CCU040 study protocol – replication study

# Study Title

The factors associated with increased risk of hospitalisation and death in people with diabetes following SARS-CoV-2 infection: A national replication study.

# Version history

|  |  |  |
| --- | --- | --- |
| v0.1 | 2022-02-16 | Initial draft |
| v0.2 | 2022-02-24 | Incorporated feedback from Naveed Sattar and Kamlesh Khunti |
| v1.0 | 2022-02-25 | Version bumped to 1.0 for submission to consortium |
| v1.1 | 2023-05-05 | Added the section on proposed outputs |

# Notes

# This protocol covers the three planned outputs listed in the section below, which have been assigned the subproject references CCU040\_01, CCU040\_02 and CCU040\_03.

# Lay summary

People with type 1 or type 2 diabetes are more likely to be admitted to hospital, and more likely to die, after getting infected with COVID-19. We have recently used data from Greater Manchester (GM) to discover why this might be. We found several things that increased the risk of poor outcomes, such as age, being male, being socially deprived, ethnicity, certain medications and certain health conditions.

We will now try to repeat these results using the national COVID-IMPACT data. Studies such as this (a replication study) where people attempt to reproduce results from other studies are important. If the results are the same, then it strengthens the findings. If different then that is important as well as it shows the initial results might not be as generalisable as previously thought and may be related to factors that may not have been considered in some of the studies. This work will also help us to understand the role of regional and national datasets and how they might be best suited to different research questions.

We also plan to extend the analysis to take advantage of any extra data that is available nationally, but that was not available in GM.

# Research hypothesis, aims and questions

The aim of this research is to attempt a replication, and extension, of two studies previously carried out using the Greater Manchester Care Record. Both papers are currently in submission, one of which has been accepted. Links to the papers will be provided when available.

The first study examined the risk of hospitalisation in patients with type 1 and type 2 diabetes. Various factors were shown to contribute to a person’s risk of admission including: age, being male, high BMI, having hypertension or COPD, low hdl cholesterol, low eGFR, higher deprivation and being of African ethnicity.

The second study looked at the risk of mortality in patients with type 2 diabetes. There were too few patients with type 1 diabetes who died following a COVID-19 infection which is why this study only focussed on type 2 diabetes. The factors that were shown to contribute to a person’s risk of death included: a low eGFR, having COPD or severe mental illness, older age, being male, being deprived and smoking. A protective association was seen for patients on metformin, SGLT-2i and GLP-1.

One of the largest limitations of observational studies is that of reproducibility. Goodman et al. define three terms for discussing research reproducibility: methods reproducibility, results reproducibility and inferential reproducibility [1]. 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. Inferential reproducibility is the degree to which different researchers would reach the same conclusion based on similar results.

For this study, methods reproducibility is trivial as we performed the original analysis. For assessing the results and inferential reproducibility we will conduct 3 analyses as shown in Table 1. The analyses are differentiated by the way in which the population is defined, and how a positive COVID test is defined. The reasons for these differences are explained below.

Table 1 - Population and COVID test definitions for the 3 replication analyses

|  |  |  |
| --- | --- | --- |
| Analysis # | Population | COVID test |
| 1 | Patients registered with a GP in Greater Manchester on 30 June 21 | Positive tests as entered in the primary care record |
| 2 | Patients registered with a GP in Greater Manchester on 30 June 21 | Positive tests from either the pillar 2 dataset or the SGSS dataset |
| 3 | Patients registered with a GP in England on 30 June 21 | Positive tests from either the pillar 2 dataset or the SGSS dataset |

Our initial study relied on COVID test results that appear in the primary care record. We also were limited to the population of Greater Manchester. Analysis 1 is therefore an attempt to reproduce the results of our original study by attempting to filter the data from COVID-IMPACT so that it is equivalent to the data that we used. Any discrepancies will be investigated to discover the cause.

Analysis 2 retains the filter to Greater Manchester patients, but uses the improved sources of COVID tests. Whether or not the results can be reproduced, will help show to what extent the COVID test results in primary care records can be relied upon.

Analysis 3 will then extend to the entire population of England. If results and inferential reproducibility can be achieved then this will provide some evidence that, under certain circumstances, scientific conclusions drawn from regional datasets can be extrapolated nationally. We will also have more power, so some results that were not statistically significant, may now become so.

*NB – 30 June 2021 is when the data in our previous analysis was examined up to. Therefore in order to replicate precisely we will use this date as the cut-off for this analysis.*

# Objectives

There are many data items and phenotypes available in the COVID-IMPACT database that were not available in the GM data that we have previously used. Therefore in addition to a like for like replication, we will also extend the analysis to include these extra covariates. Our objectives are therefore:

1. Attempt to replicate the previous GM studies using the COVID-IMPACT database
2. Use the same methods but extend the analysis to include additional covariates and update it with the most recent data

# Data

Table 2 - Required datasets and justification

|  |  |  |
| --- | --- | --- |
| Dataset | Data required from | Justification |
| GDPPR | All | The primary care record will be the location of the majority of the requested data, such as the diagnoses, medications, lab tests, and demographic data |
| HES   * Admitted patient care | 2020 | Unplanned hospital admission is an outcome of interest. |
| COVID-19 SGSS | ALL | Our cohort is patients with a positive COVID-19 test |
| Pillar 2 | ALL | Our cohort is patients with a positive COVID-19 test |
| Civil Registration – Deaths | 2020 | Death is an outcome |

# Study design

This is a retrospective case-controlled study.

For all 3 analyses the main cohort is defined as:

Patients with a diagnosis of type 1 or type 2 diabetes mellitus who also have a positive COVID-19 test result.

For each analysis, we will attempt to generate a matched cohort (1:3) of patients without diabetes, but who have a positive COVID-19 test result. The matching will be done on age (+/- 5 years), sex, and date of positive COVID-19 test (+/- 2 weeks). If possible the matching will also include deprivation quintile and ethnicity – but only if there are sufficient matches for each patient. The matching on date of COVID test is to take account of the fact that patient outcomes are likely different depending on the particular wave or variant of COVID that they contracted.

For objective #1, the covariates to be used for the replication are the same as used in the previous GM papers. They are:

* Year of birth
* Sex
* Lower level super output area
* Ethnicity
* Townsend score
* Latest values for: BMI, Hba1c, cholesterol, ldl, hdl, vitamin d, testosterone, egfr and shbg
* Smoking status
  + Has a passive smoking code
  + “Worst” smoking status in patient’s record
  + Most recent smoking status in patient’s record
* Whether the patient has: COPD, asthma, a severe mental illness or hypertension
* Whether the patient is currently prescribed: ACE inhibitor or ARB, aspirin, Clopidogrel, metformin, insulin, SGLT, GLP1a or sulphonylurea

The outcomes are:

* Hospitalisation within 28 days of first COVID-19 infection
* Death within 28 days of first positive COVID-19 test

For objective #2 we will also include the following covariates in addition to the above:

* Latest values for: systolic blood pressure, pulse, triglycerides
* Whether the patients has: CKD, AF, heart failure, cancer, previous stroke, liver disease, or a previous MI
* Whether the patient is currently prescribed: statin, ezetimibe, DPP4i, diuretic, CCB

These lists are not necessarily exhaustive and will depend on precisely what data is available, and reliable, within the database.

# Statistical methods

The methods will be an exact replication of the previous 2 studies. A brief overview is as follows.

Modelling will be conducted using logistic regression. One model will have death as the binary outcome variable (up to 30 June 2021 for objective #1, and up to the latest reliable date in the COVID-IMPACT database for objective #2), the other model will have hospitalisation within 28 days of a positive COVID test as the outcome. Both models will use the covariates described above. Specifically, we will do the following;

Firstly, to investigate potential factors associated with admission and mortality in people with diabetes, we analyse the individuals with T1DM and T2DM, without the matched individuals, separately. To do so, we will use univariate logistic regression, considering each possible factor in turn, for the two groups separately.

Next we will study whether the difference in risk between patients with diabetes and patients without diabetes was explained by other measured factors. To do so we will analyse each diabetes group with their matched cohort and compared the OR of diabetes in a model including only age, sex and diabetes with the OR of diabetes when adjusted for each additional factors in turn. Attenuation of the diabetes OR in this comparison will be interpreted as the additional factor explaining part of the association between diabetes and hospitalisation/mortality (for example, through confounding or mediation).

Finally, we will fit a fully adjusted multivariable model, separately in each matched cohort, to measure the extent of attenuation of the diabetes OR when all additional factors are accounted for.

Following these analyses we will compare the effect sizes and ORs to our previous work from the GMCR dataset.

# Planned outputs

The results from the analyses described in this paper are wide ranging and will likely appeal to multiple audiences. The results that are diabetes-specific and that arise as a result of the increased statistical power, will be of interest to people in the field of diabetes and should be targeted at diabetes journals. Whereas the methodological discussions and the impacts of the success or failure of the replication are more suited in health informatics journals. Papers that arise from the results of this analysis will be listed here:

CCU040\_01: Sars-Cov-2 infection in people with Type 1 Diabetes and Hospital Admission: An Analysis of Risk Factors for England

* The original GMCR paper did not have sufficient power to report significance for Type 1 diabetic patients. With the larger sample size most factors have now achieved significance. The results from this analysis are best suited to a diabetes journal.

CCU040\_02: The challenges of reproducibility: A worked example

* In performing the replication study described above, as expected, several challenges were encountered. This paper will describe the differences between the environments, comment on the challenges of reproducibility, and provide suggestions for the improvement of future secure data environments.

CCU040\_03: The extent to which results from a regional healthcare database scale nationally: A replication study

* This is the main thrust of the above protocol.

# Outstanding questions and responses

|  |  |  |
| --- | --- | --- |
| Raised by | Query | Response |
| Naveed Sattar | You look at people known to have COVID. Might there be a bias in diabetes vs non-DM in who got positive testing? | Perhaps. There is certainly going to be a bias in the first few months of the pandemic where people in hospital will be more likely to be tested than the general population. However it seems plausible that this bias would be the same for diabetes and non-diabetes patients. |
|  |  |  |
|  |  |  |

# References

[1] 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.