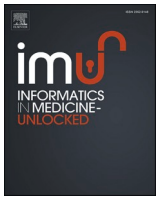
Informatics in Medicine Unlocked 40 (2023) 101291

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Curator – A data curation tool for clinical real-world evidence

Antonella Delmestri *, Daniel Prieto-Alhambra

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Centre for Statistics in Medicine, BOTNAR Research Centre, Windmill Road, University of Oxford, Oxford OX3 7LD, UK

ARTICLE INFO

Keywords:
Data wrangling
Real-world data
RWD
Electronic health records
EHR

ABSTRACT

Objective: This research aims to establish an efficient, systematic, reproducible, and transparent solution for advanced curation of real-world data, which are highly complex and represent an invaluable source of information for academia and industry.

Materials and methods: We propose a novel software solution that splits the statistical analytical pipeline into two phases. The first phase is implemented through Curator, which performs data engineering and data modelling on deidentified real-world data to achieve advanced curation and provides selected information ready to be analyzed in the second phase by statistical packages. Curator is made of a suite of Python programs and uses MySQL as its database management system. Curator has been utilised with several UK primary and secondary care data sources.

Results: Curator has been used in 25 completed clinical and health economics research studies. Their output has been published in 2 NIHR-funded reports and 33 prestigious international peer-reviewed journals and presented at 38 global conferences. Curator has consistently reduced research time and costs by over 36% and made research more reproducible and transparent.

Discussion: Curator fits in well with recent UK governmental guidelines that recognise health data curation as a complex standalone technical challenge. Curator has been used extensively on UK real-world data and can handle several linked datasets. However, for Curator to be accessed by a wider audience, it needs to become more user-friendly.

Conclusion: Curator has proven to be a cost-effective and trustworthy data curation tool, which should be developed further and made available to third parties.

1. Introduction

The rapid progress of IT technologies has made it possible to produce and routinely collect an increasing amount of health data from a variety of digital sources, such as electronic health records (EHR), billing and claims activities, disease and product registries, and personal-use electronic devices. Health data used in observational research are called real-world data (RWD) [1,2] to differentiate them from data produced by traditional randomised controlled trials (RCT) used in experimental research [3]. Due to the growing use of RWD and advances in statistical methods and data science, observational studies are now increasingly recognised as complementary to RCT in supporting patient care, health services, and decision-making [4–6]. For example, during the COVID-19 pandemic, RWD were key when producing governmental policies and recommendations for healthcare strategies, investigating hospitalisation, death, and new scopes for existing medicines, and testing potential rare vaccination adverse events [7–9].

The current use of RWD is possible due to digital primary care data collection, pioneered by the US since the 1970s [10] and the UK soon after [11,12]. In the UK, this long tradition has been facilitated by the gatekeeping role that GPs play, where the NHS provides freely available health care for 98% of the population [13] from cradle to grave [14]. However, the vast amounts of RWD, their disaggregated format, multiple coded nature, longitudinality, and multi-dimensionality present serious challenges to their use in research. Completeness, correctness, concordance, plausibility, and currency are the five common dimensions of data quality that are under scrutiny [15,16].

2. Objectives

Our primary objective was to make observational research using RWD more efficient. We identified data curation as a key area where

* Corresponding author.
E-mail address: antonella.delmestri@ndorms.ox.ac.uk (A. Delmestri).

https://doi.org/10.1016/j.imu.2023.101291
Received 29 March 2023; Received in revised form 5 June 2023; Accepted 5 June 2023
Available online 7 June 2023
2352-9148/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

statisticians spend most of their time and focused on finding a systematic solution to optimise this process.

We also aimed to make observational research more reproducible and transparent. The complexity of RWD increases the likelihood of errors, oversights, and unforeseen issues, which can result in damage to public health and scientific reputations.

3. Material and methods

We propose a specialised software solution: a two-phase statistical analytical pipeline (Fig. 1) for deidentified RWD. The first statistical phase (SP1) is implemented through Curator, which performs data engineering and data modelling on RWD to achieve advanced curation. SP1 generates selected information that can be analyzed in the second phase (SP2) by traditional statistical packages.

The resulting research output is comparable to that of other studies but is generated more quickly, lower-priced, and more accurate, transparent, and reproducible. Curator is the result of more than nine years of extensive experience and technical know-how using several complex, heterogenous UK datasets utilised internationally for medical research.

Curator is made of a set of command line PowerShell Python 3 [17] programs. It uses MySQL 8.0 [18] as its database management system, InnoDB as the engine for data storage and modelling, and 7-Zip [19] as the file archiver. Curator was developed and tested in a Microsoft Windows 10 64-bit environment with modest hardware and software requirements, as we aimed to use only free, open-source third-party software.

Curator's input is made of many hundreds of CSV-like files with no free text or narrative content. Their tabular formats depend on the specific data source/s requested via a data provider. Despite the many differences in the data structures of the data sources managed by Curator, all RWD are deidentified, patient-centric, and coded. They all require lookups, dictionaries, and specialised know-how to be queried and interpreted. They are generally GByte-sized. For example, 200–300 Gb would be sufficient for a population of 600,000 patients, which is the current default upper limit per approved project set by our data provider, the Clinical Practice Research Datalink (CPRD) [20]. MySQL scales well up to this volume of data. It would be interesting to see if and how its performance changes if a much higher volume of data became available.

Curator's output for this input size is usually a single MByte-sized CSV structured file containing one row of selected deidentified data for each patient, ready-to-use in SP2 by statisticians and epidemiologists. In statistics, epidemiology, and hereinafter, these output data are called *variables*. They can be demographics, outcomes, exposures, or covariates, for example, depending on the observational study design. The number of variables depends on the study design and research question/s and is defined by the researchers. From Curator's point of view, there is no upper limit to the number of variables. One obvious limitation is the MySQL storage threshold on table column count, which is currently 1,017 for InnoDB, the storage engine used by Curator. However, MySQL scales well up to this number of variables.

It is the researchers' and data provider's responsibility to ensure they are legally entitled to make use of the data and to run their study in a safe environment that complies with authorised and ethical constraints, including the General Data Protection Regulation (GDPR). Curator therefore does not require any other legal agreement before use.

In 2020, Curator's IP rights were assigned to Oxford University Innovation (OUI) [21].

3.1. Data-driven pipeline

Curator works on RWD that contain medical information for a group of people defined in the study design, and obtained securely from a data provider. Curator's main components are:

A. The Source Database Builder automatically creates and populates a MySQL relational database based on the downloaded structured and disaggregated raw RWD files. These files contain inconsistencies, redundancies, and information unrelated to research. The Source Database Builder addresses these issues. When loading the data using its tailored metadata, it automatically selects only research-meaningful fields and bins any unsuitable record in twin tables, where an extra column reports the reason for binning. A record might be rejected because the event date is missing or refers to an impossible moment in time for the patient (e.g. before birth, after death, or a future date), an event is a duplication, or the information is missing or impossible (e.g. the clinical or product code used is not part of the dictionary or a measurement value is clinically meaningless), etcetera. When the data processing rationale permits, the Source Database Builder implements a concurrent loading strategy using ProcessPoolExecutor, a Python multiprocessing class [22] that allows for the simultaneous loading of data files into different tables. The files are loaded into each table sequentially to avoid tables locks. At the end of the source database building, a summary table is automatically created using concurrent processing. It contains the numbers of acceptable, unacceptable, and total records per source table. These numbers are displayed for comparison with the log file provided by the data provider. Detailed log files are also automatically created to facilitate testing and increase transparency.

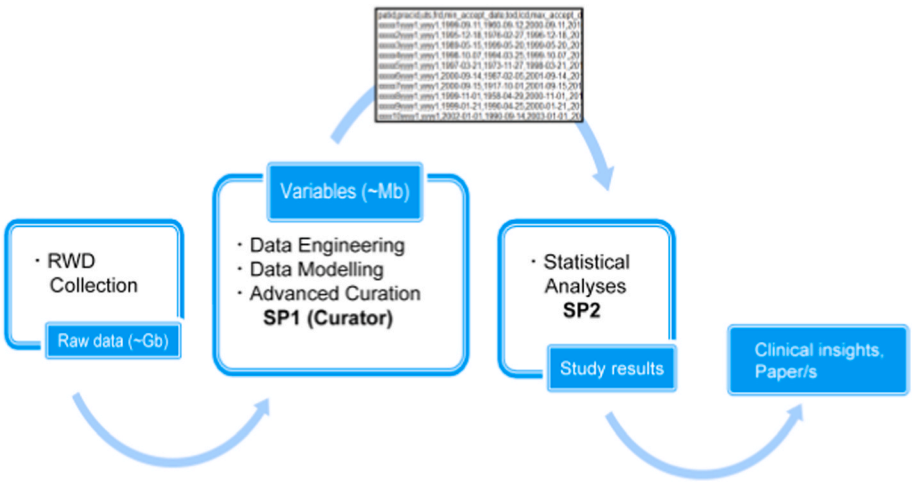


Fig. 1. Statistical phases SP1 (Curator) and SP2. RWD: real-world data.

B. The Variable Designer allows researchers to specify the variables they want by choosing a number of parameters for each variable, such as event type or *domain* (e.g. conditions, medications, immunisations, interventions, measurements, and visits identified by phenotyping algorithms and related code lists [23]); event time direction (e.g. before or on-after patient index date, which is the date of the first event of interest for the patient in the study period); event time window (e.g. 2 years, 3 months, 14 days, or none); event timepoint (e.g. first event, last event, or all events), etcetera. Curator provides comparable phenotyping code lists for each of the supported data sources based on its specific clinical coding dictionaries. We have developed these concept lists over the years, using existing literature and public repositories when available and in close collaboration with clinicians and device- and pharmaco-epidemiologists. For each data source, researchers can choose from highly granular domains. For example, in the conditions domain, researchers can choose from "Diabetes", "Diabetes type1", "Diabetes type2", and "Diabetes unspecified", and from the medications domain "Statin intake", "Statin low intake", "Statin medium intake", and "Statin high intake". The Variable Designer is implemented via an Excel workbook equipped with content controls and connected to MySQL via an ODBC Driver. Planned improvements for this component are discussed in section 5.2.
C. The Variable Extractor identifies and extracts the set of variables described by the researchers in the Variable Designer. With the exception of patient demographic characteristics such as gender or year of birth, each of the extracted variables is the result of complex epidemiological reasoning implemented by MySQL code. Several source tables, lookups, and dictionaries must be joined, and multiple specific conditions applied to identify the information, if present, specified by the variable's parameters. MySQL queries are constructed programmatically based on these parameters. To make this task as efficient as possible, supporting domain-specific tables are automatically created, from which the requested variables are concurrently extracted using ProcessPoolExecutor [22]. Log files reporting extensive details are also created to promote testing and increase transparency. To further enhance clarity and transparency, we have developed guidelines for clinical RWD processing. Some of these have been published [24], whereas for others publication work is in progress.

Each project managed by Curator requires the Source Database Builder to be run once, but several variable sets can be designed, each requiring a variable extraction (Fig. 2). Although the Source Database Builder and Variable Designer must both be completed for the corresponding Variable Extractor to start, they can progress in parallel to make the process more efficient.

At the end of its workflow for each variable extraction, Curator encrypts the output to Advanced Encryption Standard (AES) 256, creating a password-protected zipped CSV file that can easily be imported into common statistical software (e.g. R, Stata, SAS, or SPSS) after the password is securely supplied to the researcher.

3.2. Metadata

Curator automatically builds the source database using its ad-hoc optimised metadata [25] – stored in a MySQL database – that describe the supported RWD and the modelled database structures, keys, and indexes. These metadata ensure that only data that are useful for clinical research or health economics are loaded from the raw files into the source database, significantly decreasing storage needs. The metadata also enhance the source data structure by adding pivotal inferred fields to improve SQL query design, which reduces processing time. Fig. 3 shows an example of Curator's metadata for a table where *db_id* identifies the database source and its version, *tbl_id* the table to be created, *col_position* the position of the column in the table, and *field_position* the position of the corresponding field in the source file. A *col_position* equal to zero means that the corresponding field is not

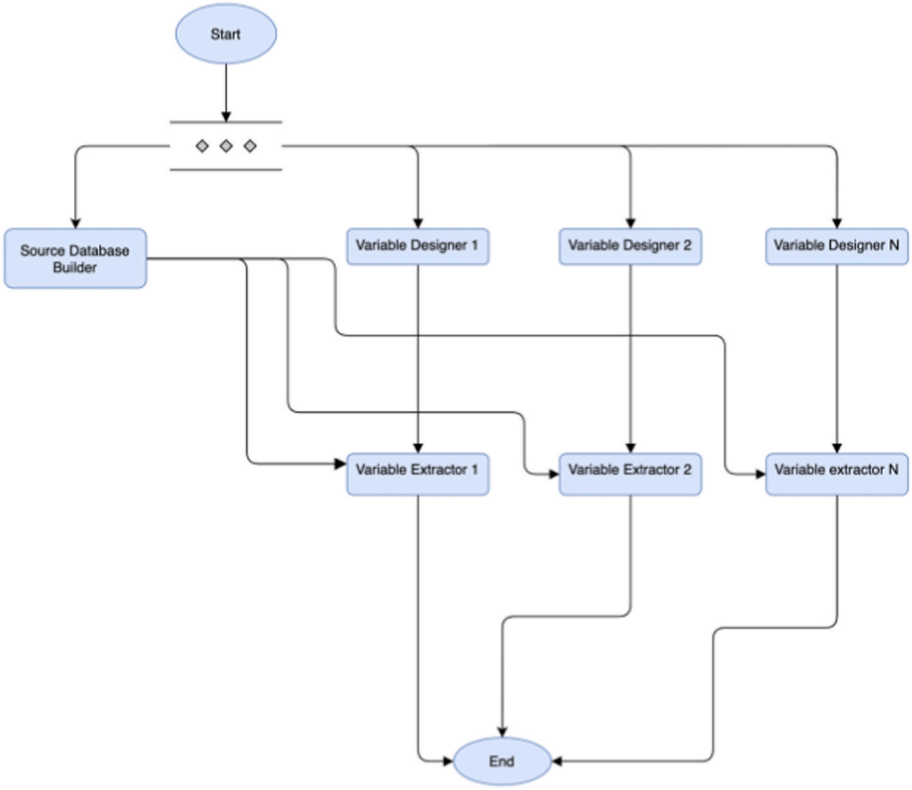


Fig. 2. Curator's data-driven pipeline.

db_id	tbl_id	col_position	col_name	field_position	col_datatype	tbl_pk	check_date	col_unsigned	col_nullable	col_default	col_autoincrement	col_unaccept_values
15	3	2	pracid	0	INT	0	NULL	1	0	NULL	NULL	NULL
15	3	1	patid	1	BIGINT	1	NULL	1	0	NULL	NULL	NULL
15	3	0	vmid	2	BIGINT	0	NULL	1	0	NULL	NULL	NULL
15	3	3	gender	3	TINYINT	0	NULL	1	0	NULL	NULL	NULL
15	3	4	yob	4	SMALLINT	0	NULL	1	0	NULL	NULL	NULL
15	3	5	mob	5	TINYINT	0	NULL	1	1	NULL	NULL	NULL
15	3	6	marital	6	TINYINT	0	NULL	1	1	NULL	NULL	NULL
15	3	7	famnum	7	MEDIUMINT	0	NULL	1	1	NULL	NULL	NULL
15	3	0	chsreg	8	TINYINT	0	NULL	1	0	NULL	NULL	NULL
15	3	0	chsdate	9	DATE	0	NULL	NULL	1	NULL	NULL	NULL
15	3	0	prescr	10	TINYINT	0	NULL	1	0	NULL	NULL	NULL
15	3	0	capsup	11	TINYINT	0	NULL	1	0	NULL	NULL	NULL
15	3	8	frd	12	DATE	0	NULL	NULL	0	NULL	NULL	NULL
15	3	9	crd	13	DATE	0	NULL	NULL	1	NULL	NULL	NULL
15	3	0	regstat	14	TINYINT	0	NULL	1	1	NULL	NULL	NULL
15	3	0	reggap	15	MEDIUMINT	0	NULL	1	1	NULL	NULL	NULL
15	3	0	internal	16	TINYINT	0	NULL	1	1	NULL	NULL	NULL
15	3	10	tod	17	DATE	0	NULL	NULL	1	NULL	NULL	NULL
15	3	11	toreason	18	TINYINT	0	NULL	1	1	NULL	NULL	NULL
15	3	12	deathdate	19	DATE	0	NULL	NULL	1	NULL	NULL	NULL
15	3	13	accept	20	TINYINT	0	NULL	1	0	NULL	NULL	0

Fig. 3. Example of Curator's metadata for a patient table.

imported into table. A *field_position* equal to zero means that the field does not exist in the source file, but is inferred from other fields for SQL code optimisation. Primary keys and secondary indexes are added after data are fully loaded to increase efficiency.

Based on this programming model, source databases are created and populated efficiently with no human investment or intervention beyond the creation of the one-off metadata, which are maintained by Curator's developers. Data providers rarely change their source data structures. Any changes are easily managed by Curator's developers via a process of cloning, updating, and versioning for backward compatibility.

3.3. Data modelling

Curator data modelling ensures that research methods can be systematically applied to supported data sources to produce trustworthy, comparable, reproducible results. The data engineering is based on bespoke built-in organising data modelling (DM) designed to store observational data based on the following principles:

A Patient-centred structure: The DM reflects the patient-centred nature of RWD.
B Minimum storage: The DM only stores data useful for research.
C Maximum performance: The DM structures and indexes the data to minimize processing time.
D Domain management: All domains use a minimal core structure made up of the patient identifier, event date, and domain-specific code used to identify the event. When necessary, other domain-specific information can be added, such as the duration for prescriptions or hospitalisations, or a value for measurements (e.g. systolic blood pressure).
E Scalability: The DM is optimised for advanced curation of clinical sources that vary in size. Thresholds are only imposed by data providers and researchers' storage capabilities.
F Code lists: The DM relies on our validated repository of RWD code lists [26, 27], introduced in section 3.1. These lists are created by phenotyping algorithms. They are made of standardised, computer-processable collections of concepts that can identify patient characteristics, such as symptoms, conditions, interventions, measurements, immunisations, and prescriptions. These code lists are tailored subsets of national or international clinical, drug, and device classification systems, such as the Systematized Nomenclature of Medicine (Snomed) [28], Read [29], International Classification of Diseases (ICD, v.10) [30], Office of Population Censuses and Surveys (OPCS, Classification of Surgical Operations v.4) [31], and Dictionary of Medicines and Devices (DM+D) [32]. Although there is no consensus yet on phenotype algorithms and code lists, researchers are working to create clinical standards [33]. Examples of well-known concept libraries are the HDRUK Phenotype Library [34], Caliber [35,36], ClinicalCode [37,38], and Phenotype Knowledgebase (PheKB) [39]. If phenotyping algorithms and related code lists are not available, they should be defined, developed, validated, shared, and reused. Phenotype definition, development, consistency, validation, and transfer, including algorithms to automate computable phenotypes and their integration with genotypes [40], are beyond the scope of this paper.

To make variable extraction as efficient as possible, Curator creates intermediate tables containing only those data relevant to the requested variables, which represent a fraction of the whole dataset. Fig. 4 shows an example of these tables for conditions, medications, and the measurements smoking status, drinking status, and body mass index (BMI).

The Variable Extractor uses these supporting tables to systematically capture the specific data indicated in the Variable Designer by the parameters, such as the domain, event time direction, event time window, or event timepoint.

3.4. Data sources

At the time of writing, Curator handles UK clinical RWD that are routinely collected in primary and secondary care settings and some registry data provided by the Clinical Practice Research Datalink (CPRD) [20].

- Primary care data (electronic health records from GP practices)
  - ➢ CPRD GOLD collected by Vision software [13,41].
  - ➢ CPRD AURUM collected by Emis software [42,43].
- Secondary care (electronic health records from Hospital Episode Statistics - HES)
  - ➢ CPRD HES Admitted Patients Care (APC) [44].
  - ➢ CPRD HES Accidents & Emergency (A&E) [45].
- Registries
  - ➢ CPRD Office of National Statistics (ONS) Mortality [46].
  - ➢ CPRD Index of Multiple Deprivation (IMD) at patient and practice level [47,48].

All supported datasets are supplied already deidentified and, if the study requires, linked, with no free text or narrative content to ensure patient confidentiality and data protection. To access RWD, each clinical study must submit a research protocol and receive ethical approval. CPRD data are obtained via its Research Data Governance process, previously the Independent Scientific Advisory Committee (ISAC). When the protocol is approved, nominated researchers are allowed to

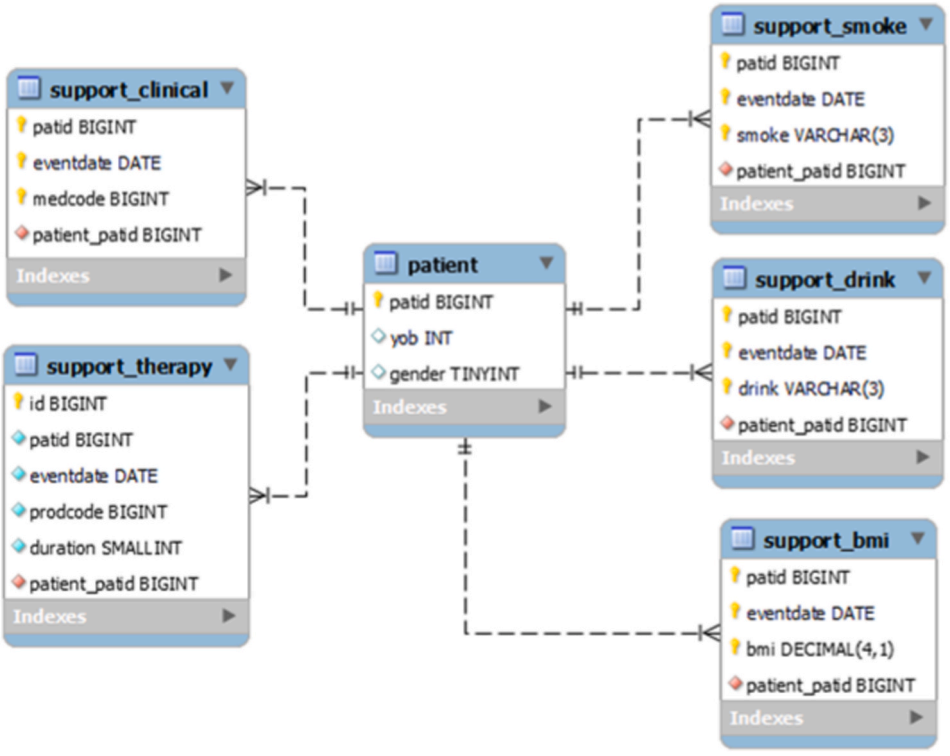


Fig. 4. Example of Curator's supporting tables.

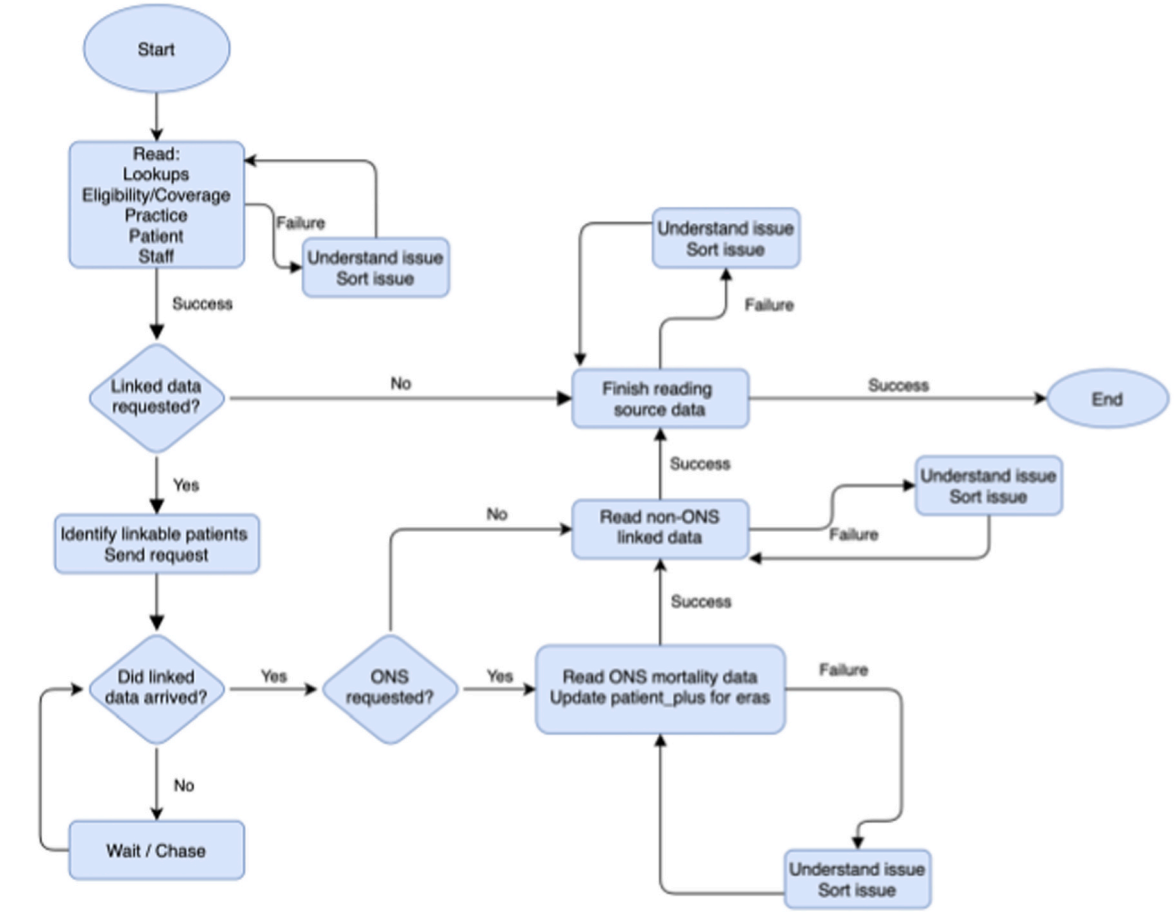


Fig. 5. Curator's data processing flow.

download primary care data for the selected individuals and request linked data, if applicable. The provision of linked data involves a trusted third party (e.g. NHS Digital) to ensure that neither the researchers nor the data provider have access to patient identifiable information.

The whole data loading schedule in Curator is complex (Fig. 5). Some linked information (e.g. ONS mortality registry) may invalidate other data and therefore must be processed before any other time-dependent data to enhance data quality. Whenever possible, processes are run in parallel to increase efficiency (e.g. loading RWD into different tables).

4. Results

Curator's application in clinical research and health economics has consistently grown since its initial development. By May 2023, it had been used in 25 studies whose output had been published in 2 NIHR-funded reports and 33 prestigious international peer-reviewed journals and presented at 38 global conferences, while other 3 studies are close to presenting their results. Curator's research output covers common chronic conditions [49–59], rare diseases [60–69], hospital interventions [70–88], drug safety and pharmacovigilance [89–110], and indexes of frailty and multi-morbidities in the elderly [111–121]. Curator has an extensive clinical and pharmacologic scope, ranging from widespread diseases (e.g. cardiovascular disease, cancer, diabetes, chronic lower respiratory disease, osteoporosis, peptic ulcer, and rheumatoid arthritis) to rare illnesses (e.g. X-linked hypophosphataemia and myeloma), from surgical procedures (e.g. hip and knee replacement, bariatric surgery, shoulder dislocation, kidney dialysis, and blood transfusion) to pharmacoepidemiology (e.g. statins, antihypertensives, antidepressants, systemic hormone replacement therapy, opioids, systemic corticosteroids, oral bisphosphonates, and antibiotics).

Curator has been used for clinical research and health economics in the Medical Sciences Division of the University of Oxford, where it was developed. It has also been used by a network of collaborators, including the Medical School of the University of Bristol, the School of Health Sciences of the University of Southampton, and the School of Medicine of the University of Leeds. Curator has supported the research of 5 DPhil and 5 MSc students. Most of Curator's studies have been funded by the NIHR, three by two industry partners (UCB Celltech and Amgen Limited), one by Versus Arthritis Charity, and one by The College of Podiatry.

Curator consistently reduces resource use, time, and costs. We are making secondary use of data that were originally collected for administrative and claims purposes and therefore contain lots of information that is not related to research. Curator only loads into the source database data that are relevant for research, compared with the standard approach of loading the full dataset. Curator therefore uses 44% less storage per record than the standard approach, when considering its metadata for each data source and MySQL data type storage requirements [122]. For example, "col_position" in Fig. 3 shows which fields are imported in the Patient table from the Patient files of a data source. The original 20 fields would use 49 bytes per record, based on MySQL data type storage requirements. Curator only uses 34 bytes: it loads 12 fields (30 bytes), adds 1 integer field (4 bytes), then infers its content to enhance query performance. In this Patient table, Curator uses around 30% less storage per record than the standard approach. The processing time is also reduced by a similar proportion as less data are loaded.

Curator automatically identifies records with unacceptable data quality at loading time and stores them in separate twin tables named <table_name>+"_bin", so that they can be tested and measured during development. Examples of unacceptable quality are records where the event date is missing or impossible (e.g. future time), records that are duplicated (e.g. two diagnoses of the same condition for the same patient on the same date), records reporting measurement values that are clearly mistaken (i.e. completely out of clinical scales), or records that conflict in a non-reconcilable way (e.g. person is reported as *smoking* and *not smoking* on the same day). The bin tables include a column for the reason for binning. Binning these records removes on average 8% of all records, resulting in a proportional decrease in variable extraction processing time than when using all available data.

A recent study calculated that data curation accounts for 80% of the work done in data science projects using RWD [123]. We considered all of our team's applicable RWD projects, 8 completed before Curator was developed and 28 managed with Curator's help, of which 25 have been published and 3 are close to publication. We classified the complexity of these 36 projects using three dimensions, each categorised as low, medium, or high complexity: number of patients (low: 1–300,000, medium: 300,001–400,000, high: 400,001–600,000), number of variables (low: 1–250, medium: 251–500, high: 501–1017), and number of linked datasets (low: 0–1, medium: 2–4, high: $\geq$ 5). 22 projects had medium complexity in all three dimensions (medium-complexity projects) and no more than 3 projects fell into any of the other 26 dimension combinations. As a proxy for the time spent on these studies by statisticians, senior statisticians, and Curator's users, whose job titles are database programmers, we used the time allocated for them in the awarded research grant applications plus any applicable time extensions. These professionals were identified via their University of Oxford employment grade (grade 7, grade 8 and grade 7, respectively), which serves as a proxy for their expertise and experience. Statisticians and database programmers are in the same employment grade, and therefore their salaries are comparable. Other universities may use different employment grade numbers or names, but would use comparable job descriptions.

Table 1 shows the typical time allocated for these professionals in the awarded grant applications for our 22 medium-complexity projects (between 300,001 and 400,000 patients, between 2 and 4 linked datasets, and between 251 and 500 curated variables), before and after Curator's development.

SP1 was allocated between 3 and 6 months of statistician time before Curator and between 0.8 and 1.6 months of a database programmer's time. Curator saved 2.2–4.4 months during SP1. Before Curator, SP2 was allocated between 9.6 and 12 months of statistician time, supervised by a senior statistician. Once Curator could provide the variables fully curated and ready for analysis, SP2 was allocated between 7.2 and 9.6

Table 1
Typical time allocated for staff in medium-complexity projects.

	SP1	SP2		SP1 + SP2
Months Without Curator	[3 – 6] *100% statistician	12 * [70 – 80]% statistician + 12 * [10 – 20]% senior statistician	[8.4 – 9.6] + [1.2 – 2.4] = [9.6 – 12]	[12.6 – 18]
Months With Curator	[1 – 2] * 80% Curator's user = [0.8 – 1.6]	12 * [50 – 60]% statistician + 12 * [10 – 20]% senior statistician	[6.0 – 7.2] + [1.2 – 2.4] = [7.2 – 9.6]	[8.0 – 11.2]
Months saved	[2.2 – 4.4]	12 * 20% statistician = 2.4		[4.6 – 6.8]

months. Curator saved an average of 2.4 months during SP2. It saved an average total of 4.6–6.8 months per project, between 36.5% and 37.8%. Based on UK University average wages for a statistician or database programmer of £38,000 per year [124], the staff time reduced by using Curator saves between £14,600 and £21,500 per medium-complexity project.

Creating advanced data curation software that operates separately to the other statistical tasks allows for a systematic approach to RWD statistical analyses that enhances transparency and reproducibility [125]. Transparency is increased in the whole statistical pipeline, reproducibility is boosted in data curation, and significant improvements are made in the reproducibility of the subsequent statistical tasks, as it becomes much easier to code them.

Curator's transparency is supported by the following features:

- Source data are never altered, and data provenance is maintained to the point of data delivery. Source data can always be accessed retrospectively to verify curated data, which are delivered in a newly created and tailored database.
- Records that are discarded are saved in twin bin tables for auditing, where an extra column reports the reason for binning.
- The metadata that drive the source database building are stored in a relational database, which is easy to understand and interpret.
- The code is written in Python, which is an interpreted (vs compiled) programming language, well regarded for its code readability. Curator's code writing style is consistent and very comprehensible, also in view of its long-term maintenance.
- Algorithms used to reconcile conflicting information have been published [24] or are in the process of being published.
- Log files are automatically created reporting the complete workflow, including the start and exit points for methods, queries, and row counts for SQL statements.

Results reproducibility is strengthened by the following characteristics:

- Curator's code is deterministic. It automates complex and repetitive tasks and minimises human intervention.
- Curator's code is written in Python, which is a cross-platform programming language that can be run on different platforms, producing consistent results with the same input data.
- Curator's code is developed with a hard coded Python version (currently v. 3.9) and without the use of any framework, to minimize the possibility of dependency conflicts.

However, transparency and reproducibility are complex and evolving research topics, and we will continue to investigate the potential for further improvements.

5. Discussion

Recently, many initiatives have appeared to evaluate and facilitate the secondary use of RWD. Fig. 6 summarises some relevant approaches, but is by no means comprehensive.

We call "*Service providers*" (e.g. trusted research environments (TREs) such as OpenSafely [126], SAIL databank [127], SRS [128], DAE [129], IQVIA E360 [130], and Aetion Evidence Platform [131]) those companies that offer users ways to analyse RWD remotely. We refer to "*Data providers*" (e.g. CPRD [20], THIN [132], and QResearch [133]) as those bodies that allow users to download deidentified RWD for analysis and work on those data locally with their own solutions. Curator is one of these solutions.

Service providers are more expensive than data providers, as remote access requires the development and maintenance of expensive informatic infrastructures, continuing data management and curation, and user training on platforms that are heterogeneous and change over time.

When using data providers, a variety of approaches can be taken to perform statistical analyses on RWD, from eEHR R package [134] and DExtER extract, transform and load (ETL) based tool [135] to massive analytics infrastructures based on common data models (e.g. Sentinel [136], i2B2 [137], and OMOP [138]) to enable network studies and federated analyses among collaborators. The classification and comparison of these solutions are beyond the scope of this article. However, several of them require big investments and informatics capabilities. They all use an embedded analytics layer that requires users to have substantial epidemiological understanding and familiarity with the specific data model and/or interface.

In this complex and fast-changing climate, Curator has a more specific, focused objective that aligns with recent UK governmental guidelines [123] that recognise "*health data curation as a complex, standalone, high status technical challenge*" and encourage the provision of systematic curation as an individual, reproducible step in the analytical pipeline.

5.1. Benefits of curator

Curator benefits from years of embedded data source knowledge and epidemiological expertise, which are not easy or fast to gain. Curator also reports a significant number of peer-reviewed publications, which makes it reliable and trustworthy. By choosing Curator, researchers can progress quickly and safely, as they do not need to acquire or apply data-specific knowledge, can avoid mistakes, and can focus on clinical questions and analyses designed to answer them.

Curator does not need expensive IT infrastructures and can be sustainably used by research groups with limited funding, which is frequently the case in academia. When compared to a traditional statistical approach, Curator decreases research time and costs for medium-complexity projects by more than 36% and improves research

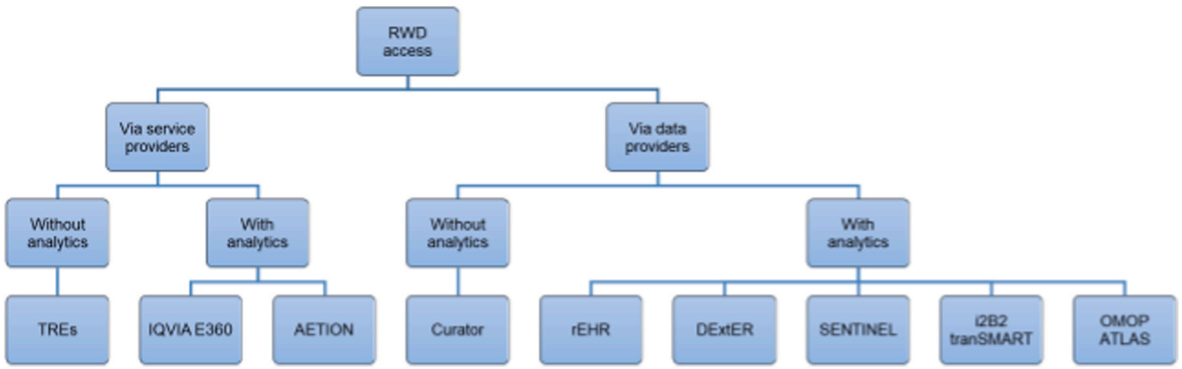


Fig. 6. Approaches to RWD

reproducibility and transparency. Curator can extract both primary care and linked data (e.g. secondary care, death registry), when the latter are part of an approved research protocol. It separates data curations from statistical analysis, generating an output that can be taken forward for analysis. This full task separation allows for enhanced control of the pipeline, the development of novel statistical methods, statistical training, and student practice.

5.2. Limitations

Some limitations should be recognised with Curator. The tool currently requires researchers with computing skills to configure some of the programs. We aim to fully automate the software in future development, as explained in section 5.3. We also aim to add more features to enhance Curator's efficacy, such as more phenotyping algorithms, related code lists, and measurement tests.

Curator has only been used with RWD from the UK thus far. However, these data are rich and have been used globally for clinical research. Considering Curator's metadata approach, we expect that minimal adaptation should be required to accommodate other data sources in the future.

The tool has also mostly been used for orthopaedical, rheumatological, and musculoskeletal science research. However, epidemiological covariates in studies conducted with Curator have widely spread through the clinical and pharmaceutical spectrum, making the clinical and pharmacologic scope of Curator substantial.

As the data science landscape is moving fast, with involvement from powerful stakeholders, Curator might struggle to hold its place. It is difficult to predict long-term national and international data policy developments. However, we believe that in the short- and medium-term Curator should continue to empower observational researchers, especially in academic environments where education and training are particularly valued, and funding is limited.

5.3. Future work

Curator was originally created as a set of command line PowerShell Python programs that required technical computing skills and specific know-how to use. Curator is now undergoing a significant upgrade to become more user-friendly. We aim to equip Curator with a graphical user interface that is sophisticated enough to cope with the Variable Designer and its detailed definitions of analysis variables. Curator can then be used by third parties and a wider range of nominated researchers with various expertise from diverse backgrounds, including epidemiologists and clinicians. Curator's interface will be able to manage several projects and multiple variable designs per project at the same time, allowing different work packages to have their own tailored and curated data extract, if appropriate. Each study will have one superuser in charge of the Source Database Builder and Variable Extractor and a flexible number of nominated collaborators working on the Variable Designer, increasing efficiency and study quality. We also aim to provide Curator with an MSI installation for a Windows 64-bit environment, a technical manual, and a user guide. The installer will include a Python distribution and a MySQL release to enhance reproducibility.

6. Conclusions

This paper has introduced Curator, a RWD curation tool that decreases research time and costs for medium-complexity projects by more than 36%. It is particularly useful when funding and IT resources are limited, as is true for many academic research groups. Curator increases reproducibility and transparency in observational research. It has contributed to 25 completed observational studies so far, whose output has been peer-reviewed and widely published. Future work will focus on making Curator available to third parties by equipping it with a graphical user interface, MSI installer, technical manual, and user guide.

Funding

The research was supported by the National Institute for Health and Care Research (NIHR) Oxford Biomedical Research Centre (BRC). Improvements for Curator were supported by the University of Oxford's Engineering and Physical Sciences Research Council (EPSRC) Impact Acceleration Account (IAA) Award EP/R511742/1. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The funders did not have any role in the study design, collection, analysis and interpretation of data, the writing of the report, or in the decision to submit the article for publication.

Contributors

Antonella Delmestri: Funding acquisition, Conceptualisation, Methodology, Software, Validation, Supervision, Writing- Original draft preparation, Writing- Reviewing and Editing. Daniel Prieto-Alhambra: Funding acquisition, Writing- Reviewing and Editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Daniel Prieto-Alhambra reports administrative support provided by NIHR Oxford Biomedical Research Centre. Antonella Delmestri reports financial support provided by the Engineering and Physical Sciences Research Council.

Curator's IP rights are assigned to Oxford University Innovation, which is supportive of this manuscript.

Acknowledgments

We gratefully thank Mrs Paloma O'Dogherty Cordero for her contribution to creating code lists for Curator's repository and Ms Wai Man for her assistance on the paper's images. We also thank Dr Danielle Robinson and Dr Annika Jödicke for their expert opinion on observational study designs. We are grateful to Dr Jennifer de Beyer for English language editing of this manuscript.

References

[1] U.S. Food & Drug Administration. Real-world evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. [Accessed 4 May 2023].
[2] Li X, Lo Re V, Toh S. Profiling real-world data sources for pharmacoepidemiologic research: a call for papers. Pharmacoepidemiol Drug Saf 2022;31(9):929–31. https://doi.org/10.1002/pds.5481.
[3] Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet 2002;359(9300):57–61. https://doi.org/10.1016/S0140-6736(02)07283-5.
[4] Deaton A, Cartwright N. Understanding and misunderstanding randomized controlled trials. Soc Sci Med 2018;210:2–21. https://doi.org/10.1016/j.socscimed.2017.12.005.
[5] Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021;374:n2061. https://doi.org/10.1136/bmj.n2061.
[6] Terry AL, Chevendra V, Thind A, Stewart M, Marshall JN, Cejic S. Using your electronic medical record for research: a primer for avoiding pitfalls. Fam Pract 2010;27(1):121–6. https://doi.org/10.1093/fampra/cmp068.
[7] Menni C, Klaser K, May A, Polidori L, Capdevila J, Louca P, et al. Vaccine side-effects and SARS-CoV-2 infection after vaccination in users of the COVID Symptom Study app in the UK: a prospective observational study. Lancet Infect Dis 2021;21(7):939–49. https://doi.org/10.1016/s1473-3099(21)00224-3.
[8] Li X, Ostropolets A, Makadia R, Shoaibi A, Rao G, Sena AG, et al. Characterising the background incidence rates of adverse events of special interest for covid-19 vaccines in eight countries: multinational network cohort study. BMJ 2021;373. https://doi.org/10.1136/bmj.n1435.
[9] Li X, Raventos B, Roel E, Pistillo A, Martinez-Hernandez E, Delmestri A, et al. Association between covid-19 vaccination, SARS-CoV-2 infection, and risk of immune mediated neurological events: population based cohort and self-controlled case series analysis. BMJ 2022;376:e068373. https://doi.org/10.1136/bmj-2021-068373.

[10] McDonald CJ, Murray R, Jeris D, Bhargava B, Seeger J, Blevins L. A computer-based record and clinical monitoring system for ambulatory care. Am J Publ Health 1977;67(3):240–5. https://doi.org/10.2105/ajph.67.3.240.
[11] Tyrer F, Hambleton I, Lawrenson R, Pierce M. Building a research database from computerised general practice records. J Infor Prim Care. 1996;September:8-13.
[12] Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2016;Suppl 1(Suppl 1):S48–61. https://doi.org/10.15265/IYS-2016-s006.
[13] Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data resource profile: clinical practice research datalink (CPRD). Int J Epidemiol 2015;44(3):827–36. https://doi.org/10.1093/ije/dyv098.
[14] Wood L, Martinez C. The general practice research database. Drug Saf 2004;27(12):871–81. https://doi.org/10.2165/00002018-200427120-00004.
[15] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inf Assoc 2013;20(1):144–51. https://doi.org/10.1136/amiajnl-2011-000681.
[16] Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev 2010;67(5):503–27. https://doi.org/10.1177/1077558709359007.
[17] Python Software Foundation. Python Documentation contents. 2023. https://docs.python.org/3/contents.html. [Accessed 4 May 2023].
[18] Corporation Oracle. MySQL 8.0 release notes. 2023. https://dev.mysql.com/doc/relnotes/mysql/8.0/en/. [Accessed 4 May 2023].
[19] Pavlov I. 7-Zip. 2023. https://www.7-zip.org/. [Accessed 4 May 2023].
[20] Medicines and healthcare products regulatory agency (MHRA). Clinical practice research Datalink (CPRD). 2023. https://cprd.com/. [Accessed 4 May 2023].
[21] Oxford University Innovation. Oxford University Innovation; 2023. https://innovation.ox.ac.uk/. [Accessed 4 May 2023].
[22] Brownlee J. Python multiprocessing: the complete guide. 2022. https://superfastpython.com/multiprocessing-in-python/.
[23] Davé S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. Pharmacoepidemiol Drug Saf 2009;18(8):704–7. https://doi.org/10.1002/pds.1770.
[24] Delmestri A, Prieto-Alhambra D. CPRD GOLD and linked ONS mortality records: reconciling guidelines. Internet J Med Inf 2020;136:104038. https://doi.org/10.1016/j.ijmedinf.2019.104038.
[25] Ulrich H, Kock-Schoppenhauer AK, Deppenwiese N, Gött R, Kern J, Lablans M, et al. Understanding the nature of metadata: systematic review. J Med Internet Res 2022;24(1):e25440. https://doi.org/10.2196/25440.
[26] Watson J, Nicholson BD, Hamilton W, Price S. Identifying clinical features in primary care electronic health record studies: methods for codelist development. BMJ Open 2017;7(11):e019637. https://doi.org/10.1136/bmjopen-2017-019637.
[27] Williams R, Brown B, Kontopantelis E, van Staa T, Peek N. Term sets: a transparent and reproducible representation of clinical code sets. PLoS One 2019;14(2):e0212291. https://doi.org/10.1371/journal.pone.0212291.
[28] SNOMED International. SNOMED CT. https://www.snomed.org/. [Accessed 4 May 2023].
[29] NHS Digital. Read Codes Version 3, https://isd.digital.nhs.uk/trud/users/guest/filters/2/categories/9/items/19/releases. [Accessed 4 May 2023].
[30] World Health Organization. International classification of diseases (ICD-10). https://icd.who.int/browse10/2019/en. [Accessed 4 May 2023].
[31] NHS Digital. NHS classifications OPCS-4. https://isd.digital.nhs.uk/trud/user/guest/group/0/pack/10. [Accessed 4 May 2023].
[32] NHS. Dictionary of medicines and devices (dm+d). https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/dictionary-medicines-and-devices-dmd. [Accessed 4 May 2023].
[33] Almowil ZA, Zhou S-M, Brophy S. Concept libraries for automatic electronic health record based phenotyping: a review. Int J Popul Data Sci 2021;6(1):1362. https://pubmed.ncbi.nlm.nih.gov/34189274/.
[34] HDR-UK. Phenotype library. https://phenotypes.healthdatagateway.org/phenotypes/. [Accessed 4 May 2023].
[35] HDR-UK. Caliber. https://www.hdruk.ac.uk/case-studies/caliber/. [Accessed 4 May 2023].
[36] Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inf Assoc 2019;26(12):1545–59. https://doi.org/10.1093/jamia/ocz105.
[37] The University of Manchester. Institute of population health UK. ClinicalCodes.org. https://clinicalcodes.rss.mhs.man.ac.uk/. [Accessed 4 May 2023].
[38] Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, 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. https://doi.org/10.1371/journal.pone.0099825.
[39] Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inf Assoc 2016;23(6):1046–52. https://doi.org/10.1093/jamia/ocv202.
[40] Frey LJ. Artificial intelligence and integrated genotype–phenotype identification. Genes 2019;10(1):18. https://doi.org/10.3390/genes10010018.
[41] Medicines and Healthcare products Regulatory Agency (MHRA). CPRD GOLD data specification. 2023. https://cprd.com/sites/default/files/2022-10/CPRD%20GOLD%20Full%20Data%20Specification%20v2.5%20update.pdf. [Accessed 4 May 2023].
[42] Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, et al. Data resource profile: clinical practice research datalink (CPRD) aurum. Int J Epidemiol 2019:1–8. https://doi.org/10.1093/ije/dyz034.
[43] Medicines and Healthcare products Regulatory Agency (MHRA). CPRD aurum data specification. 2023. https://cprd.com/sites/default/files/2023-04/CPRD%20Aurum%20Data%20Specification%20v2.9.pdf. [Accessed 4 May 2023].
[44] Medicines and Healthcare products Regulatory Agency (MHRA). Hospital Episode statistics (HES) admitted patient care and CPRD primary care data documentation. 2023. https://cprd.com/sites/default/files/2022-02/Documentation_HES_APC_set22.pdf. [Accessed 4 May 2023].
[45] Medicines and Healthcare products Regulatory Agency (MHRA). Hospital Episode statistics (HES) accident & emergency and CPRD primary care data documentation. 2023. https://cprd.com/sites/default/files/2022-02/Documentation_HES_AE_set21.pdf. [Accessed 4 May 2023].
[46] Medicines and Healthcare products Regulatory Agency (MHRA). ONS death registration data and CPRD primary care data Documentation. 2023. https://cprd.com/sites/default/files/2022-02/Documentation_Death_set22_v2.6.pdf. [Accessed 4 May 2023].
[47] Medicines and Healthcare products Regulatory Agency (MHRA). Small area level data based on patient postcode. 2023. https://cprd.com/sites/default/files/2022-05/Documentation_SmallAreaData_Patient_set22_v3.3.pdf. [Accessed 4 May 2023].
[48] Medicines and Healthcare products Regulatory Agency (MHRA). Small area level data based on practice postcode. 2023. https://cprd.com/sites/default/files/2022-05/Documentation_SmallAreaData_Practice_set22_v3.4_1.pdf. [Accessed 4 May 2023].
[49] Judge A, Javaid MK, Leal J, Hawley S, Drew S, Sheard S, et al. Models of care for the delivery of secondary fracture prevention after hip fracture: a health service cost, clinical outcomes and cost-effectiveness study within a region of England, 4. Health Services and Delivery Research; 2016. https://doi.org/10.3310/hsdr04280.
[50] Leal J, Gray AM, Hawley S, Prieto-Alhambra D, Delmestri A, Arden NK, et al. Cost-Effectiveness of orthogeriatric and fracture liaison service models of care for hip fracture patients: a population-based study. J Bone Miner Res 2017;32(2):203–11. https://doi.org/10.1002/jbmr.2995.
[51] Shah A, Prieto-Alhambra D, Hawley S, Delmestri A, Lippett J, Cooper C, et al. Geographic variation in secondary fracture prevention after a hip fracture during 1999-2013: a UK study. Osteoporos Int 2017;28(1):169–78. https://doi.org/10.1007/s00198-016-3811-4.
[52] Metcalfe D, Masters J, Delmestri A, Judge A, Perry D, Zogg C, et al. Coding algorithms for defining Charlson and Elixhauser co-morbidities in Read-coded databases. BMC Med Res Methodol 2019;19(1):115. https://doi.org/10.1186/s12874-019-0753-5.
[53] Ferguson R, Culliford D, Prieto-Alhambra D, Pinedo-Villanueva R, Delmestri A, Arden N, et al. Encounters for foot and ankle pain in UK primary care: a population-based cohort study of CPRD data. Br J Gen Pract 2019;69(683):e422–9. https://doi.org/10.3399/bjgp19X703817.
[54] Khalid S, Ernst MT, Javaid M, Libanti C, Cooper C, Delmestri A, et al. Imminent (1- and 2-year) fracture risk following a first (sentinel) fracture: a multi-national European cohort study including over 700,000 participants from Denmark, Spain, and the UK. Osteoporos Int 2019;30(S2):S166. https://doi.org/10.1007/s00198-019-04990-z.
[55] Ferguson RJ, Prieto-Alhambra D, Walker C, Yu D, Valderas JM, Judge A, et al. Validation of hip osteoarthritis diagnosis recording in the UK clinical practice research datalink. Pharmacoepidemiol Drug Saf 2019;28(2):187–93. https://doi.org/10.1002/pds.4673.
[56] Lowe KE, Mansfield KE, Delmestri A, Smeeth L, Roberts A, Abuabara K, et al. Atopic eczema and fracture risk in adults: a population-based cohort study. J Allergy Clin Immunol 2020;145(2):563–571 e8. https://doi.org/10.1016/j.jaci.2019.09.015.
[57] Khalid S, Pineda-Moncusi M, El-Hussein L, Delmestri A, Ernst M, Smith C, et al. Predicting imminent fractures in patients with a recent fracture or starting oral bisphosphonate therapy: development and international validation of prognostic models. J Bone Miner Res 2021;36(11):2162–76. https://doi.org/10.1002/jbmr.4414.
[58] Khalid S, Reyes C, Ernst M, Delmestri A, Toth E, Libanati C, et al. One- and 2-year incidence of osteoporotic fracture: a multi-cohort observational study using routinely collected real-world data. Osteoporos Int 2022;33(1):123–37. https://doi.org/10.1007/s00198-021-06077-0.
[59] Pineda-Moncusí M, El-Hussein L, Delmestri A, Cooper C, Moayyeri A, Libanati C, et al. Estimating the incidence and key risk factors of cardiovascular disease in patients at high risk of imminent fracture using routinely collected real-world data from the UK. J Bone Miner Res 2022;37(10):1986–96. https://doi.org/10.1002/jbmr.4648.
[60] Javaid MK, Delmestri A, Shaw N, Prieto-Alhambra D, Cooper C, Pinedo-Villanueva R. X-Linked hypophosphataemia: burden of disease using UK primary care data. Osteoporos Int 2018;29(S1):S277. https://doi.org/10.1007/s00198-018-4465-1.
[61] Hawley S, Delmestri A, Prieto-Alhambra D, Pinedo-Villanueva R, Javaid M, Shaw N, et al. X-Linked hypophosphataemia: prevalence and mortality within the United Kingdom. J Bone Miner Res 2019;34(S1):373–4. https://doi.org/10.1002/jbmr.3936.
[62] Hawley S, Shaw NJ, Delmestri A, Prieto Alhambra D, Cooper C, Pinedo Villanueva R, et al. X-linked hypophosphataemia: prevalence and mortality rate within the United Kingdom Clinical Practice Research Datalink. Osteoporos Int 2019;30(S2):S406–7. https://doi.org/10.1007/s00198-019-04993-w.
[63] Hawley S, Shaw NJ, Delmestri A, Prieto-Alhambra D, Cooper C, Pinedo-Villanueva R, et al. Prevalence and mortality of individuals with X-linked

hypophosphatemia: a United Kingdom real-world data analysis. J Clin Endocrinol Metab 2020;105(3):e871. https://doi.org/10.1210/clinem/dgz203. e78.
[64] Hawley S, Kishore B, Rabin NK, Yong K, Ashcroft J, Bowcock S, et al. Overall fracture incidence in multiple myeloma patients relative to general population controls: a parallel cohort study of the UK CPRD. Osteoporos Int 2020;31(S1): S483. https://doi.org/10.1007/s00198-020-05696-3.
[65] Hawley S, Kishore B, Rabin NK, Yong K, Ashcroft J, Bowcock S, et al. Imminent fracture incidence among index fracture patients with multiple myeloma relative to general population controls: a parallel cohort study of the UK CPRD. Bone Research Society 2020;31(S1):S484. https://doi.org/10.1007/s00198-020-05696-3.
[66] Hawley S, Shaw N, Delmestri A, Prieto-Alhambra D, Cooper C, Pinedo-Villanueva R, et al. Characterization of comorbidity in X-linked hypophosphatemia: a prospective parallel cohort study using the UK CPRD. J Bone Miner Res 2020;35(S1):286. https://doi.org/10.1002/jbmr.4206.
[67] Hawley S, Shaw NJ, Delmestri A, Prieto-Alhambra D, Cooper C, Pinedo-Villanueva R, et al. Characterization of comorbidity in X-linked hypophosphataemia: a prospective parallel cohort study using the UK CPRD. Osteoporos Int 2020;31(S1):S546–7. https://doi.org/10.1007/s00198-020-05696-3.
[68] Hawley S, Shaw NJ, Delmestri A, Prieto-Alhambra D, Cooper C, Pinedo-Villanueva R, et al. Higher prevalence of non-skeletal comorbidity related to X-linked hypophosphataemia: a UK parallel cohort study using CPRD. Rheumatology 2021;60(9):4055–62. https://doi.org/10.1093/rheumatology/keaa859.
[69] Maronga C, Javaid MK, Pinedo-Villanueva R. Prevalence and risk factors of hypertension among patients with X-linked hypophosphatemia. ECTS Congress featuring BRS Annual Meeting 2023;7(S3):e10738. https://doi.org/10.1002/jbm4.10738.
[70] Hawley S, Delmestri A, Judge A, Edwards CJ, Cooper C, Arden NK, et al. Total hip and knee replacement among incident osteoarthritis and rheumatoid arthritis patients within the UK Clinical Practice Research Datalink (CPRD) compared to Hospital Episode Statistics (HES): a validation study. Pharmacoepidemiol Drug Saf 2016;25(S3):251. https://doi.org/10.1002/pds.4070.
[71] Shah A, Judge A, Delmestri A, Edwards K, Arden NK, Prieto-Alhambra D, et al. Incidence of shoulder dislocations in the UK, 1995-2015: a population-based cohort study. BMJ Open 2017;7(11):e016112. https://doi.org/10.1136/bmjopen-2017-016112.
[72] Hawley SJ, Edwards CJ, Arden NK, Delmestri A, Cooper C, Judge A, et al. Descriptive epidemiology of hip and knee replacement among rheumatoid arthritis patients in England and Wales: variation by age, sex, geography and socioeconomic status. Rheumatology 2017;56(S2):ii111–i112. https://doi.org/10.1093/rheumatology/kex062.157.
[73] Rees JL, Shah A, Edwards K, Sanchez-Santos MT, Robinson DE, Delmestri A, et al. Treatment of first-time traumatic anterior shoulder dislocation: the UK TASH-D cohort study. Health Technol Assess 2019;23:1–104. https://doi.org/10.3310/hta23180.
[74] Matharu GS, Mouchti S, Twigg S, Delmestri A, Murray DW, Judge A, et al. The effect of smoking on outcomes following primary total hip and knee arthroplasty: a population-based cohort study of 117,024 patients. Acta Orthop 2019;90(6): 559–67. https://doi.org/10.1080/17453674.2019.1649510.
[75] Robinson DE, Douglas I, Tan GD, Cooper C, Delmestri A, Javaid MK, et al. Bariatric surgery increases the risk of major osteoporotic fracture: a self-controlled case series patients from the CPRD GOLD database linked to HES. Pharmacoepidemiol Drug Saf 2019;28(S2):492. https://doi.org/10.1002/pds.4864.
[76] Robinson DE, Douglas I, Tan GD, Cooper C, Delmestri A, Javaid MK, et al. Fracture risk following bariatric surgery: a self-controlled case series and risk prediction algorithm. Ann Rheum Dis 2019;78(S2):933. https://doi.org/10.1136/annrheumdis-2019-eular.2437.
[77] Robinson DE, Douglas I, Tan GD, Cooper C, Delmestri A, Javaid MK, et al. Major osteoporotic fracture risk following bariatric surgery: a self-controlled case series including 5492 people from the UK CPRD and linked HES database. Osteoporos Int 2019;30(S2):S541–2. https://doi.org/10.1007/s00198-019-04993-w.
[78] Hawley S, Edwards CJ, Arden NK, Delmestri A, Cooper C, Judge A, et al. Descriptive epidemiology of hip and knee replacement in rheumatoid arthritis: an analysis of UK electronic medical records. Semin Arthritis Rheum 2020;50(2): 237–44. https://doi.org/10.1016/j.semarthrit.2019.08.008.
[79] Kolovos S, Burn E, Delmestri A, Smith L, Kingsbury S, Stone M, et al. Impact of long-term follow-up on revision of knee and hip replacement surgery. Osteoporos Int 2021;31(S1):S364–5. https://doi.org/10.1007/s00198-020-05696-3.
[80] Pinedo-Villanueva R, Delmestri A, Smith L, Kingsbury S, Stone M, Conaghan P, et al. Primary care consultations after long-term follow-up of unrevised hip and knee replacements: findings from routinely collected healthcare records in the UK. Osteoporos Int 2020;31(S1):S331–2. https://doi.org/10.1007/s00198-020-05696-3.
[81] Robinson DE, Douglas I, Tan GD, Delmestri A, Judge A, Cooper C, et al. Bariatric surgery increases the rate of major fracture: self-controlled case series study in UK Clinical Practice Research Datalink. J Bone Miner Res 2021;36(11):2153–61. https://doi.org/10.1002/jbmr.4405.
[82] Mohammad HR, Gooberman-Hill R, Delmestri A, Broomfield J, Patel R, Huber J, et al. Risk factors associated with poor pain outcomes following primary knee replacement surgery: analysis of data from the clinical practice research datalink, hospital episode statistics and patient reported outcomes as part of the STAR research programme. PLoS One 2021;16(12):e0261850. https://doi.org/10.1371/journal.pone.0261850.
[83] Pinedo-Villanueva R, Delmestri A, Smith L, Kingsbury S, Stone M, Conaghan P, et al. Primary care consultations after long-term follow-up of unrevised hip and knee replacements: findings from routinely collected healthcare records in the UK. Osteoporos Int 2020;31(S1):S331–2. https://doi.org/10.1007/s00198-020-05696-3.
[84] Leal J, Murphy J, Garriga C, Delmestri A, Rangan A, Price A, et al. Costs of joint replacement in osteoarthritis: a study using the national joint registry and clinical practice research datalink data sets. Arthritis Care Res 2022;74(3):392–402. https://doi.org/10.1002/acr.24470.
[85] Smith LK, Garriga C, Kingsbury SR, Pinedo-Villanueva R, Delmestri A, Arden NK, et al. UK poSt Arthroplasty Follow-up rEcommendations (UK SAFE): what does analysis of linked, routinely collected national data sets tell us about mid-late term revision risk after hip replacement? Retrospective cohort study. BMJ Open 2022;12(3):e050877. https://doi.org/10.1136/bmjopen-2021-050877.
[86] Smith LK, Garriga C, Kingsbury SR, Pinedo-Villanueva R, Delmestri A, Arden NK, et al. UK poSt Arthroplasty Follow-up rEcommendations (UK SAFE): what does analysis of linked, routinely collected national datasets tell us about mid-late term revision risk after knee replacement? BMJ Open 2022;12(3):e046900. https://doi.org/10.1136/bmjopen-2020-046900.
[87] Pinedo-Villanueva R, Kolovos S, Maronga C, Delmestri A, Howells N, Judge A, et al. Primary care consultations and pain medicine prescriptions: a comparison between patients with and without chronic pain after total knee replacement. BMC Muscoskel Disord 2022;23(548):1–9. https://doi.org/10.1186/s12891-022-05492-6.
[88] Pinedo-Villanueva R, Kolovos S, Burn E, Delmestri A, Smith LK, Judge A, et al. Association between outpatient follow-up and incidence of revision after knee and hip replacements: a population-based cohort study. BMC Muscoskel Disord 2023;24(106):1–9. https://doi.org/10.1186/s12891-023-06190-7.
[89] Hawley S, Leal J, Delmestri A, Prieto-Alhambra D, Arden NK, Cooper C, et al. Anti-osteoporosis medication prescriptions and incidence of subsequent fracture among primary hip fracture patients in england and wales: an interrupted time-series analysis. J Bone Miner Res 2016;31(11):2008–15. https://doi.org/10.1002/jbmr.2882.
[90] Hawley S, Cordtz R, Dreyer L, Edwards CJ, Arden NK, Delmestri A, et al. The Impact of biologic therapy introduction on hip and knee replacement among rheumatoid arthritis patients: an interrupted time series analysis using the clinical practice research datalink. Arthritis Rheumatol 2016;68(S10). https://acrabstracts.org/abstract/the-impact-of-biologic-therapy-introduction-on-hip-and-knee-replacement-among-rheumatoid-arthritis-patients-an-interrupted-time-series-analysis-using-the-clinical-practice-research-datalink/.
[91] Hawley S, Leal J, Delmestri A, Prieto-Alhambra D, Arden NK, Cooper C, et al. Anti-osteoporosis medication prescriptions and incidence of secondary fracture amongst primary hip fracture patients in england and wales: an interrupted time series and economic analysis. Osteoporos Int 2016;27(S1):S60–1. https://doi.org/10.1007/s00198-016-3520-z.
[92] Shah A, Prieto-Alhambra D, Hawley S, Delmestri A, Lippett J, Cooper C, et al. Two-fold regional variation in initiation of anti-osteoporosis medication after hip fracture in the UK. Osteoporos Int 2016;27(S2):S628. https://doi.org/10.1007/s00198-016-3743-z.
[93] Martín-Merino E, Petersen I, Hawley S, Álvarez-Gutierrez A, Delmestri A, Llorente-Garcia A, et al. Risk of venous thromboembolism amongst users of different antiosteoporosis drugs: a multinational population-based cohort study. Osteoporos Int 2016;27(S1):S158. https://doi.org/10.1007/s00198-016-3530-x.
[94] Ali MS, Caskey FJ, Delmestri A, Dedman D, Arden N, Ben-Shlomo Y, et al. Oral bisphosphonate use and risk of acute kidney injury, gastrointestinal events and hypocalcaemia in patients with moderate-advanced chronic kidney disease: a population-based cohort study. J Bone Miner Res 2017;32(S1):S180. https://doi.org/10.1002/jbmr.3363.
[95] Garriga C, Judge A, Hawley S, Delmestri A, Prieto-Alhambra D, Cooper C, et al. Bisphosphonates and age-related macular degeneration: a propensity-matched cohort and nested case-control analysis. J Bone Miner Res 2017;32(S1):S364. https://doi.org/10.1002/jbmr.3363.
[96] Martín-Merino E, Petersen I, Hawley S, Alvarez-Gutierrez A, Khalid S, Llorente-Garcia A, et al. Risk of venous thromboembolism among users of different anti-osteoporosis drugs: a population-based cohort analysis including over 200,000 participants from Spain and the UK. Osteoporos Int 2018;29(2):467–78. https://doi.org/10.1007/s00198-017-4308-5.
[97] Garriga C, Pazianas M, Hawley S, Delmestri A, Prieto-Alhambra D, Cooper C, et al. Oral bisphosphonate use and age-related macular degeneration: retrospective cohort and nested case-control study. Ann N Y Acad Sci 2018;1415 (1):34–46. https://doi.org/10.1111/nyas.13589.
[98] Hawley S, Cordtz R, Dreyer L, Edwards CJ, Arden NK, Delmestri A, et al. Association between NICE guidance on biologic therapies with rates of hip and knee replacement among rheumatoid arthritis patients in England and Wales: an interrupted time-series analysis. Semin Arthritis Rheum 2018;47(5):605–10. https://doi.org/10.1016/j.semarthrit.2017.09.006.
[99] Ali MS, Robinson DE, Pallares N, Tebe C, Cooper C, Abrahamsen B, et al. The effect of oral bisphosphonates on acute kidney injury, gastrointestinal events and hypocalcaemia in patients with chronic kidney disease. Pharmacoepidemiol Drug Saf 2018;27(S2):184. https://doi.org/10.1002/pds.4629.
[100] Garriga C, Pazianas M, Hawley S, Delmestri A, PrietoAlhambra D, Cooper C, et al. Impact of bisphosphonates in age-related macular degeneration: retrospective cohort and nested case-control study. Osteoporos Int 2018;29(S1):S181–2. https://doi.org/10.1007/s00198-018-4465-1.
[101] Khalid S, Ernst M, Rubin KH, Martinez-Laguna D, Delmestri A, Javaid KM, et al. Secular trends in the initiation of therapy in secondary fracture prevention in

Europe: a multi-national study including data from Denmark, Spain, and the UK. Value Health 2018;21(S3):S299–300. https://doi.org/10.1016/j.jval.2018.09.1785.
[102] Nagra NS, Robinson DE, Douglas I, Delmestri A, Dakin SG, Snelling SJB, et al. Antibiotic treatment and flares of rheumatoid arthritis: a self-controlled case series study analysis using CPRD GOLD. Sci Rep 2019;9(1):8941. https://doi.org/10.1038/s41598-019-45435-1.
[103] Hawley S, Prieto-Alhambra D, Delmestri A, Arden N, Cooper C, Judge A, et al. Anti-osteoporosis medication prescriptions and incidence of secondary fracture amongst hip fracture patients in England and Wales: an age stratified interrupted time series analysis. Osteoporos Int 2018;29(S1):S302. https://doi.org/10.1007/s00198-018-4465-1.
[104] Robinson DE, Ali MS, Delmestri A, Prieto-Alhambra D. Negative control time window to identify residual confounding in comparative effectiveness study: an example of bisphosphonate use and risk of fracture. Pharmacoepidemiol Drug Saf 2018;27(S2):161. https://doi.org/10.1002/pds.4629.
[105] Strauss VY, Robinson DE, Ali MS, Tomlinson L, Cooper C, Caskey F, et al. Oral bisphosphonate use and latent class trajectories of kidney function: a cohort study. Pharmacoepidemiol Drug Saf 2018;27(S2):186. https://doi.org/10.1002/pds.4629.
[106] Khalid S, Ernst MT, Javaid MK, Libanati C, Cooper C, Delmestri A, et al. Imminent fracture risk amongst new users of oral bisphosphonates in actual practice settings: a multinational European cohort stydy from Denmank, Spain and the UK. Osteoporos Int 2019;30(S2):S487. https://doi.org/10.1007/s00198-019-04993-w.
[107] Nagra N, Robinson DE, Delmestri A, Dakin S, Snelling S, Carr AJ, et al. Antibiotic use and the development of rheumatoid arthritis (RA) and risk of RA flares: case-control and self-controlled case series studies in two national electronic patient databases (SIDIAP and CPRD). Arthritis Rheumatol 2019;71(S10). https://doi.org/10.1002/art.41108.
[108] Skjodt MK, Khalid S, Ernst M, Rubin KH, Martinez-Laguna D, Delmestri A, et al. Secular trends in the initiation of therapy in secondary fracture prevention in Europe: a multi-national cohort study including data from Denmark, Catalonia, and the United Kingdom. Osteoporos Int 2020;31(8):1535–44. https://doi.org/10.1007/s00198-020-05358-4.
[109] Alarkawi D, Ali MS, Bliuc D, Pallares N, Tebe C, Elhussein L, et al. Oral bisphosphonate use and all-cause mortality in patients with moderate-severe (grade 3B-5D) chronic kidney disease: a population-based cohort study. J Bone Miner Res 2020;35(5):894–900. https://doi.org/10.1002/jbmr.3961.
[110] Robinson DE, Ali MS, Pallares N, Tebe C, Elhussein L, Abrahamsen B, et al. Safety of oral bisphosphonates in moderate-to-severe chronic kidney disease: a binational cohort analysis. J Bone Miner Res 2021;36(5):820–32. https://doi.org/10.1002/jbmr.4235.
[111] Elhussein L, Ying HE, Robinson DE, Delmestri A, Strauss VY, Prieto-Alhambra D. Three strategies to measure frailty and/or complex health needs in real world data: an analysis of UK primary care and hospital linked data. Pharmacoepidemiol Drug Saf 2020;29(S3):283–4. https://doi.org/10.1002/pds.5114.
[112] He Y, Elhussein L, Delmestri A, Robinson DE, Strauss VY, Prieto-Alhambra D. The use of preventative treatments (statins, bisphosphates and anti-hypertensives) in older patients with complex health needs: an analysis of UK primary care and linked hospital data. Pharmacoepidemiol Drug Saf 2020;29(S3):297. https://doi.org/10.1002/pds.5114.
[113] Strauss VY, Robinson DE, Delmestri A, Prieto-Alhambra D, Silman A. Longitudinal trajectories of frailty in older populations: latent curve models and mortality. Pharmacoepidemiol Drug Saf 2020;29(S3):288. https://doi.org/10.1002/pds.5114.
[114] Ferguson R, Prieto-Alhambra D, Peat G, Delmestri A, Jordan KP, Strauss VY, et al. Influence of pre-existing multimorbidity on receiving a hip arthroplasty: cohort study of 28 025 elderly subjects from UK primary care. BMJ Open 2021;11(9):e046713. https://doi.org/10.1136/bmjopen-2020-046713.
[115] Ferguson R, Prieto-Alhambra D, Peat G, Delmestri A, Jordan KP, Strauss VY, et al. Does pre-existing morbidity influences risks and benefits of total hip replacement for osteoarthritis: a prospective study of 6682 patients from linked national datasets in England. BMJ Open 2021;11(9):e046712. https://doi.org/10.1136/bmjopen-2020-046712.
[116] Elhussein L, Jödicke AM, Delmestri A, Robinson DE, Strauss VY, Prieto-Alhambra D. Use of and adherence to bisphosphonates in elderly patients with complex health needs: an analysis of UK primary care records. Pharmacoepidemiol Drug Saf 2021;30(S1):383. https://doi.org/10.1002/pds.5305.
[117] Jödicke AM, Tan EH, Robinson DE, Delmestri A, Prieto-Alhambra D. Risk of falls following the initiation of antihypertensives in the elderly: a self-controlled case series in the UK. Pharmacoepidemiol Drug Saf 2021;30(S1):20. https://doi.org/10.1002/pds.5305.
[118] Oda T, Jödicke AM, Robinson DE, Delmestri A, Keogh RH, Prieto-Alhambra D. Bisphosphonate use and risk of severe acute kidney injury: a self-controlled case series in frail elderly patients in the UK. Pharmacoepidemiol Drug Saf 2021;30(S1):164. https://doi.org/10.1002/pds.5305.
[119] Jödicke AM, Tan EH, Robinson DE, Delmestri A, Prieto-Alhambra D. Fracture risk following the initiation of antihypertensive therapy amongst frail elderly people: a self-controlled case series analysis. Pharmacoepidemiol Drug Saf 2021;30(S1):232–3. https://doi.org/10.1002/pds.5305.
[120] Oda T, Jödicke AM, Robinson DE, Delmestri A, Keogh RH, Prieto-Alhambra D. Oral bisphosphonates are associated with increased risk of severe acute kidney injury in elderly patients with complex health needs: a self-controlled case series in the UK. J Bone Miner Res 2022;37(7):1270–8. https://doi.org/10.1002/jbmr.4573.
[121] Elhussein L, Jödicke AM, He Y, Delmestri A, Robinson DE, Strauss VY, et al. Characterising complex health needs and the use of preventive therapies in the older population: a population-based cohort analysis of UK primary care and hospital linked data. BMC Geriatr 2023;23(58):1–9. https://doi.org/10.1186/s12877-023-03770-z.
[122] Corporation Oracle. MySQL 8.0 data type storage requirements. 2023. https://dev.mysql.com/doc/refman/8.0/en/storage-requirements.html. [Accessed 4 May 2023].
[123] Goldacre B, Better Morley J, 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.
[124] University of Warwick. jobs.ac.UK. 2023. https://www.jobs.ac.uk/. [Accessed 4 May 2023].
[125] McPhillips T, Willis C, Gryk MR, Nuñez-Corrales S, Ludäscher B. Reproducibility by other means: transparent research objects, 502–09; 2019. https://doi.org/10.1109/eScience.2019.00066.
[126] Bennett Institute for. Applied data science. OpenSAFELY. 2023. https://www.opensafely.org/. [Accessed 4 May 2023].
[127] Welsh Government's Health and Care Research Wales. SAIL databank. 2023. https://saildatabank.com/. [Accessed 4 May 2023].
[128] Office for National Statistics. Secure research service (SRS). 2023. https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/secureresearchservice. [Accessed 4 May 2023].
[129] NHS Digital. Data access environment (DAE). https://digital.nhs.uk/services/data-access-environment-dae. [Accessed 4 May 2023].
[130] IQVIA Inc. E360™ factsheet. https://www.iqvia.com/library/fact-sheets/iqvia-e360-double-page-factsheet. [Accessed 4 May 2023].
[131] Aetion Inc. Aetion evidence platform. 2023. https://aetion.com/platform/. [Accessed 4 May 2023].
[132] The Health Improvement Network. THIN. 2023. https://www.the-health-improvement-network.com/?hsLang=en. [Accessed 4 May 2023].
[133] Nuffield Department of Primary Care Health Sciences UoO. QResearch. 2023. https://www.qresearch.org/. [Accessed 4 May 2023].
[134] Springate DA, Parisi R, Olier I, Reeves D, Kontopantelis E. rEHR: an R package for manipulating and analysing Electronic Health Record data. PLoS One 2017;12(2):e0171784. https://doi.org/10.1371/journal.pone.0171784.
[135] Gokhale KM, Chandan JS, Toulis K, Gkoutos G, Tino P, Nirantharakumar K. Data extraction for epidemiological research (DExtER): a novel tool for automated clinical epidemiology studies. Eur J Epidemiol 2021;36(2):165–78. https://doi.org/10.1007/s10654-020-00677-6.
[136] Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, et al. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. Pharmacoepidemiol Drug Saf 2012;21(Suppl 1):23–31. https://doi.org/10.1002/pds.2336.
[137] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inf Assoc 2010;17(2):124–30. https://doi.org/10.1136/jamia.2009.000893.
[138] Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inf Assoc 2015;22(3):553–64. https://doi.org/10.1093/jamia/ocu023.