Plan Overview

A Data Management Plan created using DMPonline

Title: Copy of Long-term kidney outcomes after COVID-19

Creator: Viyaasan Mahalingasivam

Affiliation: London School of Hygiene and Tropical Medicine

Template: LSHTM DM Plan for research projects

Project abstract:

The long-term consequences on the health of COVID-19 survivors remains largely unknown but there are justifiable concerns that it may lead to an increase in the burden of chronic disease.

Through the use of large population electronic health records, my doctoral research will investigate the impact of SARS-CoV-2 infection on long-term kidney outcomes in the general population, as well as a range of long-term clinical outcomes on people with pre-existing chronic kidney disease (CKD). This research will be important in helping health systems plan configuration of services in the ongoing pandemic-era and beyond. Part of this research will contribute to the COVID-19 Longitudinal Health and Wellbeing National Core Study and I am member of the OpenSAFELY collaboration.

My final thesis will bring together several workstreams:

1. A literature review on epidemiological research on COVID-19 and kidney disease to date,
2. An investigation into the long-term kidney outcomes after SARS-CoV-2 infection in the general population (Study 1),
3. An investigation into long-term mortality and complications after SARS-CoV-2 infection in people with pre-existing CKD (Study 2).

This research is being continually shaped through collaboration with patient and public involvement (PPI) partners and my training and development programme. This will include an additional study to be conducted in Year 3, which I will plan by early 2023. I will also undertake economic evaluation of my research after completing training courses as mandated by NIHR (with additional appropriate supervision through my supervisory panel).

ID: 98512

Start date: 01-04-2021

End date: 31-07-2024

Last modified: 14-04-2022

Grant number / URL: NIHR301535

the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Copy of Long-term kidney outcomes after COVID-19

Introduction

Project Title

Long-term kidney outcomes after COVID-19

Name of the Principal Investigator

Viyaasan Mahalingasivam

Funder(s)

National Institute for Health Research

State the end date of your project

31/07/2024

Data Description

What primary data will you collect or create?

We will be using routinely-collected electronic health record data on the OpenSAFELY platform. This will comprise primary care data with additional linkage to other data sources (e.g. NHS Secondary Uses Service).

What data formats or standards will you use to store data produced by your project?

OpenSAFELY aims to substantially exceed, by design, the current requirements on securing sensitive healthcare data.
OpenSAFELY does not move patient data outside of the secure environments where it already resides: instead, trusted analysts can run large scale computation across pseudonymised patient records in situ, and in near-real-time.
We do not have unconstrained access to view and manipulate raw data on a remote machine: instead, we will work on the data at arm’s length using OpenSAFELY services.

What documentation or metadata is needed to understand your data?

Data management will be undertaken using Python, with analysis carried out using Stata 17. The study protocol, code for data management and analysis, and code lists used to define study measures (exposures, outcomes, covariates) are available at https://github.com/opensafely/post-covid-kidney-outcomes.
All code created for data management and analysis will be shared, informatively, for review and re-use by all subsequent users. In most settings for NHS patient data analysis the same data management tasks are achieved by a huge range of bespoke and duplicative methods, in a huge range of different tools, with single tasks often spread between platforms or programming languages. In OpenSAFELY the data management is always done the same way, using the same OpenSAFELY tools, so code is created in a form where it can be quickly read, understood, adapted, and re-used by any user for any other data science project.
All code ever executed against the patient data can be shared, as an informative public log, because none of that code is disclosive of patient data. Normally Trusted Research Environments that execute code against real patient data try to keep a log of activity in the platform, but they cannot share every action in each user session, because the code for data management and analysis was generated while working directly with the real data, and so there is a substantial risk that the code itself might contain some disclosive information about individual patients. Some platforms log screen recordings, or keystrokes, for later review, but these can also never be shared openly, because of the disclosure risk (they are also laborious to review). OpenSAFELY code is only generated by working with dummy data, so we can be certain that it is non-disclosive. This means it can be shared, and so we do share it: all of it, automatically, in public, and by default. This means that every interested stakeholder can see what we have done with patients' data inside OpenSAFELY, and all patients, professionals and policymakers can be confident that data has only been used for the purpose for which access was granted: this is crucial for building trust, and a substantial improvement on the current paradigm

whereby Trusted Research Environments only share a list of projects with permissions. The removal of privacy risks from data management and analysis code also frees up OpenSAFELY code for sharing and re-use under open licenses: it means that there is no information governance or privacy barrier to users sharing code for others to review, critically evaluate, improve, and re-use, wherever they wish.

What codes of practice, if any, will you follow for creating and handling data?

OpenSAFELY contains a range of flexible, pragmatic, but broadly standardised tools that users work with to convert raw patient data into “research ready” datasets, and to then execute code across those datasets. Standardising the data management pathway in this way brings numerous benefits around re-usability, efficiency, security, and transparency.
After completion of each analysis, only minimally disclosive summary data is released outside the secure environment, such as summary tables or figures, after strict disclosivity checks and redactions, to ensure safe data, safe settings and safe outputs. When access to any TRE, including one using OpenSAFELY code, is considered to be appropriate, the dataset described by the study definition should also be justified and proportionate, in accordance with the Caldicott principles and the DCMS data ethics framework.

What approach will you take to handling rights associated with the data?

OpenSAFELY code is only generated by working with dummy data, so we can be certain that it is non-disclosive. This means it can be shared, and so we do share it: all of it, automatically, in public, and by default. This means that every interested stakeholder can see what we have done with patients' data inside OpenSAFELY, and all patients, professionals and policymakers can be confident that data has only been used for the purpose for which access was granted: this is crucial for building trust, and a substantial improvement on the current paradigm whereby Trusted Research Environments only share a list of projects with permissions. The removal of privacy risks from data management and analysis code also frees up OpenSAFELY code for sharing and re-use under open licenses: it means that there is no information governance or privacy barrier to users sharing code for others to review, critically evaluate, improve, and re-use, wherever they wish.

Data Storage and Management

Where will you store data during the project lifetime?

- Dedicated server maintained at partner institution

We will be using data on OpenSAFELY-TPP. OpenSAFELY-TPP has been implemented inside the data centres of TPP, one of the largest providers of GP electronic health record software in England, in the locations where patients' records already reside. This means that the data never moves location.

What security measures, if any, will you apply to protect data?

- Remove identifiable information (e.g. anonymisation)
- Password protection
- Avoid use of third party storage, such as Dropbox
- Controlled access limited to authorized users only

OpenSAFELY aims to substantially exceed, by design, the current requirements on securing sensitive healthcare data.
OpenSAFELY does not move patient data outside of the secure environments where it already resides: instead, trusted analysts can run large scale computation across pseudonymised patient records in situ, and in near-real-time.
In the case of OpenSAFELY-TPP and OpenSAFELY-EMIS, we have implemented OpenSAFELY inside the data centres of the largest providers of GP electronic health record software in England, in the locations where patients' records already reside. This means that the data never moves location. (It also means that we get to work closely with EHR software developers in these companies, who know their own data extremely well).
However this is not the only privacy safeguard. In addition, we do not give users unconstrained access to view and manipulate raw data on a remote machine: instead, users work on the data at arm’s length using OpenSAFELY services.
OpenSAFELY contains a range of flexible, pragmatic, but broadly standardised tools that users work with to convert raw patient data into “research ready” datasets, and to then execute code across those datasets. Standardising the data management pathway in this way brings numerous benefits around re-usability, efficiency, security, and transparency.

1. *All code created for data management and analysis can be shared, informatively, for review and re-use by all subsequent users.* In most settings for NHS patient data analysis the same data management tasks are achieved by a huge range of

bespoke and duplicative methods, in a huge range of different tools, with single tasks often spread between platforms or programming languages. In OpenSAFELY the data management is always done the same way, using the same OpenSAFELY tools, so code is created in a form where it can be quickly read, understood, adapted, and re-used by any user for any other data science project.

2. *Users are blocked from directly viewing the raw patient data or the research ready datasets, but still write code as if they were in a live data environment*. In most other settings analysts write their code, which converts raw data into finished graphs and tables, by working directly with (and seeing) the real data, iterating and testing as they go. In OpenSAFELY, the data management tools used to produce their research-ready datasets also produce simulated, randomly generated “dummy data” that has the same structure as the real data, but none of the disclosive risks. Every researcher is therefore provided with a full, offline development environment where they can build all their data management and analysis code quickly, but only against dummy patient data. This minimises needless interaction with disclosive patient records and allows anyone with technical skills to swifty check and reproduce the methods. Researchers develop all their code for statistical analysis, dashboards, graphs and tables against this dummy data, using open tools and services like GitHub. Their code is then tested automatically by the OpenSAFELY tools, using the dummy data. When it is capable of running to completion, it is packaged up inside a “container”, using a tool called “Docker”. All their data management and analysis code is then sent securely into the live data environment to be executed against the real patient data: researchers can only view their results tables and graphs, but no researcher ever needs to enter the real patient-data environment, or see the real patient data. It is useful to contrast this against other settings which work with synthetic data (real data, but with statistical noise added in an effort to preserve privacy): they typically require researchers to also use that synthetic data to run their analyses, which can undermine the reliability of the results: in OpenSAFELY the synthetic dummy data is *only* used for code development, not code execution. In this way we get all the privacy preserving benefits of completely random synthetic data, but also retain all of the analytic benefits that come from executing code against real patient data.
3. *All code ever executed against the patient data can be shared, as an informative public log, because none of that code is disclosive of patient data.* Normally TREs that execute code against real patient data try to keep a log of activity in the platform, but they cannot share every action in each user session, because the code for data management and analysis was generated while working directly with the real data, and so there is a substantial risk that the code itself might contain some disclosive information about individual patients. Some platforms log screen recordings, or keystrokes, for later review, but these can also never be shared openly, because of the disclosure risk (they are also laborious to review). OpenSAFELY code is only generated by working with dummy data, so we can be certain that it is non-disclosive. This means it can be shared, and so we do share it: all of it, automatically, in public, and by default. This means that every interested stakeholder can see what every analyst has done with patients' data inside OpenSAFELY, and all patients, professionals and policymakers can be confident that data has only been used for the purpose for which access was granted: this is crucial for building trust, and a substantial improvement on the current paradigm whereby TREs only share a list of projects with permissions. The removal of privacy risks from data management and analysis code also frees up OpenSAFELY code for sharing and re-use under open licenses: it means that there is no information governance or privacy barrier to users sharing code for others to review, critically evaluate, improve, and re-use, wherever they wish.

These working methods, and the code in which they are embodied, mean that OpenSAFELY substantially exceeds current best practice around secure execution of analysis code on pseudonymised patient data, when combined with the other governance features of a strong TRE. After completion of each analysis, only minimally disclosive summary data is released outside the secure environment, such as summary tables or figures, after strict disclosivity checks and redactions, to ensure safe data, safe settings and safe outputs. When access to any TRE, including one using OpenSAFELY code, is considered to be appropriate, the dataset described by the study definition should also be justified and proportionate, in accordance with the Caldicott principles and the DCMS data ethics framework.

Data Sharing and Reuse

Is any of your data *NOT* suitable for sharing beyond the current project? If so, state the reason

Because the data is updated in real-time, it would not be possible to re-analyse exactly the same data. However, data management code is always shared so it can be re-executed on OpenSAFELY for the purposes of repeat analyses.

Can data be made available to anyone? If not state the reason it needs to be restricted and criteria for gaining access

Researchers who hold honorary contracts with NHS England and have signed Data Access Agreements relevant to level 4 access for the purposes of checking and redacting data prior to release.

What actions will be performed to prepare your data for sharing?

- Develop an access agreement

I have approved Level 4 access for the purposes of checking and redacting data prior to release with a signed agreement for access

to data for COVID19 purposes.

When are you likely to make it available?

- During project lifetime

How do you intend to make it available?

- Via a Collaborator-maintained system

All data management and analysis code is sent securely into the live data environment to be executed against the real patient data: researchers can only view their results tables and graphs, but no researcher ever needs to enter the real patient-data environment, or see the real patient data.

How will potential users learn of the data's existence and how it may be accessed?

- Direct contact with data creator/project lead
- Description on project website
- Citation in publications
- Described in publications (e.g. in an access statement)

Resourcing

What do you consider to be the primary data management challenges in your project?

All outputs for release will need to be checked according to OpenSAFELY policy by both the researcher and a trained colleague. I have attended an output checking course run by the Office for National Statistics.

What resources would it be helpful for the School to provide to help deliver your plan?