespigny Park, London SE5 8AF, UK; [6]European Society of Cardiology (ESC), 2035 Route des Colles, Les Templiers - CS 80179 Biot, 06903 Sophia Antipolis, France; [7]Vifor Pharma Ltd, lughofstrasse 61, 8152 Glattbrugg, Zurich, Switzerland; [8]Department of Epidemiology, Bayer Pharma AG, Müllerstrasse 178, 13353 Berlin, Germany; [9]Division of Cardiology and Metabolism—Heart Failure, Cachexia & Sarcopenia; Department of Cardiology (CVK), Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité University Medicine, Charitépl. 1, 10117 Berlin, Germany; [10]Department of Cardiology and Pneumology, University Medicine Göttingen (UMG), Robert-Koch-Strasse 40, 37099, Göttingen, Germany; and [11]Julius Centre for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands

Received 1 June 2017; revised 19 July 2017; editorial decision 2 August 2017; accepted 8 August 2017; online publish-ahead-of-print 29 August 2017

Aims Cohorts of millions of people's health records, whole genome sequencing, imaging, sensor, societal and publicly available data present a rapidly expanding digital trace of health. We aimed to critically review, for the first time, the challenges and potential of big data across early and late stages of translational cardiovascular disease research.

Methods and results We sought exemplars based on literature reviews and expertise across the BigData@Heart Consortium. We identified formidable challenges including: data quality, knowing what data exist, the legal and ethical framework for their use, data sharing, building and maintaining public trust, developing standards for defining disease, developing tools for scalable, replicable science and equipping the clinical and scientific work force with new inter-disciplinary skills. Opportunities claimed for big health record data include: richer profiles of health and disease from birth to death and from the molecular to the societal scale; accelerated understanding of disease causation and progression, discovery of new mechanisms and treatment-relevant disease sub-phenotypes, understanding health and diseases in whole populations and whole health systems and returning actionable feedback loops to improve (and potentially disrupt) existing models of research and care, with greater efficiency. In early translational research we identified exemplars including: discovery of fundamental biological processes e.g. linking exome sequences to lifelong electronic health records (EHR) (e.g. human knockout experiments); drug development: genomic approaches to drug target validation; precision medicine: e.g. DNA integrated into hospital EHR for pre-emptive pharmacogenomics. In late translational research we identified exemplars including: learning health systems with outcome trials integrated into clinical care; citizen driven health with 24/7 multi-parameter patient monitoring to improve outcomes and population-based linkages of multiple EHR sources for higher resolution clinical epidemiology and public health.

* Corresponding author. Tel: +44 20 3549 5329, Fax: +44 20 7813 0242, E-mail: h.hemingway@ucl.ac.uk

† Details are listed at the end of acknowledgement section.

Conclusion High volumes of inherently diverse ('big') EHR data are beginning to disrupt the nature of cardiovascular research and care. Such big data have the potential to improve our understanding of disease causation and classification relevant for early translation and to contribute actionable analytics to improve health and healthcare.

Keywords Electronic health records • Health informatics • Bio-informatics • e-Health • Precision medicine • Translational research

Introduction

Electronic records relevant to the understanding of health and disease are found in diverse sources including not only the formal electronic health records (EHR) used in a growing number of healthcare organizations but also in omic, imaging, wearable and other data. These record data are increasingly being used for research, beyond the primary purpose for which they were collected. 'A new era of data-based and more precise medical treatment'[1] is envisaged in which the practice of medicine becomes 'evidence generating'.[2] One emerging prospect is the use of big record data to traverse the translational pathways from early discovery phases of translation to later implementation phases. Previous reviews on mining EHR have not had a focus on cardiovascular disease[3] or have focused on cardiovascular care[4,5] without a consideration of the translational pathways. We provide, for the first time, a critical review of big health record data for cardiovascular disease research across the translational spectrum, including early phases of discovery science, drug development and repurposing, and precision medicine, and later translational phases of learning health care systems, real world evidence, citizen-centred, and public health.

We review four areas in relation to big health record data:

(i) What data *resources* exist for cardiovascular disease research?
(ii) What are the *challenges* and *barriers* to realizing these opportunities?
(iii) What is the potential of such data in *early translational* research including discovery science, drug development and repurposing, precision medicine?
(iv) What is the potential of such data in *late translational* research including learning health care systems, real world evidence, citizen-centred and public health?

Big health record data resources

'Big data' are usefully characterized by 'variety, volume, velocity, and value' (a fifth V, veracity, relating to data quality is dealt with below in the challenges section). EHR are intrinsically 'big' due to their complexity ('variety') and numbers of patients and amount of information on each patient ('volume') and are collected for a variety of purposes (such as clinical care, billing, auditing, and quality monitoring).[6–9]

Tradeoffs between scale and depth

Figure 1 illustrates the variety and volume of data showing the relation between scale (number of people) and depth of phenotypic and omics information in different settings: national population-based, hospital-based, and disease or procedure based registries. The amount of phenotypic information in hospital EHR is much greater than any single registry; but such deeper hospital EHR data, has been challenging for researchers to access at scale.[10] Hospital EHR potentially provide phenotypically detailed data on all diseases including clinical blood laboratory values, imaging, clinically used device data, and text.[11–14] EHR comprise both structured and unstructured electronic data generated and captured during routine clinical care. Structured EHR data are recorded using controlled clinical terminologies [such as Systematized Nomenclature of Medicine - Clinical terms (SMOMED-CT)] or statistical classification systems (such as ICD-9, ICD-9-CM, or ICD-10). Unstructured clinical data such as patient medical histories, discharge summaries, handover notes, and imaging reports are captured and recorded in patient's health records as raw unformatted text. Such varied data, from different sources, has been likened to a tapestry[15] which can be woven together using data linkage and integration techniques into a fine-grained longitudinal picture of health over time (the 'human phenome sequence'). Such diverse data may offer higher resolution of clinically relevant clusters of diseases, causes, and classifiers.

Figure 1 makes an important distinction between those record resources with, and without genomic information. *Boxes 1* and *2* provide further details of these resources which may be accessed for translational collaborative research. Biobanks and genomics consortia increasingly rely on EHR linkages for the ascertainment, validation, and phenotyping of not only specific disease outcomes but also the entire longitudinal phenome, as captured by an growing array of digital sources.[16] Thus any one data resource may include combinations of researcher-generated data (such as omics) and researcher-harnessed data from EHR. Recent initiatives, such as the Innovative Medicines Initiative Big Data for Better Outcomes 'Big Data@Heart',[17,18] and the American Heart Association (AHA) Verily AstraZeneca 'One Brave Idea' initiative.[19] seek to exploit different sources of records and omics data, across multiple consented and anonymized sources—using the human as the 'new model organism'.

Digital trace of health, outwith healthcare

The resources illustrated in *Boxes 1* and *2* are making increasing use of such data sources including the physical environment, consumer information, socioeconomic and behavioural factors[20,21] and user-generated data from mobile health apps, wearables, sensors and social media.[22–24] In particular the 'always on' aspects of mobile and wearables provides major opportunities.

In order to exploit these resources for translational research there is an increasing use of computer science approaches to harness publicly available curated knowledge in different fields including: the

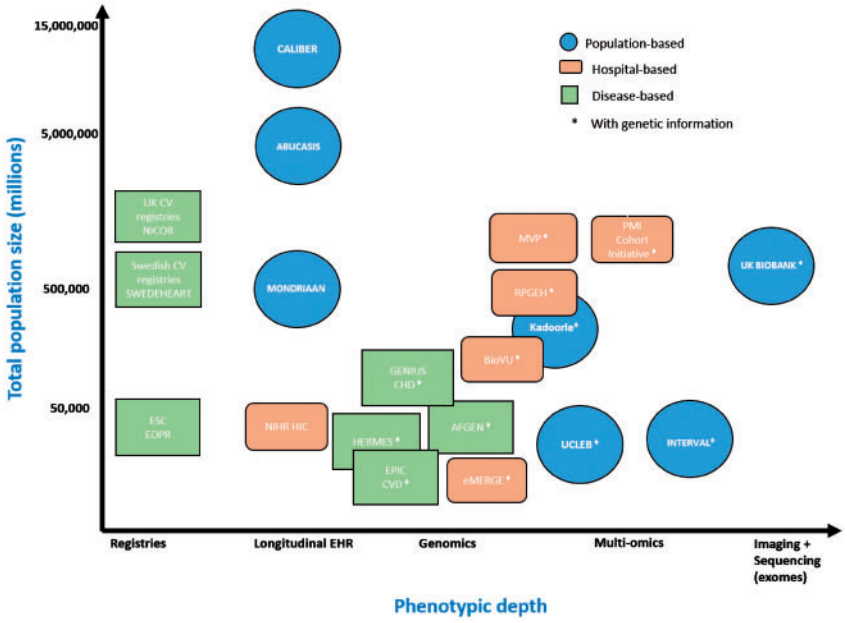

Figure 1 Scale (*N* people), phenotypic and omic resolution of population-based, hospital-based and disease-based exemplar data resources relevant to cardiovascular disease research (for further details see *Boxes 1* and *2*). AF, atrial fibrillation; AFGen, AF Consortium; CHD, coronary heart disease; ESC, European Society of Cardiology; EPIC, European Prospective Investigation into Cancer and Nutrition; ERFC, Emerging Risk Factors Collaboration; eMERGE, Electronic Medical Records and Genomics; HF, heart failure; PMI, precision medicine initiative; MVP, Million Veterans Programme; NICOR, National Institute for Cardiovascular Outcomes Research; NIHR, National Institute for Health Research; RPGEH, Research Programme on Genes, Environment, and Health; UCLEB, University College, London School of Hygiene and Tropical Medicine, Edinburgh, Bristol.

medical literature (e.g. PubMed), catalogues of genetic variant-phenotype associations (PhenoScanner[25,26]), disease-agnostic drug targets (e.g. DrugBank[27]), drug compounds (CHEMBL,[28] and adverse drug reactions (e.g. IMI PROTECT[29]).

Volume: scale with cohorts of millions of participants

Higher resolution enquiry of common and rare diseases (or rare outcomes of common diseases, including drug side effects), demands higher sample sizes: 5000 people in the Framingham cohort, 500 000 in UK Biobank,[16] 15 000 000 in curated, linked EHR cohorts such as CALIBER,[30–32] (*Figure 3*) and cross-national collections of EHR cohorts in 100 000 000.[33–35] An individual's interactions with the healthcare system may also generate big data; in the general population on average one person accumulates 1000 health events over 3 years in national coded data; a single cardiac MR scan has 10^8 voxels and a clinical grade (×30) whole genome sequence provides 15 Gb of data.[36,37]

Value: opportunity to disrupt current models of research and care

The value of diverse, high volume data is already changing the way that health care is delivered and is yielding insights in early and late translation (see Potential for early translational research section). There are many sources of value in big data, beyond the immediate scientific dimensions of scale and longitudinal phenotypic resolution. These include the *whole-system relevance* when population and healthcare system records are used: for example, in countries with nationwide health record systems, EHR are the only way of obtaining large scale representative samples. The *velocity* of big data is an opportunity for real time analytics with intelligent feedback loops to improve healthcare systems and individual decision making. The exploitation of such rich big record data sources is *more efficient and cost-effective* compared with traditional researcher-led approaches since for example, in EHR cohorts the cost to research funders of baseline and follow up data *collection* is zero (the data exist as part of healthcare systems). The costs however of collating, cleaning and curating these data and meeting the challenges outlined below are substantial and are further elaborated below.

Big health data challenges

In realizing the opportunities of such diverse, large volume data there are formidable challenges. These include: knowing what data are potentially available, information governance, models of data access (responsible data sharing), building and maintaining public trust, developing standards for defining disease, and developing tools for scalable, replicable science and equipping the clinical and scientific work force with new inter-disciplinary skills.

Are the data of sufficient quality for a given research question?

Challenge: The quality of EHR data can be said to be 'in the eye of the researcher'. In any given dataset the amount of missing data, often not

Box 1 Examples of large scale genomic—electronic health record resources

A. *Population-based*

UK-Biobank (UK, *n* = 500 000) Custom exome array, and exome sequencing, panel of 32 biomarkers, activity monitors and behaviours in 500k and cardiac MR underway in 100k, linkage to hospital and primary care EHR (http://www.ukbiobank.ac.uk/).[36]

UCLEB (UK, *n* = 30 000) 14 consented British cohorts (12 population-based and 2 randomized trials) including ∼30k participants Genomics, quantitative NMR metabolomics (∼18k), Somalogic-Proteomics (2k), digitalized ECGs at baseline and in some studies with multiple follow-ups, and imaging (cardiac, carotid and brain) in a sub-set and linkage to hospital EHR (http://datacompass.lshtm.ac.uk/40/).

INTERVAL (UK, *n* = 50 000) 'multi-omics' bioresource, includes whole-genome sequencing (×20 depth), genome- wide genotypes, lipidomics, proteomics, metabolomics, accelerometry (100 Hz, 7 days). Participants are linked to electronic health records and are aged 18–80 years. This study involves the largest experiment to date using SomaLogic's proteomics assay. (http://www.intervalstudy.org.uk/)

EPIC-CVD (EU, *n* = 520 000) case-cohort study, embedded in the 10-country, 22-centre pan-European EPIC cohort, involving >25 000 incident CVD cases. Multiple gene arrays (GWAS, exomechip, metabochip), >75 circulating biomarkers, extensive lifestyle profiling, and large subsets with serial measurements and linkage with electronic health records.(http://www.epiccvd.eu/)

China Kadoorie Biobank (China, *n* = 510 000) large Biobanked cohort investigating genetic and environmental causes of common chronic diseases in the Chinese population across 10 geographic regions (http://www.ckbiobank.org/site/).

B. *Hospital-based*

DiscovEHR project of the Regeneron Genetics Center and the Geisinger Health System (US, *n* = 42k): enrolees with whole exome sequencing and linkage to electronic health records over 15 years of clinical care (http://www.discovehrshare.com).[78,165,166]

US Department of Veteran Affairs—Million Veteran Program (US, *n* = 500k): aiming to recruit 1 million users of the VA healthcare system and collect DNA specimens, tissue samples, electronic health records from VA and survey data (https://www.research.va.gov/mvp/)

Kaiser Permanente—Research Program on Genes, Environment and Health (US, *n* = 500k): Based on the over six million-member Kaiser Permanente Medical Care Plan of Northern California (KPNC) and Southern California (KPSC), the completed resource will link together comprehensive electronic medical records, data on relevant behavioural and environmental factors, and biobank data (genetic information from saliva and blood) from 500 000 consenting health plan members (http://www.rpgeh.kaiser.org/).

Vanderbilt BioVU (US): BioVU is Vanderbilt's biorepository of DNA extracted from discarded blood collected during routine clinical testing and linked to de-identified medical records in the Synthetic Derivative. The goal of BioVU is to provide a resource to Vanderbilt investigators for studies of genotype-phenotype associations (https://victr.vanderbilt.edu/pub/biovu/).

eMERGE (US, *n* = 105k): consists of nine study sites, two central sequencing and genotyping facilities, and a coordinating centre. eMERGE aims to continue to develop and validate electronic phenotyping algorithms for large-scale, high-throughput genomics research; to discover genetic variants related to complex traits; to disseminate results and lessons learned to the scientific community; and to deliver state-of-the-art genomic knowledge, methods, and approaches to clinical decision support and clinical care. Specifically: (i) sequence and assess the phenotypic implication of rare variants in ∼100 clinically relevant genes presumed to affect gene function in about 25 000 individuals; (ii) assess the phenotypic implications of these variants, (iii) integrate genetic variants into EMRs for clinical care; and (iv) create community resources. (https://emerge.mc.vanderbilt.edu/)

Precision Medicine Initiative Cohort Program (US): longitudinal research effort that aims to engage one million or more US participants to enable research that will, over time, improve the ability to prevent and treat disease based on individual differences in lifestyle, environment and genetics. Participants will be invited to contribute a range of data about themselves by completing questionnaires, granting access to their electronic health records, providing blood and urine samples, undergoing physical evaluations and sharing real-time information via smartphones or wearable devices.(https://allofus.nih.gov/)

C. *Disease-based*

GENIUS-CHD (global, *n* = 250k) Coronary Heart Disease (CHD) patients (of which 129k are Acute Coronary Syndrome, ACS patients) with genotyping and biobanked samples and longitudinal follow up, from over 60 studies (including observational and randomized trials); central goal is to identify genetic and non-genetic determinants of subsequent or recurrent event risk, to facilitate discovery and validation of novel molecular pathways and drug targets for CHD secondary prevention. (http://www.genius-chd.com/)

HERMES Consortium (global, *n* = 11k) International consortium whose main aim is to identify the genetic determinants for incident HF and recurrent events with over 11 000 heart failure cases. (http://www.hermesconsortium.org/)

AFGen Consortium (US & EU) 30 studies with Exome-chips and GWAS array. Further phenotyping underway for 40k AF cases. (https://www.afgen.org/)

Box 2 Examples of large scale electronic health record resources without genomic information

A. *Population-based:*

CALIBER (UK, *n* = 10 M) CALIBER is a population based research platform of linked EHR and administrative health data from primary care (Clinical Practice Research Datalink), secondary care (Hospital Episode Statistics), disease (Myocardial Ischaemia National Audit Project) and death (Office for National Statistics) registries with longitudinal data on all prescribed medicines, diagnoses and blood values (>300k cases of AF, HF, and ACS) and a set of computational tools and research-ready phenotyping algorithms. (https://www.caliberresearch.org)

ABUCASIS (ES, *n* = 5M) EHR of entire Valencia population, including SIA (whole data from ∼5.100.000 subjects when they attend the physicians´ office in primary care; hospital morbidity, CMBD diagnostics of all the hospital admissions; mortality; GAIA (all prescriptions); vaccination; visits to the Health Care Centre (in 2014 >60M); laboratory tests; Clinical Risk Groups (CRG) classification in each subject.

Mondriaan (NL, *n* = 15M) Pharmacy and claims data, harmonized with >500k GP data and subset with genetics data. 15 years follow-up. (http://mondriaanfoundation.org/)

B. *Hospital-based:*

National Institute for Health Research Health Informatics Collaborative (HIC) (UK): platform for extracting phenotypically rich clinical data for research from hospital care across give major NHS trusts (Oxford, Cambridge, UCLH, Guy's and Imperial) and five disease areas (acute coronary syndrome, viral hepatitis, critical care, ovarian cancer, renal transplantation). (http://www.hic.nihr.ac.uk/)

C. *Disease-based: National quality of care and outcome registries*

SWEDEHEART (SE, *n* = 2M) SWEDEHEART is a national registry, including all patients undergoing coronary angiography, percutaneous coronary intervention, heart surgery and TAVI since 1990 s, and almost all patients with acute myocardial infarction. Patients with MI, <75 years of age, are also followed for 1 year regarding secondary prevention. SWEDEHEART collects more than 500 variables. The database is regularly validated and the agreement between the registry and the electronic health records is 95–96%. (http://www.ucr.uu.se/swedeheart/)

European Society of Cardiology European Research Programme (EOPR) AF (multinational) Multinational (31 countries) observational study including patients with atrial fibrillation since 2012 (https://www.escardio.org/Research/Registries-&-surveys/Observational-registry-programme)

National Institute for Cardiovascular Outcomes Research (NCOR) (UK) National cardiovascular disease and procedure registries, with data on >2 m individuals; ACS- through the Myocardial Ischaemia National Audit Project registry (MINAP) >1 m patients; HF Registry >100k patients; Arrhythmia registry. (https://www.ucl.ac.uk/nicor)

missing at random, or inaccurate data, may prohibit valid inference for some but not all research questions. Linked EHR, subject to robust pre-processing and cleaning, have been shown to provide valid measures of risk factors and a wide range of diseases, and therefore offer a common scaffold on which to build specific research questions.[30] *Solution*: A data mantra is 'collect once, use many times': and there are calls to make good quality clinical record keeping, as 'research grade data'. It should be noted that accurate and complete recording, though desirable, does not replace appropriate study design or resolve limitations such as confounding by indication. Validity and data quality may be assessed in multiple ways including:

- *Cross referencing multiple sources of data in the same individuals* (each with their own strengths and limitations): e.g. for acute myocardial infarction linking four national population based sources (primary care, hospital, heart attack and death registries) (CALIBER) shows the positive predictive value and prognostic validity of cases defined in different sources, and allows development and sharing of phenotypic algorithms.[38,39] Comparisons of trial adjudicated and medical claims data have been shown to be poor for some endpoints (e.g. bleeding[40]), a comparison of adjudicated endpoints and administrative data showed good agreement.[41]
- *International comparisons*: for example, EHR cohorts in heart attack survivors using ICD codes from different versions (ICD-9-CM, ICD-9, ICD-10) and different countries (US, Sweden, France and England) demonstrated for 12 risk factors consistent relative risks associations with fatal and non-fatal long term outcomes.[42] In general populations the Emerging Risk Factors Collaboration (ERFC) has shown consistency across continents of risk factor associations with CHD incidence.[43]
- *Genomic approaches* to validating case definitions: across 1000s of hospital ICD codes ('phenome-wide'), reproduce associations from genome wide association studies obtained one phenotype at a time[25,44,45] (*Table 1*, Denny *et al.*[44] *Figure 2*).

What data exist?

Challenge: 'Genome browsers' facilitate discovery in biological sciences, but currently the contents of the big data tapestry and whether they are suitable for a particular research purpose are hard to uncover within a researcher's own country, let alone across different countries (see *Figure 1* and *Boxes 1* and *2*). *Solution*: If big data are to disrupt current research models then there is a need for searchable catalogues of data, metadata, feasibility counts (and ideally sample data) and access arrangements. The creation of public, standards-driven metadata and data portals can assist researchers in locating the right dataset for their research question and obtaining up to date details on data availability and accessibility. For example, the IMI-funded European Medical Information Framework (EMIF) data catalogue contains

Table 1 Early translation exemplars of big health record data research: discovery of disease mechanism, drug development, and precision medicine

Health challenges	Example	Author/ year	*N* patients	*N* and type of sources*	Phenotype at baseline	Longitudinal phenotypes, omics and imaging	Analysis approaches
Discovery							
Human knockouts and health	Population based resource for experimental medicine in 'human knockouts'[62]	Narasimham *et al.* 2016	3.222k	3 Consented cohort Recall by genotype, EHR-1°,	Parentally related Pakistani adults recruited from antenatal clinic	Exome sequencing: III rare variant genotype; predicts loss of gene function Result 1358 phenotypes	Genetics Experimental medicine Informatics
Discovery approaches agnostic to disease and biology	GWAS and phenome wide association studies (PheWAS)[44]	Denny *et al.* 2017	13.835k	EHR	1358 phenotypes in Hospital treated patients	Genotyped 3144 SNPs	Informatics GWAS
Discovering new disease sub-types	Heart failure with preserved ejection fraction divided into two groups with differing outcomes[65]	Shah *et al.* 2015	0.397k	5 67 parameters physical characteristics, blood labor, ECG, echocardiography	Heart failure (preserved ejection fraction, HFpEF)	HF hospitalization	Machine learning: unbiased hierarchical cluster analysis on continuous values
Developing models of disease networks	Networks of more than 1000 longitudinal disease trajectories: gout important for cardiovascular disease progression[149]	Jensen *et al.* 2014	6.2 m	1 EHR-2° (admissions, outpatients, casualty)	All diagnosed diseases	All diagnosed diseases (14.9 years)	Trajectory/network analysis
Drug discovery and repurposing							
Drug target validation	Inactivating mutation in gene (NPC1L1) mimicking drug (ezetimibe) and effect on LDL cholesterol and coronary disease [78]	Stitziel *et al.* 2014	7.364k	Various, incl. Biovu & GoDarts	CHD cases	Exon sequencing genetics of NPC1L1	Genetics for drug target validation
Repurposing existing drugs	Mapping GWAS catalogues to druggable genome and 3 tiers of compounds 8 drug target gene associations concordant, 19 discordant[85] e.g. Tocilizumab licensed for rheumatoid arthritis, being tested for use in coronary disease	Finan *et al.* 2017	>100k in 84 GWAS relevant to CVDs	3 All GWAS, All compounds	84 GWAS in 39 CVDs 388 associations in 670 genes, of which 135 genes drugabble	All compounds with bioactivity against targets 18 844 in CHEMBL Druggable genome	Bioinformatics

Continued

Table 1 *Continued*

Health challenges	Example	Author/ year	*N* patients	*N* and type of sources*	Phenotype at baseline	Longitudinal phenotypes, omics and imaging	Analysis approaches
Trial endpoint optimization	Heart failure and peripheral arterial disease are comments diseases, seldom prominent in trial endpoints[32]	Rapsomaniki *et al.* 2014	2000k	4 EHR1°, QR, A, M	Healthy, free from diagnosed CVDs at baseline	12 incident CVDs over follow up	Cohort epidemiology
Precision medicine							
Pre-emptive pharmacogenomics in care	Genotypes used to select anti-platelet drug, or dosing in warfarin[99]	Van Driest *et al.* 2014	10k	EHR structured and text	Hospital patients at high risk of subsequent receipt of antithrombotics	Clopidogrel CYP2C19; Simva SLC01B1; Warfarin VK0RC1; CYP2C9; thiopurine TPMT; tacrolimus CYP3A5	Demonstration project
Tailoring drug treatment decisions to a patients risk of benefit and harm	Prolonged dual anti-platelet therapy: Development and validation of risk prediction models for benefits (CVD death, MI and stroke) and harms (bleeding) [100]	Pasea *et al.* 2017	18.307k	4 EHR-1° QR, A, M	Stable CAD 12 months post-AMI	CVD, MI, stroke Bleeding	Multiple prognostic risk models and net benefit

*EHR, electronic health records; QR, quality registry; A, Administrative data; M, mortality; GWAS, genome wide association study.

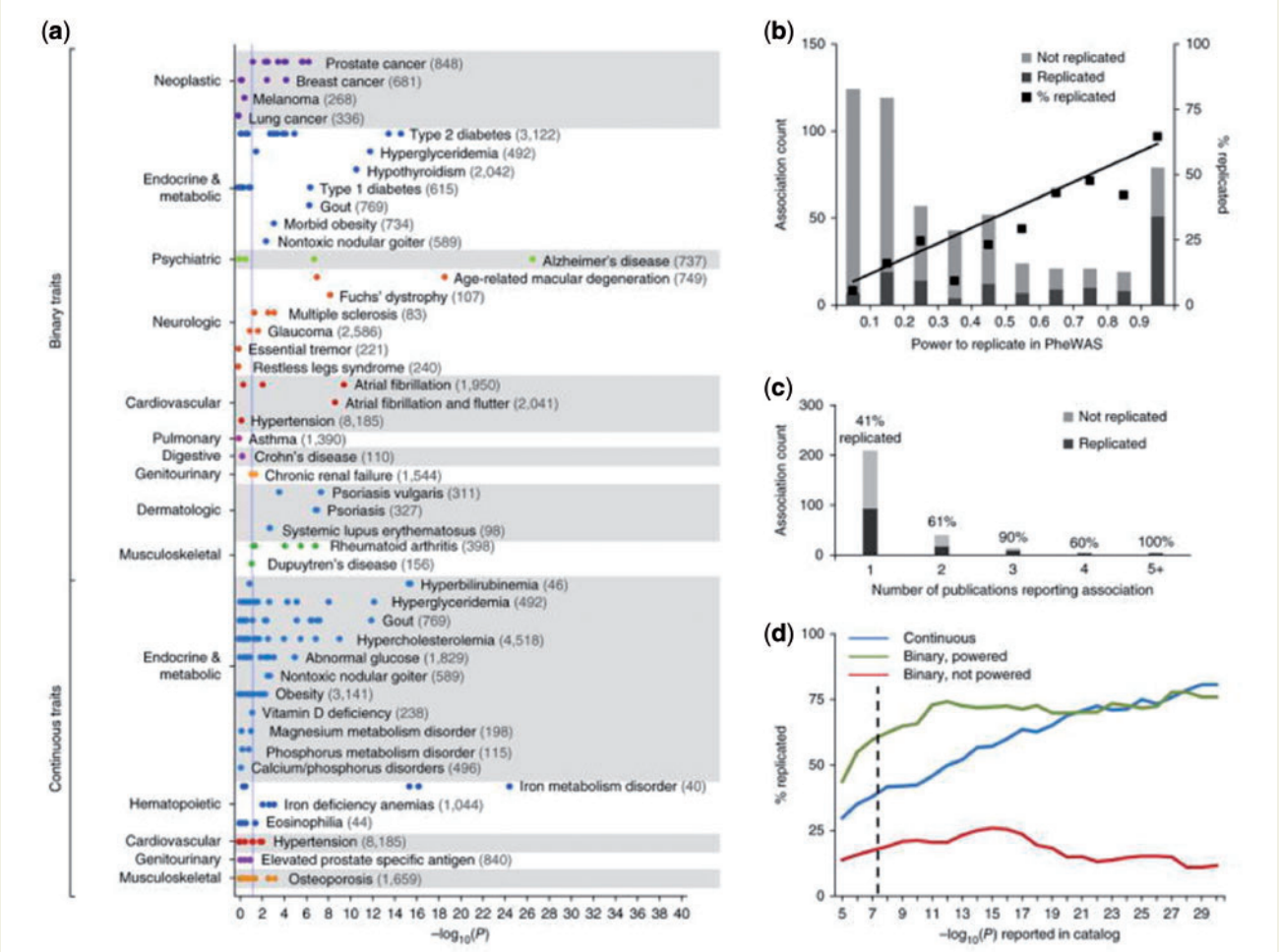

Figure 2 Electronic health record (EHR) Phenome Wide association studies (PheWAS). Source: Denny *et al.*[44] (reproduced by kind permission). Each point represents the $-\log_{10}(P)$ of a single SNP-phenotype association tested with PheWAS. This study is restricted to SNP-phenotype associations that achieved genome-wide significance ($P \leq 5 \times 10^{-8}$) in at least one prior genome wide association study (GWAS) study that included individuals of European ancestry. Numbers in parentheses beside each phenotype represent the sample size within the PheWAS data set. The vertical blue line represents $P = 0.05$. Binary traits refer to all adequately powered, binary traits in the NHGRI Catalog with exact matches to a PheWAS phenotype. For example, 5/5 catalog SNPs associated with rheumatoid arthritis were replicated at $P < 0.05$ in PheWAS, and 9/15 SNPs associated with type 2 diabetes were replicated. Continuous traits are those numerically defined traits in the NHGRI Catalog that are related to PheWAS diseases (e.g. 'iron deficiency anaemia' was the PheWAS trait paired with the 'serum iron level' catalog trait).

information on over 300 data sources ranging from EHR, consented cohort studies, and surveillance datasets.

What is the legal and ethical framework for using such data?

Challenge: The information governance of big health data resources presents major challenges. The need for protecting privacy, confidentiality, discrimination and other potential harms is vital. However how the regulatory environment proportionately balances these concerns with the potential benefits of data sharing (or, indeed, the harms by not sharing) is evolving.

Solution: Broad consent models, such as those in UK Biobank, have an important role, recognizing that it is not possible to stipulate all the potential research uses of data, nor how they will change. Some have argued that a new social contract is required with trusted use of data under innovative, proportionate governance delivering benefits to patients and public.[46–48]

How are data shared?

Challenge: Despite exhortation from funders, journals and the public to share data, all too often this does not happen. Once researchers have permissions to access data, the mode of data sharing may pose challenges to the researcher.

Solution: Data sharing may involve: (i) material transfer agreements with data being physically shared e.g. UK Biobank; (ii) role-based secure remote access; (iii) distributed analyses where data remain

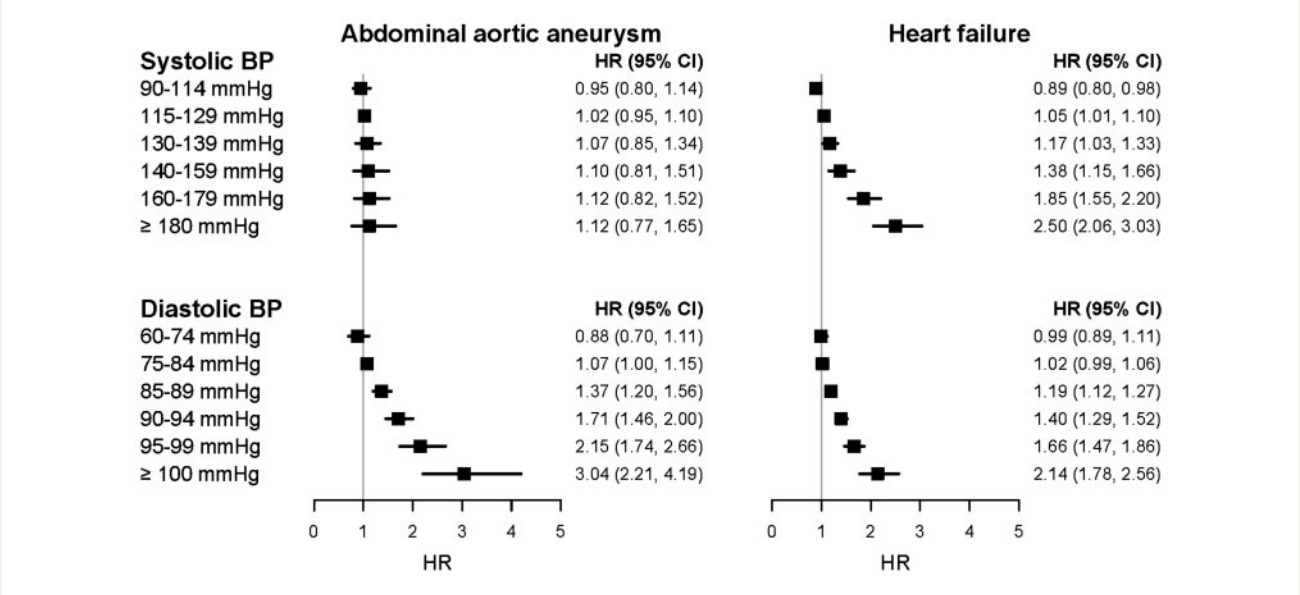

Figure 3 Resolution across a range of risk factor levels (systolic and diastolic blood pressure) and range of different initial presentations of cardiovascular disease (abdominal aortic aneurysm and heart failure only shown here): discovery of heterogeneous associations in a cohort of >1m adults initially free from diagnosed cardiovascular disease using national structured linked electronic health records from the CALIBER resource, in which EHR phenotyping algorithms are created, validated and shared using a robust methodology.[32,50]

stored in individual sources. The Global Alliance for Genomics and Health[49] is establishing a common framework for harmonized sharing of genomic and clinical data. Distributed analytical tools (e.g. DataSHIELD, i2b2) and common data models [e.g. Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)] can facilitate the remote and sequestered processing of complex datasets without the direct need to transfer data directly.

How are disease and trait phenotypes defined and shared?

Challenge: There is a lack of an international framework for defining, phenotyping, sub-phenotyping and discovering disease phenotypes in the context of health records. There are multiple controlled clinical terminologies and ontologies (including SNOMED-CT, ICD-10, and the Human Phenotype Ontology), but how these terms should be combined to define meaningful entities, let alone how they should be combined with research data is unclear. Currently many diseases lack internationally agreed criteria (preferably in a machine-readable format) for defining cases and non-cases; acute myocardial infarction, type 2 diabetes are exceptions. Current definitions of many diseases such as HF, AF and ACS span heterogeneous groups of patients and describe syndromes only rather than definitions based on understanding of molecular mechanism.

Solution: Sharing, validating and refining replicable, scalable EHR phenotypic algorithms requires international efforts.[50] e.g. PheKB in hospital EHR (codes and text) and national structured records e.g. CALIBER. For example, defining atrial fibrillation using structured national health records may involve several hundred codes for diagnoses, drugs, procedures in a phenotyping algorithm. Clinical information standards such as openEHR[51] or semantic web technologies[52,53] can enable researchers to create computational representations of phenotyping algorithms which facilitate their sharing across the research community.

What are the tools, methods and analytic approaches?

Challenge: There is a wide array of relevant approaches from quantitative disciplines (mathematics, computer science, statistics, software engineering) and from biological disciplines: until recently these have seldom been focused on big health record data.

Solution: While there are 7 million hits per day on the European Bioinformatics Institute website; such national and international resources for health informatics are lacking. There is a need for organizations to be established which provide the analogous reference data, tools and methods in health informatics in general[54] as well as integration across cardiovascular efforts[55,56] in order to scale the science.

What skills and training are required?

Challenge: Few clinicians and health care professionals have had formal training in informatics, data science, (computer) coding, software development or other increasingly relevant skills. In many countries there are large shortfalls in the number of data scientists that have been trained.

Solution: National efforts are likely to be important to substantially increase the number, and change the kind, of people required to deliver data-based medicine: hybrid professionals,(for example subspecialty physician accreditation in informatics), data scientists, data wranglers, and data-savvy health care professionals.[57] The 10×10 ('ten by ten') program was launched in 2005 by the American Medical Informatics Association (AMIA) and Oregon Health & Science University (OHSU). The genesis for the program came when then-President of AMIA, Dr Charles Safran, called for at least one

physician and one nurse in each of the 6000 hospitals in the US to have some training in medical informatics. The National Academy of Science has recommended the importance of agile assembly and rewarding of scientific teams across diverse disciplines including genomics, basic biology, mathematics, computer science, statistics, engineering.

Potential for early translational research

In this section we provide selected exemplars of the potential of big health record data arising from the variety, volume and value of the data being realized and how big data are contributing to scientific advance in cardiovascular medicine from discovery of underlying disease mechanisms, disease taxonomy, of treatment relevant sub-types of disease which underpin drug development, and precision medicine.[58,59]

Discovery in genetic and EHR data

It is important to note that it is challenging to provide deep mechanistic insight in large scale EHR data resources given the limited availability of genetic information in sufficient depth. Bespoke, recallable investigator-led studies such as East London Genes & Health (ELGH[60]) and the NIHR BioResource[61] enable the coupling of EHR data with extreme genotypes (or phenotypes) and enable their in-depth study using bespoke experimental protocols.[62,63] Complete gene knockouts are highly informative about gene function with a recent study of 3222 British Pakistani-heritage exome-sequenced adults with high parental relatedness, discovered 1111 rare-variant homozygous likely loss of function (rhLOF) genotypes predicted to disrupt (knockout) 781 genes. Linking to EHR, investigators observed no association of rhLOF genotypes with prescription- or doctor-consultation rate, and no disease-related phenotypes in 33 of 42 individuals with rhLOF genotypes in recessive Mendelian disease genes. Phased genome sequencing of a healthy PRDM9 knockout mother, her child and controls, showed meiotic recombination sites localized away from PRDM9-dependent hotspots, demonstrating PRDM9 redundancy in humans. Genomic approaches to validating case definitions: across 1000 s of hospital ICD codes ('phenome-wide'), reproduce associations from genome wide association studies obtained one phenotype at a time (*Table 1*, Denny *et al.*,[44] *Figure 2*).

Discovery in larger scale epidemiology

Big health record data can contribute to the discovery of new associations, which would be hard to generate from traditional consented cohorts without record linkage. For example, *Figure 3* and *Table 1*, Rapsomaniki *et al.*[32] illustrates how the power of large scale health records allows enquiry into less common cardiovascular diseases such as abdominal aortic aneurysm: Here there is a marked discordance between the strong association of diastolic blood pressure with abdominal aortic aneurysm compared with the lack of association with systolic blood pressure. These findings have implications for understanding the aetiology of abdominal aortic aneurysms, screening and prevention and understanding the underlying molecular mechanisms of disease for creating interventions.

A key prerequisite for precision medicine is the estimation of disease progression from the current patient state. Disease correlations and temporal disease progression (trajectories) have mainly been analysed with focus on a small number of diseases or using large-scale approaches without time consideration, exceeding a few years. Investigators performed a discovery-driven analysis of temporal disease progression patterns using data from an electronic health registry covering the whole population of Denmark. Utilizing the entire spectrum of diseases, they convert 14.9 years of registry data on 6.2 million patients into 1171 significant trajectories. Key diagnoses such as gout and chronic obstructive pulmonary disease (COPD) were identified as central to disease progression across many of these trajectories and hence important to diagnose earlier. Such data-driven trajectory analyses may be useful for predicting and preventing future diseases of individual patients.

Discovery with deep phenotypic data

Most cardiovascular diseases (including acute myocardial infarction) have syndromic descriptions and labels, which may span multiple underlying pathological disease processes.[64] One approach to discovering mechanistically relevant disease types is to phenomap disease. For example, *Table 1*, Shah *et al.*,[65] in heart failure with preserved ejection fraction machine learning on 46 continuous clinical, laboratory, electrocardiographic, and echocardiographic findings has been used to define mutually exclusive groups, which relate to subsequent outcomes.[65–67] The cardiac atlas project (of normal and diseased hearts) is an example of large scale collaborations on feature extraction in imaging[68,69] using data sharing in standard formats Digital Imaging and Communications in Medicine (DICOM) of pixel and non-pixel data.[70] Personalization using physiological simulations[71] for example for cardiac resynchronization therapy[71,72] is proposed. Unstructured free-text data in EHR may add further resolution for patient stratification and disease co-occurrence estimation, which subsequently can be mapped to systems biology frameworks.[67]

Drug development and repurposing

More drugs are required to prevent and treat cardiovascular diseases. Since 2000, the FDA has approved only two new classes of cardiac drugs with widespread application: P2Y12 receptor inhibitors (such as clopidogrel, ticagrelor, prasugrel) and novel oral anti-coagulants (such as dagibatran, apixaban, rivaroxaban, edoxaban). Costly, late drug failures occurring within phase III trials have been recently seen for CETP inhibitors which raise HDL-cholesterol (HDL-C),[73–75] ivabradine which lowers the heart rate[76] and darapladib, a selective oral inhibitor of lipoprotein-associated phospholipase A_2.[77]

Discovering and validating drug targets

EHR-DNA resources may play an increasingly important role in drug discovery, genomic drug target validation, marker validation and in drug repurposing. For example, NPC1L1 (*Table 1*, Stitziel *et al.*[78]) demonstrates the strategy that human mutations that inactivate a gene encoding a drug target can mimic the action of an inhibitory drug—here ezetemibe—and thus can be used to infer potential effects of that drug. Ezetemibe is known to affect the marker (LDL cholesterol) but, until recently, not the disease (myocardial infarction). Among the largest sources of cases of MI and controls in this study was a DNA resource integrated into a health system with rich EHR.[78] The discovery of PCSK9 as a drug target to lower cholesterol,[79] which could in principle have been made in EHR-DNA resources, illustrates the importance of rare variants in identification of pathways relevant to the whole population. Mendelian randomization studies are

important in evaluating whether markers—such as heart rate and HDL cholesterol—are causal for the disease of interest. Such genetic studies have questioned the role of heart rate[80,81] and HDL cholesterol[82] in the aetiology of heart attack.

Drug repurposing and PheWAS

Identifying novel disease indications for already approved drugs (repositioning or repurposing) has been successful for sildenafil,[83] and beta blockers (repurposed for heart failure). The discovery that IL-6 is causally related to myocardial infarction[43] has led to proposals for repurposing tocilizumab, which is currently licensed for rheumatoid arthritis. Here the question is what other phenotypes are associated with the drug-relevant genetic variant?' (*Figure 2*) For example, examining 778 disease phenotypes based on ICD codes in the EHR[84] identified potential novel pleiotropic associations with a variant in the sodium channel gene *SCN10A*. This variant is associated not only with the anticipated arrhythmias, but (possibly) also with unanticipated diseases, here cholecystitis. Recent interest has been to scale this approach to systematically evaluate drugs against a wide range of untested diseases. To be successful this would require substantially larger EHR-DNA resources incorporating longitudinal disease trajectories from big record data[85] and might aid drug repurposing efforts.

Trial endpoint optimization

Drugs may fail in phase III trials because of the composition of primary endpoints. For example, the inclusion of myocardial infarction—which is not causally related to heart rate—in the trial of the heart rate lowering drug ivabradine. In trials of treatments in type 2 diabetes the primary endpoint often includes non-fatal MI, non-fatal stroke and death from cardiovascular diseases. Large scale record cohorts however demonstrate that the initial presentation of cardiovascular disease is commonly heart failure and peripheral arterial disease[86]—neither of which are prominent components of primary trial endpoints. Moreover, inclusion of some diseases might dilute the trial endpoint since type 2 diabetes is associated with a lower risk of aneurysms.[86] In CALIBER, the ability to reliably resolve 12 different CVDs demonstrates that the majority of incident cases of CVD are neither heart attack nor stroke[86] and that risk factor associations are heterogeneous across different diseases.[86–89]

Trials of new drugs

Once the 'right drug, the right target and right endpoints' have been evaluated, the next and most costly hurdle is to carry out the definitive experiment—the phase III trial. Twenty years ago the West of Scotland Coronary Prevention Study (WOSCOPS) statin trial study demonstrated the value of EHR linkage for long-term follow-up of clinical outcomes.[41,90] Underpinning regulatory and data standards and interoperability issues[91] are the focus of international initiatives,[92–94] but in cardiovascular disease there has not yet been a pragmatic phase III trial of a pre-licence drug. The Salford Lung Study (GSK, relovair) is the world's first such trial and is set in a regional 'whole health system' EHR.[95,96]

Integrating pharmacogenomics

Multi-scale biological data, when combined with these deeper phenotypes, underpin further dissection of disease. Whole genome sequencing is beginning to be implemented in clinical care, for molecular diagnosis, identification of risk of subsequent wide range of diseases, reproductive considerations and drug response.[36,97] It is in drug response that precision medicine is finding early application. Here the goal is to identify biologically relevant subgroups in which either the benefit is greater, or, more commonly, the harms are fewer (interaction on the relative risk scale). Pre-emptive genomic testing, in which actionable genetic variants have already been assessed prior to drug exposure, is beginning to be implemented in the EHR for the care of patients[98] (*Table 1*, Van Driest *et al.*[99]).

Personalized estimates of benefits and harms

One example of the need to individualize risk comes from prolonged dual anti-platelet therapy among patients who have survived 1 year after acute myocardial infarction. For example, *Table 1*, Pasea *et al.*,[100] in prognostic models for risk of atherothrombotic and bleeding events have recently been developed and validated and allow an updatable estimation of net clinical benefits for each patient to guide the decision for prolonged dual anti-platelet therapy.

Clinical record data are highly effective in distinguishing risk groups, for diverse diseases and in diverse settings[101–103] and higher risk patients usually have more absolute benefit than those in lower risk groups (i.e. without biologic interaction). Clinical risk prediction algorithms and decision support are rapidly proliferating in CVD and many tools can be envisaged in the management of a single patient, spanning benefits and harms at different time points. Clinical data can outperform the Framingham risk score,[102] and can flexibly model start point populations and endpoints and be easily updated in the light of new imaging, genetic information, and implemented in clinical practice. Predictions may be improved by incorporating clinical trajectories.[103] For example patients in whom blood pressure declines over time, without diagnosed heart failure, have a worse survival than those whose blood pressure remains stable.[104] Using all available data points across data modalities combined with machine learning or Bayesian network models may further add to prediction.[105–107]

Potential for late translational research

Learning health care systems

Increasing costs, complexity of patients and fragmentation of healthcare systems are challenges to delivering high quality care with better outcomes and value. Far from a data-based health care system, all too often there is a largely data free (or data silo'd) approach where the benefits of science and evidence, and experience of care are characterized by missed opportunities, waste and harm.[108–110] The state of 'digital maturity' in hospitals and health eco-systems, varies hugely. Arguably, more people die from lack of use of data than misuse of any other technology.[111] The concept of learning health systems puts informatics and big data as a central driver of quality, not only seeking to put what is known to work into practice (closing the 'second translational gap') but also contributing in new ways to understanding what is effective.[112–114] It is worth noting that however 'big' the data are observational analyses will not replace the need for randomized intervention studies due to the inherent limitations of observational studies to evaluate reliably any modest effect of interventions.

Table 2 **Late translation exemplars of big health record data research: learning health systems, citizen driven health, and public health**

Health challenges	Example	Author/ year	N patients (000's)	N and type of sources*	Phenotype at baseline	Longitudinal phenotypes, omics and imaging	Design/Analysis/ Disciplines
Learning health systems							
Integrating trials in clinical care	Thrombus aspiration at the time of primary coronary intervention (TASTE trial) has no impact on short or long term outcomes[115]	Fröbert *et al.* 2014	7.244k	3 QR, A, M	STEMI, Angio findings	Follow up for ACM, stent restem, uf MI (1 yr)	RCT: Point of care, registry embedded, pragmatic
Comparing effectiveness of whole health systems	Large differences in care and outcomes between UK and Sweden (all hospitals)[137]	Chung *et al.* 2014	500k	2 QR, M	NSTEMI STEMI	ACM (30 d)	Survival analysis
Vigilance for safety	Mining text could have detected the Vioxx – acute MI signal earlier than conventional pharmacoepidemology approaches[13]	Lependu *et al.* 2013	1.8k	1 hospital records structured and text	All diagnosed diseases rofecexib	All drug safety signals including (acute MI) 11 m clinical notes	Text mining
Targeting cost effective care	Cost effectiveness decision models provide willingness to pay estimates in different risk groups, and different treatment benefits for stable coronary disease[138]	Asaria *et al.* 2016	100k	4 EHR-primary care, QR, A, M	Stable CAD, stable angina (xxx AMI)	All hospitals, procedures, drug use, resource use	Health economics
Citizen driven health							
Real time real world 24/7 monitoring: the 'sensed self'	Pacemaker monitoring might lower event rates[141]	Hindricks *et al.* 2014	0.716k	4 Implantable monitor	Heart failure Multi-parameter monitoring	Primary outcome was: ACM, hospitalization for heart failure, worsening of NYHA class.	RCT of a detailed monitoring intervention
Delivering individualized interventions through mobile phones	Texts might increase smoking cessation[142]	Free *et al.* 2011	5.8k		Smoker	Cessation Text messaging	RCT of behavioural intervention
Understanding the public through social media	Twitter language might predict community heart disease rates[23]	Eichstaedt *et al.* 2015	148 m country mapped	CDC atherosclerosis Twitter	Mapping of words used in tweets to psychological constructs	N/A	Ecological correlations at county level
Public health							
Epidemiology of all CVDs and clinically relevant sub-types of disease	Incidence and survival of NSTEMI and STEM ICD9-CM[145]	Yeh *et al.* 2010	3000k	2 HMO, M	46 086 hospitalizations STEMI/NSTEMI	All cause mortality30 day	Cohort

Continued

Table 2 *Continued*

Health challenges	Example	Author/ year	N patients (000's)	N and type of sources*	Phenotype at baseline	Longitudinal phenotypes, omics and imaging	Design/Analysis/ Disciplines
Rare disease epidemiology	Rare disease: valid EHR phenotypes & new associations with coronary disease [HCM][151]	Pujades-Rodriguez *et al.* 2016	1.16k	4 EHR-primary care A, QR, M	Hypertrophic cardiomyopathy	Coronary, stroke, HF, arrhythmia, bleeding, DVT/PE at 4 years follow up	Cohort
Evaluating population impact of interventions	Introduction of smoke free legislation in different countries at different times: impact on admissions to hospital with heart attack England smoke-free 1 July 2007[160]	Sims *et al.* 2010	millions	1 HES	MI admission 1 July 2002–30 September 2008	N/A	Natural experiment Time series analysis

EHR, electronic health records; QR, quality registry; A, Administrative data; M, mortality; GWAS, genome wide association study; HCM, Hypertrophic Cardiomyopathy; MI, myocardial infarction; STEMI, ST-segment elevation MI; NSTEMI, Non ST-segment elevation MI; CAD, Coronary Artery Disease; ACM, All-Cause Mortality; RCT, Randomized Clinical Trial; NYHA, New York Heart Association; ICD9-CM, International Classification of Diseases 9th revision – Clinical Modifications; CVD, Cardiovascular Disease; HMO, Health Management Organisation; CDC, Centres for Disease Control and Prevention; HES, Hospital Episode Statistics; DVT, Deep Vein Thrombosis; PE, Pulmonary Embolism.

Building trials into health systems

A trial of thrombus aspiration demonstrated the feasibility of randomizing a high proportion of patients at point of care in the setting of a national quality registry[115,116] (*Table 2*, Fröbert and James[115]). These findings and the growing evidence that EHR can provide a platform for assessing feasibility, refining protocols and recruiting patients[41,90,117] have stimulated major interest because of the lower cost and higher speed of trial delivery. Pragmatic point-of-care EHR based trials are underway e.g. of high vs. low dose aspirin trial among people with stable coronary disease.[118–120]

Building quality into healthcare delivery: decision support and data based medicine

Early examples of data-based medicine are already here, with clinical data providing both the 'brain' to understand what needs fixing and the 'spinal cord' to help fix it. For example, analysis of health record cohorts provides understanding of the patient journey and cumulative missed opportunities of cardiovascular care over time[121,122] and may provide risk prediction tools which are derived from clinical data, and used in practice to support healthcare decision making.[102,123]

A small but growing number of hospitals have a suite of readily modifiable information feedback loops to improve care.[124] There is a need for more empirical demonstration of the impact on outcomes of these systems. A key challenge lies in intelligent real time systems.[125–127] Practice-based medicine[128,129] involves large-scale, real time studies (based on a health system's own data) to generate evidence directly relevant to the patient in front of the clinician. Sometimes this observation is sufficient, sometimes it allows systematic identification of the need for trials. These trials may exploit the efficiency of big data in point-of-care individual patient randomized trials embedded in a learning health system or may involve randomizing clusters of health care professionals, for example to evaluate complex interventions, such as decision support.[130,131]

Big data for safety vigilance

Mining EHR in real time with both coded and text data is an important source of safety information. For example, *Table 2*, Lependu *et al.*,[13] the excess myocardial infarction risk associated with rofecoxib (Vioxx) could have been detected 1–2 years earlier had records. There are international initiatives to achieve the vast scale required to evaluate drug safety in up to 150 million patients.[34,132–135] Using the Medicare Patient Safety Monitoring System there was a decline in adverse events following heart attack and heart failure, but not for pneumonia or conditions requiring surgery,[136] possibly as a result of more organized quality initiatives in the cardiovascular diseases.

International comparisons of whole system care and outcomes

Nationwide, policy relevant comparisons of care and outcomes among people with CVDs across health systems have only recently been reported. For example, *Table 2*, Chung *et al.*,[137] using data from ongoing quality registries from all hospitals in Sweden and the UK,

including more than half a million patients, demonstrates that 30-day MI mortality was higher in the UK than in Sweden. Politicians, policy makers and health care professionals seek to make claims that their health systems deliver world class care and outcomes—ongoing, even semi-automated comparisons across countries might be used to evaluate whether such claims are 'data-based'.

Cost effectiveness of innovation

Big data provide new opportunities in understanding the cost effectiveness of existing and new interventions. Because of the ability to assess baseline risks in unselected general populations (commonly higher risk than those reported in trials), such 'real world evidence' is increasingly required by payers and the regulators. As more data sources are linked, greater granularity of the care data (e.g. 67 different types of primary care 'consultation') may provide more accurate and more complete resource use data. For example, *Table 2*, Asaria *et al.*,[138] cost-effectiveness decision models can be developed before trials report to estimate the willingness to pay and pricing of a drug according to different trial benefits (relative risk reductions) applied to patients at different strata of risk.

Citizen-centred health

People increasingly have more and different information than their doctor or researcher raising new possibilities of 'disintermediation', potentially disrupting current models of health care and research.[139]

The heart and circulation are increasingly observable as a 'sensed self' with novel wireless devices for mobile monitoring, with huge new data streams.[22,140] Smartphone apps and sensors are available to record and transmit to physician, electrocardiograms (e.g. to screen for atrial fibrillation), heart rate, blood pressure, radial artery waveforms, respiratory rate, oxygen saturation, temperature, even ultrasound.[22] These may provide deeper, naturalistic phenotyping in areas often lacking in the clinical record, including: physical activity, weight, diet, sleep, quality of life, and symptoms and medication compliance. For accelerometry questions remain about how best to analyse and present such data.

Implantable devices such as pacemakers provide tele-monitoring data which might reduce the risk of fatal and non-fatal outcomes in patients with heart failure (*Table 2*, Hindricks *et al.*[141]). Interventions can be delivered through mobile means and text messaging may increase smoking cessation rates (*Table 2*, Free *et al.*[142]). Apple ResearchKit provides new ways to recruit people rapidly into studies.

Open, publicly available data donated and shared by citizens is becoming increasingly available. User generated content in social media are inherently public and the language used in twitter can be used to predict community heart disease rates (*Table 2*, Eichstaedt *et al.*[23]) and it is plausible that Google searches[24] might give clues to environmental pollution triggers of acute cardiovascular events. As patients increasingly access, own and control their health records[143] they may share their clinical records, genetic and other data through initiatives like 'Patients like me' and '23 And Me', offering networks of individuals to develop communities of interest e.g. in rare diseases for orphan drugs. Citizens may do their own science; with schoolchildren exploiting publically available data to develop diagnostic tools using artificial neural networks.[144]

Public health

There are major gaps in our ability to prevent the onset of and prolong life in, many of the most common cardiovascular diseases in the 21st century including atrial fibrillation, heart failure, peripheral arterial disease. There are also gaps in our ability to measure disease and model the impact of interventions in populations. Clinicians diagnose more specific entities than 'heart attack', 'CHD' or 'CVD' yet conventional consented cohorts have lacked the statistical size or the phenotypic resolution to measure clinically relevant sub-types of disease. Big data can study the diseases that clinicians diagnose to provide scalable, population based, updatable measurements of modern disease burden vital for the evaluation of alternative strategies of prevention. For example, big data can be used to estimate the incidence and survival of the treatment-relevant sub-types of MI (ST elevation and non-ST elevation) (*Table 2*, Exemplar Yeh *et al.*[145] or stable angina).[146]

Meaningfully complex models of public health

Existing models of disease prevention are simple and often focus on one disease or one risk factor at a time. Big data invite a richer understanding of the importance of: multiple diseases co-occurring[147]; networks of risk factors (obesity[20] and smoking[148] and diseases[149]; fine-grained geospatial resolution; rare[150] *Table 2*, Pujades-Rodriguez *et al.*[151] and common diseases; diseases as causes or triggers of cardiovascular events[152]; diseases of developing[101] and developed countries, and across multiple biological scales through to societal influences on health). In order to understand weather and climate big data, with appropriately complex mathematical models, are used in national institutes,[153] but no such analogue exists for public health.

Big socio-economic data

Unlike many technological advances, big data may have a role in actionable understanding of, and reductions in, inequalities in health and healthcare in rich and poor countries. The opportunity to move to a neighbourhood with lower poverty may reduce obesity and diabetes.[154] The data in this trial were collected through traditional means, but such data could have been captured in part with cross-government record linkages. Big data are important for achieving sustainable development goals[155] and recommendations have been made for the recording social and behavioural determinants in the clinical record.[156] Linking health record data to an individual's lifelong tax contributions may provide new policy relevant insights into the relations between wealth and health.[157,158] Cross-government approaches to big data might open up enquiry into neglected populations with insights to improve the cardiovascular health of those on social welfare benefits, the homeless, refugee, and prison populations.

Population impact of interventions

Big data can be used to evaluate the population impact of healthcare or public health interventions.[159] For example, *Table 2*, Sims *et al.*,[160] shows how health records have been used to demonstrate the impact of the public smoking ban on hospital admissions for heart attack.[160–162] Importantly, big health data are a means to evaluate the impact on population health of primary care[163] the state of digital maturity of a hospital or health system[164] or the existence of quality and outcome registries.

Conclusion

Big health record data are beginning to disrupt the nature of cardiovascular research as well as models of care. Exploiting such data is beginning to improve understanding of cardiovascular disease causation and classification, and contributing actionable analytics to improve health and healthcare, but major challenges need to be addressed to realize more fully their potential.

Acknowledgements

Harry Hemingway and Spiros Denaxas conceived the idea and wrote the first draft and co-authors provided further content and critical comment. Participating Institutions and people are listed in full on the website (http://www.bigdata4betterhearts.eu). University Medical Center Utrecht (UMCU): Diederick E. Grobbee, Folkert Asselbergs, Arno Hoes, Ghislaine van Thiel, Hans van Delden, René Eijkemans, Rolf Groenwold, Pieter Stolk, Olaf Klungel. University Medicine Göttingen (UMG); Stefan D. Anker, Gerd P Hasenfuss, Stephan von Haehling. European Society of Cardiology (ESC): Panos Vardas, Aldo Maggioni, Gerhard Hindricks, Nikolaos Maniadakis, John Camm, Isabel Bardinet, Christina Dimopoulou. European Heart Network: Susanne Løgstrup. University College London (UCL): Harry Hemingway, Richard Dobson, Aroon Hingorani, Spiros Denaxas, Folkert Asselbergs, JP Casas, Amitava Banerjee, Bernard de Bono, Tom Lumbers, Dan Swerdlow, Riyaz Patel, Claudia Langenberg. University of Cambridge (CAM): John Danesh, Adam Butterworth, Angela Wood. International Consortium for Health Outcomes Measurement (ICHOM): Jason Arora. Fundación para la investigación del Hospital Clinico de la Comunidad Valenciana (INCLIVA): Josep Redon, F.J. Chorro, Juan Carlos Pérez Cortés, J. Sanchis, J. Nuñez, V. Bodi, F. Martinez, Daniel Saez-Domingo, Jose L. Trillo. Centro de Investigacion Cardiovascular de Barcelona/Institut Catala de Ciències Cardiovasculars (ICCC): Lina Badimon. Karolinska Institutet (KI): Tomas Jernberg, Henrik Olsson, Jonas Faxén, Tonje Thorvaldsen, Lars H Lund, Gianluigi Savarese, Stefan James, Bodil Svennblead. Universitätsklinikum Hamburg-Eppendorf (UKE): Stefan Blankenberg, Tanja Zeller, Renate Schnabel. MedLaw Consult (MedLaw): Evert Ben van Veen. University of Birmingham (BHAM): Paulus Kirchhof, Dipak Kotecha. Uppsala Clinical Research Center, Uppsala University (UPPS): Bodil Svennblead, Stefan James, Jonas Oldgren. Actelion: Andrea Bayer. Bayer: Gunnar Brobert, Tomasz Dyszynski, Christoph Gerlinger, Daniel Freitag, John Edward Butler-Ransohoff, Alex Asiimwe, Alexander Michel, Kiliana Suzart-Woischnik, Sebastian Kloss. Novartis: Frederico Calado, Anders Gabrielsen. Institut de Recherches Internatonales Servier: Vanessa Blanc, Nicolas Boisseau, Fabrice Couvelard, Willy Gosgnach, Olivier Gryson, Julien Hervouet, Weiwei Li-Bertheau, Benoît Tyl. Somalogic: Jessica Williams. Jessica Ash, Steve Williams, Rachel Ostroff, Alan Williams. Vifor Pharma: Maureen Cronin, Vincent Fabian, Ong Siew Hwa, Avi Leaf. Hyve: Kees van Bochove, Marinel Cavelaars, Maxim Moinat, Stefan Payralbe.

Funding

The Farr Institute is funded from the Medical Research Council, Arthritis Research UK, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Economic and Social Research Council, Engineering and Physical Sciences Research Council, NIHR, National Institute for Social Care and Health Research, and Wellcome Trust (grant MR/K006584/1). The BigData@Heart Consortium is funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No. 116074. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA; it is chaired, by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and ESC.

Conflict of interest: J.D. reports grants from UK Medical Research Council, grants from British Heart Foundation, grants from UK National Institute of Health Research, grants from European Commission, during the conduct of the study; personal fees and non-financial support from Merck Sharpe & Dohme UK Atherosclerosis, personal fees and non-financial support from Novartis Cardiovascular and Metabolic Advisory Board, personal fees and non-financial support from Pfizer Population Research Advisory Panel, grants from British Heart Foundation, grants from European Research Council, grants from Merck, grants from National Institute of Health Research, grants from NHS Blood and Transplant, grants from Novartis, grants from Pfizer, grants from UK Medical Research Council, grants from Wellcome Trust, outside the submitted work. A.M. reports personal fees from Novartis, personal fees from Bayer, personal fees from Cardiorentis, personal fees from Fresenius, outside the submitted work. M.C. reports personal fees from Vifor Int. G.B. reports personal fees from Bayer. P.V. reports grants and personal fees from Bayer, grants and personal fees from Servier, grants and personal fees from Pfeizer, grants and personal fees from Menarini, grants and personal fees from Boehringer, during the conduct of the study. S.D.A. receives honoraria for speaking or consultancy from Bayer, Boehringer Ingelheim, Novartis, Servier and Vifor Int, as well as grant support for clinical trial research from Abbott Vascular and Vifor Int.

References

1. 2016 State of the Union | whitehouse.gov [Internet]. https://www.whitehouse.gov/joint-address (11 August 2017).
2. Sim I. Two ways of knowing: big data and evidence-based medicine. *Ann Intern Med* 2016;**164**:562.
3. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012;**13**:395–405.
4. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol* 2016;**13**:350–359.
5. Cowie MR, Bax J, Bruining N, Cleland JGF, Koehler F, Malik M, Pinto F, van der Velde E, Vardas P. EHJ POSITION STATEMENT e-Health: a position statement of the European Society of Cardiology. *Eur Heart J* 2016;**37**:63–66.
6. NICOR (National Institute for Cardiovascular Outcomes Research) [Internet]. http://www.ucl.ac.uk/nicor (11 August 2017).
7. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, Fonarow GC, Geraci SA, Horwich T, Januzzi JL, Johnson MR, Kasper EK, Levy WC, Masoudi FA, McBride PE, McMurray JJV, Mitchell JE, Peterson PN, Riegel B, Sam F, Stevenson LW, Tang WHW, Tsai EJ, Wilkoff BL. 2013 ACCF/AHA guideline for the management of heart failure: executive summary: a report of the American college of cardiology foundation/American Heart Association task force on practice guidelines. *Circulation* 2013;**128**:1810–1852.
8. Jernberg T, Johanson P, Held C, Svennblad B, Lindbäck J, Wallentin L. Association between adoption of evidence-based treatment and survival for patients with ST-elevated myocardial infarction. *JAMA* 2011;**305**:1677–1684.
9. Huffman MD, Prabhakaran D, Abraham AK, Krishnan MN, Nambiar AC, Mohanan PP. Optminal in-hospital and discharge medical therapy in acute coronary syndromes in Kerala. *Circ Cardiovasc Qual Outcomes* 2013;**6**:436–443.
10. Denaxas SC, Asselbergs FW, Moore JH. The tip of the iceberg: challenges of accessing hospital electronic health record data for biological data mining. *BioData Min* 2016;**9**:29.
11. Pakhomov SSV, Hemingway H, Weston SA, Jacobsen SJ, Rodeheffer R, Roger VL. Epidemiology of angina pectoris: role of natural language processing of the medical record. *Am Heart J* 2007;**153**:666–673.
12. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. Brusic V, editor. *PLoS One* 2012;**7**:e30412.
13. Lependu P, Iyer SV, Bauer-Mehren A, Harpaz R, Ghebremariam YT, Cooke JP, Shah NH. Pharmacovigilance using clinical text. *AMIA Jt Summits Transl Sci Proc* 2013;**2013**:109.

14. NIHR Health Informatics Collaborative [Internet]. http://www.hic.nihr.ac.uk/ (8 May 2017).
15. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *JAMA* 2014;**311**:2479–2480.
16. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**:e1001779.
17. Anker S, Asselbergs FW, Brobert G, Vardas P, Grobbee DE, Cronin M. Big data in cardiovascular disease. *Eur Heart J* 2017;**38**:1863–1865.
18. Big Data for Better Hearts [Internet]. http://www.bigdata-heart.eu/ (13 July 2017).
19. One Brave Idea. http://www.onebraveidea.com/ (11 August 2017).
20. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. *N Engl J Med* 2007;**357**:370–379.
21. Barabási A-L. Network medicine—from obesity to the "Diseasome". *N Engl J Med* 2007;**357**:404–407.
22. Walsh JA, Topol EJ, Steinhubl SR. Novel wireless devices for cardiac monitoring. *Circulation* 2014;**130**:573–581.
23. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, Jha S, Agrawal M, Dziurzynski LA, Sap M, Weeg C, Larson EE, Ungar LH, Seligman MEP. Psychological language on Twitter predicts county-level heart disease mortality. *Psychol Sci* 2015;**26**:159–169.
24. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. *Science* 2014;**343**:1203–1205.
25. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, Paul DS, Freitag D, Burgess S, Danesh J, Young R, Butterworth AS. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* 2016;**32**:3207–3209.
26. PhenoScanner Home Page [Internet]. http://www.phenoscanner.medschl.cam.ac.uk/phenoscanner (13 July 2017).
27. DrugBank [Internet]. https://www.drugbank.ca/ (13 July 2017).
28. ChEMBL [Internet]. https://www.ebi.ac.uk/chembl/ (13 July 2017).
29. Hayward AC, Wang L, Goonetilleke N, Fragaszy EB, Bermingham A, Copas A, Dukes O, Millett ERC, Nazareth I, Nguyen-Van-Tam JS, Watson JM, Zambon M, Johnson AM, McMichael AJ. Natural T cell-mediated protection against seasonal and pandemic influenza. Results of the flu watch cohort study. *Am J Respir Crit Care Med* 2015;**191**:1422–1431.
30. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, Kivimaki M, Timmis AD, Smeeth L, Hemingway H. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol* 2012;**41**:1625–1638.
31. Rapsomaniki E, Stogiannis D, Chung S-C, Pujades-Rodriguez M, Shah AD, Pasea L, Denaxas S, Timmis A, Emmas C, Hemingway H. Health outcomes in patients with stable coronary artery disease following myocardial infarction; construction of a PEGASUS-TIMI-54 like population in UK linked electronic health records. Poster. *Eur Heart J* 2014;**35**:363–363.
32. Rapsomaniki E, Timmis A, George J, Pujades-Rodriguez M, Shah AD, Denaxas S, White IR, Caulfield MJ, Deanfield JE, Smeeth L, Williams B, Hingorani A, Hemingway H. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1·25 million people. *Lancet* 2014;**383**:1899–1911.
33. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, Melamed R, Rabadan R, Bernstam EV, Brunak S, Jensen LJ, Nicolae D, Shah NH, Grossman RL, Cox NJ, White KP, Rzhetsky A. A nondegenerate code of deleterious variants in mendelian loci contributes to complex disease risk. *Cell* 2013;**155**:70–80.
34. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, Morris JA. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *J Am Med Inform Assoc* 2010;**17**:652–662.
35. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li Y-C, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015;**216**:574–578.
36. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, Pushkarev D, Neff NF, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M, Sagreiya H, Whaley R, Knowles JW, Chou MF, Thakuria JV, Rosenbaum AM, Zaranek AW, Church GM, Greely HT, Quake SR, Altman RB. Clinical assessment incorporating a personal genome. *Lancet* 2010;**375**:1525–1535.
37. Genomics England [Internet]. https://www.genomicsengland.co.uk/the-100000-genomes-project/ (11 August 2017).
38. Herrett E, Bhaskaran K, Timmis A, Denaxas S, Hemingway H, Smeeth L. Association between clinical presentations before myocardial infarction and coronary mortality: a prospective population-based study using linked electronic records. *Eur Heart J* 2014;**35**:2363–2371.
39. Rubbo B, Fitzpatrick NK, Denaxas S, Daskalopoulou M, Yu N, Patel RS, Hemingway H. Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: a systematic review and recommendations. *Int J Cardiol* 2015;**187**:705–711.
40. Guimarães PO, Krishnamoorthy A, Kaltenbach LA, Anstrom KJ, Effron MB, Mark DB, McCollam PL, Davidson-Ray L, Peterson ED, Wang TY. Accuracy of medical claims for identifying cardiovascular and bleeding events after myocardial infarction. *JAMA Cardiol* 2017;**2**:750.
41. Ford I, Murray H, Packard CJ, Shepherd J, Macfarlane PW, Cobbe SM. Long-term follow-up of the West of Scotland Coronary Prevention Study. *N Engl J Med* 2007;**357**:1477–1486.
42. Rapsomaniki E, Thuresson M, Yang E, Blin P, Hunt P, Chung S-C, Stogiannis D, Pujades-Rodriguez M, Timmis A, Denaxas SC, Danchin N, Stokes M, Thomas-Delecourt F, Emmas C, Hasvold P, Jennings E, Johansson S, Cohen DJ, Jernberg T, Moore N, Janzon M, Hemingway H. Using big data from health records from four countries to evaluate chronic disease outcomes: a study in 114 364 survivors of myocardial infarction. *Eur Heart J Qual Care Clin Outcomes* 2016;**2**:172–183.
43. Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium. Interleukind-6 receptor as a target for prevention. *Lancet* 2012;**379**:1214–1224.
44. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, Basford MA, Carrell DS, Peissig PL, Kho AN, Pacheco JA, Rasmussen LV, Crosslin DR, Crane PK, Pathak J, Bielinski SJ, Pendergrass SA, Xu H, Hindorff LA, Li R, Manolio TA, Chute CG, Chisholm RL, Larson EB, Jarvik GP, Brilliant MH, McCarty CA, Kullo IJ, Haines JL, Crawford DC, Masys DR, Roden DM. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;**31**:1102–1111.
45. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, Riveros-Mckay F, Kostadima MA, Lambourne JJ, Sivapalaratnam S, Downes K, Kundu K, Bomba L, Berentsen K, Bradley JR, Daugherty LC, Delaneau O, Freson K, Garner SF, Grassi L, Guerrero J, Haimel M, Janssen-Megens EM, Kaan A, Kamat M, Kim B, Mandoli A, Marchini J, Martens JHA, Meacham S, Megy K, O'Connell J, Petersen R, Sharifi N, Sheard SM, Staley JR, Tuna S, van der Ent M, Walter K, Wang S-Y, Wheeler E, Wilder SP, Iotchkova V, Moore C, Sambrook J, Stunnenberg HG, Di Angelantonio E, Kaptoge S, Kuijpers TW, Carrillo-de-Santa-Pau E, Juan D, Rico D, Valencia A, Chen L, Ge B, Vasquez L, Kwan T, Garrido-Martín D, Watt S, Yang Y, Guigo R, Beck S, Paul DS, Pastinen T, Bujold D, Bourque G, Frontini M, Danesh J, Roberts DJ, Ouwehand WH, Butterworth AS, Soranzo N. The allelic landscape of human blood cell trait variation and links to common complex disease resource the Allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* 2016;**167**:1415–1429.
46. Nuffield Council on Bioethics. *The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues*. Nuffield Council on Bioethics; 2015.
47. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data ran into trouble. *J Med Ethics* 2015;**41**:404–409.
48. Sethi N, Laurie GT. Delivering proportionate governance in the era of eHealth: making linkage and privacy work together. *Med Law Int* 2013;**13**:168–204.
49. Global Alliance for Genomics and Health [Internet]. http://genomicsandhealth.org/ (21 March 2017).
50. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, Shah AD, Timmis AD, Schilling RJ, Hemingway H, Kiechl S. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. *PLoS One* 2014;**9**:e110900.
51. Papež V, Denaxas S, Hemingway H. Evaluating openEHR for storing computable representations of electronic health record phenotyping algorithms. In: *2017 IEEE 30th Int Symp Comput Med Syst*. 2017. p509–514.
52. Papež V, Denaxas S, Hemingway H. Evaluation of semantic web technologies for storing computable definitions of electronic health records phenotyping algorithms. In: AMIA Annual Symposium.2017. https://arxiv.org/abs/1707.07673 (11 August 2017).
53. Tapuria A, Evans M, Austin T, Lea N, Kalra D. Development and evaluation of a memory clinic information system. *Stud Health Technol Inform* 2014;**205**:106–110.
54. The House of Lords S and TC. *Genomic Medicine*. Authority of the House of Lords; 2009;I.
55. Denaxas SC, Morley KI. Big biomedical data and cardiovascular disease research: opportunities and challenges. *Eur Heart J Qual Care Clin Outcomes* 2015;**1**:9–16.

56. Winslow RL, Saltz J, Foster I, Carr JJ, Ge Y, Miller MI, Younes L, Geman D, Graniote S, Kurc T, Madduri R, Ratnanather T, Larkin J, Ardekani S, Brown T, Kolasny A, Reynolds K, Shipway M, Toerper M. The CardioVascular Research Grid (CVRG) Project. In: *Proceedings of the AMIA Summit on Translational Bioinformatics* 2011. pp. 77–81.
57. Detmer DE, Shortliffe EH. Clinical Informatics Prospects for a New Medical Subspecialty. *JAMA* 2014;**311**:2067–2068.
58. Salari K, Watkins H, Ashley EA. Personalized medicine: hope or hype? *Eur Heart J* 2012;**33**:1564–1570.
59. Hingorani AD, Windt D. A V D, Riley RD, Abrams K, Moons KGM, Steyerberg EW, Schroter S, Sauerbrei W, Altman DG, Hemingway H. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ* 2013;**346**:e5793.
60. East London Genes & Health [Internet]. http://www.genesandhealth.org/ (13 July 2017).
61. NIHR BioResource [Internet]. https://bioresource.nihr.ac.uk/(13 July 2017).
62. Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR. Health and population effects of rare gene knockouts in adult humans with related parents. *Science* 2016;**352**:474–477.
63. Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, Won H-H, Karczewski KJ, O'Donnell-Luria AH, Samocha KE, Weisburd B, Gupta N, Zaidi M, Samuel M, Imran A, Abbas S, Majeed F, Ishaq M, Akhtar S, Trindade K, Mucksavage M, Qamar N, Zaman KS, Yaqoob Z, Saghir T, Rizvi SNH, Memon A, Hayyat Mallick N, Ishaq M, Rasheed SZ, Memon F-U-R, Mahmood K, Ahmed N, Do R, Krauss RM, MacArthur DG, Gabriel S, Lander ES, Daly MJ, Frossard P, Danesh J, Rader DJ, Kathiresan S. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity A major goal of biomedicine is to understand the function of every gene in the human genome. *Nature* 2017;**544**:235–239.
64. Monaco C, Mathur A, Martin JF. What causes acute coronary syndromes? Applying Koch's postulates. *Atherosclerosis* 2005;**179**:1–15.
65. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, Bonow RO, Huang C-C, Deo RC. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation* 2015;**131**:269–279.
66. Altman RB, Ashley EA. Using "big data" to dissect clinical heterogeneity. *Circulation* 2015;**131**:232–234.
67. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søeby K, Bredkjær S, Juul A, Werge T, Jensen LJ, Brunak S, Ritchie MD. Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol* 2011;**7**:e1002141.
68. Dilsizian SE, Siegel EL. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr Cardiol Rep* 2014;**16**:441.
69. Wu P-Y, Chandramohan R, Phan JH, Mahle WT, Gaynor JW, Maher KO, Wang MD. Cardiovascular transcriptomics and epigenomics using next-generation sequencing: challenges, progress, and opportunities. *Circ Cardiovasc Genet* 2014; **7**:701–710.
70. Suinesiaputra A, Medrano-Gracia P, Cowan BR, Young AA, Medrano-Gracia P, Young AA. Big heart data: advancing health informatics through data sharing in cardiovascular imaging. *IEEE J Biomed Health Inform* 2014;**19**:1283–1290.
71. Weese J, Groth A, Nickisch H, Barschdorf H, Weber FM, Velut J, Castro M, Toumoulin C, Coatrieux JL, De Craene M, Piella G, Tobón-Gomez C, Frangi AF, Barber DC, Valverde I, Shi Y, Staicu C, Brown A, Beerbaum P, Hose DR. Generating anatomical models of the heart and the aorta from medical images for personalized physiological simulations. *Med Biol Eng Comput* 2013 ;**51**: 1209–1219.
72. Sermesant M, Chabiniok R, Chinchapatnam P, Mansi T, Billet F, Moireau P, Peyrat JM, Wong K, Relan J, Rhode K, Ginks M, Lambiase P, Delingette H, Sorine M, Rinaldi CA, Chapelle D, Razavi R, Ayache N. Patient-specific electromechanical models of the heart for the prediction of pacing acute effects in CRT: a preliminary clinical validation. *Med Image Anal* 2012;**16**:201–215.
73. Barter PJ, Caulfield M, Eriksson M, Grundy SM, Kastelein JJP, Komajda M, Lopez-Sendon J, Mosca L, Tardif J-C, Waters DD, Shear CL, Revkin JH, Buhr KA, Fisher MR, Tall AR, Brewer B. Effects of torcetrapib in patients at high risk for coronary events. *N Engl J Med* 2007;**357**:2109–2122.
74. Schwartz GG, Olsson AG, Abt M, Ballantyne CM, Barter PJ, Brumm J, Chaitman BR, Holme IM, Kallend D, Leiter LA, Leitersdorf E, McMurray JJV, Mundl H, Nicholls SJ, Shah PK, Tardif J-C, Wright RS. Effects of dalcetrapib in patients with a recent acute coronary syndrome. *N Engl J Med* 2012;**367**: 2089–2099.
75. Keene D, Price C, Shun-Shin MJ, Francis DP. Effect on cardiovascular risk of high density lipoprotein targeted drug treatments niacin, fibrates, and CETP inhibitors: meta-analysis of randomised controlled trials including 117,411 patients. *BMJ* 2014;**349**:g4379.
76. Fox K, Ford I, Ferrari R. Ivabradine in stable coronary artery disease. *N Engl J Med* 2014;**371**:2435.
77. The Stability Investigators. Darapladib for preventing ischemic events in stable coronary heart disease. *N Engl J Med* 2014;**370**:1702–1711.
78. Stitziel NO, Won H-H, Morrison AC, Peloso GM, Do R, Lange LA, Fontanillas P, Gupta N, Duga S, Goel A, Farrall M, Saleheen D, Ferrario P, König I, Asselta R, Merlini PA, Marziliano N, Notarangelo MF, Schick U, Auer P, Assimes TL, Reilly M, Wilensky R, Rader DJ, Hovingh GK, Meitinger T, Kessler T, Kastrati A, Laugwitz K-L, Siscovick D, Rotter JI, Hazen SL, Tracy R, Cresci S, Spertus J, Jackson R, Schwartz SM, Natarajan P, Crosby J, Muzny D, Ballantyne C, Rich SS, O'Donnell CJ, Abecasis G, Sunaev S, Nickerson DA, Buring JE, Ridker PM, Chasman DI, Austin E, Kullo IJ, Weeke PE, Shaffer CM, Bastarache LA, Denny JC, Roden DM, Palmer C, Deloukas P, Lin D-Y, Tang Z-Z, Erdmann J, Schunkert H, Danesh J, Marrugat J, Elosua R, Ardissino D, McPherson R, Watkins H, Reiner AP, Wilson JG, Altshuler D, Gibbs RA, Lander ES, Boerwinkle E, Gabriel S, Kathiresan S. Inactivating mutations in NPC1L1 and protection from coronary heart disease. *N Engl J Med* 2014;**371**:2072–2082.
79. Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* 2006;**354**:1264–1272.
80. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med* 2015;**12**:e1001885.
81. den Hoed M, Eijgelsheim M, Esko T, Brundel BJJM, Peal DS, Evans DM, Nolte IM, Segrè AV, Holm H, Handsaker RE, Westra H-J, Johnson T, Isaacs A, Yang J, Lundby A, Zhao JH, Kim YJ, Go MJ, Almgren P, Bochud M, Boucher G, Cornelis MC, Gudbjartsson D, Hadley D, van der Harst P, Hayward C, den Heijer M, Igl W, Jackson AU, Kutalik Z, Luan J, Kemp JP, Kristiansson K, Ladenvall C, Lorentzon M, Montasser ME, Njajou OT, O'Reilly PF, Padmanabhan S, St Pourcain B, Rankinen T, Salo P, Tanaka T, Timpson NJ, Vitart V, Waite L, Wheeler W, Zhang W, Draisma HHM, Feitosa MF, Kerr KF, Lind PA, Mihailov E, Onland-Moret NC, Song C, Weedon MN, Xie W, Yengo L, Absher D, Albert CM, Alonso A, Arking DE, de Bakker PIW, Balkau B, Barlassina C, Benaglio P, Bis JC, Bouatia-Naji N, Brage S, Chanock SJ, Chines PS, Chung M, Darbar D, Dina C, Dörr M, Elliott P, Felix SB, Fischer K, Fuchsberger C, de Geus EJC, Goyette P, Gudnason V, Harris TB, Hartikainen A-L, Havulinna AS, Heckbert SR, Hicks AA, Hofman A, Holewijn S, Hoogstra-Berends F, Hottenga J-J, Jensen MK, Johansson A, Junttila J, Kääb S, Kanon B, Ketkar S, Khaw K-T, Knowles JW, Kooner AS, Kors JA, Kumari M, Milani L, Laiho P, Lakatta EG, Langenberg C, Leusink M, Liu Y, Luben RN, Lunetta KL, Lynch SN, Markus MRP, Marques-Vidal P, Mateo Leach I, McArdle WL, McCarroll SA, Medland SE, Miller KA, Montgomery GW, Morrison AC, Müller-Nurasyid M, Navarro P, Nelis M, O'Connell JR, O'Donnell CJ, Ong KK, Newman AB, Peters A, Polasek O, Pouta A, Pramstaller PP, Psaty BM, Rao DC, Ring SM, Rossin EJ, Rudan D, Sanna S, Scott RA, Sehmi JS, Sharp S, Shin JT, Singleton AB, Smith AV, Soranzo N, Spector TD, Stewart C, Stringham HM, Tarasov KV, Uitterlinden AG, Vandenput L, Hwang S-J, Whitfield JB, Wijmenga C, Wild SH, Willemsen G, Wilson JF, Witteman JCM, Wong A, Wong Q, Jamshidi Y, Zitting P, Boer JMA, Boomsma DI, Borecki IB, van Duijn CM, Ekelund U, Forouhi NG, Froguel P, Hingorani A, Ingelsson E, Kivimaki M, Kronmal RA, Kuh D, Lind L, Martin NG, Oostra BA, Pedersen NL, Quertermous T, Rotter JI, van der Schouw YT, Verschuren WMM, Walker M, Albanes D, Arnar DO, Assimes TL, Bandinelli S, Boehnke M, de Boer RA, Bouchard C, Caulfield WLM, Chambers JC, Curhan G, Cusi D, Eriksson J, Ferrucci L, van Gilst WH, Glorioso N, de Graaf J, Groop L, Gyllensten U, Hsueh W-C, Hu FB, Huikuri HV, Hunter DJ, Iribarren C, Isomaa B, Jarvelin M-R, Jula A, Kähönen M, Kiemeney LA, van der Klauw MM, Kooner JS, Kraft P, Iacoviello L, Lehtimäki T, Lokki M-LL, Mitchell BD, Navis G, Nieminen MS, Ohlsson C, Poulter NR, Qi L, Raitakari OT, Rimm EB, Rioux JD, Rizzi F, Rudan I, Salomaa V, Sever PS, Shields DC, Shuldiner AR, Sinisalo J, Stanton AV, Stolk RP, Strachan DP, Tardif J-C, Thorsteinsdottir U, Tuomilehto J, van Veldhuisen DJ, Virtamo J, Viikari J, Vollenweider P, Waeber G, Widen E, Cho YS, Olsen JV, Visscher PM, Willer C, Franke L, Erdmann J, Thompson JR, Pfeufer A, Sotoodehnia N, Newton-Cheh C, Ellinor PT, Stricker BHC, Metspalu A, Perola M, Beckmann JS, Smith GD, Stefansson K, Wareham NJ, Munroe PB, Sibon OCM, Milan DJ, Snieder H, Samani NJ, Loos RJF. Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat Genet* 2013;**45**:621–631.
82. Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, Hindy G, Hólm H, Ding EL, Johnson T, Schunkert H, Samani NJ, Clarke R, Hopewell JC, Thompson JF, Li M, Thorleifsson G, Newton-Cheh C, Musunuru K, Pirruccello JP, Saleheen D, Chen L, Stewart AFR, Schillert A, Thorsteinsdottir U, Thorgeirsson G, Anand S, Engert JC, Morgan T, Spertus J, Stoll M, Berger K, Martinelli N, Girelli D, McKeown PP, Patterson CC, Epstein SE, Devaney J, Burnett M-S, Mooser V, Ripatti S, Surakka I, Nieminen MS, Sinisalo J, Lokki M-L, Perola M, Havulinna A, de Faire U, Gigante B, Ingelsson E, Zeller T, Wild P, de Bakker PIW, Klungel OH, Maitland-van der Zee A-H, Peters BJM, de Boer A, Grobbee DE, Kamphuisen PW, Deneer VHM, Elbers CC, Onland-Moret NC, Hofker MH, Wijmenga C, Verschuren WMM, Boer JMA, van der Schouw YT, Rasheed A, Frossard P, Demissie S, Willer C, Do R, Ordovas JM, Abecasis GR,

Boehnke M, Mohlke KL, Daly MJ, Guiducci C, Burtt NP, Surti A, Gonzalez E, Purcell S, Gabriel S, Marrugat J, Peden J, Erdmann J, Diemert P, Willenborg C, König IR, Fischer M, Hengstenberg C, Ziegler A, Buysschaert I, Lambrechts D, Van de Werf F, Fox KA, El Mokhtari NE, Rubin D, Schrezenmeir J, Schreiber S, Schäfer A, Danesh J, Blankenberg S, Roberts R, McPherson R, Watkins H, Hall AS, Overvad K, Rimm E, Boerwinkle E, Tybjaerg-Hansen A, Cupples LA, Reilly MP, Melander O, Mannucci PM, Ardissino D, Siscovick D, Elosua R, Stefansson K, O'Donnell CJ, Salomaa V, Rader DJ, Peltonen L, Schwartz SM, Altshuler D, Kathiresan S. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* 2012;**380**:572–580.

83. Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbring SJ, Lin SM. Opportunities for drug repositioning from phenome-wide association studies. *Nat Biotechnol* 2015;**33**:342–345.
84. Ritchie MD, Denny JC, Zuvich RL, Crawford DC, Schildcrout JS, Bastarache L, Ramirez AH, Mosley JD, Pulley JM, Basford MA, Bradford Y, Rasmussen LV, Pathak J, Chute CG, Kullo IJ, McCarty CA, Chisholm RL, Kho AN, Carlson CS, Larson EB, Jarvik GP, Sotoodehnia N, Manolio TA, Li R, Masys DR, Haines JL, Roden DM. Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 2013;**127**:1377–1385.
85. Finan C, Gaulton A, Kruger F, Lumbers T, Shah T, Engmann J. The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* 2017;**9**:eaag1166.
86. Shah AD, Langenberg C, Rapsomaniki E, Denaxas S, Pujades-Rodriguez M, Gale CP, Deanfield J, Smeeth L, Timmis A, Hemingway H. Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1·9 million people. *Lancet Diabetes Endocrinol* 2015;**3**:105–113.
87. George J, Rapsomaniki E, Pujades-Rodriguez M, Shah AD, Denaxas S, Herrett E, Smeeth L, Timmis A, Hemingway H. How Does cardiovascular disease first present in women and men? Incidence of 12 cardiovascular diseases in a contemporary cohort of 1,937,360 people. *Circulation* 2015;**132**:1320–1328.
88. Pujades-Rodriguez M, Timmis A, Stogiannis D, Rapsomaniki E, Denaxas S, Shah A, Feder G, Kivimaki M, Hemingway H. Socioeconomic deprivation and the incidence of 12 cardiovascular diseases in 1.9 million women and men: implications for risk prediction and prevention. *PLoS One* 2014;**9**:e104671.
89. Pujades-Rodriguez M, George J, Shah AD, Rapsomaniki E, Denaxas S, West R, Smeeth L, Timmis A, Hemingway H. Heterogeneous associations between smoking and a wide range of initial presentations of cardiovascular disease in 1 937 360 people in England: lifetime risks and implications for risk prediction. *Int J Epidemiol* 2015;**44**:129–141.
90. Group TWOSCPS. Computerised record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. The West of Scotland Coronary Prevention Study Group. *J Clin Epidemiol* 1995;**48**:1441–1452.
91. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, Bauck AE, Cifelli D, Smerek MM, Dickerson J, Laws RL, Madigan RA, Rusincovitch SA, Kluchar C, Califf RM. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc* 2013;**20**:e226–e231.
92. Doods J, Botteri F, Dugas M, Fritz F, Ehr4cr WP7. A European inventory of common electronic health record data elements for clinical trial feasibility. *Trials* 2014;**15**:18.
93. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, Dugas M, Dupont D, Schmidt A, Singleton P, De Moor G, Kalra D. Electronic health records: new opportunities for clinical research. *J Intern Med* 2013;**274**:547–560.
94. Ethier J-F, Curcin V, Barton A, McGilchrist MM, Bastiaens H, Andreasson A, Rossiter J, Zhao L, Arvanitis TN, Taweel A, Delaney BC, Burgun A. Clinical data integration model. Core interoperability ontology for research using primary care data. *Methods Inf Med* 2015;**54**:16–23.
95. New JP, Bakerly ND, Leather D, Woodcock A. Obtaining real-world evidence: the Salford Lung Study. *Thorax* 2014;**69**:1152–1154.
96. Elkhenini HF, Davis KJ, Stein ND, New JP, Delderfield MR, Gibson M, Vestbo J, Woodcock A, Bakerly ND. Using an electronic medical record (EMR) to conduct clinical trials: Salford Lung Study feasibility. *BMC Med Inform Decis Mak* 2015;**15**:8.
97. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT, Stacey SN, Frigge ML, Holm H, Saemundsdottir J, Helgadottir HT, Johannsdottir H, Sigfusson G, Thorgeirsson G, Sverrisson JT, Gretarsdottir S, Walters GB, Rafnar T, Thjodleifsson B, Bjornsson ES, Olafsson S, Thorarinsdottir H, Steingrimsdottir T, Gudmundsdottir TS, Theodors A, Jonasson JG, Sigurdsson A, Bjornsdottir G, Jonsson JJ, Thorarensen O, Ludvigsson P, Gudbjartsson H, Eyjolfsson GI, Sigurdardottir O, Olafsson I, Arnar DO, Magnusson OT, Kong A, Masson G, Thorsteinsdottir U, Helgason A, Sulem P, Stefansson K. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 2015;**47**:435–444.
98. Roden DM, Wilke RA, Kroemer HK, Stein CM. Pharmacogenomics: the genetics of variable drug responses. *Circulation* 2011;**123**:1661–1670.
99. Van Driest SL, Shi Y, Bowton EA, Schildcrout JS, Peterson JF, Pulley J, Denny JC, Roden DM. Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing. *Clin Pharmacol Ther* 2014;**95**:423–431.
100. Pasea L, Chung S-C, Pujades-Rodriguez M, Moayyeri A, Denaxas S, Fox KAA, Wallentin L, Pocock SJ, Timmis A, Banerjee A, Patel R, Hemingway H. Personalising the decision for prolonged dual antiplated therapy. *Eur Heart J* 2017;**38**:1048–1055.
101. Rassi A, Rassi A, Little WC, Xavier SS, Rassi SG, Rassi AG, Rassi GG, Hasslocher-Moreno A, Sousa AS, Scanavacca MI. Development and validation of a risk score for predicting death in Chagas' heart disease. *N Engl J Med* 2006;**355**:799–808.
102. Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Med Care* 2013;**51**:251–258.
103. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;**10**:e1001381.
104. Hemingway H, Feder G, Fitzpatrick N, Denaxas S, Shah A, Timmis A. Using nationwide 'big data' from linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the ClinicAI disease research using LInked Bespoke studies and Electronic health Records (CALIBER) programme. *Program Grants Appl Res* 2017;**5**:1–330.
105. Mora A, Sicari R, Cortigiani L, Carpeggiani C, Picano E, Capobianco E. Prognostic models in coronary artery disease: cox and network approaches. *R Soc Open Sci* 2015;**2**:140270.
106. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Syst Appl* 2008;**34**:366–368.
107. Oztekin A, Delen D, Kong Z. Predicting the graft survival for heart-lung transplantation patients: an integrated data mining methodology. *Int J Med Inform* 2009;**78**:e84–e96.
108. Institute of Medicine. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Choice Reviews Online*. The National Academies Press; 2013.
109. Gabriel SE, Normand S-LT. Getting the methods right–the foundation of patient-centered outcomes research. *N Engl J Med* 2012;**367**:787–790.
110. Gallego B, Dunn AG, Coiera E. Role of electronic health records in comparative effectiveness research. *J Comp Eff Res* 2013;**2**:529–532.
111. Google: 100,000 lives a year lost through fear of data-mining | Technology | The Guardian [Internet]. https://www.theguardian.com/technology/2014/jun/26/google-healthcare-data-mining-larry-page (11 August 2017).
112. Schneeweiss S. Learning from big health care data. *N Engl J Med* 2014;**370**:2161–2163.
113. Khoury MJ, Bradley LA. Why should genomic medicine become more evidence-based? *Genomic Med* 2007;**1**:91–93.
114. Sutton M, Nikolova S, Boaden R, Lester H, McDonald R, Roland M. Reduced mortality with hospital pay for performance in England. *N Engl J Med* 2012;**367**:1821–1828.
115. Fröbert O, James SK. Thrombus aspiration during myocardial infarction. *N Engl J Med* 2014;**370**:675–676.
116. Lagerqvist B, Fröbert O, Olivecrona GK, Gudnason T, Maeng M, Alström P, Andersson J, Calais F, Carlsson J, Collste O, Götberg M, Hårdhammar P, Ioanes D, Kallryd A, Linder R, Lundin A, Odenstedt J, Omerovic E, Puskar V, Tödt T, Zelleroth E, Östlund O, James SK. Outcomes 1 year after thrombus aspiration for myocardial infarction. *N Engl J Med* 2014;**371**:1111–1120.
117. The SCOT-HEART Investigators. CT coronary angiography in patients with suspected angina due to coronary heart disease (SCOT-HEART). *Lancet* 2015;**385**:2383–2391.
118. Optimal Aspirin Dose for Patients with Coronary Artery Disease Approved as Topic for First PCORnet Research Trial | PCORI [Internet]. http://www.pcori.org/news-release/optimal-aspirin-dose-patients-coronary-artery-disease-approved-topic-first-pcornet (21 March 2017).
119. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inf Assoc* 2014;**21**:578–582.
120. D'Avolio L, Ferguson R, Goryachev S, Woods P, Sabin T, O'Neil J. Implementation of the Department of Veterans Affairs' first point-of-care clinical trial. *J Am Med Inform Assoc* 2012;**19**:e170–e176.
121. Simms AD, Weston CF, West RM, Hall AS, Batin PD, Timmis A, Hemingway H, Fox KAA, Gale CP. Mortality and missed opportunities along the pathway of care for ST-elevation myocardial infarction: a national cohort study. *Eur Heart J Acute Cardiovasc Care* 2015;**4**:241–253.
122. Herrett E, George J, Denaxas S, Bhaskaran K, Timmis A, Hemingway H, Smeeth L. Type and timing of heralding in ST-elevation and non-ST-elevation myocardial infarction: an analysis of prospectively collected electronic healthcare records

linked to the national registry of acute coronary syndromes. *Eur Heart J Acute Cardiovasc Care* 2013;**2**:235–245.
123. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;**335**:136.
124. Rosser D, Cowley NJ, Ray D, Nightingale PG, Jones T, Moore J, Coleman JJ. Quality improvement programme, focusing on error reduction: a single center naturalistic study. *JRSM Short Rep* 2012;**3**:36.
125. Brown B, Williams R, Ainsworth J, Buchan I. Missed opportunities mapping: computable healthcare quality improvement. In: *Studies in Health Technology and Informatics*. IOS Press; 2013. p387–391.
126. Shah NH. Mining the ultimate phenome repository. *Nat Biotechnol* 2013;**31**:1095.
127. Holmes AB, Hawson A, Liu F, Friedman C, Khiabanian H, Rabadan R, Rzhetsky A. Discovering disease associations by integrating electronic clinical data and medical literature. *PLoS One* 2011;**6**:e21132.
128. Longhurst CA, Harrington RA, Shah NH. A "green button" for using aggregate patient data at the point of care. *Health Aff (Millwood)* 2014;**33**:1229–1235.
129. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. *N Engl J Med* 2011;**365**:1758–1759.
130. van Staa T-P, Dyson L, McCann G, Padmanabhan S, Belatri R, Goldacre B, Cassell J, Pirmohamed M, Torgerson D, Ronaldson S, Adamson J, Taweel A, Delaney B, Mahmood S, Baracaia S, Round T, Fox R, Hunter T, Gulliford M, Smeeth L. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. *Health Technol Assess* 2014;**18**:1.
131. Dregan A, van Staa T, Mcdermott L, McCann G, Ashworth M, Charlton J. Cluster randomized trial in the general practice research database: 2. Secondary prevention after first stroke (eCRT study): study protocol for a randomized controlled trial. *Trials* 2012;**13**:181.
132. Psaty BM, Breckenridge AM. Mini-sentinel and regulatory science - big data rendered fit and functional. *N Engl J Med* 2014;**370**:2165–2167.
133. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, Nelson JC, Racoosin JA, Robb M, Schneeweiss S, Toh S, Weiner MG. The U.S. Food and Drug Administration's mini-sentinel program: status and direction. *Pharmacoepidemiol Drug Saf* 2012;**21**:1.
134. Coloma PM, Schuemie MJ, Trifirò G, Gini R, Herings R, Hippisley-Cox J, Mazzaglia G, Giaquinto C, Corrao G, Pedersen L, van der Lei J, Sturkenboom M. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf* 2011;**20**:1–11.
135. Hunter A. The innovative medicines initiative: a pre-competitive initiative to enhance the biomedical science base of Europe to expedite the development of new medicines for patients. *Drug Discov Today* 2008;**13**:371–373.
136. Wang Y, Eldridge N, Metersky ML, Verzier NR, Meehan TP, Pandolfi MM, Foody JAnne. M, Ho S-Y, Galusha D, Kliman RE, Sonnenfeld N, Krumholz HM, Battles J. National trends in patient safety for four common conditions, 2005–2011. *N Engl J Med* 2014;**370**:341–351.
137. Chung S-C, Gedeborg R, Nicholas O, James S, Jeppsson A, Wolfe C, Heuschmann P, Wallentin L, Deanfield J, Timmis A, Jernberg T, Hemingway H. Acute myocardial infarction: a comparison of short-term survival in national outcome registries in Sweden and the UK. *Lancet* 2014;**383**:1305–1312.
138. Asaria M, Walker S, Palmer S, Gale CP, Shah AD, Abrams KR, Crowther M, Manca A, Timmis A, Hemingway H, Sculpher M. Using electronic health records to predict costs and outcomes in chronic disease using the example of stable coronary artery disease. *Heart* 2016;**102**:8.
139. The Creative Destruction of Medicine by Eric Topol [Internet]. http://creativedestructionofmedicine.com/ (11 August 2017).
140. Health eHeart study [Internet]. https://www.health-eheartstudy.org/ (11 August 2017).
141. Hindricks G, Taborsky M, Glikson M, Heinrich U, Schumacher B, Katz A, Brachmann J, Lewalter T, Goette A, Block M, Kautzner J, Sack S, Husser D, Piorkowski C, Søgaard P. Implant-based multiparameter telemonitoring of patients with heart failure (IN-TIME): a randomised controlled trial. *Lancet* 2014;**384**:583–590.
142. Free C, Knight R, Robertson S, Whittaker R, Edwards P, Zhou W, Rodgers A, Cairns J, Kenward MG, Roberts I. Smoking cessation support delivered via mobile phone text messaging (txt2stop): a single-blind, randomised trial. *Lancet* 2011;**378**:49–55.
143. HealthVault [Internet]. https://www.healthvault.com/gb/en (11 August 2017).
144. Teen Develops Computer Algorithm to Diagnose Leukemia [Internet]. http://cloud4cancer.appspot.com/ (11 August 2017).
145. Yeh RW, Sidney S, Chandra M, Sorel M, Selby JV, Go AS. Population trends in the incidence and outcomes of acute myocardial infarction. *N Engl J Med* 2010;**362**:2155–2165.
146. Hemingway H, Mccallum A, Shipley M, Manderbacka K, Martikainen P, Keskimäki I. Incidence and prognostic implications of stable angina pectoris among women and men. *JAMA* 2006;**295**:1404–1411.
147. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 2012;**380**:37–43.
148. Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. *N Engl J Med* 2008;**358**:2249–2258.
149. Jensen AB, Moseley PL, Oprea TI, Gade Ellesøe S, Eriksson R, Schmock H. ARTICLE Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun* 2014;**5**:4022.
150. Øyen N, Poulsen G, Boyd HA, Wohlfahrt J, Jensen PKA, Melbye M. Recurrence of congenital heart defects in families. *Circulation* 2009;**120**:295–301.
151. Pujades-Rodriguez M, Guttmann OP, Gonzalez-Izquierdo A, Duyx B, O'Mahony C, Elliott P, Hemingway H. Prognosis of patients with hypertrophic cardiomiopathy: a contemporary population record linkage cohort in England. *European Heart Journal* 2016;**37**(Abstract Suppl),162.
152. Smeeth L, Thomas S, Hall A, Hubbard R, Farrington P, Vallance P. Risk of myocardial infarction and stroke after acute infection or vaccination. *N Engl J Med* 2004;**351**:2611–2618.
153. Met Office Hadley Centre for Climate Science and Services [Internet]. *Met Office, FitzRoy Road, Exeter, Devon, EX1 3PB, United Kingdom*. http://www.metoffice.gov.uk/publicsector/climate-programme (26 August 2016).
154. Ludwig J, Sanbonmatsu L, Gennetian L, Adam E, Duncan GJ, Katz LF, Kessler RC, Kling JR, Lindau ST, Whitaker RC, McDade TW. Neighborhoods, obesity, and diabetes—a randomized social experiment. *N Engl J Med* 2011;**365**:1509–1519.
155. Wyber R, Vaillancourt S, Perry W, Mannava P, Folaranmi T, Celi LA. Big data in global health: improving health in low- and middle-income countries. *Bull World Heal Organ* 2015;**93**:203–208.
156. Adler NE, Stead WW. Patients in context—EHR capture of social and behavioral determinants of health. *N Engl J Med* 2015;**372**:698–701.
157. Bozio A, Crawford R, Emmerson C, Tetlow G. *Retirement Outcomes and Lifetime Earnings: Descriptive Evidence from Linked ELSA—NI Data*. 2010.
158. Administrative Data Research Network (ADRN) [Internet]. https://adrn.ac.uk/ (11 August 2017).
159. Ford ES, Ajani UA, Croft JB, Critchley JA, Labarthe DR, Kottke TE, Giles WH, Capewell S. Explaining the Decrease in U.S. Deaths from Coronary Disease, 1980–2000. *N Engl J Med* 2007;**356**:2388–2398.
160. Sims M, Maxwell R, Bauld L, Gilmore A. Short term impact of smoke-free legislation in England: retrospective analysis of hospital admissions for myocardial infarction. *BMJ* 2010;**340**:c2161.
161. Juster HR, Loomis BR, Hinman TM, Farrelly MC, Hyland A, Bauer UE, Birkhead GS. Declines in hospital admissions for acute myocardial infarction in New York state after implementation of a comprehensive smoking ban. *Am J Public Health* 2007;**97**:2035–2039.
162. Pell JP, Haw S, Cobbe S, Newby DE, Pell ACH, Fischbacher C, McConnachie A, Pringle S, Murdoch D, Dunn F, Oldroyd K, Macintyre P, O'Rourke B, Borland W. Smoke-free legislation and hospitalizations for acute coronary syndrome. *N Engl J Med* 2008;**359**:482–491.
163. Rasella D, Harhay MO, Pamponet ML, Aquino R, Barreto ML. Impact of primary health care on mortality from heart and cerebrovascular diseases in Brazil: a nationwide. *BMJ* 2014;**349**:g4014.
164. HIMSS Europe [Internet]. http://www.himss.eu/ (11 August 2017).
165. Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, Fetterolf SN, O'Dushlaine C, Van Hout CV, Staples J, Gonzaga-Jauregui C, Metpally R, Pendergrass SA, Giovanni MA, Kirchner HL, Balasubramanian S, Abul-Husn NS, Hartzel DN, Lavage DR, Kost KA, Packer JS, Lopez AE, Penn J, Mukherjee S, Gosalia N, Kanagaraj M, Li AH, Mitnaul LJ, Adams LJ, Person TN, Praveen K, Marcketta A, Lebo MS, Austin-Tse CA, Mason-Suares HM, Bruse S, Mellis S, Phillips R, Stahl N, Murphy A, Economides A, Skelding KA, Still CD, Elmore JR, Borecki IB, Yancopoulos GD, Davis FD, Faucett WA, Gottesman O, Ritchie MD, Shuldiner AR, Reid JG, Ledbetter DH, Baras A, Carey DJ. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR Study. *Science* 2016;**354**:aaf6814.
166. Abul-Husn NS, Manickam K, Jones LK, Wright EA, Hartzel DN, Gonzaga-Jauregui C, O'Dushlaine C, Leader JB, Lester Kirchner H, Lindbuchler DAndra. M, Barr ML, Giovanni MA, Ritchie MD, Overton JD, Reid JG, Metpally RPR, Wardeh AH, Borecki IB, Yancopoulos GD, Baras A, Shuldiner AR, Gottesman O, Ledbetter DH, Carey DJ, Dewey FE, Murray MF. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* 2016;**354**:aaf7000.