UK research data resources based on primary care electronic health records: review and summary for potential users

Lara Edwards[1], James Pickett[1], Darren M Ashcroft[2], Hajira Dambha-Miller[3], Azeem Majeed[4], Christian Mallen[5], Irene Petersen[6], Nadeem Qureshi[7], Tjeerd van Staa[8], Gary Abel[9], Chris Carvalho[10], Rachel Denholm[11,12,13,14,15], Evangelos Kontopantelis[16], Ayoyemi Macaulay[1], John Macleod[11,15]*

[1]Health Data Research UK (HDR UK), London, UK; [2]Centre for Pharmacoepidemiology and Drug Safety, NIHR Greater Manchester Patient Safety Translational Research Centre, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK; [3]Primary Care Research Centre, University of Southampton, Southampton, UK; [4]Primary Care and Public Health, Imperial College London, London, UK; [5]Institute for Global Health, Keele University, Keele, UK; [6]Department of Primary Care & Population Health, Institute of Epidemiology & Health, University College London, London, UK; [7]Centre for Academic Primary Care, University of Nottingham, Nottingham, UK; [8]Health eResearch Centre, University of Manchester, Manchester, UK; [9]Department of Health and Community Sciences (Medical School), Faculty of Health and Life Sciences, University of Exeter, Exeter, UK; [10]Clinical Effectiveness Group, Queen Mary University of London, London, UK; [11]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK; [12]Centre for Academic Primary Care, University of Bristol, Bristol, UK; [13]NIHR Bristol Biomedical Research Centre, Bristol, UK; [14]Health Data Research UK South-West, Bristol, UK; [15]NIHR Applied Research Collaboration (ARC) West, Bristol, UK; [16]Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, UK

***For correspondence:** John.Macleod@bristol.ac.uk

Competing interest: See page 7

Received: 02 April 2023
Accepted: 07 July 2023
Published: 09 August 2023

Author Keywords: electronic health records, primary care databases, population level linked data, population, primary health care

DOI:10.3399/BJGPO.2023.0057

Abstract

Background: The range and scope of electronic health record (EHR) data assets in the UK has recently increased, which has been mainly in response to the COVID-19 pandemic. Summarising and comparing the large primary care resources will help researchers to choose the data resources most suited to their needs.

Aim: To describe the current landscape of UK EHR databases and considerations of access and use of these resources relevant to researchers.

Design & setting: Narrative review of EHR databases in the UK.

Method: Information was collected from the Health Data Research Innovation Gateway, publicly available websites and other published data, and from key informants. The eligibility criteria were population-based open-access databases sampling EHRs across the whole population of one or more countries in the UK. Published database characteristics were extracted and summarised, and these were corroborated with resource providers. Results were synthesised narratively.

Results: Nine large national primary care EHR data resources were identified and summarised. These resources are enhanced by linkage to other administrative data to a varying extent. Resources are mainly intended to support observational research, although some can support experimental studies. There is considerable overlap of populations covered. While all resources are accessible to bona fide researchers, access mechanisms, costs, timescales, and other considerations vary across databases.

Conclusion: Researchers are currently able to access primary care EHR data from several sources. Choice of data resource is likely to be driven by project needs and access considerations. The landscape of data resources based on primary care EHRs in the UK continues to evolve.

How this fits in

This narrative review is intended to provide an update on the continually evolving UK landscape of primary care EHR-linked databases available for research purposes. Similar reviews have been conducted previously; however, with the emergence of newer linked data assets, this update provides a current view of these different data assets, providing detail on scale, scope, and data sources within each, as well as how researchers can access them, costing models across each, and the training and accreditation required.

Introduction

Information held in EHRs is a valuable research resource, particularly where the source data systems have near universal, longitudinal population coverage, as is the case with UK primary care EHRs. Given that the main purpose of EHRs is for clinical management, great care on interpretation is needed when data are used for research. Many issues of data completeness and quality, alongside the biases inherent in observational epidemiology, attach to analyses based on them; these are discussed below. This notwithstanding, EHRs have supported observational research for several decades.[1,2]

The range and scope of EHR-based data assets in the UK has recently increased, which has been primarily in response to the COVID-19 pandemic. Newer data assets may be less familiar to researchers, making their choice of the data resource most appropriate for their intended study difficult. This review aimed to summarise the current major sources of primary care EHRs data resources in the UK, alongside key characteristics of these relevant to potential users. It is hoped this information will help researchers choose the data resource most suited to their needs.

The review focused exclusively on UK EHR resources. Global resources, their development, and their uses are discussed elsewhere.[3,4] Similarly, discussion of important issues, such as controversy around data sharing and patient perspectives, is beyond the scope of this article but these are discussed elsewhere.[5]

Historical context

In the UK, primary medical care moved progressively from paper-based to electronic records from the late 1980s. Record-keeping in UK primary care is now almost exclusively electronic.[6] A variety of commercially supplied clinical software systems are used in primary care. Currently, the following three vendors dominate the UK market: EMIS Health; SystmOne (provided by The Phoenix Partnership; TPP) and Vision (Cegedim Healthcare Solutions). Partnerships between practices, system vendors, academics, and for-profit companies subsequently made subsets of electronic primary care health records available for research.

These partnerships led to the formation of the General Practice Research Database now known as the Clinical Practice Research Datalink (CPRD),[7,8] QResearch,[9] The Health Improvement Network (THIN)[10] database, and Optimum Patient Care Research Database (OPCRD).[11] The Royal College of General Practitioners (RCGP) has supported practice-based infectious disease surveillance since 1957.[12] This system is now electronic and supports a broader Research and Surveillance Centre (RCGP RSC).[13] More recently, other partnerships have arisen (see below).

The population coverage of each database reflects the popularity and geographical reach of the parent systems[6] as well as the practices that opt into them. EMIS Health is the most common provider to practices across the UK, and EMIS Health and TPP cover more than 90% of practices in England.

Initially, the major focus of EHR research was pharmaco-epidemiology but their research use now encompasses most aspects of observational epidemiology, including risk prediction,[14–17] health services research,[18–20] and clinical trials.[21] This expansion has been facilitated by enhancement of EHR resources through linkage to other administrative data and to data collected in research studies and clinical audit.

Whole population coverage

Statistical power in EHR research reflects sample size, whereas external validity is related to sample representativeness of the target population. Whole population coverage of an EHR database in a single nation has proved difficult to achieve for technical, socio-political, and legal reasons. Under the EU General Data Protection Regulation (GDPR) and the UK Data Protection Act (DPA) 2018, the legal data controller of primary care is the GP practice, which is responsible for the legal use of data and can decide whether data from practice patients may be processed for research purposes.[22]

Pre-COVID-19, Wales was the only UK nation to achieve near full-population coverage in a primary care EHR research database. The Welsh Longitudinal General Practice Dataset (WLGP),[23] hosted by SAIL Databank,[24,25] provides coverage of 83% of the population of Wales and 80% of Welsh GP practices. It is linked to other routine health and administrative datasets.[26]

COVID-19 pandemic response

The COVID-19 pandemic created a situation where observational research based on EHR data at scale became a public health and policy priority, to identify risk factors for and sequelae of infection, and to investigate the effects of treatment and prevention measures. To enable this, a Notice under Regulation 3(4) of the Health Service (Control of Patient Information) Regulations 2002 (COPI) was introduced, covering England and Wales, by the Secretary of State for Health in March 2020, which directed general practices to provide primary care information deemed essential for the COVID-19 response.[27]

New EHR-based UK data resources have been enabled by the pandemic response, including a minimised primary care data extract, GP Data for Pandemic Planning and Research (GDPPR). A partnership between Health Data Research UK (HDR UK), NHS Digital, and the British Heart Foundation (BHF) formed the BHF Data Science Centre-led CVD-COVID-UK/COVID-IMPACT Consortium.[28] This project resulted in the NHS Digital Trusted Research Environment (TRE), now NHS England Secure Data Environment (SDE); and enabled research relevant to COVID-19 with linkage to other datasets held by NHS England. The Consortium also includes other national TREs; SAIL Databank and the Scottish National Data Safe Haven.[29] OpenSAFELY is a new TRE project created in collaboration across the Bennett Institute at the University of Oxford, the EHR Research Group at the London School of Hygiene and Tropical Medicine (LSHTM), the EHR suppliers TPP SystmOne and EMIS Health, and NHS England. The open source OpenSAFELY software tools are implemented inside the data centres of TPP and EMIS to enable secure and federated analysis of all structured GP data without the need for raw data to be extracted and disseminated.[30,31]

Other UK nations established large EHR-based data resources to support COVID-19-related research, including the Early Pandemic Evaluation and Enhanced Surveillance of COVID-19 (EAVE II) database in Scotland.[32]

Given this evolving landscape, the review aims to provide a summary and comparison of the current UK-based large primary care EHR data resources, as a guide to researchers.

Method

The Health Data Research Innovation Gateway[33] was searched with the term 'primary care'. This search was supplemented with information from key informants in the National Institute for Health and Care Research (NIHR) School for Primary Care Research[34] and the wider primary care research community. Consideration was restricted to datasets openly accessible to external researchers.

Sources of primary care data for research purposes

National resources

Nine data resources were identified, which are described in Supplementary Tables S1 and S2. Each includes patients resident in one or more of the UK nations. The summary characteristics tabulated were obtained via publicly accessible websites and published data. Data providers were contacted to confirm accuracy and completeness of information.

Regional data sources

Some UK regions have developed local EHR databases, with linkage to primary care data, to support care delivery and planning; NHS business intelligence; and research. Some of these resources are accessible to researchers, although this has been generally restricted, to date, to local analysts. Because of this, these resources are not described in detail. Examples include a regional network of TREs in Scotland[35] such as DataLoch;[36] and others across England such as Combined Intelligence for Population Health (CIPHA),[37] HDR UK hub Discover-NOW,[38] the Bristol, North Somerset and South Gloucestershire systemwide dataset,[39] and the Connected Bradford database.[40] Regional EHR data will eventually become more accessible for research through the current NHS England Data for Research & Development (R&D) Programme to develop an interoperable network of NHS-owned subnational SDEs across England.[41]

Discussion

National primary care EHR data resources

Researcher-relevant characteristics of the nine data resources identified are described below.

1. Scope, scale, and data source

CPRD, QResearch, and THIN work with software suppliers to aggregate EHRs from practices that opt in. RCGP RSC and OPCRD hold agreements at the practice level to provide data and create resources that include records from different EHR vendors. Individuals can opt out of data sharing through contacting their practice.

OpenSAFELY provides secure access to full de-identified EHR records held by TPP and EMIS (>99% of patients in England, combined),[31] and enables consistent, federated analysis across the two. A GDPPR extract is available from NHS England Data Access Request Service,[42] in addition to access via the NHS England SDE. Use of these resources is currently enabled by COPI transitional provision. General use beyond the pandemic is under negotiation.

These large data resources include records from between 3 and 70 million individuals with varying person follow-up time (see Supplementary Tables S1 and S2). Reported size of the data resource may include historic patients now deceased or embarked (that is, patients who have left the geographical catchment area of the resource) such that the number of live, registered patients may be lower than total numbers reported. For example, as of November 2022 CPRD reports 60 million patients, of which 18 million are currently registered active patients, with at least 20 years of follow-up for 25% of the patients.[43] There is substantial overlap of patients represented between data resources.

All resources identified have been enhanced through linkage to other administrative data to a varying extent. Typically linkage is to secondary care records, death records, cancer registrations, and census-derived sociodemographic measures. More recently, linkage has been expanded to other datasets such as COVID-19 testing, immunisation, and intensive care.

Users typically must demonstrate a level of skills and experience appropriate to their intended research before gaining access, and may have to evidence completion of specific training, in addition to information governance and data security training.

In addition to supporting observational research, some resources offer extra research services; for example, to facilitate data-enabled trials.[21]

Refer to Supplementary Table S1: *Scope, scale and restrictions on use of UK primary care data resources*, for a detailed description of the scope, scale, and data sources of data assets.

Mechanisms of data access

Access models

Across these resources, data are accessed either through provision of a study-specific extract with assurances around security, appropriate handling, and data deletion or via a TRE or SDE. In both cases, the process typically involves several steps.

Steps and timescales

Typically, potential users are required to submit a proposal to an oversight committee. 'Access times' often describe time to this approval rather than time to data access, which can be misleading. Time to data access depends on multiple considerations that can incur considerable delays, these include the following:

- Ethical and other approvals: access to some resources requires prior ethical and R&D approvals to be in place. Some data resources have pre-approval from research ethics committees for particular types of research. Complex linked data applications and non-observational studies are more likely to require prior ethical approvals.
- Accreditation: this may be at the institutional or individual level. Some resources require organisations to have the NHS Data Security and Protection Toolkit[44] in place, in line with GDPR. Individual users may be required to complete specific training such as Safe Researcher Training offered by the UK Data Service. Some resources provide training for main users of the data, with the expectation that knowledge is passed on within the user institute. Some resources do not specify particular training requirements but expect applicants to evidence specific competencies.
- Application process: beyond completion of an application form, the application process may necessitate engagement with the data provider to discuss the proposed research; for example, to estimate feasibility and statistical power. The more elaborate this process, the greater time required.
- Linked data: this typically requires additional permissions, causing delays particularly when the linkages sought are new rather than established. New linkages, where available, will generally incur greater costs and delays.
- Data preparation and processing: depending on the data resource and project, preparation of a suitable extract or pre-processing of data made available through a TRE or SDE may incur further delays.

2. Funding models for access

Data are made accessible through the following three main funding models: (a) an annual licence (some negotiated at an organisation level); (b) per project, which may include a base cost with additional charges representing resources in preparing bespoke or complex data requests or linkages; (c) on an academic collaboration basis.

Refer to Supplementary Table S2: *Access processes and requirements for primary care data resources*, for detailed description of data access mechanisms and processes across data assets in scope.

3. Analysis of primary care EHRs

Once data have been accessed as above, several considerations apply to the analysis process.

Data wrangling and curation

Data wrangling and curation describe the processes of preparing the data before they can be analysed. The readiness of data for analysis varies depending on the data resource. Resources generally provide some form of data dictionary or data notes describing metadata and provenance of the data. Clinical and prescription data is commonly provided in a structured clinical vocabulary agnostic to the source system. Common formats include SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms), Read Codes, ICD-10 (International Classification of Diseases, Tenth Revision) codes, as well as local codes (which may be less interoperable). Sometimes a combination of these is used.

The extent of curation needed varies with study design, but may include manipulating tables, deriving variables, linking data sources, and identifying study cohorts. Where several studies require similar manipulation of data, reusing a common code is helpful. Some resources require users to share code, using repositories such as GitHub.[45] OpenSAFELY requires all code to be posted on GitHub before execution and publishes links to all executed code automatically at jobs.opensafely.org; analysts use standardised OpenSAFELY dataset building tools, which are integrated with the codelist development and sharing tools at OpenCodeLists.org.[46]

The CVD-COVID-UK/COVID IMPACT Consortium publishes protocols, code, and phenotype code lists via the HDR UK Gateway and GitHub.

Another common step required of analysts is to create EHR phenotypes that describe clinical concepts. Phenotype libraries and other resources to support standardisation and reproducibility have also been developed.[47–51] Publishers may expect authors to provide code lists, algorithms, and programme files as supplements in published articles.

Using a Trusted Research Environment (TRE) or Secure Data Environment (SDE)

Several data resources provide access via a TRE or SDE. Models vary in several ways, including the following:

- the prepackaged tools and software available in the analytical environment;
- the ability to import a user's own code or software;
- availability of code for common data management tasks;
- the degree to which previous users' data curation, variable derivation, and documentation is available to new users;
- threshold of small number suppression to protect against risk of patient reidentification;
- the level of user support available;
- ease of use;
- cost of use.

Some models allow curation, documentation, novel variable derivation, and associated documentation to be stored beyond the life of a single project or analysis and made available to future users, increasing the value of the resource. A UK Health Data Research Alliance White Paper[52] has set out guidelines and principles for TRE and SDE good practice structured around the 'Five Safes' framework,[53] and the Goldacre Review recommended use of TREs and SDE as the norm for analysis of health data.[54]

Methodological and other considerations for working with primary care EHR data resources

Clinical context

Primary care EHRs are created primarily to support continuity in clinical care, as a medico-legal document, and to support payment systems. Their use in research needs to take into consideration why and how the data were collected. Because of this, experience of creating EHRs can help in guiding and interpreting analysis of them. Data recording and coding is influenced by many considerations. Understanding these, how they influence the content of the record, and the potential for bias to be introduced is essential to making valid inferences.[55]

Analytic and epidemiological considerations

Working with these data requires considerable epidemiological and analytical experience, including knowledge of common analytical tools and experience in handling large data resources. Access may be contingent on evidencing these competencies.

Population-level data also have characteristics that can make them challenging to use.[56] Missing data and misclassification are key issues. Data are unlikely to be missing completely at random. Multiple imputation can be used to address this; however, it may introduce additional bias if used inappropriately.[57] Sometimes missingness can be addressed through linkage to other data, facilitating the assessment of the extent of potential bias.[57] Research questions must be evaluated for feasibility against the quality of the available data. For example, recording and management of many chronic conditions, risk markers, and other aspects of care have been incentivised in UK primary care, potentially introducing variations in data quality between information whose recording is or is not incentivised.[58]

Other epidemiological considerations are those attached to the difficulty of making valid causal inference in observational data where exposure allocation is non-random. The main issue is confounding by indication, where risk of exposure is associated with risk of outcome through a pathway independent of exposure.[59] Collider bias[60] and immortal time bias[61] are also frequently important. The nature of causes, causal inference, and addressing bias attached to this endeavour have been discussed elsewhere, both in general terms[62] and in the context of EHRs.[63]

Future work and future developments

Models and mechanisms for accessing primary care EHRs, enhanced through linkage to other information, continue to evolve. This information is likely to include non-health administrative data, research data, patient-reported data, and data from patient-based and other sensors. Eventually this evolution may lead to near-whole population, real-time data from across the health and care system, linked to multimodal data from other sources being readily, securely, and acceptably available for analysis. Multiple biases will attach to these analyses and appreciation of their possible influence is important, particularly when analysis is genuinely intended to inform policy choices. Strategies to address these biases will also evolve. The broad term 'artificial intelligence' is currently applied to a variety of automated analytical approaches (including machine learning and deep learning) intended to make the extraction of useful inference from multimodal data more efficient and reliable.[64,65] Linkage-enhanced data from health and care systems is likely to increasingly provide the substrate for such methods. Ultimately, this may lead to better understanding of the forces shaping human health and wellbeing, both in individuals and between social groups. This may support action to reduce inequities in these outcomes.

This article summarises major UK primary care data resources in terms of their strengths, weaknesses, and the opportunities they provide for researchers. Securing access to an appropriate dataset for research is often a complex transaction, for reasons described above. This article is intended to help researchers navigate that complexity. This is also a rapidly evolving landscape, shaped by multiple social, technical, and political considerations. In general, the trend is towards more streamlined, secure, and transparent access to better data, with the ambition that this will ultimately lead to health improvement for individuals and populations.

Funding

No specific funding was awarded to complete this work. However, LE and JP acknowledge funding from the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (MC_PC_20058). DMA is funded by the National Institute for Health and Care Research (NIHR) through the Greater Manchester Patient Safety Translational Research Centre (NIHR Greater Manchester PSTRC, Grant number: PSTRC-2016-003). CM is funded by the NIHR School for Primary Care Research and NIHR Applied Research Collaboration (ARC) West Midlands. GA is supported by the NIHR ARC South West Peninsula. AM is supported by the NIHR ARC NW London. JM is supported by the NIHR Health Research ARC West and the NIHR Bristol Biomedical Research Centre. RD is supported by the NIHR Bristol Biomedical Research Centre and Health Data Research UK South West. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care, or UK Research and Innovation. All stated funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethical approval

Not applicable for this review.

Provenance

Freely submitted; externally peer reviewed.

Acknowledgements

The authors would like to thank the representatives of those data custodians that validated the summary information in Supplementary Tables S1 and S2; Chris Orton, Ashley Akbari (WLGP, SAIL Databank), Rouven Priedon and John Nolan (CVD-COVID-UK/COVID-IMPACT), Kimberley Watson (GDPPR, NHS England), Catheryn Evans (CPRD), Pete Stokes, Brian MacKenna, Ben Goldacre (OpenSAFELY), Julia Hippisley-Cox, Rebekah Burrow (QResearch), Simon de Lusignan (RCGP RSC), Samir Dhalla (THIN database).

Competing interests

Hajira Dambha-Miller is the Editor-in-Chief of BJGP Open, but had no involvement in the peer review process or decision on this manuscript.

References

1. Chaudhry Z, Mannan F, Gibson-White A, *et al*. Outputs and growth of primary care databases in the United kingdom: Bibliometric analysis. *J Innov Health Inform* 2017; **24**(3): 942. DOI: https://doi.org/10.14236/jhi.v24i3.942
2. McDonnell L, Delaney BC, Sullivan F. Finding and using routine clinical datasets for observational research and quality improvement. *Br J Gen Pract* 2018; **68**(668): 147–148. DOI: https://doi.org/10.3399/bjgp18X695237
3. Aminpour F, Sadoughi F, Ahamdi M. Utilization of open source electronic health record around the world: a systematic review. *J Res Med Sci* 2014; **19**(1): 57–64.
4. Celi LA, Majumder MS, Ordóñez P, Osorio JS, *et al*. Leveraging Data Science for Global Health. In: *Leveraging data science for global health*. 1st edn. Cham: Springer Nature; 2020. http://link.springer.com/10.1007/978-3-030-47994-7 DOI: https://doi.org/10.1007/978-3-030-47994-7
5. Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.data ran into trouble. *J Med Ethics* 2015; **41**(5): 404–409. https://jme.bmj.com/content/41/5/404.info DOI: https://doi.org/10.1136/medethics-2014-102374
6. Kontopantelis E, Stevens RJ, Helms PJ, *et al*. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ Open* 2018; **8**(2): e020738. DOI: https://doi.org/10.1136/bmjopen-2017-020738
7. Herrett E, Gallagher AM, Bhaskaran K, *et al*. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol* 2015; **44**(3): 827–836. DOI: https://doi.org/10.1093/ije/dyv098
8. Clinical Practice Research Datalink (CPRD). Digital object identifiers (DOIs) for datasets. https://cprd.com/digital-object-identifiers-dois-datasets (accessed 13 Jul 2023).
9. QResearch. Generating new knowledge to improve patient care. https://www.QResearch.org/ (accessed 13 Jul 2023).
10. The Health Improvement Network (THIN). Data to build better population health outcomes and a foundation for research. https://www.the-health-improvement-network.com/ (accessed 13 Jul 2023).
11. NHS Health Research Authority. Optimum Patient Care Research Database. https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/optimum-patient-care-research-database/ (accessed 1 Aug 2023).
12. Royal College of General Practitioners. RCGP Research and Surveillance Centre (RSC). https://www.rcgp.org.uk/clinical-and-research/our-programmes/research-and-surveillance-centre (accessed 13 Jul 2023).
13. Leston M, Elson WH, Watson C, *et al*. Representativeness, vaccination uptake, and COVID-19 clinical outcomes 2020–2021 in the UK Oxford-Royal College of General Practitioners Research and Surveillance Network: cohort profile summary. *JMIR Public Health Surveill* 2022; **8**(12): e39141. DOI: https://doi.org/10.2196/39141
14. Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008; **336**(7659): 1475–1482. DOI: https://doi.org/10.1136/bmj.39609.449676.25
15. Cornish R, Macleod J, Strang J, *et al*. Risk of death during and after opiate substitution treatment in primary care: prospective observational study in UK General Practice Research Database. *BMJ* 2010; **341**: c5475. DOI: https://doi.org/10.1136/bmj.c5475
16. Macleod J, Steer C, Tilling K, *et al*. Prescription of benzodiazepines, Z-drugs, and gabapentinoids and mortality risk in people receiving opioid agonist treatment: observational study based on the UK Clinical Practice Research Datalink and Office for National Statistics death records. *PLoS Med* 2019; **16**(11): e1002965. DOI: https://doi.org/10.1371/journal.pmed.1002965
17. Osborn DPJ, Hardoon S, Omar RZ, *et al*. Cardiovascular risk prediction models for people with severe mental illness: results from the prediction and management of cardiovascular risk in people with severe mental illnesses (PRIMROSE) research program. *JAMA Psychiatry* 2015; **72**(2): 143–151. DOI: https://doi.org/10.1001/jamapsychiatry.2014.2133
18. Jick H, Jick SS, Myers MW, Vasilakis C. Third-generation oral contraceptives and venous thrombosis. *Lancet* 1997; **349**(9053): 731–732. DOI: https://doi.org/10.1016/S0140-6736(05)60173-0
19. Bhanu C, Jones ME, Walters K, *et al*. Physical health monitoring in dementia and associations with Ethnicity: a descriptive study using electronic health records. *BJGP Open* 2020; **4**(4): bjgpopen20X101080. DOI: https://doi.org/10.3399/bjgpopen20X101080
20. Pham TM, Petersen I, Walters K, *et al*. Trends in dementia diagnosis rates in UK ethnic groups: analysis of UK primary care data. *Clin Epidemiol* 2018; **10**: 949–960. DOI: https://doi.org/10.2147/CLEP.S152647
21. 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 Technol Assess* 2014; **18**(43): 1–146. DOI: https://doi.org/10.3310/hta18430
22. British Medical Association. GPs as data controllers under the General Data Protection Regulation. 2018. https://www.bma.org.uk/advice-and-support/ethics/confidentiality-and-health-records/gps-as-data-controllers-under-gdpr (accessed 13 Jul 2023).
23. SAIL Databank. Welsh Longitudinal General Practice Dataset — (WLGP). https://web.www.healthdatagateway.org/dataset/33fc3ffd-aa4c-4a16-a32f-0c900aaea3d2 (accessed 13 Jul 2023).
24. SAIL Databank. SAIL Databank is a rich and trusted population databank. https://saildatabank.com/ (accessed 13 Jul 2023).
25. Ford DV, Jones KH, Verplancke J-P, *et al*. The SAIL Databank: building a national architecture for E-health research and evaluation. *BMC Health Serv Res* 2009; **9**: 157. DOI: https://doi.org/10.1186/1472-6963-9-157

26. Akbari A, Lyons R, Bandyopadhyay A, *et al.* Analysis of factors associated with changing general practice in the first 14 years of life in Wales using linked cohort and primary care records: implications for using primary care Databanks for life course research. *Int J Popul Data Sci* 2018; **3**(4). DOI: https://doi.org/10.23889/ijpds.v3i4.818
27. NHS Digital. Control of patient information (COPI) notice. https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/control-of-patient-information-copi-notice [last updated 15 Feb 2022] (accessed 13 Jul 2023).
28. Wood A, Denholm R, Hollings S, *et al.* Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021; **373**: n826. DOI: https://doi.org/10.1136/bmj.n826
29. Public Health Scotland. Electronic Data Research and Innovation Service (eDRIS). https://www.isdscotland.org/products-and-services/edris/ (accessed 13 Jul 2023).
30. Andrews C, Schultze A, Curtis H, *et al.* OpenSAFELY: Representativeness of electronic health record platform OpenSAFELY-TPP data compared to the population of England. *Wellcome Open Res* 2022; **7**: 191. DOI: https://doi.org/10.12688/wellcomeopenres.18010.1
31. Walker AJ, MacKenna B, Inglesby P, *et al.* Clinical coding of long COVID in English primary care: a federated analysis of 58 million patient records in situ using OpenSAFELY. *Br J Gen Pract* 2021; **71**(712): e806–e814. DOI: https://doi.org/10.3399/BJGP.2021.0301
32. Simpson CR, Robertson C, Vasileiou E, *et al.* Early pandemic evaluation and enhanced surveillance of COVID-19 (EAVE II): protocol for an observational study using linked Scottish national data. *BMJ Open* 2020; **10**(6): e039097. DOI: https://doi.org/10.1136/bmjopen-2020-039097
33. Health Data Research UK. Health Data Research Innovation Gateway. Gateway to health data and tools for research. https://www.healthdatagateway.org/ (accessed 13 Jul 2023).
34. National Institute for Health and Care Research. NIHR School for Primary Care Research. https://www.spcr.nihr.ac.uk/ (accessed 13 Jul 2023).
35. Gao C, McGilchrist M, Mumtaz S, *et al.* A national network of safe havens: Scottish perspective. *J Med Internet Res* 2022; **24**(3): e31684. DOI: https://doi.org/10.2196/31684
36. University of Edinburgh. DataLoch. https://dataloch.org/ (accessed 13 Jul 2023).
37. Combined Intelligence for Population Health. CIPHA data platform. https://www.cipha.nhs.uk/ (accessed 13 Jul 2023).
38. Imperial College Healthcare Partners. Discover NOW Health Data Research Hub. https://discover-now.co.uk/ (accessed 13 Jul 2023).
39. Healthier Together: Bristol, North Somerset and South Gloucestershire Integrated Care Board. BNSSG system wide dataset. https://bnssghealthiertogether.org.uk/population-health-management/ (accessed 13 Jul 2023).
40. Sohal K, Mason D, Birkinshaw J, *et al.* Connected Bradford: a whole system data linkage accelerator. *Wellcome Open Res* 2022; **7**: 26. DOI: https://doi.org/10.12688/wellcomeopenres.17526.2
41. Bloomfield C. Investing in the future of health research: secure, accessible and life-saving (NHS England). 2022. https://www.england.nhs.uk/blog/investing-in-the-future-of-health-research-secure-accessible-and-life-saving/ (accessed 13 Jul 2023).
42. NHS England. Data Access Request Service (DARS). https://digital.nhs.uk/services/data-access-request-service-dars (accessed 13 Jul 2023).
43. Medicines & Healthcare Products Regulatory Agency. Clinical Practice Research Datalink. https://cprd.com/ (accessed 13 Jul 2023).
44. NHS England. Data Security and Protection Toolkit. https://www.dsptoolkit.nhs.uk/ (accessed 13 Jul 2023).
45. GitHub. GitHub website. https://github.com/ (accessed 13 Jul 2023).
46. OpenSAFELY. OpenCodelists. https://www.opencodelists.org (accessed 13 Jul 2023).
47. Chapman M, Mumtaz S, Rasmussen LV, *et al.* Desiderata for the development of next-generation electronic health record phenotype libraries. *Gigascience* 2021; **10**(9): giab059. DOI: https://doi.org/10.1093/gigascience/giab059
48. Springate DA, Kontopantelis E, Ashcroft DM, *et al.* ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PLoS One* 2014; **9**(6): e99825. DOI: https://doi.org/10.1371/journal.pone.0099825
49. Health Data Research UK. HDR Phenotype Library. https://phenotypes.healthdatagateway.org/ (accessed 13 Jul 2023).
50. SAIL DataBank. Concept Library. https://conceptlibrary.saildatabank.com (accessed 13 Jul 2023).
51. Sharma M, Petersen I, Nazareth I, Coton SJ. An algorithm for identification and classification of individuals with type 1 and type 2 diabetes mellitus in a large primary care database. *Clin Epidemiol* 2016; **8**: 373–380. DOI: https://doi.org/10.2147/CLEP.S113415
52. UK Health Data Research Alliance, NHSX. Building trusted research environments — principles and best practices; towards TRE ecosystems. 2021. https://zenodo.org/record/5767586#.ZATOu3bP02w (accessed 13 Jul 2023). 10.5281/zenodo.5767586 DOI: https://doi.org/10.5281/zenodo.5767586
53. Desai T, Ritchie F, Welpton R. Five safes: designing data access for research. 2016. https://www2.uwe.ac.uk/faculties/bbs/Documents/1601.pdf (accessed 13 Jul 2023).
54. Goldacre B, Morley J. Better, broader, safer: using health data for research and analysis. A review commissioned by the Secretary of State for Health and Social Care (Department of Health and Social Care). 2022. https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis (accessed 13 Jul 2023).
55. Petersen I, Welch CA, Nazareth I, *et al.* Health indicator recording in UK primary care electronic health records: key implications for handling missing data. *Clin Epidemiol* 2019; **11**: 157–167. DOI: https://doi.org/10.2147/CLEP.S191437

56. Christen P, Schnell R. Thirty-three myths and misconceptions about population data: from data capture and processing to linkage. *Int J Popul Data Sci* 2023; **8**(1). DOI: https://doi.org/10.23889/ijpds.v8i1.2115
57. Cornish RP, Tilling K, Boyd A, *et al*. Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years. *Int J Epidemiol* 2015; **44**(3): 937–945. DOI: https://doi.org/10.1093/ije/dyv035
58. Gulliford MC, Charlton J, Ashworth M, *et al*. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. *PLoS One* 2009; **4**(9): e7168. DOI: https://doi.org/10.1371/journal.pone.0007168
59. Freemantle N, Marston L, Walters K, *et al*. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013; **347**: f6409. DOI: https://doi.org/10.1136/bmj.f6409
60. Cole SR, Platt RW, Schisterman EF, *et al*. Illustrating bias due to conditioning on a collider. *Int J Epidemiol* 2010; **39**(2): 417–420. DOI: https://doi.org/10.1093/ije/dyp334
61. Mansournia MA, Nazemipour M, Etminan M. Causal diagrams for immortal time bias. *Int J Epidemiol* 2021; **50**(5): 1405–1409. DOI: https://doi.org/10.1093/ije/dyab157
62. Krieger N, Davey Smith G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *Int J Epidemiol* 2016; **45**(6): 1787–1808. DOI: https://doi.org/10.1093/ije/dyw114
63. Kotz D, O’Donnell A, McPherson S, Thomas KH. Using primary care databases for addiction research: an introduction and overview of strengths and weaknesses. *Addict Behav Rep* 2022; **15**: 100407. DOI: https://doi.org/10.1016/j.abrep.2022.100407
64. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. *Am J Epidemiol* 2019; **188**(12): 2222–2239. DOI: https://doi.org/10.1093/aje/kwz189
65. Jorm LR. Commentary: towards machine learning-enabled epidemiology. *Int J Epidemiol* 2021; **49**(6): 1770–1773. DOI: https://doi.org/10.1093/ije/dyaa242