Background

Following the end of the COPI notice, the legal basis of the OpenSAFELY COVID-19 Service transitioned to a Direction, which requires that OpenSAFELY to respect Type 1 opt-outs (T1OO), by making the data of anyone who has registered a T1OO unavailable to users. This is easily solvable for all of the currently registered population in either TPP or EMIS, as the T1OO code will be recorded in their patient data.

OpenSAFELY (in TPP and EMIS) retains patient data from deregistered patients (i.e. people with an ongoing deregistered status). In the scenario of a patient moving from a TPP practice to an EMIS practice and subsequently requesting a T1OO in the new EMIS practice, their records remaining in the TPP practice would not include that T1OO request. Currently, TPP and EMIS do not share a list of T1OOs between their systems, as there has never been a requirement to do so. Therefore, it is possible that OpenSAFELY could process the deregistered patient data in TPP *after* the patient has requested a T1OO in their new EMIS practice (or vice versa).

This paper discusses how an urgently developed interim mitigation, to ensure OpenSAFELY upholds the T1OO for deregistered patients, will result in significant methodological issues for epidemiological studies. These issues may then cause consequential erroneous study conclusions and risk causing harm to patients. Fortunately, several options exist to resolve these methodological issues.

For clarity, this paper describes the methodological issues from the perspective of running studies in the TPP side of OpenSAFELY, but these issues could equally be applied to studies run in the EMIS side.

The current interim mitigation

The urgently developed mitigation, to allow the T1OO to be upheld for deregistered patient data, given the current absence of sharing of the T1OO list between TPP and EMIS, is for OpenSAFELY to enforce a blunt assumption that any person who has deregistered from a TPP practice (i.e. not died whilst a patient in TPP) has a small chance of having requested a T1OO that we don’t know about (i.e. by requesting this T1OO if they register with a new EMIS practice).

By excluding all the people who have deregistered from a TPP practice, then we would be sure to exclude all the people who could have requested a T1OO out after deregistering. We do not need to exclude people who died on or before the date of their deregistration, since it is not possible for someone to opt out after death.

This solution makes us confident of excluding people with a T1OO.

Methodological issues with excluding all deregistered people

Excluding all deregistered patients means we would also be excluding large numbers of people whose data we would be allowed to use if we could accurately determine their T1OO status. Epidemiological studies typically run over several years; OpenSAFELY holds GP patient data from patients who were registered since 2009, which means there are 15 years of patient deregistrations. Our analysis shows that 6.1 million patients have an ongoing deregistered status in our TPP dataset since 2009, all of whom would need to be excluded. This will lead to studies encountering substantial epidemiological problems, which we describe below.

A substantially less representative population

If we were excluding 6.1 million people from the OpenSAFELY population *at random*, we would not face serious issues with using the data. However, it is expected that the deregistered patient population will be substantially different to that of the population as a whole, and excluding those people will substantially affect the overall makeup of the population. It is not possible to *quantify* these effects without a detailed study comparing the included/excluded populations, but the following are examples of the types of effects we would expect to see.

Certain demographic groups are more likely to move around than other groups, making them therefore more likely to deregister from any given practice and therefore be excluded by our interim measure. Examples include:

- Young people - more likely to be leaving parental homes, going to university, changing jobs, which would cause them to register at a different practice
- People moving into a care home - might be moving geographically to do so, and have health needs that are very different to the population as a whole
- Specific geographic regions - for example people in London be more likely than average to change GPs due to younger demographics, a high proportion of people renting etc

Excluding a higher proportion of people such as the list above will make any study using the data substantially less representative of the population as a whole.

Worked example

We want to study the incidence of acute kidney injury each month over the course of the pandemic.

- Our study time window would begin at the start of the pandemic, e.g. Feb 2020.
- To be eligible for inclusion in the study people must be registered with a TPP practice at the start of the study time window.
    - This is so that we can be sure that diagnoses of acute kidney injury are recorded reliably. If we were to include people who were not registered in Feb

    - 2020 but instead registered later, we can’t be sure that such events would be transferred to their record.
    - Also, including people registered after the study start date would also constitute using information from “the future” (i.e. after the study start date). This is a common pitfall in epidemiological research, and can lead to issues such as immortal time bias.
- Our study is only interested in the first recorded acute kidney injury diagnosis, so conventionally, patients would be followed up from the start date (Feb 2020), until they had an acute kidney injury diagnosis, died, or deregistered.
- The rate of acute kidney injury in the population would be calculated by dividing the total number of events in the month of interest, by the total “person time” (i.e. total number of days for all patients in that time window) contributed by the eligible population in that month.
- If we had to exclude patients who were deregistered at any time, we’d be more likely to exclude younger people. As younger people are overall at less risk of acute kidney injury, we would end up measuring an older population in OpenSAFELY. This would lead to an overestimation of the rate of acute kidney injury.

The effect changes over time

Electronic health record studies are retrospective in nature and therefore always have a start date in the past. OpenSAFELY removing the 6.1 million deregistered patients means that the further back in time a study starts, the smaller the available registered population size that study would have access to. This is because when a study has an start date that is further in the past, the patients in that study have a longer opportunity to deregister, at which point they must be excluded by our proposed mitigation.

Moreover, this smaller available study population is not compensated by the influx of newly registered patients after the study start date ***because*** to be included in a study, each patient will have to have been registered at a TPP practice at the start date of the study.

The consequence of this for any follow up study is that the available study population will always be lower the earlier the study starts This means the biases described above that are introduced by excluding deregistered people will be greater the further back you look. This is especially an issue for studies that require a lot of follow-up time, as they would need to start as early in the data as possible. It is also problematic for studies that are interested in how something changes over time.

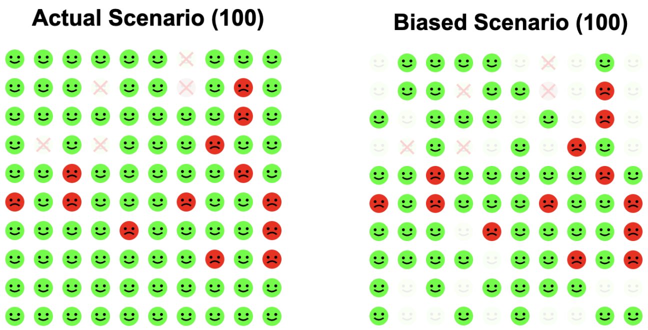
Mortality rate (and other outcomes) will be artificially inflated

To calculate the rate of death for a population in any given time period, we’d take the count of deaths in that time period (the numerator), and divide it by the number of people in the population (the denominator). By excluding people who are deregistered, the available study population at the start of the study will always be smaller than the true study population. This effect will increase as you look further back in time. As we’re only removing people who did not die while registered the measured number of deaths will be unaffected. We are therefore decreasing the denominator without changing the numerator, which will increase the apparent mortality rate the further you go back in time.

As we’re excluding more people the further back we look, this artificial increase will be greater going further back in time.

The rate of other diseases/outcomes will similarly be increased the further back we look because the population we’re left with after excluding deregistered people are more likely to die, and are therefore sicker on average than the whole population. Outcomes with a strong association with death, such as cancer, heart attack, stroke, will be especially strongly affected.

The overall effect of this is that less recent mortality rates (and other outcomes) will have a mortality rate that is higher than it actually is.



Actual Scenario (100)

Biased Scenario (100)

Worked examples

We want to study the mortality rate in the whole population for each year over the last 15 years.

- For each year we’re interested in, we take the population of people who were alive and currently registered in a TPP practice at the start of each year.
- We count the total deaths that occur in each year.
- The annual death rate for each year is calculated by dividing the total deaths each year (the numerator) by the population at the start of each year (the denominator).

We would expect the numerator (number of deaths) to be unaffected by this issue, as we’re able to keep all patients who died. However, the cohort of patients from 2009 would have a much higher proportion of deregistered patients than the cohort of patients from 2023, because much more time has passed for them to deregister. This could cause as much as X% artificial inflation of the mortality rate in 2009.

We want to study the long term mortality rate for cancer patients

Doing a study that follows a cohort of patients up for a long time, say 10 years, means that the start date of the study must necessarily be at least 10 years in the past. This means that studies requiring long follow-up will be strongly affected by the issues described above as the study denominator, the available study population, will be significantly less than the true study population, **overestimating the mortality rates**.

Sample size

The most obvious effect of excluding a large proportion of the OpenSAFELY population is a reduced sample size for any studies that use the data. While there would still be a large remaining population in OpenSAFELY TPP, having a large sample size is important for most studies. Having a smaller sample size will make every study conducted a bit worse, and will make some types of study entirely unviable, for example:

- Reduce the precision of all studies, making us less certain of any results that are obtained.
- Reduce the ability to look at small sub-divisions of the population, for example determining the incidence of a disease in a specific age group, sex, and ethnic group.
- Limit the ability of studies that have a borderline level of statistical power to detect an effect, such as those looking at rare diseases.

Working on a fix

We hope that this interim arrangement will be short lived, and our teams are working energetically and in broad collaboration on a better approach that we hope will be implemented soon. Any code you write and test now will be able to run without issue on a more representative population that becomes accessible. Our advice is therefore to continue working, be aware of the issues above, but seriously consider re-executing your analysis prior to release.