Statistical Analysis Plan

# Rationale and Overview

This document extends the Diabetes UK Grant Statistical Analysis Plan (SAP), specifically focusing on element **2c** which will be performed by The University of Edinburgh.

The hypothesis for analysis 2c:

Risk ratios for CVD death following non-COVID hospitalised pneumonias are not substantially different to the risk ratios following a COVID related hospitalised pneumonia.

# Study Design

Study 2c will utilise a longitudinal prospective and retrospective cohort study design.

# Study Period

The ‘post-exposure’ study period for analyses investigating the association of exposure to COVID-19 with bacterial or viral pneumonia with and without COVID-19 will be 1st January 2020 – 31st December 2021.

# Study Population

Individuals with type 2 diabetes who are present in the SCI-Diabetes dataset and alive and observable 1st January 2016 for the pre-COVID-19 cohort and on 1st January 2020 for the COVID-19 cohort will be included in the study populations. Type 2 diabetes will be defined using the SDRN-epi type assignation algorithm.

# Definition of Exposure

Exposure will be defined as a hospital admission with bacterial or viral pneumonia. Pneumonia will be subset as:

1. Admitted for any bacterial or viral pneumonia
2. Pneumonia excluding those prior COVID-19 positive
3. Pneumonia with a positive COVID-19 test within 8 weeks of pneumonia admission
4. Pneumonia and COVID-19 positive concurrent with the admission.

The following ICD-10 condition codes will be used to calculate pneumonia and COVID-19.

* Pneumonia – J10.0, J11.0, J12-J18
* COVID-19 – U07.1, U07.2 or a positive PCR test within a hospital admission window.

# Definition of Outcomes

For the analysis, we will consider the following outcomes:

* Cardiovascular mortality in the study window

The following ICD-10 condition codes will be used for cardiovascular related mortality

I20-I25, I26-I28, I30-I39, I40-I49 I50-I52, I60-I69 and I46.

* Cardiovascular disease prior to and during the study window

I20-I25, I50, I61-I64, G45, I70.2, I70.8, I70.9, I73.9, E10.5, E11.5. E13.5 and E14.5

Hospital admission data will derived from Scottish Morbidity Record (SMR) 01. Mortality data will be obtained from Medical Certificate of Cause of Death (MCCD) data provided by National Records of Scotland (NRS). Cause specific information for cause of death will be defined by ICD-10 coding present at any point on an MCCD. Cause specific SMR01 information will be defined by ICD-10 coding in any position on any SMR01 record. In-hospital mortality will be defined as any death which occurs during a period of hospital admission, as defined by the start and end date of an SMR01 record. Where possible we shall re-use covariate definitions from previous COVID-19 analyses performed by the research group:

(https://dx.doi.org/10.1016%2FS2213-8587(20)30405-8; https://doi.org/10.1371/journal.pmed.1003374)

# Observability

The observability status of individuals will be defined using attendance of routine observations and receipt of prescriptions during the study period. If Individuals become unobservable during the study period they will be censored on the date at which they first become unobservable.

# Derivation of Dataset for Analysis

Individuals will contribute person time from the latest of: study start date, diagnosis of diabetes. To be included in the study population, individuals must be observable at the start of the study period. Individuals will stop contributing person time to the analyses on the earliest of: the end of the study period, date of death, first date of at which the individual ceases to be observable, the date at which the individual experiences the specific outcome of interest for the analysis (e.g. hospital admission, cause-specific hospital admission).

# Statistical Analyses

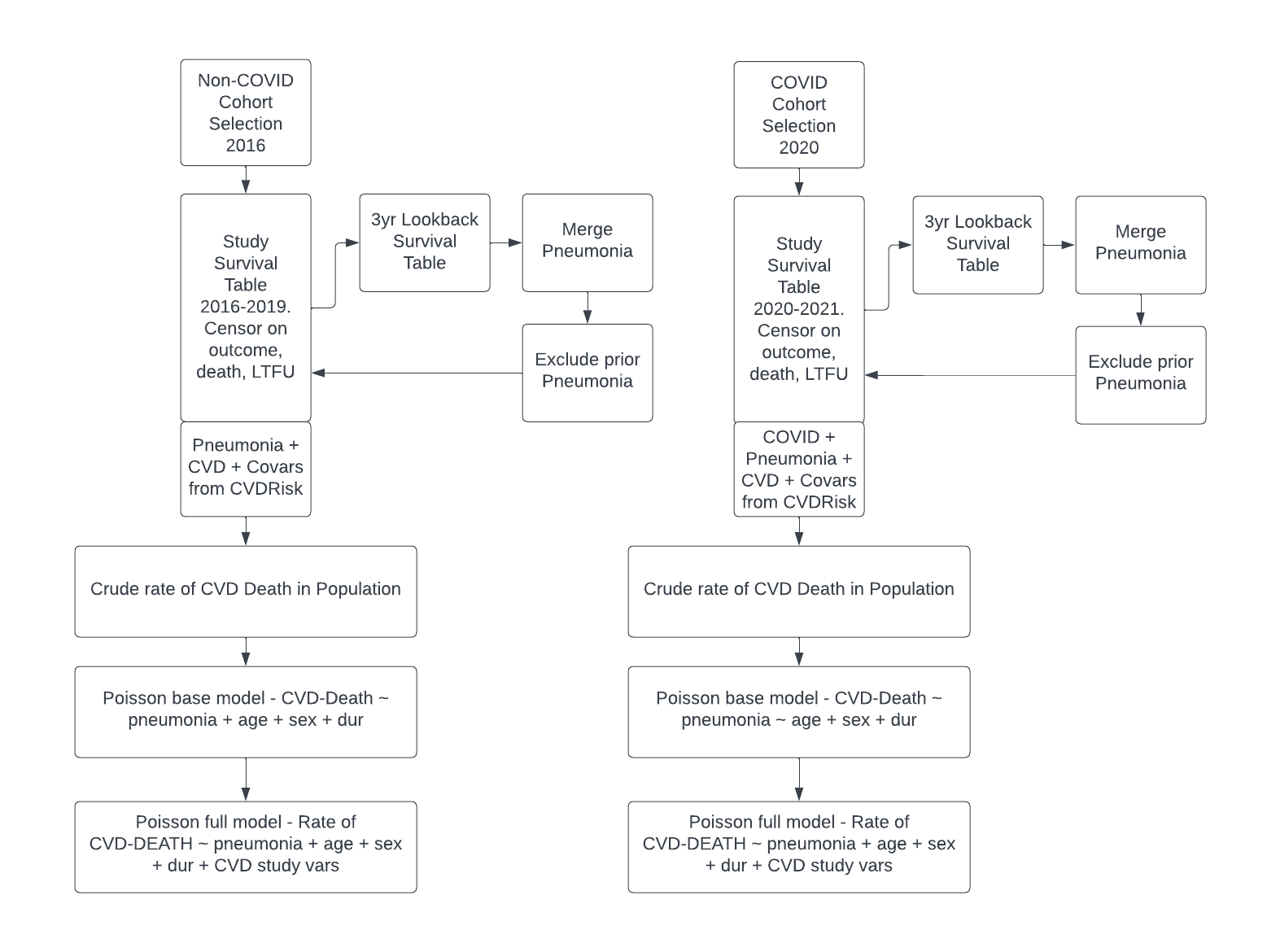
For all analyses, missing data shall be imputed using a multiple imputation procedure as employed in the R package Amelia. Imputation will be performed to obtain 5 imputed datasets, each model will be applied to the 5 imputed datasets and the resulting parameter estimates will be aggregated across all 5 datasets.

A pseudo analysis flow is provided in Figure 1. These analyses will consist of a Poisson regression with data organised into a longitudinal survival table format. The survival table format will utilise time slices of 28 days in length. Interval censoring will be applied where there is loss of observability or death.

The association of the various pneumonia criteria with cardiovascular death will be estimated as a regression coefficient in each model. The exposure variable will be constructed as a time updated binary exposure. Adjustment will be performed firstly with a simple model including age, sex and diabetes duration and secondly with a more complex model including several covariates expected to confound the association between pneumonia exposure and cardiovascular death. The adjustment covariates will not be time updated and will enter the model at baseline only. These adjustment covariates will be as defined in the following T1CVD risk modelling manuscript, with the addition of Ethnicity and number of prescribed medications determined when modelling cardiovascular risk in type 2 diabetes.https://doi.org/10.1007/s00125-021-05478-4

1. Current Age (squared and cubed terms also)
2. Sex
3. Diabetes duration
4. Deprivation
5. HbA1c
6. HbA1c (mean over last 3 years)
7. BMI (log)
8. Height
9. Weight
10. Systolic BP
11. Tchol:HDL Ratio (log)
12. eGFR (log)
13. Albuminuric grade (Norm/Micro/Macro)
14. Retinopathy status
15. Smoking status
16. Treated for dyslipidaemia
17. Treated for hypertension
18. Ever atrial fibrillation
19. Number of prescribed meds (from Mellor T2 model)
20. Ethnicity (from Mellor T2 model)

Figure 1. Pseudo analysis flow



# Sensitivity Analysis

The following adjustments will be applied to the analysis

1. Drop influenza from pneumonia code list.
2. Limit CVD related death to the underlying cause of death as CVD and not as any mention on the death certificate.
3. Age stratification will be applied to determine if any differences in results are related to specific age strata.