Big biomedical data and cardiovascular disease research: opportunities and challenges

Spiros C. Denaxas (1,2,*), Katherine I. Morley (1,3)

1. Farr Institute of Health Informatics Research, University College London, United Kingdom
2. Institute of Health Informatics, University College London, United Kingdom
3. National Addiction Centre, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, United Kingdom.

* Corresponding author:
Farr Institute of Health Informatics Research, University College London
222 Euston Road, London, NW1 2DA
Email: s.denaxas@ucl.ac.uk

Over few past years, "*big data"* has become a frequently used catchall phrase for research approaches involving the use of complex, large-scale data sets [1 2]. There are many types of data that may fit this description, but within the sphere of clinically-oriented research this term is often considered synonymous to Electronic Health Record (EHR) data, or Electronic Medical Record (EMR) data. The powerful potential of these data for advancing biomedical research has been recognised by many [3-5]. Funders in both the USA and the UK have recently made substantial investments in the area, such as the USD$215 million Precision Medicine Initiative [6] announced by the US government and Genomics England, aiming to sequence 100,000 whole genomes during routine clinical care[7]. Additionally, funding organizations are actively encouraging research utilizing large-scale biomedical data through specific initiatives. The Big Data To Knowledge (BD2K) programme[8] was established by the National Institutes of Health (NIH) in 2012 to address the challenges and opportunities presented by big biomedical data through the provision of seed funding for biomedical data-science based research, methods and training material development. In the UK, a consortium of 10 UK government and charity

funders, led by the Medical Research Council (MRC) have committed over £90 million across several initiatives that are aimed at supporting translational research using big data such as the national Farr Institute of Health Informatics Research[9], the UK Health Informatics Research Network (UKHIRN) and the Medical Bioinformatics Initiative. The amount of EHR data being digitally generated and collected is vast and rapidly expanding, and presents multiple opportunities that have the potential to transform medical practice and cardiovascular research across all stages of translation.

However, big data is not a panacea for all research problems, and for many researchers the path from big data to clinical impact for a specific research question is unclear. There are many factors that must be considered when planning to use EHR data for research related not just to ethical and policy issues raised by combining data sources [10 11] but also the logistical and analytical decisions the process entails. One of the major impediments to the use of EHRs for research is that is the data they contain differs from data collected in a conventional cohort study or randomised controlled trial (RCT) in terms of both why and how it is recorded, and requires substantial processing before it can be statistically analysed. These data are generated and recorded throughout the patient pathway during interactions with primary, secondary and tertiary healthcare providers. Data from specialised disease registries, which were originally set up for auditing clinical standards and benchmarking quality improvement initiatives, may also be incorporated. These different sources also record information in different ways. EHR data can be structured (e.g. diagnosis recorded using medical classification systems such as the International Classification of Diseases 10th revision (ICD-10) [12] or SNOMED-CT[13]) or unstructured (e.g. textual narrative in clinical notes or coronary angiography reports in hospital information systems[14]). EHR are also increasingly including cardiovascular imaging data from procedures such as echocardiography, angiography, magnetic resonance imaging or computed tomography[15]. For all sources of information, data collection will have been motivated by clinical care, administrative, or other reasons, and will be recorded using a variety of ways. The research-user is faced with substantial missing or incomplete information, data collected at irregular time-points, information that may be temporally inconsistent, and potentially the task of integrating and harmonising information contained in multiple sources.

These challenges are not insurmountable and do not mean that EHR data cannot be widely used for research, but do require a clear identification of research areas that can best leverage EHR data, and the development of tools that smooth the path from research question to research result.

Research opportunities well-placed to leverage EHR data

High-resolution observational cohort studies

Linkage of multiple EHR data sources permits the creation of large-scale cohorts of patients for whom extensive follow-up data is already available. This allows researchers to answer questions that reliance upon traditional investigator-led cohort studies would otherwise make impossible due to the scale, diagnostic resolution, timeframe, or cost. In addition, it allows researchers to define and examine the entire patient journey, from early presentations of non-acute manifestations through the various syndrome transitions to cardiac (or non-cardiac) death. This enables them to resolve the time sequence, examine and understand, the aetiological and prognostic differences between different coronary disease phenotypes[18].

Chung *et al*. [19] were able to take advantage of available EHR data in this way to conduct a comparative effectiveness study of acute coronary care on an international scale. Currently, Sweden and the UK are the only two countries in the world with ongoing, national registries for acute coronary syndrome events that cover all hospital care. Using these data, the authors showed that 30-day mortality following acute myocardial infarction was substantially higher in the UK, and that uptake of effective treatment was slower in the UK. The richness of the data meant that a substantial amount of clinical information could be incorporated into the casemix, including demography, risk factor comorbidity, and pre-hospital treatment. The researchers were also able to determine that diagnoses made in the two countries were comparable by examining troponin values and propensity to make a diagnosis. The results from this study are thus more robust than those based on a simple comparison of mortality rates, or focused on data from bespoke studies undertaken in hospitals that may not be representative of the broader healthcare system.

EHR cohorts can also be used to make timely contributions to debates of clinical importance, such as the controversy over the relationship between varenicline

and adverse cardiovascular events. In 2011 a meta-analysis of 14 RCTs raised concerns that use of varenicline for smoking cessation may increase risk for adverse cardiovascular events (ischemia, arrhythmia, congestive heart failure, sudden death or cardiovascular-related death)[20]. Three subsequent meta-analyses of RCTs did not find a significant association[21-23]. However, the question remains controversial, partly due to disagreements over analytical methods used in these studies, but also because meta-analyses are limited to the analysis of existing studies[22 24 25]. Svanström and colleagues were able to rapidly contribute new data to the debate by investigating the question in a cohort made up of the EHR data of over 35,000 Danish individuals who used either varenicline or bupropion for smoking cessation [26]. In this observational study, published in 2012, there was no evidence for a higher number of adverse events in patients using varenicline (acute coronary syndrome, ischaemic stroke, and cardiovascular death). It would not have been feasible to take a comparable traditional cohort study from study design to publication within a similar timeframe, especially as very large number of patients would be required to ensure sufficient outcome numbers (only 117 were observed amongst the 35,000 patients in the EHR study).

The capacity to investigate novel research questions has generally been limited by available data and funding for obtaining new data, but EHR data can potentially be used to address this problem. The relationship between auto-immune inflammatory conditions and atrial fibrillation is one example where EHR data have been able to fill a research niche. Although there is substantial research interest in this area [27], many of the large cardiovascular cohort studies (e.g. Framingham[28]) have limited data available on inflammatory conditions as this was not part of the original study design. However, researchers in the UK, USA, and Denmark have been able to use EHR resources to explore this research area using very large samples, finding associations with increased risk of AF and a range of conditions including rheumatoid arthritis and psoriasis[29-32]. Other researchers have taken an even broader, non-hypothesis-driven approach, using advanced computation techniques that consider any and all disease information available in EHR data to identify novel associations between diseases [33]. The costs associated with using EHR data for these studies would have been much lower than comparable data-collection, making them a cost-effective entry point into new areas of cardiovascular research.

Enhanced clinical trials

There is growing concern that current model of discovering new interventions, evaluating them through clinical trials and implementing the findings as part of clinical care is significantly inefficient. The translation process itself is taking too long, with an average figure of 17 years reported in some cases[34]. Additionally, the number of new drugs introduced to the market pet year has been broadly flat since the 1950s yet the costs have steadily grown[35] and the cost of bringing a new licensed drug to the market has been estimated between 5 and 11 billion USD$[36].

In cardiovascular diseases, the problem is more acutely manifested through problems observed in the current clinical trials pipeline. There is a lack of contemporary and representative population data that can be utilized to draw accurate estimates of events and inform the selection of appropriate primary and secondary endpoints for clinical trial. Clinical trials are often conducted in highly selected populations that are not necessarily representative of the populations presented in routine clinical care and as such, results obtained have limited generalizability and external validity[37]. For example, the clinical characteristics, treatments and inpatient outcomes of patients enrolled in a large trial of acute heart failure (Acute Study of Nesiritide in Decompensated Heart Failure) were found to be significantly different that those found in a contemporary disease registry[38]. Furthermore, despite their growing importance in CVD research, non-drug interventions such as interventions based on clinical algorithms and decision support tools are not systematically evaluated through clinical trials since the process of randomization and outcome ascertainment is not seamlessly integrated into the clinical care pathway.

This has had a significant negative impact on clinical trial conduct and findings. For example, recently there have been several late drug failures occurring within phase III clinical trials of therapeutic agents each costing several hundred million USD$. High-Density Lipoprotein Cholesterol (HDL-C) raising agents such as niacin, fibrates and cholesteryl ester transfer protein (CETP) failed to reduce all cause mortality, coronary heart disease mortality and myocardial infarction event rates in patients treated with statins[39]. Likewise, heart rate lowering agents such as ivabradine when introduced to patients with stable coronary artery disease without clinical heart

failure failed to improve cardiovascular mortality and non-fatal myocardial infarctions rates [40].

There is growing optimism that EHR can enrich RCT design, delivery and follow up.. EHR data can offer real-world phenotype-rich data that can directly inform trial design, enable the identification of optimal target populations and offer accurate event rate estimates similar to those encountered in clinical care. The entire trial conduct pipeline, from recruitment at the point of care to randomization and adverse event capture can be integrated with routine clinical care enabling the cost-effective and efficient trialling of non-drug interventions. Additionally, EHR can provide richer contemporary data on trial participants at a fraction of the cost thus enabling the generalization of trial results to external populations [41]. For example, the Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction (TASTE) trial[42] for assessing the clinical effect of routine intracoronary thrombus aspiration before primary percutaneous coronary intervention in patients with ST-segment elevation myocardial infarction recruited patients by enrolling patients though the Swedish Coronary Angiography and Angioplasty Registry and utilizing national EHR and registry data for defining trial endpoints. Finally, EHR data provide valid, complete, long-term follow-up of phase III trials that would otherwise be too costly and complex to establish and too narrow in focus [43]. While EHR offer a rich data-scaffolding for designing and implementing clinical trials, significant challenges still exist, mainly around information governance and recruitment of clinicians as outlined in the evaluation by van Staa and colleagues[44 45].

Challenges in the pathway from EHR data to research results

Although the benefits of using EHR data for research are potentially large, the widespread use of EHR data is hampered by the fact that there are currently a number of additional steps, and many associated queries, in the pathway from research question to results and publication. As an example, consider a research project using existing data to investigate whether there is a relationship between gender and onset of atrial fibrillation (AF). Most projects would involve applying standard analytical techniques to a bespoke investigator-led cohort of healthy individuals followed-up for cardiovascular conditions including AF (e.g. The Framingham Heart Study). For an

existing data set, only relatively minimal data preparation would be required before analyses could be conducted and data are often provided with detailed documentation. However, using existing EHR data to answer the same question would require a number of additional preparatory steps before statistical analyses could be conducted. Broadly speaking, these relate to: (i) identifying the EHR source(s) that contain the data needed for the research question; (ii) developing strategies for extracting the required information from the data source(s), and combining it where necessary; (iii) creating a data set that is ready for analysis using standard statistical techniques (see Figure 1).

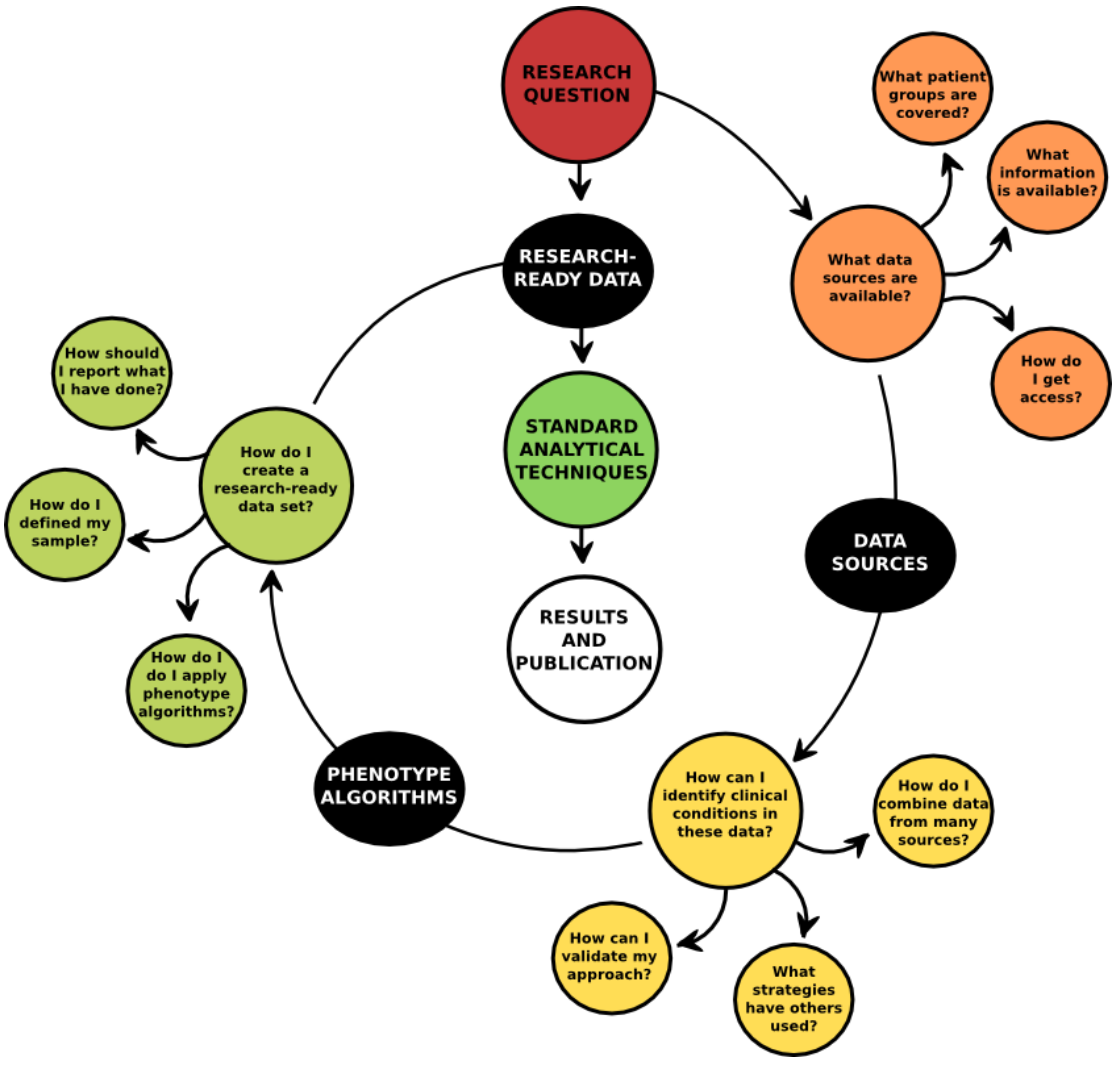

Figure 1: Diagram of steps from research question to results and publication. The four central circles show the path from research question to results for a conventional study using existing data. Circles on outside of the spiral indicate the additional steps needed to conduct a research project using electronic health record data.

What EHR data sources are available?

The availability of diverse data sources, including EHR, is rapidly expanding, making the identification of relevant sources for a single project overwhelming. Selecting appropriate data sources for research is dependent upon knowledge of the patients included (e.g. in-patients, ambulatory care, specialist treatment), the types of data recorded (e.g. diagnosis, prescriptions, test results, procedures), and the format of those data (e.g. diagnostic codes, imaging, free text), but often much of this can be difficult to determine in detail until data access has been granted. A recent Wellcome Trust report on the discoverability of EHR and other biomedical datasets for research[46] found that for the vast majority of sources, no systematic method is used to capture, curate, and display information about the data contained in each source, or to provide guidance on the information governance restrictions attached to them which determine how they can be accessed and used for research. The limited use of standardised methods (e.g. metadata) for describing such information hinders recognition of the limitations and opportunities these data sources present, and potentially results in under-utilisation of data sources due to lack of knowledge about what they contain.

However, overcoming this challenge is worth the additional effort, as combining data from multiple sources strengthens EHR-based cardiovascular research. For example, Herrett *et al.* explored the completeness of recording for acute myocardial infarction (AMI) in four EHR sources: primary care (Clinical Practice Research Datalink; CPRD), hospital admissions (Hospital Episode Statistics; HES); a MI disease registry (Myocardial Ischaemia National Audit Project; MINAP), and national mortality data (Office of National Statistics; ONS). Compared to the disease registry, which was treated as the gold standard data source, none of the other data sources captured all MI events and consequently incidence rates based on data from a single source were underestimated by 25-50% [47]. This finding is not limited to AMI; a similar investigation of AF diagnoses found that only about 40% of the 72,793 AF patients identified had a diagnosis recorded in both primary and secondary care [48].

Thus, for our example research question regarding gender and AF, we would likely decide to combine multiple EHR data sources, such as CPRD, HES, and ONS. This would enable us to use a sample of individuals broadly representative of the UK general population, and would include a more representative set of AF cases as

diagnoses made in both primary and secondary care would be identified. However, individual access applications would need to be made for each EHR source prior to linkage of the different data sources, and information about what is contained within each would currently be limited to knowledge of the clinical coding systems used.

How can I define clinical conditions in EHR data?

Once relevant the data source(s) have been identified, researchers face another challenge: how to determine which patients have been diagnosed with a particular condition. Extracting phenotypic information (i.e. disease status), a process known as *phenotyping*, is a time-consuming and challenging task even in relation to a single data source, as multiple diagnosis codes may be used to describe similar or related conditions and their data. This challenge is amplified when data from multiple sources, recorded using different coding systems, are combined. Figure 2 illustrates this, using as an example data for one individual from the three EHR sources in our hypothetical research question. In this example, an AF diagnosis is recorded at three different time-points: as a secondary diagnosis during a hospital admission, in the primary care record after hospital admission information is transferred to their GP, and as a primary diagnosis when the patient is admitted to hospital for an AF-related surgical procedure. This information needs to be reconciled in order to determine not only if, but also when, a diagnosis occurred.

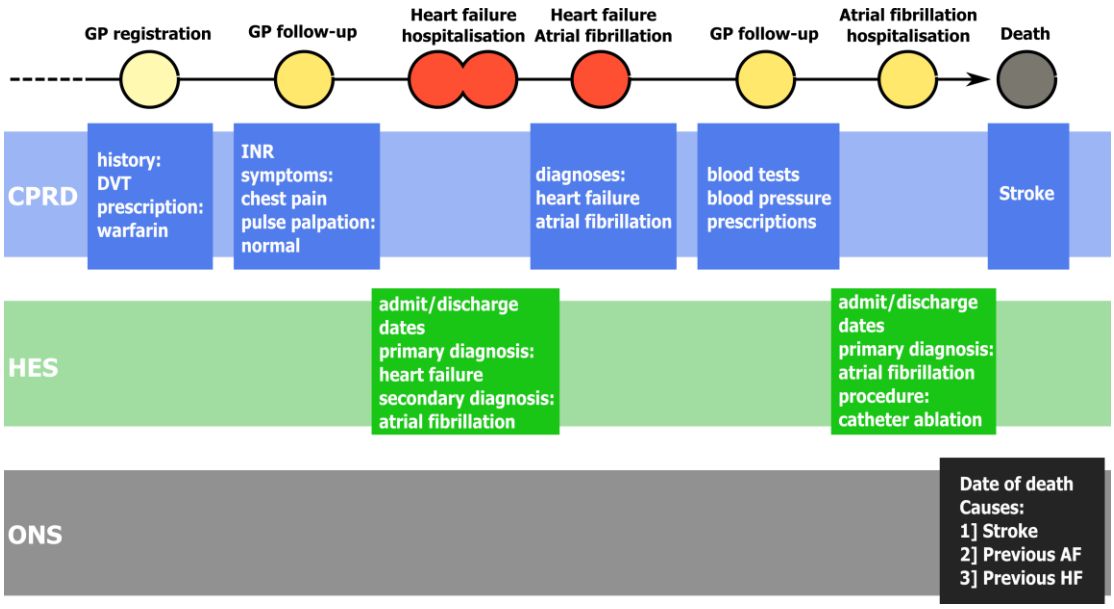

Figure 2: Illustration of linked primary care data (Clinical Practice Research Datalink; CPRD), secondary care data (Hospital Episode Statistics; HES), and mortality data (Office of National Statistics; ONS) for a single patient. Circles on the top line show events recorded in one or more sources; red circles indicate a

diagnosis. DVT indicates deep vein thrombosis; INR indicates International Normalisation Ratio; AF indicates atrial fibrillation; HF indicates heart failure.

Reconciling coded information from multiple sources is made more challenging by the different medical classification systems that are used by each source. For example, in the UK, primary care sources use Read codes, a subset of the Systematic Nomenclature of Medicine – Clinical Terms (SNOMED-CT) clinical terminology, whereas secondary care and mortality sources use the International Classification of Disease – 10th Revision (ICD-10). Combining data recorded using these systems for a single condition, such as AF, is not straightforward as the clinical resolution they offer can vary substantially; there are 23 Read codes relating to AF, including disease subtype classification, but only one ICD-10 code. Data-driven computational methodologies, such as support vector machines (SVM), can be applied on unstructured data (e.g. clinical text, electrocardiographic (ECG) monitoring data) to further enhance and fine-tune the accuracy of algorithms utilizing coded data [4 49 50]. For example, Mohebbi *et al.* created an algorithm which consistsed of a linear discriminant analysis based feature reduction scheme and a SVM-based classifier and were able to accurately (sensitivity 99.07%, specificity 100%, positive predictive value 100%) detect AF cases using RR intervals extracted from ECG signals[51].

No standardised methodologies and mechanisms exist to help research-users define, share and evaluate EHR-derived phenotypes in a consistent way, or to apply algorithms for creating these phenotypes to their own data, although development of tools for this is very active [52-54]. The USA-based eMERGE Consortium have developed an AF phenotype algorithm [55] which focuses on clinical notes and electrocardiogram impression data. These data are not available in CPRD, HES, or ONS, although there is a UK-oriented EHR phenotype resource called CALIBER that does contain an AF phenotype based on coded data from primary and secondary care [48], which could be applied in this situation. However, if no phenotype algorithm existed, we would need to go through the process of developing a new phenotype algorithm for AF, and we would need to repeat this process for every other variable we wanted to include in our final data set such as gender and any covariates such as other cardiovascular diseases, smoking status, or hypertension.

Validation, preferably against a gold standard, is a key step of defining disease phenotyping algorithms[56]. The goal of the validation exercise is to evaluate the accuracy of the algorithm: is the phenotyping algorithm including all patients that are eligible and excluding all patients that are ineligible, thus accurately allocating them in the case and control groups. Some phenotypes, such as type 2 diabetes[57], are inherently complex as they make use of multiple data elements (e.g. diagnostic codes, medication information, laboratory measurements, clinical text) and should ideally be validated through manual review of case notes in primary or secondary healthcare providers in order to understand the information the physician had available at the time of diagnosis. Clinical notes however are not available at scale due to information governance restrictions and scaling this process for large cohorts of patients is challenging and time-consuming. An alternative approach is to validate the developed phenotyping algorithms by conducting epidemiological analyses of the association of known risk factors and the phenotype in question and compare associations found in other studies. Other phenotypes, such as white blood cell count, the goal of the validation exercise is to ensure that the algorithm included all eligible patients and discarded outliers and incorrect values.

How do I create a research-ready EHR data set?

The process of applying phenotype algorithms to raw EHR data and creating a data set that is ready to be statistically analysed requires several data transformations that are challenging due to data heterogeneity and complexity. Description of the process is rarely provided as part of academic outputs, and there is increasing recognition of the weaknesses that pervade the current landscape of EHR research in relation to sharing and standardisation of data transformation methods [58]. The prevalent scientific culture does not promote or reward sharing of standardised and re-usable data transformation libraries, which leads to substantial duplication of effort and increases the potential for a lack of reproducible results from EHR-based studies.

As for a conventional study, an EHR-based study requires a clear definition including the population from which individuals are sampled, inclusion and exclusion criteria, follow-up, and handling of missing data. For our example question, we may need to specify the age range of our patients, whether we are including individuals

with prior cardiovascular conditions such as heart failure, and how missing data were handled, but there is additional information that should be reported for EHR data including: the data sources included, the end date of our follow-up data, whether there are exclusion/inclusion criteria based on data quality or other administrative information, details of new phenotype algorithms, and how data were multiply imputed if applicable. While this information can be described to some extent in the Methods section of a scientific paper, the associated computational manipulation and analyses are not standardised for EHR data, and there is currently no provision in scientific papers for detailed explanations of these methods or distribution of associated phenotype algorithms, computer software, or scripts.

Recommendations for advancing EHR research

Many countries in Europe, and internationally, have EHR systems that could be utilised for research; national, centralised resources that facilitate the steps from research question to research data set would substantially enhance the research potential of these data sources. Initiatives are already underway to achieve this in some countries, but few tackle all aspects of this process.

The UK-based CALIBER platform [59] combines a repository of EHR phenotypes with curated record linkages combining primary care (Clinical Practice Research Datalink), hospital discharge (Hospital Episode Statistics), disease registry (Myocardial Ischaemia National Audit Project[60]) and death registry (Office of National Statistics) data in over 2 million adults with 10 million person years of follow-up. However, this resource does not provide any tools for bidirectional interactions with EHR data sources. In contrast, the Clinical Record Interactive Search (CRIS) system (based at the NIHR Mental Health Biomedical Research Centre and Dementia Unit at the South London and Maudsley NHS Foundation Trust) allows researchers to investigate anonymised secondary care data, including clinical notes and other text, via novel user-friendly tools that facilitate identification of patients meeting certain criteria and development of text-mining algorithms [61]. The Electronic Medical Records and Genomics (eMERGE) Network [52], a US National Human Genome Research Institute-funded consortium, combines a phenotype repository with

EHR data from multiple secondary healthcare providers, including imaging and text, linked to genotypic data for all participants.

National EHR portals could combine the strengths of all these projects by including: (i) a national catalogue of contemporary EHR sources curated using metadata standards; (ii) an interactive thesaurus of EHR-derived phenotype algorithms; (iii) standards-driven tools that will enable researchers to visually create observational and interventional research studies (population, inclusion/exclusion criteria, sources, phenotypes). The national catalogue should support the harvesting and integration of metadata from external sources, and manual curation by researchers within a standardised and reproducible framework, as well as providing guidance on data access and data content. This will allow users to identify data sources that can provide information both within and across disease areas. The EHR phenotype algorithms and data set creation tools need to be implemented in a fashion that supports reuse and modification by other users, as well as appropriate academic credit and/or citation. Creating this type of resource will help to foster an "open source" approach to EHR research in which researchers can collaborate and learn from each other, and this will ultimately produce a greater advance in EHR research than could be achieved by any research group in isolation.

Electronic Health Records: Electronic Health Records (EHR) are data generated and recorded during routine clinical care. EHRs are diverse and encompass nationally and regionally available structured and unstructured data from primary care, hospitals, administrative data, and disease, procedure and death registries; increasingly including genomic, imaging and patient sensor data.

Medical ontology: a structured controlled vocabulary of medical concepts and their semantic relations used to record, store and transmit medical knowledge and patient-related clinical information efficiently

Metadata: is data that describes aspects around a particular data element. For an EHR source metadata can include information about the manner in which the data get generated and recorded, the medical ontologies used to record information and the methods by which researchers can access the data for research.

Phenotyping: In the context of EHR, phenotyping is defined as the process of creating algorithms that define an observable trait (physical or biochemical) such as a clinical condition within EHR data.

Box 1. Definitions

References

1. Community cleverness required. Nature 2008;**455**(7209):1-1.
2. Challenges and Opportunities. 02/11/ 2011. http://dx.doi.org/10.1126/science.331.6018.692.
3. Weber G, Mandl K, Kohane I. Finding the Missing Link for Big Biomedical Data. JAMA 2014.
4. Jensen P, Jensen L, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;**13**(6):395-405.
5. Khoury M, Lam TK, Ioannidis J, et al. Transforming epidemiology for 21st century medicine and public health. Cancer epidemiology, biomarkers & prevention 2013;**22**(4):508-16.
6. Collins F, Varmus H. A New Initiative on Precision Medicine. N Engl J Med 2015;**372**(9):793-95.
7. 100,000 Genomes Project. http://www.genomicsengland.co.uk/.
8. Margolis R, Derr L, Dunn M, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. Journal of the American Medical Informatics Association 2014;**21**(6):957-58.
9. Farr Institute for Health Informatics Research. http://www.farrinstitute.org.
10. Richards N, King J. Big Data Ethics. Social Science Research Network Working Paper Series 2014.
11. Boyd D, Crawford K. CRITICAL QUESTIONS FOR BIG DATA. Information, Communication & Society 2012;**15**(5):662-79.
12. World Health Organization, International Classification of Diseases (ICD). http://apps.who.int/classifications/icd10/browse/2015/en.
13. Stearns MQ, Price C, Spackman KA, et al. SNOMED clinical terms: overview of the development process and project status. Proceedings of the Americal Medical Informatics Association Symposium 2001:662-66.
14. Wang Z, Shah A, Tate R, et al. Extracting Diagnoses and Investigation Results from Unstructured Text in Electronic Health Records by Semi-Supervised Machine Learning. PLoS ONE 2012;**7**(1):e30412.
15. Petersen S, Selvanayagam J, Wiesmann F, et al. Left ventricular non-compaction: insights from cardiovascular magnetic resonance imaging. Journal of the American College of Cardiology 2005;**46**(1):101-05.

16. Gymrek M, McGuire A, Golan D, et al. Identifying personal genomes by surname inference. Science (New York, NY) 2013;**339**(6117):321-24.
17. Sheather J, Brannan S. Patient confidentiality in a time of care.data. BMJ 2013;**347**:f7042.
18. Timmis A, Feder G, Hemingway H. Prognosis of stable angina pectoris: why we need larger population studies with higher endpoint resolution. Heart (British Cardiac Society) 2007;**93**(7):786-91.
19. Chung S-C, Gedeborg R, Nicholas O, et al. Acute myocardial infarction: a comparison of short-term survival in national outcome registries in Sweden and the UK. Lancet 2014;**383**(9925):1305-12.
20. Singh S, Loke Y, Spangler J, et al. Risk of serious adverse cardiovascular events associated with varenicline: a systematic review and meta-analysis. Canadian Medical Association journal 2011;**183**(12):1359-66.
21. Mills E, Thorlund K, Eapen S, et al. Cardiovascular events associated with smoking cessation pharmacotherapies: a network meta-analysis. Circulation 2014;**129**(1):28-41.
22. Prochaska J, Hilton J. Risk of cardiovascular serious adverse events associated with varenicline use for tobacco cessation: systematic review and meta-analysis. BMJ 2012;**344**.
23. Ware J, Vetrovec G, Miller A, et al. Cardiovascular safety of varenicline: patient-level meta-analysis of randomized, blinded, placebo-controlled trials. American journal of therapeutics 2013;**20**(3):235-46.
24. Prochaska J, Hilton J. Varenicline's adverse events. Choice of summary statistics: relative and absolute measures. BMJ 2013;**346**.
25. Krebs P, Sherman S. ACP Journal Club: review: varenicline for tobacco cessation does not increase CV serious adverse events. Annals of internal medicine 2012;**157**(4).
26. Svanström H, Pasternak B, Hviid A. Use of varenicline for smoking cessation and risk of serious cardiovascular events: nationwide cohort study. BMJ (Clinical research ed) 2012;**345**.
27. Hu Y-F, Chen Y-J, Lin Y-J, et al. Inflammation and the pathogenesis of atrial fibrillation. Nature reviews Cardiology 2015;**12**(4):230-43.
28. Kannel WB, McGee DL. Diabetes and cardiovascular disease. The Framingham study. JAMA 1979;**241**(19):2035-38.
29. Kim S, Liu J, Solomon D. The risk of atrial fibrillation in patients with rheumatoid arthritis. Annals of the rheumatic diseases 2014;**73**(6):1091-95.
30. Lindhardsen J, Ahlehoff O, Gislason GH, et al. Risk of atrial fibrillation and stroke in rheumatoid arthritis: Danish nationwide cohort study. BMJ (Clinical research ed) 2012;**344**.
31. Parisi R, Rutter M, Lunt M, et al. Psoriasis and the Risk of Major Cardiovascular Events: Cohort Study Using the Clinical Practice Research Datalink. The Journal of investigative dermatology 2015.
32. Ahlehoff O, Gislason G, Jørgensen C, et al. Psoriasis and risk of atrial fibrillation and ischaemic stroke: a Danish Nationwide Cohort Study. European heart journal 2012;**33**(16):2054-64.
33. Jensen AB, Moseley P, Oprea T, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature communications 2014;**5**.

34. Morris Z, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. Journal of the Royal Society of Medicine 2011;**104**(12):510-20.
35. Scannell J, Blanckley A, Boldon H, et al. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov 2012;**11**(3):191-200.
36. The Truly Staggering Cost Of Inventing New Drugs - Forbes. http://www.forbes.com/sites/matthewherper/2012/02/10/the-truly-staggering-cost-of-inventing-new-drugs/.
37. Stuart E, Cole S, Bradshaw C, et al. The use of propensity scores to assess the generalizability of results from randomized trials. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2011;**174**(2):369-86.
38. Ezekowitz J, Hu J, Delgado D, et al. Acute Heart Failure. Circulation: Heart Failure 2012;**5**(6):735-41.
39. Keene D, Price C, Shun-Shin M, et al. 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.
40. Fox K, Ford I, Steg P, et al. Ivabradine in Stable Coronary Artery Disease without Clinical Heart Failure. N Engl J Med 2014;**371**(12):1091-99.
41. New J, Bakerly N, Leather D, et al. Obtaining real-world evidence: the Salford Lung Study. Thorax 2014:thoraxjnl-2014-205259.
42. Fröbert O, Lagerqvist B, Olivecrona G, et al. Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction. N Engl J Med 2013;**369**(17):1587-97.
43. Ford I, Murray H, Packard C, et al. Long-Term Follow-up of the West of Scotland Coronary Prevention Study. N Engl J Med 2007;**357**(15):1477-86.
44. van Staa T-P, Dyson L, McCann G, et al. The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: evaluations of two exemplar trials. Health technology assessment (Winchester, England) 2014;**18**(43):1-146.
45. van Staa T-P, Goldacre B, Gulliford M, et al. Pragmatic randomised trials using routine electronic health records: putting them to the test. BMJ 2012;**344**:e55.
46. Castillo T, Arofan G, Moore S, et al. Enhancing discoverability of public health and epidemiology, 2014.
47. Herrett E, Shah A, Boggon R, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. BMJ 2013;**346**.
48. Morley K, Wallace J, Denaxas S, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PloS one 2014;**9**(11).
49. Pathak J, Kho A, Denny J. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. Journal of the American Medical Informatics Association : JAMIA 2013;**20**(e2).

50. Chen Y, Carroll R, Hinz EM, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. Journal of the American Medical Informatics Association : JAMIA 2013;**20**(e2).
51. Detection of atrial fibrillation episodes using SVM. Engineering in Medicine and Biology Society, 2008 EMBS 2008 30th Annual International Conference of the IEEE; 2008. IEEE.
52. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genetics in medicine 2013;**15**(10):761-71.
53. Kho AN, Pacheco Ja, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Science translational medicine 2011;**3**(79):79re1-79re1.
54. Köhler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research 2014;**42**(Database issue):D966-74.
55. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. American journal of human genetics 2010;**86**(4):560-72.
56. Newton K, Peissig P, Kho A, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association 2013;**20**(e1):e147-e54.
57. Shah AD, Langenberg C, Rapsomaniki E, et al. Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1·9 million people. The lancet Diabetes & endocrinology 2015;**3**(2):105-13.
58. Khoury M, Gwinn M, Ioannidis J. The emergence of translational epidemiology: from scientific discovery to population health impact. American journal of epidemiology 2010;**172**(5):517-24.
59. Denaxas S, George J, Herrett E, et al. Data Resource Profile: Cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). International Journal of Epidemiology 2012;**41**(6):1625-38.
60. Herrett E, Smeeth L, Walker L, et al. The Myocardial Ischaemia National Audit Project (MINAP). Heart 2010;**96**(16):1264-67.
61. Stewart R, Soremekun M, Perera G, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry 2009;**9**:51.