The challenges of replication: a worked example of methods reproducibility using electronic health record data

Richard Williams[*1,2], Thomas Bolton[3], David Jenkins[1], Mehrdad A Mizani[3,4], Matthew Sperrin[1], Cathie Sudlow FMedSci[3], Angela Wood[3,5,6,7], Adrian Heald[8], Niels Peek[1,9], on behalf of the CVD-COVID-UK/COVID-IMPACT Consortium

1. Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK.
2. NIHR Applied Research Collaboration Greater Manchester, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK.
3. British Heart Foundation Data Science Centre, Health Data Research UK, London, UK.
4. Institute of Health Informatics, University College London, London NW1 2DA, UK.
5. British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge UK.
6. Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge UK
7. Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK
8. Department of Diabetes and Endocrinology, Salford Royal NHS Foundation Trust, Salford, UK.
9. The Healthcare Improvement Studies Institute (THIS Institute), Department of Public Health and Primary Care, University of Cambridge, UK

*Correspondence to Richard Williams, richard.williams@manchester.ac.uk

Keywords: Replication study, methods reproducibility, routinely collected healthcare data, electronic health record data, trusted research environment, secure data environment.

Word count: 7153

Abstract

The ability to reproduce the work of others is an essential part of the scientific disciplines. However, in practice it is hard, with several authors describing a “replication crisis” in research. For observational studies using electronic health record (EHR) data, replication is also important. However, replicating observational studies using EHR data can be challenging for many reasons, including complexities in data access, variations in EHR systems across institutions, and the potential for confounding variables that may not be fully accounted for. Observational research is typically given less weight in systematic reviews and clinical guidelines, in favour of more conclusive research such as randomised control trials. Observational research that is replicable has more impact.

In this study we aimed to replicate a previous study that had examined the risk of hospitalisation following a positive COVID-19 test in individuals with diabetes. Using EHR data from the NHS England’s Secure Data Environment covering the whole of England, UK (population 57m), we sought to replicate findings from the original study, which used data from Greater Manchester (a large urban region in the UK, population 2.9m). Both analyses were conducted in Trusted Research Environments (TREs) or Secure Data Environments (SDEs), containing linked primary and secondary

care data. However, the small differences between the environments and the data sources led to several challenges in assessing reproducibility. In this paper we describe the differences between the environments, reflect on the challenges faced, and produce a list of recommendations for TREs and SDEs to assist future replication studies.

Introduction

There is a replication 'crisis' in research. A Nature survey of 1,576 researchers found that 52% of respondents thought that there was a significant 'crisis' of reproducibility, and 72% had tried and failed to reproduce someone else's work [1]. A narrative review focusing specifically on health informatics literature found an increasing interest in replication, but a lack of replication studies [2].

There are many reasons why the replication crisis exists. There is pressure on academics to publish novel research, with replication studies viewed as second rate when compared with primary studies [1,2]. There is publication bias, where statistically significant results are more likely to be submitted by authors, and more likely to be published by journals [3]. There are also specific reasons related to the domain of observational research with electronic health record (EHR) data. The sensitive nature of healthcare data means that there are often barriers to obtaining the same data. When alternative sources of data are used, the collection methods or data structure can often be different, leading to opposite results in many cases [4]. Replication of observational EHR studies, with perhaps more complex methods and data than a typical randomised control trial, can be hampered by high-impact journals' strict word limits, hindering researchers' ability to fully detail their workflows and potentially limit reproducibility.

Replication studies rely on the ability to reproduce several aspects of the original work. Goodman et al. define three terms for discussing research reproducibility: methods reproducibility, results reproducibility and inferential reproducibility [5]. Methods reproducibility is the degree to which a publication includes sufficient information such that other researchers could repeat the analysis. Results reproducibility is the degree to which other researchers can achieve the same results using the same or different datasets. Inferential reproducibility is the degree to which different researchers would reach the same conclusion based on similar results.

There are several barriers to methods reproducibility. Authors often describe their methods without providing their data curation or data analysis code. In the absence of this code, it is hard to reproduce the methods reliably and with confidence. Even when it is provided, the code may not be well documented, or it may be difficult to implement in a different environment. For retrospective observational studies using EHR data there are further challenges. The raw data cannot be analysed immediately as it has been collected for patient care or for billing purposes and so is not in the correct format for analysis. It must therefore first be cleaned, curated and transformed before it is in a suitable format for analysis. This is different to other types of study where the data are collected primarily for analysis and are structured accordingly. Analysts also have little to no control over the raw data and must make do with what is available which may not be comparable to previous studies. It is therefore important to examine methods reproducibility, and how it can be improved, in retrospective studies using EHR data.

Our team has previously examined the risk factors for hospitalisation following COVID-19 in individuals with diabetes [6]. The study was performed in a regional database of primary and secondary care data covering Greater Manchester (GM), a large conurbation (population 2.9m) in the northwest of England (UK). We attempted to replicate this study using a database covering the whole of England. We had the same data engineers and analysts working on the project, and access to the original code, yet the methods reproducibility was not straightforward. Therefore, the objective for this paper is to provide a step-by-step breakdown of the methodological and compute-environment differences between the studies, and their implications for researchers working with

heterogeneous databases of EHR data. A separate paper provides the actual results of the replication, for this same study, and focuses on the degree to which those results, and inferential reproducibility, can be achieved in a regional vs a national database [7].

Methods

The original study applied a combination of univariable and multivariable analyses to a dataset of individuals with diabetes and matched individuals without diabetes, in order to determine risk factors for hospital admission following a COVID-19 diagnosis [6]. Once the data had been cleaned and structured into a wide format, where each row corresponded to a single individual, a series of R scripts produced the outputs [8–10]. For the replication study, once the data was transformed into the same format and variables names comforted to those in the original study, replicating the methods was straightforward and the original R scripts were applied unchanged to the new dataset. Therefore, this paper will focus on the dimension of methods reproducibility related to cleaning and transforming data from a different environment into the format used in our previous study. The statistical methods are therefore not further described but are available in the original paper [6].

This analysis was performed according to a pre-specified analysis plan published on GitHub, along with the phenotyping and analysis code (https://github.com/BHFDSC/CCU040_01).

Data sources

The original study used data from the Greater Manchester Care Record (GMCR). The GMCR is an integrated shared care record containing primary and secondary care data for the residents of Greater Manchester (population 2.9m). It is used for direct care by clinicians who, with patient consent, can access parts of a patient's health record that might otherwise be unavailable because they are held by other care organisations within the National Health Service (NHS) or Social Care. A de-identified copy of the data was made available to researchers during the COVID-19 pandemic and coordinated by the University of Manchester with a view to enabling research focussed on the determinants of health outcomes following COVID-19 infection and in relation to the consequences of the COVID-19 pandemic on health care system as a whole.

This study was performed in NHS England's Secure Data Environment (NHS England SDE), which provides access to a range of national data sets relating to healthcare for approved research programmes. In this case, and wherever the NHS England SDE is mentioned in this paper, data were accessed in an instance of the NHSE England SDE made available for COVID-19 research to the BHF Data Science Centre's CVD-COVID-UK/COVID-IMPACT Consortium (which is coordinated by the BHF Data Science Centre, part of Health Data Research UK). For the avoidance of any doubt, the policies and procedures supporting the SDE are further complemented and supplemented by the Consortium's own detailed "ways of working" document.

For this study we relied on: primary care data from the General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR) [11]; secondary care data from Hospital Episode Statistics (HES) Admitted Patient Care (APC); and COVID-19 test data from the Second-Generation Surveillance System (SGSS) data set which includes almost all community COVID-19 test results in England. Data have been linked by NHS Digital using the NHS Number, a unique ten-digit health identifier in the NHS.

This study consisted of two analyses, each focusing on a primary cohort comprising patients with a coded diagnosis of diabetes, identified prior to their first positive COVID-19 test. A diagnosis of diabetes was defined by the presence of a SNOMED code from the diabetes diagnosis code set in a

patient's primary care record, as taken from the GDPPR dataset. The difference between the two analyses was the data source of the positive COVID-19 test. For the first analysis, the positive test results were exclusively obtained from the primary care record. This was because that was the only source of COVID-19 test results in the GMCR. The first analysis attempted to reproduce the data to be as similar as possible to the original. The second analysis used COVID-19 test and diagnosis data from the primary care record, and test data from the SGSS. This is summarised in Table 1.

Differences between the GMCR and the NHS England SDE were recorded during the execution of the replication study. On completion, we collaborated with the coordinating team of the CVD-COVID-UK/COVID-IMPACT consortium (authors TB, MM, AJ and CS) to ensure a balanced view of both environments.

Themes

The results are structured around the following four themes: access and governance; the research environment; data feeds; and data management, curation and sharing. These themes are inspired by the Goldacre review into the safe use of data for health research [12].

Results

Access and governance

How researchers obtain access

In the GMCR, researchers complete an application form providing details about: their research question; what specific data they require in order to answer the question; and the statistical methods they will be applying. Approval is via the Secondary Uses and Research Governance group who ensure that the data requested is proportionate to the research question, and that the statistical approach is suitable. Any researchers can be given access, but the principal investigator must be from a university in Greater Manchester (Manchester Metropolitan University, University of Bolton, University of Manchester and University of Salford).

At the time of our original study, the legal basis for processing the patient data was via a notice from the UK Secretary of State for Health under section 5b of The Health Service (Control of Patient Information) Regulations 2002. Under this notice there was a restriction to COVID-19 research. Following the expiry of the notice in 2022, we obtained approval from the UK Health Research Authority to conduct any health research and are no longer restricted to COVID-19 studies.

For access to data in the NHS England SDE, the CVD-COVID-UK/COVID-IMPACT research programme received approval to access data from the Independent Group Advising on the Release of Data via an application made in NHS England's Data Access Request Service (DARS) Online system (ref. DARS-NIC-381078-Y9C5K). Furthermore, the North East - Newcastle and North Tyneside 2 research ethics committee provided ethical approval for the CVD-COVID-UK/COVID-IMPACT research programme (REC No 20/NE/0161) to access, within secure trusted research environments (TRE), unconsented, whole-population, de-identified data from EHRs collected as part of patients' routine healthcare. The CVD-COVID-UK/COVID-IMPACT Approvals & Oversight Board, comprising representatives from data custodians, data controllers, researchers and public contributors, reviews all project proposals to ensure that they fall within the scope of the regulatory and ethical approvals.

During the project proposal and approval process, project teams request access to the available datasets required for their project, with justification for how each dataset will be used. Following review and approvals, accredited analysts are onboarded to the NHS England SDE and provided with data access. Analysts require their institution (here, University of Manchester) to be named as a joint

data controller in the data sharing agreement (DSA) with NHS England. In signing the DSA, each partner institution agrees to ensure that analysts from that institution have appropriate information governance and data protection training in relation to the use and storage of health data and have a confident understanding of their responsibilities. Project leads and their named analysts are also expected to undertake their research in line with the CVD-COVID-UK/COVID-IMPACT consortium's "ways of working" document, which details how projects should be run, promotes cross-institutional collaboration and requires all outputs to be published in open-access publications, and for the analysis code (usually a combination of R, PySpark, and SQL scripts), code lists, phenotyping algorithms and protocol to be made publicly available in a GitHub repository.

Data discovery

Analysts need to know in advance of an application whether the data source is likely to contain the necessary information to enable their project. Metadata catalogues are a standard way to present this information.

The BHF Data Science Centre's Health Data Science team provides metadata from the NHS England SDE via table and value level data dictionaries in their Data Summary Dashboard (https://bhfdatasciencecentre.org/dashboard/). This includes the data dictionary for each dataset, along with a summary of the content and reports on the completeness and coverage of the data. During the application review process, other consortium members are encouraged to provide input on where best to find certain data items.

In addition to the metadata, both environments have an iterative process where applications are reviewed by people with expertise in the underlying data. For the NHS England SDE this is the BHF DSC Health Data Science Team, and for the GMCR it is a group of research data engineers (RDEs). The feasibility of projects is assessed, recommendations for alterations are suggested, and advice is provided on the most suitable datasets and data items.

Publications

In the GMCR, publications are checked by the Secondary Uses & Research Governance group (SURG) in order to ensure the validity and credibility of the results and to preserve the reputation of the GMCR from substandard research. In the NHS England SDE, publications undergo an initial check by the BHF Data Science Centre's coordinating team for any factual inaccuracies, to ensure that the project's outcomes remain within scope of the CVD-COVID-UK/COVID-IMPACT research programme's regulatory and ethical approval, the appropriate acknowledgment statements have been included, and the required content has been uploaded to the project's GitHub repository. Draft manuscripts are then shared with all members of the CVD-COVID-UK/COVID-IMPACT Consortium for peer review. The BHF Data Science Centre also provides opportunities for Patient and Public Involvement and Engagement (PPIE).

The difference here is that while both environments would block publications, at the time of the study, only the GMCR had documentation detailing how and why this might occur. The writing of this publication highlighted the inconsistency and the next version of the CVD-COVID-UK/COVID-IMPACT ways of working document was updated accordingly.

Research environment

Technology

In the GMCR, data are stored in a Microsoft SQL Server database with column-store indexes. The data are transformed with SQL, extracted into flat file format, and made available to analysts via a secure file share system. The study teams access the environment via a connection to a remote

virtual desktop environment running Windows. They then analyse the data extracts using R (via RStudio) or Stata. Users have database read-only access and cannot create permanent database tables.

In the NHS England SDE, data are stored in as column-oriented tables (Delta tables in Amazon Web Services (AWS)) and is accessed through Hive metastore in a Databricks analytics platform, Apache Spark, RStudio Pro IDE, RStudio Server, or AWS virtual desktop solution for Stata. Databricks currently supports Spark SQL, PySpark, SparkR, and Python interfaces. Analysts log into the NHS England SDE via a portal to access a Windows-based Virtual Desktop Interface using supported browsers and two-factor authentication. The data are also read-only in this environment, but analysts can create tables in a collaboration database with both read and write permissions.

Import control

At the time of the studies, neither the GMCR, nor the NHS England SDE, enforced import checking. In both environments it was possible to copy code snippets, scripts and small text-based reference data such as code lists directly into the environment. The NHS England SDE now enforces input checking with files up to 1MB going through the NHS England's input checking service. Files larger than 1MB are also allowed but must go via a special request.

Export control

In the GMCR, all results and aggregate data must first be checked by another analyst for disclosure risk prior to exporting. However, there is currently no technical mechanism to enforce this, and it instead relies on user training. The next version of the GMCR will have an independent output checking process that is enforced technically.

The NHS England SDE operates a Safe Output Service to ensure that disclosure control rules are always maintained in the aggregated results, and no data elements are visible in any code that are exported from the environment. The independent output checking team ensure that aggregate/summary-level results are appropriately disclosure controlled and justified by supporting contextual information. For example, counts under ten are suppressed, and counts over ten are rounded to the nearest five. If output submissions are approved, then the files are made available to download. For further info see https://digital.nhs.uk/services/data-access-environment-dae/user-guides/using-databricks-in-the-data-access-environment#safe-output-service.

Execution time

In the GMCR, queries run relatively quickly, which is likely due to the smaller population size. In the NHS England SDE, due to the much larger population size (57m vs 2.9m), some queries take longer to run. However, analysts have the permission to save intermediate database tables, which means that they can run some of the time-consuming queries once before caching the results for use in future queries.

For analysts attempting to repeat analyses with larger datasets this can potentially be an obstacle. Any statistical or machine learning methods that are computationally intensive or that scale in a non-linear way, for example multiple imputation or resampling methods such as bootstrapping, may run quickly in one environment, but take an unreasonable amount of time, or fail completely, in another with a substantially larger dataset.

Data feeds

The GMCR and the NHS England SDE, at the time of each study, were databases containing linked primary and secondary care data for the purpose of COVID-19 research.

GP data

In the NHS England SDE, the GP data are sourced from the GDPPR dataset [11]. There is an updated list of SNOMED code clusters that currently include over 36,400 SNOMED concepts that are extracted. This represents a substantial amount of patient data (in particular given that the concepts included are amongst the most frequently used by GPs), but it is worth noting that there are over 900,000 SNOMED codes in the UK and international releases available to GPs (even though the majority of these are barely if ever used). In the GMCR, GP data comes from a direct feed from each practice. It includes the entire medical history of clinical codes (currently all SNOMED concepts, but at the time of the study the data also included Readv2, CTV3 codes) for each patient. Individuals who opt out of data sharing for secondary use are not included in either database.

The GDPPR dataset does not include the SNOMED concepts for the following covariates that were used in the original analysis and so we were not able to perform an exact replication study: testosterone level, vitamin D level, and sex hormone binding globulin (SHBG) level. It is also missing some, but not all, of the severe mental illness codes used in the original study, particularly symptom codes rather than diagnosis codes. Our original plan was to also replicate a study focusing on COVID-19 and mental health, but this was not possible due to the reduced set of diagnoses available in GDPPR.

COVID-19 tests

The GMCR does not provide access to a dedicated COVID-19 test database and so COVID-19 tests are taken from the GP record. In the NHS England SDE, COVID-19 tests are available from several sources (in addition to those that occur in the GDPPR dataset):

- Second Generation Surveillance System (SGSS) includes first positive pillar 1 and pillar 2 tests
- Pillar 2 Antigen (positive and negative)
- Pillar 3 Antibody (positive and negative)
- COVID-19 Hospitalisation in England Surveillance System (CHESS)
- HES data (though this contains diagnoses of COVID rather than actual test results)

Hospital admissions

The NHS England SDE has access to HES and SUS data including admitted patient care (APC) data where hospital admissions were sourced from. In the GMCR, hospital admissions came from direct data feeds from each hospital trust in GM. However, some hospital feeds did not come online within GMCR until May 2020, and historic data was not available.

Date of death

In GMCR this is redacted to month of death. In the NHS England SDE this is not redacted. Date of death comes from the REG_DATE_OF_DEATH field in the Civil Registrations of Death dataset. According to the NHS metadata catalogue this is not classed as a potentially identifiable field and so the data are not redacted. GDPPR also provides the year of death, where the date of death field is considered potentially identifiable for this dataset, highlighting a lack of consistency.

In the GMCR, we were unable to calculate the commonly used outcome of “death within 28 days of a positive COVID-19 test” because we only had access to individual’s month of death. We needed to work with the system supplier (GraphNet Health Ltd) to add a field to the database with this information which could be calculated before the data was pseudonymised.

Data management, curation and sharing

Data curation

In the GMCR, there is a clear separation between data curation and data analysis activities. The data curation, which includes processing, transforming and cleaning the data, is performed by a small team of research data engineers (RDEs) who have access to the entire database. The RDEs have expertise in EHR data, software engineering, database management and database querying. Study teams submit proposals which explain their research questions and the data required to answer them. Analysts are encouraged to request the data they would ideally like, then the Research Data Engineers (RDEs) determine feasibility and suggest alterations where appropriate. The data provided to the analysts are minimised to that which is required to answer the research questions. The data are provided by the RDEs to the study analysts in a format that is cleaned and ready to load immediately into statistical software. The data analysis is then performed by the study teams. The data analysts only have access to the data extract required for their study; they do not have access to the underlying database.

In the NHS England SDE, analysts have access to the raw and curated datasets and can undertake both data curation and data analysis. The BHF Data Science Centre Health Data Science Team provide different levels of data curation support to projects and analyst teams. This includes signposting to resources (e.g., tutorial, data summary, and data insight notebooks, common code, and curated tables), providing data curation guidance, performing exploratory data analysis to inform data curation, developing part of the data curation pipeline in collaboration with analysts, developing the full data curation pipeline on behalf of analysts (similar to the GMCR model), and reviewing data curation pipelines that have been created by the analysts. The Health Data Science Team also provides analytical support. Inductions, further technical support, and help with resolving data queries is provided by the NHS England Data Wrangler Team. This means that support can be tailored to analyst teams with varying levels of experience and also targeted to where support is most needed to speed research productivity.

Data management and sharing

In the GMCR, the RDEs use a custom SQL templating language to create extract scripts which are then compiled into raw SQL. This allows common chunks of reusable SQL, and lists of clinical codes from electronic phenotypes, to be consistently used across multiple projects without error. The compiled SQL can then be copied from the RDE's local machine to the secure VDE and executed against the database to produce flat files, usually csv, which are then provided to the analysts. The RDEs also build and maintain a public library of clinical code sets, phenotypes and reusable database queries (https://github.com/rw251/gm-idcr).

The data extraction code for each individual project within the GMCR is publicly available at the above linked GitHub repository. During compilation of the SQL for a project, any clinical code sets that are used are automatically collated into a single csv file which is available for download from the project's folder. Also, any metadata related to the various chunks of SQL that are used by the project, are also automatically extracted and made available in a single README file. This is the README file for the original GMCR study described in this paper (https://github.com/rw251/gm-idcr/blob/master/projects/020%20-%20Heald/README.md).

In the NHS England SDE, data manipulation and curation are generally performed using Databricks notebooks that are similar to Jupyter notebooks with Spark, collaboration tools, version control through an internal GitLab, and analytics capabilities. Code can be developed directly within the

notebooks, as stand-alone Python files, or in an IDE (VS Code or PyCharm) with GitLab integration. Small code snippets and text can be copied and pasted into the SDE environment.

The reproducible and reusable data curation and analysis pipelines, in addition to dataset summaries and exploratory analyses, are shared via collaboration workspaces for analysts to reuse/adapt to help reduce the amount of time spent on data preparation and to accelerate research. Similarly, following the Consortium's collaborative way of working, analysts can also make use of existing code and tables (such as cohorts) developed and shared by other analysts working on all other approved projects. In line with the CVD-COVID-UK/COVID-IMPACT Consortium's principles – based on a collaborative, transparent and inclusive ethos – all related analysis plans, protocols, code, phenotype code lists and reports are made publicly available via the BHF Data Science Centre's collection on the HDR UK Gateway and HDR UK Phenotype Library, repositories in the BHF Data Science Centre's GitHub organisation, and through open-access publications.

Discussion

Summary

In this paper we have described the differences between the two secure data environments used when attempting to replicate the results of a regional observational study of EHR data in a national database. We have shown that methods reproducibility is hard even in the scenario where there is perfect sharing of the study definition, algorithms and clinical code sets. Differences in the data, and in the environments themselves, are a barrier to quick replication of existing studies. We will now reflect on the implications for researchers and provide a series of recommendations for Trusted Research Environments (TREs) and Secure Data Environments (SDE)s to improve the ease with which replication studies can be performed. The full list of recommendations is in

Table 2.

These recommendations will be of relevance to the UK NHS sub-national and regional Secure Data Environment (SDE) programme, launched in 2022, which aims to create a network of secure data environments across England [13]. These environments, developed through partnerships between the NHS and universities, will give researchers controlled access to anonymized NHS patient data for research purposes.

Another relevant initiative is the development of a series of standards for best practice for TREs and SDEs by the UK TRE community. This includes the standardised architecture for trusted research environments (SATRE) [14], the development of a federated network of TREs (TRE-FX) [15], and software for the semi-automated checking of research outputs (SACRO) [16].

Access and governance

Different access and governance arrangements can act as a barrier to replication. The ambition of most SDE programmes is that researchers will be able to federate their analysis across multiple environments. If a separate application form needs completing for each environment in a federated analysis, and if the governance arrangements are different, then this will add a considerable burden to researchers, and likely mean federation would never happen in more than a couple of environments.

> **Recommendation 1:** SDEs/TREs that wish to allow federated analysis should consider unified application processes so that researchers are only required to apply once.

The governance arrangements in the GMCR are designed so that sub-standard research can be blocked from publication in the interest of preserving the reputation of the environment. This is also true for the SAIL databank [17] where research can be blocked if it is in breach of their output review policy. For example, the data are not permitted to be used for performance tracking of individual organisations. As described above, research carried out in the NHS England SDE through the BHF Data Science Centre is subject to statistical disclosure control by NHS England processes, subject to checks by the BHF Data Science Centre and then subject to peer review by the CVD-COVID-UK/COVID-IMPACT consortium. For replication we should avoid the situation where a study is possible in one environment but blocked from publication in another.

> **Recommendation 2:** SDEs/TREs should clearly define what research outputs are permitted, the process that is used for assessment, and any appeals process.

Metadata

In the GMCR, we have found that providing researchers with access to the descriptions of the fields of the database is unhelpful because additional information is missing. The typical metadata for a field, such as the name and description, does not include measures of completeness, how usage varies over time, or whether the information is available or redacted. These are things that could be addressed. However, the detail and complexity in a database of EHR data is restricted to the handful of fields containing clinical codes. These data columns will contain all medical concepts from diagnoses and procedures to medications and results. This is in contrast to data from randomised controlled trials or cohort studies where each measurement or observation will be contained in a separate field. Standard metadata catalogues are ideally suited to these controlled studies but are inappropriate for longitudinal EHR data as has been shown in a previous study [18]. Also, the data provided to the analysts is transformed into a format that is ready for research and may bear little relationship to the underlying database structure, so it is better to describe the available data in broad terms and offer a service where preliminary ideas can be checked for feasibility.

This highlights a need for improved metadata catalogues that are designed specifically for EHR data. However, several of the discrepancies encountered in the data in the two environments, such as how a hospital admission is defined, or which clinical codes are available, would likely not be documented in a data catalogue. Even if they were, the volume of information contained in the metadata catalogue would then be so large as to reduce its utility. Computable study definitions, combined with machine-readable metadata catalogues, might enable feasibility checking and automatic execution of replication studies. However, given it is difficult to find two datasets with all variables necessary for a particular study to be recorded in the same way, and given that these differences can make data reproducibility problematic, it may be some time before we can achieve this, even when utilising data within one national digital health infrastructure.

> **Recommendation 3:** Metadata catalogues designed specifically for longitudinal EHR data should be researched and developed.

> **Recommendation 4:** Research is needed to develop computable study definitions that can be executed against machine-readable metadata catalogues.

Access costs

The currently accepted way of conducting safe research is via SDEs. The sharing of code and tools available in these environments is expected to lead to an acceleration of research [12]. However, research conducted in this way has the potential to be more expensive than the “old” way of simply giving out copies of the data. The costs of the additional technical and administrative infrastructure

required for SDEs are passed onto researchers via an access fee. Previously the costs of storage and compute using local university resources may have been hidden from the study teams. If the access fees are large, then the number of research groups who will be able to afford to conduct research in the environments will reduce. This could mean that the benefits of the environment will be limited to the quality and safety of the research, rather than an increase in the quantity. The prospect of federation further adds to this problem, where the cost would quickly become prohibitive if a large access cost was required for each individual SDE in which a researcher wished to federate their analysis.

> **Recommendation 5:** SDEs/TREs should consider how automation and other efficiencies can reduce access costs. Ensuring that replication and federated analyses do not become prohibitively expensive is crucial for the advancement of research.

Research environment

Environment heterogeneity

Small differences between environments can have a big effect on how users interact with them. These differences are unlikely to become apparent until you have access to the environment. One example from this replication study, is the ability to create permanent database tables, which was possible in the NHS England SDE but not in the GMCR. Therefore, in the GMCR, all interim calculations were done via temporary tables. These tables only last for the lifetime of the query and so data caching for improved performance on subsequent queries is not possible. Queries must also be deterministic and not random. For example, where a matched cohort is required in multiple queries the ideal situation would be to define the cohort once, save it to a permanent table, and then use it in subsequent queries. The limitation means that instead the cohort must be created in every query that it is used in, and it must generate the exact same matched cohort each time. This is a seemingly trivial difference, and one that would be unknowable until access had been granted, but it has a significant effect on the interactions with the system.

> **Recommendation 6:** SDEs/TREs should be designed to be agile and adaptable, incorporating best practices as they evolve. The SATRE specification [14] is the most likely source for these best practices.

Execution time

A good example of the replication issue where methods do not scale to a larger cohort is for the matching that is required for cohort and case-control studies. Our study relied on a matched cohort of individuals. In the GMCR, matching is done via a loop in SQL. In the first pass we attempt to get a single exact match for each individual based on sex, age and date of positive COVID-19 test. The matching criteria are then progressively relaxed for individuals with no matches – e.g., date of positive test within 2 weeks, and then within 4 weeks, similarly for age. The process is then repeated for the 2nd and 3rd matches for each individual. This approach scaled in a polynomial way in terms of time and memory usage, which for the size of the cohort (diabetes + COVID-19 positive test in GM) was an acceptable solution.

However, the GMCR method did not scale well to the national population and consumed too much memory. Instead, the cohort matching was rewritten in Python and the algorithm improved by pre-sorting the data. The relevant data was extracted and loaded into Pandas data frames in Python, the matching performed, and then the results written back to the database. This rewrite took a long time to make it sufficiently performant, but using Python, rather than being restricted to SQL did have a few advantages. With SQL we matched on age, sex and date of COVID-19 test. Where there wasn't an exact match, we relaxed the age and date of COVID-19 test. With Python we could do the

same, but in the case of no exact matches we could find people with the same sex, similar age, and the nearest COVID-19 positive test. This was either not possible with SQL, or beyond our knowledge and capabilities.

> **Recommendation 7:** SDEs/TREs should ensure that data can be accessed and processed with multiple languages such as SQL, R and Python.

> **Recommendation 8:** SDEs/TREs should implement mechanisms to monitor and manage execution time variability. Providing researchers with tools to estimate and optimise execution times can improve the efficiency and reliability of data analysis.

Import controls

At the time the studies were performed, neither environment had checks on imported content such as clinical code sets or analysis code below a certain size. This leads to a good experience for analysts who can simply copy and paste content into the environment. The SATRE specification [14] has an optional requirement for "an approval process before allowing code into the technical environment" (ref 2.1.13). While this might seem a sensible approach it is perhaps not justified. The rationale would be to prevent a user from compromising the system either maliciously or accidentally. However, a malicious user denied access to copying in content, could simply write their malicious code within the environment. It would take longer but would circumvent the restrictions and be very hard to detect. Instead, a better target would be for future SDE/TREs to be safely sandboxed in such a way that malicious code is ineffective. In conjunction with adequate statistical disclosure control for outputs, this would lead to the best experience for end users while preserving the integrity of the environment. In any event, an essential requirement for replication studies is for existing code to be imported.

> **Recommendation 9:** SDEs/TREs should allow safe content to be easily imported. Where import controls are enforced, they should aim to reduce the barriers to researchers wherever possible.

Data feeds

Both of the environments in this study might reasonably be described as containing "linked primary and secondary care data from the UK". The assumption would be that studies requiring this sort of data would be possible in either environment. However, this is not the case, and until you explore the data in detail there are several hidden differences. Our original plan was to replicate another study from the GMCR. However, it could not proceed because there were certain SNOMED codes related to severe mental health that were unavailable in the GDPPR dataset. Therefore, anyone attempting to replicate from a local database with the full GP record, must evaluate the availability of equivalent or proxy variables in the datasets that are limited extracts of GP records, such as GDPPR.

Working with EHR data requires that you make the most of what you have as there is no opportunity to change the data or affect how it is collected. The two databases used in this study contain primary and secondary care data and have a common purpose. However, there are differences, which are apparent when we consider the key data items. The most important data items for these studies were: the diagnosis of diabetes, and the presence of a positive COVID-19 test, as these defined the cohort of patients; and the event of a hospital admission as this was the main outcome. Despite the similarities of the underlying data in the databases, all 3 of these components were affected. In the GMCR, the COVID-19 tests were taken from the GP record, the diagnosis of diabetes was contained in the full GP record, and the hospital admissions were from direct feeds from each hospital trust

(i.e. not HES data). In the NHS England SDE the COVID-19 tests were from the SGSS data in addition to the GP record, the diagnosis of diabetes was from the GDPPR dataset, and the hospital admissions were from HES APC data. These differences may or may not affect the results of the replication, but they further highlight that the metadata for an environment needs to be sufficiently detailed to ensure these differences are explicit as they will undoubtedly affect methods reproducibility.

Data management, curation and sharing

There are 3 distinct phases that are necessary to analyse EHR data: familiarity, engineering (or curation) and analysis. Data familiarity is the understanding of the provenance of the underlying data and the structure that it is stored in. Data engineering/curation is taking the raw data and transforming it in such a way so that it is ready for analysis. Finally, data analysis is the application of statistical methods to the transformed data to produce results.

Data engineering and data analysis are two different domains, with different tools, languages and skillsets, and it is better to have people with expertise in each rather than both. If environments are built that require individuals to have expertise in both, then that substantially slows down research in this area due to the lack of such people. If instead there is a clear distinction between the activities, then it is easier to find expertise in each area and therefore a lower barrier to research. While engineering and analysis skills can be brought to a new environment, the familiarity with the data must be developed each time for each new environment. The RDE model in the GMCR is designed to speed up research by removing the time required for external researchers to develop the data familiarity. Instead, a small team of engineers, with an in-depth understanding of the data and engineering skill, provide data analysts with bespoke datasets for their analysis. Therefore, researchers using the GMCR do not need data familiarity or data engineering skills. Researchers without these skills may struggle when moving to an environment without RDE support.

> **Recommendation 10:** SDEs/TREs need a support structure for researchers which includes people with expertise in the underlying data.

The reuse of data wrangling code, clinical code sets and phenotypes within the GMCR is completely managed by the RDEs who have full editorial control. In the national database, at the time of the study, these digital artefacts were stored within each project directory. Reuse is encouraged, but it can sometimes be hard to find the relevant cleaning or analysis code. When code is found, it can be hard to select between multiple similar options. This problem is not unique to the NHS England SDE. Sites that help users share their clinical code sets such as clinicalcodes.org [19] or the HDRUK gateway [20] suffer a similar problem. Namely that by making the sharing of clinical code sets easy, there is a proliferation of similar code sets, particularly for common long-term conditions.

> **Recommendation 11:** Libraries of code, clinical code sets and phenotypes should consider their editorial policy. If there are no barriers to uploading content, then standardised tools should be created to allow easy discovery and comparison of the digital artifacts.

While the data preparation code is shared automatically in the GMCR, the data analysis code is only shared if the study team choose to do so. The sharing is encouraged but not mandated. The NHS England SDE through the BHF Data Science Centre has an advantage here. The ways of working for the CVD-COVID-UK/COVID-IMPACT research programme mandates that all preparation and analysis code is shared. However, there are still issues because simply putting something on GitHub does not necessarily mean that it is easily reused. SDEs should consider whether to mandate or simply encourage sharing of analysis code, and effort should be taken to ensure that this code is more readily reusable, and not simply shareable, for example by using RO-Crates [21,22].

Recommendation 12: SDEs/TREs should mandate the sharing of data curation and data analysis code.

Data reproducibility

Based on our experiences with the replication study, and the discussion above, we propose that the three types of reproducibility (methods, results, inferential – Goodman [5]) should be extended to four. Methods reproducibility would remain as the ability to reproduce the statistical analysis of a study given the same data. The new "Data reproducibility" would be the ability to which the data to be analysed could be prepared, extracted and cleaned from a different database. In retrospective observational research, an author can provide all of their code, and have it extremely well documented, but if the person attempting replication is using data from a different source, then there is still a data transformation and cleaning exercise required which will affect the reproducibility. In this case the methods reproducibility would be simple, but the data reproducibility would remain hard.

In our case, once the data had been transformed into the same format as required by the R scripts from the original study, the methods reproducibility was trivial. The R code was expecting data in a tabular format with predefined columns. When applied to data in the same format, but from the NHS England SDE rather than the GMCR, the code ran without exception. There was one minor change needed, but this was quick and easy to do. The change was because GMCR admissions data are stored in such a way that patients who have not been discharged have a blank discharge date. In the HES APC data in the NHS England SDE, the discharge date field is never blank and an ancient date such as 1800-01-01 indicates an undischarged patient. This was spotted at the point of analysis as a handful of patients had very large negative lengths of hospital stay. While this could have been amended in the data curation code, we instead improved the analysis code to correctly handle records with a negative length of hospital stay. This was noticed easily because the large negative values skewed the results significantly. However, one limitation of replication studies is that it is possible that similar data changes could introduce mistakes that were not as easily detectable.

A related work has shown that while a common protocol for studies is helpful, it is not sufficient to remove all the bias of using different databases [23]. Madigan et al [4] found that the choice of database can influence findings, with 36% of the 53 drug/outcome pairs that were analysed had statistically significant decreased risk in some databases, but statistically significant increased risk in others.

Other environments and models

Our review has focused on two secure data environments containing linked primary and secondary data. There are several other such environments. OpenSAFELY [24] ensures that researchers do not access the data directly. Queries are constructed outside the environment, executed within the environment, and then the results are presented back to the researchers. OHDSI [25] require participating centres to transform their data into the OMOP common data model. This then allows researchers to execute code against multiple centres and collate the results. CPRD [26] and the SAIL Databank [17] currently implement TREs to allow researchers to access healthcare data in a secure environment.

We believe that there would be similar issues when attempting to replicate between any of the environments described above and that our recommendations would still apply.

Limitations

The findings are specific to the UK's healthcare data systems which limits the paper's applicability to countries with different healthcare data practices. Further research could explore similar replication studies in other healthcare systems to enhance the generalisability of the recommendations.

Conclusion

In the process of conducting a replication study, we have demonstrated that methods reproducibility can face major difficulties even with perfect sharing of code. It is straightforward to share the cleaned data definition, and the statistical code used to analyse it. However, data reproducibility remains challenging. Our recommendations, together with future research on making study definitions and metadata catalogues machine-readable, should reduce the barriers to replication studies, and elevate the potential of observational studies using EHR data. This is particularly relevant at a time when electronic health record data are increasingly being used to guide national and international health policy direction.

Contributions

RW processed and cleaned the data. RW and DJ performed the analysis. NP and AH provided guidance on the analysis and construction of the paper. CS, TB, MM and AJ, provided specific information on the NHS England SDE and CVD-COVID-UK/COVID-IMPACT consortium. All authors drafted and reviewed the manuscript.

Members of the wider CVD-COVID-UK/COVID-IMPACT consortium (https://bhfdatasciencecentre.org/wp-content/uploads/2024/05/010524-CVD-COVID-UK-COVID-IMPACT-Consortium-Members.pdf) also provided comments on drafts of the protocol and manuscript.

Funding

The British Heart Foundation Data Science Centre (grant No SP/19/3/34678, awarded to Health Data Research (HDR) UK) funded co-development (with NHS England) of the Secure Data Environment service for England, provision of linked datasets, data access, user software licences, computational usage, and data management and wrangling support, with additional contributions from the HDR UK Data and Connectivity component of the UK Government Chief Scientific Adviser's National Core Studies programme to coordinate national COVID-19 priority research. Consortium partner organisations funded the time of contributing data analysts, biostatisticians, epidemiologists, and clinicians.

The associated costs of accessing data in NHS England's secure data environment service for England, for analysts working on this study, were funded by the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics, which is funded by UK Research and Innovation (grant ref: MC_PC_20058).

This research was co-funded by the NIHR Manchester Biomedical Research Centre (NIHR203308) and the NIHR Applied Research Collaboration Greater Manchester (NIHR200174). AMW is supported by the BHF Data Science Centre (HDRUK2023.0239) and as an NIHR Research Professor (NIHR303137). This work was supported by core funding from the: British Heart Foundation (RG/18/13/33946), NIHR Cambridge Biomedical Research Centre (BRC-1215-20014; NIHR203312) [*], Cambridge BHF Centre of Research Excellence (RE/18/1/34212), BHF Chair Award (CH/12/2/29428) and by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council,

Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome.

The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Ethical approval

The North East - Newcastle and North Tyneside 2 research ethics committee provided ethical approval for the CVD-COVID-UK/COVID-IMPACT research programme (REC No 20/NE/0161) to access, within secure trusted research environments, unconsented, whole-population, de-identified data from electronic health records collected as part of patients' routine healthcare.

Acknowledgements

This work is carried out with the support of the BHF Data Science Centre led by HDR UK (BHF Grant no. SP/19/3/34678). This study makes use of de-identified data held in NHS England's SDE for England, and made available via the BHF Data Science Centre's CVD-COVID-UK/COVID-IMPACT consortium. This work uses data provided by patients and collected by the NHS as part of their care and support. We would also like to acknowledge all data providers who make health relevant data available for research.

Data availability

The data used in this study are available in NHS England's SDE service for England, but as restrictions apply they are not publicly available (https://digital.nhs.uk/services/secure-data-environment-service). The CVD-COVID-UK/COVID-IMPACT programme led by the BHF Data Science Centre (https://bhfdatasciencecentre.org) received approval to access data in NHS England's SDE service for England from the Independent Group Advising on the Release of Data (IGARD) (https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/independent-group-advising-on-the-release-of-data) via an application made in the Data Access Request Service (DARS) Online system (ref. DARS-NIC-381078-Y9C5K) (https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services). The CVD-COVID-UK/COVID-IMPACT Approvals & Oversight Board (https://bhfdatasciencecentre.org/areas/cvd-covid-uk-covid-impact/) subsequently granted approval to this project to access the data within NHS England's SDE service for England. The de-identified data used in this study were made available to accredited researchers only. Those wishing to gain access to the data should contact bhfdsc@hdruk.ac.uk in the first instance.

References

[1] M. Baker, 1,500 scientists lift the lid on reproducibility, Nature. 533 (2016) 452–454. https://doi.org/10.1038/533452a.

[2] E. Coiera, E. Ammenwerth, A. Georgiou, F. Magrabi, Does health informatics have a replication crisis?, J. Am. Med. Informatics Assoc. 25 (2018) 963–968. https://doi.org/10.1093/jamia/ocy028.

[3] M.J. Schuemie, P.B. Ryan, G. Hripcsak, D. Madigan, M.A. Suchard, Improving reproducibility by using high-throughput observational studies with empirical calibration, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 376 (2018). https://doi.org/10.1098/rsta.2017.0356.

[4] D. Madigan, P.B. Ryan, M. Schuemie, P.E. Stang, J.M. Overhage, A.G. Hartzema, M.A. Suchard, W. Dumouchel, J.A. Berlin, Evaluating the impact of database heterogeneity on observational study results, Am. J. Epidemiol. 178 (2013) 645–651. https://doi.org/10.1093/aje/kwt010.

[5] S.N. Goodman, D. Fanelli, J.P.A. Ioannidis, What does research reproducibility mean?, in: Get. to Good Res. Integr. Biomed. Sci., Springer International Publishing, 2018: pp. 96–102. https://doi.org/10.1126/scitranslmed.aaf5027.

[6] A.H. Heald, D.A. Jenkins, R. Williams, M. Sperrin, H. Fachim, R.N. Mudaliar, A. Syed, A. Naseem, J.M. Gibson, K.A. Bowden Davies, N. Peek, S.G. Anderson, Y. Peng, W. Ollier, The Risk Factors Potentially Influencing Hospital Admission in People with Diabetes, Following SARS-CoV-2 Infection: A Population-Level Analysis, Diabetes Ther. 13 (2022) 1007–1021. https://doi.org/10.1007/s13300-022-01230-2.

[7] R. Williams, D. Jenkins, T. Bolton, A. Heald, M.A. Mizani, M. Sperrin, N. Peek, Replicating a COVID-19 study in a national England database to assess the generalisability of research with regional electronic health record data, MedRxiv. (2024) 2024.08.06.24311538. https://doi.org/10.1101/2024.08.06.24311538.

[8] R Core Team, R: A Language and Environment for Statistical Computing, (2022). https://www.r-project.org/.

[9] H. Wickham, R. François, L. Henry, K. Müller, dplyr: A Grammar of Data Manipulation, (2022). https://cran.r-project.org/package=dplyr.

[10] T.S. Barrett, E. Brignone, Furniture for Quantitative Scientists, R J. 9 (2017) 142–148. https://doi.org/10.32614/RJ-2017-037.

[11] NHS Digital, General Practice Extraction Service (GPES) Data for pandemic planning and research: a guide for analysts and users of the data, (n.d.). https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data (accessed May 31, 2023).

[12] B. Goldacre, J. Morley, N. Hamilton, Better, broader, safer: using health data for research and analysis, 2022.

[13] N. England, How will Secure Data Environments be delivered?, (n.d.). https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/secure-data-environments/how-will-secure-data-environments-be-delivered/ (accessed March 7, 2024).

[14] D.C. Cole, H. Sood, D.S. Li, K. Oldfield, M. Craddock, N. Swanepoel, P.S. Coleman, D.M. O'Reilly, D.D. Kerr, D.C. O'Donovan, P.J. Hetherington, D.J. Madge, D. Sarmiento-Perez, E. Chalstrey, D.J. Robinson, J. Beggs, T. Machin, A. Chuter, SATRE: Standardised Architecture for Trusted Research Environments, (2023). https://doi.org/10.5281/zenodo.10055345.

[15] DARE UK, TRE-FX: Delivering a federated network of trusted research environments to enable safe data analytics, (n.d.). https://dareuk.org.uk/driver-project-tre-fx/ (accessed March 7, 2024).

[16] DARE UK, SACRO: Semi-Automated Checking of Research Outputs, (n.d.). https://dareuk.org.uk/driver-project-sacro/ (accessed March 7, 2024).

[17] SAIL, SAIL Databank, (n.d.). https://saildatabank.com/ (accessed March 5, 2024).

[18] G. Tilston, R. Williams, E. Griffiths, S. Al-Adely, S. Lawson-Tovey, W. Hulme, A. Short, J. Davies, J. Welch, N. Peek, Can Researchers Assess the Suitability of Datasets to Answer Their Research Questions, with Access to Metadata Only?, in: Stud. Health Technol. Inform., IOS

Press BV, 2022: pp. 66–70. https://doi.org/10.3233/SHTI220033.

[19] D.A. Springate, E. Kontopantelis, D.M. Ashcroft, I. Olier, R. Parisi, E. Chamapiwa, D. Reeves, ClinicalCodes: An online clinical codes repository to improve the validity and reproducibility of research using electronic medical records, PLoS One. 9 (2014) e99825. https://doi.org/10.1371/journal.pone.0099825.

[20] HDRUK, Phenotype Library, (2021). https://phenotypes.healthdatagateway.org/ (accessed March 5, 2024).

[21] P. Sefton, E. Ó Carragáin, S. Soiland-Reyes, O. Corcho, D. Garijo, R. Palma, F. Coppens, C. Goble, J.M. Fernández, K. Chard, J.M. Gomez-Perez, M.R. Crusoe, I. Eguinoa, N. Juty, K. Holmes, J.A. Clark, S. Capella-Gutierrez, A.J.G. Gray, S. Owen, A.R. Williams, G. Tartari, F. Bacall, T. Thelen, H. Ménager, L. Rodríguez-Navas, P. Walk, brandon whitehead, M. Wilkinson, P. Groth, E. Bremer, L.J. Castro, K. Sebby, A. Kanitz, A. Trisovic, G. Kennedy, M. Graves, J. Koehorst, S. Leo, M. Portier, P. Brack, M. Ojsteršek, B. Droesbeke, C. Niu, K. Tanabe, T. Miksa, M. La Rosa, C. Decruw, A. Czerniak, J. Jay, S. Serra, R. Siebes, S. de Witt, S. El Damaty, D. Lowe, X. Li, S. Gundersen, M. Radifar, RO-Crate Metadata Specification 1.1.3, (2023). https://doi.org/10.5281/zenodo.7867028.

[22] S. Peroni, S. Soiland-Reyes, P. Sefton, M. Crosas, L.J. Castro, F. Coppens, J.M. Fernández, D. Garijo, B. Grüning, M. La Rosa, S. Leo, E. Ó Carragáin, M. Portier, A. Trisovic, P. Groth, C. Goble, Packaging research artefacts with RO-Crate, Data Sci. 5 (2022) 97–138. https://doi.org/10.3233/DS-210053.

[23] A. Afonso, S. Schmiedl, C. Becker, S. Tcherny-Lessenot, P. Primatesta, E. Plana, P. Souverein, Y. Wang, J.C. Korevaar, J. Hasford, R. Reynolds, M.C.H. de Groot, R. Schlienger, O. Klungel, M. Rottenkolber, A methodological comparison of two European primary care databases and replication in a US claims database: inhaled long-acting beta-2-agonists and the risk of acute myocardial infarction, Eur. J. Clin. Pharmacol. 72 (2016) 1105–1116. https://doi.org/10.1007/s00228-016-2071-8.

[24] OpenSAFELY, OpenSAFELY, (2021). https://www.opensafely.org/ (accessed March 5, 2024).

[25] OHDSI, Observational Health Data Sciences and Informatics, (2016) 1–37. https://www.ohdsi.org/ (accessed March 5, 2024).

[26] CPRD, Clinical Practice Research Datalink | CPRD, (n.d.). https://cprd.com/ (accessed March 5, 2024).

Table 1 - Differences between the original GMCR study and the two analyses in this replication study

	Original GMCR study	**This study - 1st analysis**	**This study - 2nd analysis**
Population	Patients registered with a GP in Greater Manchester. Does not include individuals who have opted out of secondary use of their GP data.	Patients registered with a GP in England, UK, in practices that opted-in for GPES extraction*. Does not include individuals who have opted out of secondary use of their GP data.	
Primary care data	Direct feed from GP practices. Containing all events in the patient record.	Data from the GDPPR dataset. Contains a subset of records in the patient record that were both available via GPES and considered relevant to pandemic planning and research.	
Admission data	Direct feed from each hospital within GM	HES APC data	
COVID-19 test data	From GP record	From GP record	From SGSS data and GP record

** 98% of practices in England*

Table 2 - Full list of recommendations for Trusted Research Environments (TREs) and Secure Data Environments (SDEs) to facilitate replication studies

Recommendation 1	SDEs/TREs that wish to allow federated analysis should consider unified application processes so that researchers are only required to apply once.
Recommendation 2	SDEs/TREs should clearly define what research outputs are permitted, the process that is used for assessment, and any appeals process.
Recommendation 3	Metadata catalogues designed specifically for longitudinal EHR data should be researched and developed.
Recommendation 4	Research is needed to develop computable study definitions that can be executed against machine-readable metadata catalogues
Recommendation 5	SDEs/TREs should consider how automation and other efficiencies can reduce access costs. Ensuring that replication and federated analyses do not become prohibitively expensive is crucial for the advancement of research.
Recommendation 6	SDEs/TREs should be designed to be agile and adaptable, incorporating best practices as they evolve. The SATRE specification is the most likely source for these best practices.
Recommendation 7	SDEs/TREs should ensure that data can be accessed and processed with multiple languages such as SQL, R and Python.
Recommendation 8	SDEs/TREs should implement mechanisms to monitor and manage execution time variability. Providing researchers with tools to estimate and optimise execution times can improve the efficiency and reliability of data analysis.
Recommendation 9	SDEs/TREs should allow safe content to be easily imported. Where import controls are enforced, they should aim to reduce the barriers to researchers wherever possible.
Recommendation 10	SDEs/TREs need a support structure for researchers which includes people with expertise in the underlying data.

Recommendation 11	Libraries of code, clinical code sets and phenotypes should consider their editorial policy. If there are no barriers to uploading content, then standardised tools should be created to allow easy discovery and comparison of the digital artifacts.
Recommendation 12	SDEs/TREs should mandate the sharing of data curation and data analysis code.